Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Master in Data Science

Master Thesis

A Machine Learning Exploration of the Interplay between Chemical Pollutants, Geographic Distribution and Lung Cancer Incidence

Author: Laura del Valle Andara Méndez

Madrid. July, 2025

This Master Thesis has been deposited in ETSI Informáticos de la Universidad Politécnica de Madrid.

Master Thesis
Master in ***Data Science***

Title: **A Machine Learning Exploration of the Interplay between Chemical Pollutants, Geographic Distribution and Lung Cancer Incidence**
July, 2025

Author: **Laura del Valle Andara Méndez**

Tutor:
Guillermo Antonio Vigueras González
ETSI Informáticos
Departamento de Lenguajes Y Sistemas Informáticos e Ingeniería De Software
Universidad Politécnica de Madrid

Co-Tutor:
Paloma Tejera Nevado

Centro de Tecnología Biomédica
Universidad Politécnica de Madrid

Resumen

El cáncer de pulmón representa una de las principales causas de muerte a nivel mundial, siendo responsable de un alto porcentaje de diagnósticos y fallecimientos por cáncer, particularmente debido a su detección tardía y a factores de riesgo aún poco comprendidos. Si bien el tabaquismo es identificado como la causa principal, un número creciente de casos en personas no fumadoras ha impulsado la investigación sobre otros factores de riesgo, especialmente los ambientales. Entre estos, la exposición a pequeñas partículas que se encuentran en el aire con un diámetro de máximo 2,5 micrómetros (PM2.5) han demostrado tener una relación significativa con el desarrollo del cáncer pulmonar.

Este Trabajo Final de Máster propone el desarrollo de un modelo de aprendizaje automático que permita identificar y analizar correlaciones entre variables ambientales, exposición a sustancias químicas y la incidencia del cáncer de pulmón en diferentes regiones geográficas. Para ello, se realiza una revisión del estado del arte sobre los factores de riesgo del cáncer de pulmón, se recopilan datos ambientales y epidemiológicos, y se aplica una metodología que incluye el preprocesamiento de datos, el entrenamiento de modelos supervisados, y la evaluación de su rendimiento mediante métricas específicas.

De acuerdo con la evaluación realizada, aunque la regresión lineal ofrece la mayor transparencia para interpretar patrones geográficos, el modelo de Random Forest resulta más eficaz para revelar la influencia de los contaminantes ambientales gracias a su capacidad para capturar relaciones no lineales. Por otro lado, XGBoost logra el mejor equilibrio entre precisión predictiva e interpretabilidad, siendo un modelo fiable, a pesar de un leve compromiso en términos de interpretabilidad.

Gracias a los resultados obtenidos, se puede determinar el modelo más adecuado a utilizar según el objetivo del análisis, ya sea profundizar en la interpretación, evaluar el impacto de contaminantes ambientales o maximizar la precisión predictiva. Esta información facilita una selección informada de la herramienta más eficaz para cada caso, contribuyendo a optimizar la toma de decisiones en salud pública y a diseñar estrategias preventivas.

Abstract

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, primarily due to late-stage diagnoses and limited understanding of various contributing risk factors. While smoking is identified as the primary cause, a growing number of cases in non-smokers has driven research into other risk factors, especially environmental ones. Among these, exposure to fine particles in the air with a diameter of up to 2.5 micrometers (PM2.5) has been shown to have a significant relationship with the development of lung cancer.

This Masters Final Project aims to develop a machine learning model capable of identifying and analyzing correlations between environmental variables, chemical exposure, and lung cancer incidence across different geographic regions. The research involves a comprehensive review of current knowledge on lung cancer risk factors, collection and preprocessing of environmental and epidemiological data, and the implementation of supervised machine learning techniques to uncover complex interactions that traditional statistical approaches may overlook.

According to the evaluation conducted, although linear regression offers the greatest transparency for interpreting geographic patterns, the Random Forest model proves more effective in revealing the influence of environmental pollutants due to its ability to capture non-linear relationships. On the other hand, XGBoost achieves the best balance between predictive accuracy and interpretability, being a reliable model despite a slight trade-off in terms of explainability.

Thanks to the results obtained, it is possible to determine the most appropriate model to use depending on the objective of the analysis, whether it is to deepen interpretation, evaluate the impact of environmental pollutants, or maximize predictive accuracy. This information facilitates an informed selection of the most effective tool for each case, contributing to optimizing decision-making in public health and designing preventive strategies.

Table of Contents

List of Tables

List of Figures

List of Equations

1 Introduction

1.1 Problem Statement

Over time, the importance of medical research in enhancing people's quality of life has become increasingly evident, underscoring the significance of understanding the root causes of diseases. Cancer is a complex group of diseases that can affect any part of the body, collectively known as malignancies. The primary feature of cancer is the rapid, uncontrolled growth of abnormal cells, which can invade and destroy healthy tissue, including vital organs, a process referred to as metastasis, leading to severe health complications and potentially death. During their lifetimes, one in five men or women are likely to be diagnosed with cancer, while about one in nine men and one in twelve women will die from the disease [1].

Cancer can be classified based on the location in the body where it originates, which is crucial for understanding its behavior and guiding treatment. Lung cancer is the most prevalent type of cancer worldwide, followed by breast cancer, the most common among women. Colorectal cancer, which involves the colon and rectum, is third and heavily targeted by screening initiatives. Prostate cancer ranks fourth, mainly affecting men. Stomach and liver cancers are also notably common, with high mortality rates due to late diagnosis and aggressive nature, respectively [2].

According to the International Agency for Research on Cancer (IARC) and its Global Cancer Observatory (GCO) [2], in 2022, there were approximately 20 million new cancer cases and 9.7 million cancer deaths worldwide. Lung cancer remained the most common and deadliest cancer globally, with nearly 2.5 million new cases (12.4% of all cancers) and was also the leading cause of cancer death, accounting for 1.8 million deaths (18.7% of all cancer deaths). It was followed by breast cancer with 11.6% of cases, colorectal cancer with 9.6%, prostate cancer with 7.3%, and stomach cancer with 4.9%. This information emphasizes the necessity for public health initiatives and sustained research efforts aimed at reducing the burden of lung cancer worldwide.

Lung cancer arises from the excessive and uncontrolled proliferation of certain cells in the lung, causing local problems due to the occupation of space and compression of nearby structures. It is commonly acknowledged that the primary cause of lung cancer is long-term tobacco smoking. In fact, the American Cancer Society (ACS) reports that approximately 85% of lung cancer cases can be attributed to smoking [3]. However, it's concerning that 10–15% of cases occur in individuals who have never smoked, this has led researchers to intensify their exploration of additional risk factors, expanding the focus of lung cancer studies to better comprehend these less prevalent causes.

Despite extensive research on cancer, many causes remain unidentified, making it a continuing focus of scientific investigation. Research has shown that environmental factors contribute significantly to the incidence of diseases such as lung cancer [4,5]. Fine particulate matter (PM2.5) is recognized as a contributing factor to lung cancer [6]. Investigating biomarkers related to PM2.5

exposure offers a promising approach, as these can reflect the biological responses to environmental pollutants and help in understanding health impacts.

Given that lung cancer is often diagnosed at advanced stages, where treatment options are limited, identifying these early biomarkers is crucial for detecting potential risk factors and preventing high-risk situations. Moreover, targeted screening of high-risk individuals can facilitate early detection and significantly improve survival rates.

The complexity of lung cancer, influenced by the interplay of genetic, environmental, and lifestyle factors, presents considerable challenges in its research and management. Current studies have expanded our understanding of these environmental risk factors, but there remains a gap in effectively analyzing and predicting the interrelationships between these diverse exposures and lung cancer incidence. Conventional statistical approaches might not adequately capture the complex interactions and potential synergistic effects among various environmental factors.

In response to this situation, it is proposed to use machine learning techniques to determine correlations between chemicals, their incidence according to geographic distribution and lung cancer to identify and characterize the correlations and interactions that contribute to lung cancer risk. This method incorporates a variety of supervised learning techniques to enhance predictive performance and provide insights that are not readily discernible through conventional approaches.

The insights generated by this model are expected to enhance our understanding of the dynamics between environmental factors and lung cancer across different geographic areas, potentially leading to more targeted public health initiatives and preventive measures. This could significantly contribute to lowering the incidence of lung cancer, particularly among non-smokers, and improve survival rates through earlier detection and intervention strategies.

The preceding factors provide a basis to establish the objectives of this investigation:

1.2 Problem Objectives

1.2.1 General Objective

Develop a machine learning model capable of uncovering correlations among environmental variables, chemical exposures, and the incidence of lung cancer.

1.2.2 Specific Objectives

- Review the state of the art regarding the risk factors associated with lung cancer and their relationship with environmental factors.
- Collect environmental data, exposure levels across various locations, and correlate with lung cancer incidence.

- Preprocess and clean the data to ensure consistency, accuracy, and compatibility.
- Develop a machine learning model using the collected data.
- Evaluate the trained model's performance using appropriate metrics.
- Interpret the model results to identify significant correlations.

1.3 Methodology

For every data science project, defining a methodology is crucial to execute the work in a structured and efficient manner. Given the specific attributes of this task, including a clear research objective, an iterative methodology, and a focus on predictive modeling to generate practical insights for decision-making, the CRISP-DM methodology was chosen as the most suitable approach. According to the CRISP-DM framework [7], the phases are the following (Fig. 1):

- **Business understanding:** This initial phase focuses on comprehending the business goals and requirements, translating them into a clear data mining problem with a plan designed to achieve these goals.
- **Data understanding:** It begins by gathering the initial data and continues with activities to get familiar with the data, to identify quality problems, discover first insights and form hypotheses.
- **Data preparation:** This phase covers all activities to construct the final dataset that will be fed into the modeling tools, like selecting attributes, cleaning and transforming the data.
- **Modeling:** In this phase, different modeling techniques are selected and applied, also their parameters are tuned to obtain optimal values.
- **Evaluation:** Before proceeding to final deployment of the model, it is important to evaluate the model and review the steps executed to be certain it properly achieves the business objectives.
- **Deployment:** It often involves applying “live” models within an organization’s decision-making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the needs, the deployment phase might range from simply creating a report to establishing a complex, repeatable data mining process.

For this project, after understanding the objective and selecting the sources, data is extracted, cleaned, and different supervised machine learning models are used, considering previous research. Finally, specific evaluation criteria are established to select the most suitable model.

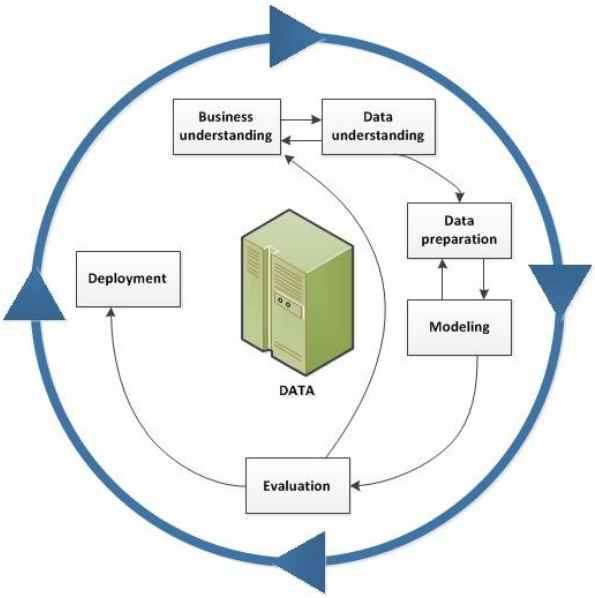


Figure 1. Steps of CRISP-DM methodology

Source: IBM, CRISP-DM Help overview, IBM SPSS Modeler SaaS Documentation.

2 State of the Art

In this research it is considered relevant to present information about cancer, specifically about lung cancer, as well as some of its causes, the role of key pollutants and their geographical distribution. On the other hand, the state of the art regarding recent trends aimed at enhancing Learning Analytics tools through new technologies is highlighted.

The current challenge is to design and propose theoretical models on how things work in the world and what effects may be occurring in real time, in addition to being able to analyze the impact of decisions derived from research and the changes they generate.

2.1 Lung Cancer

Cancer, alongside cardiovascular diseases, constitutes one of the most significant public health challenges of our time on a global scale. Among all cancer types, lung cancer remains the leading cause of cancer-related deaths worldwide, with approximately 2.5 million new cases and 1.8 million deaths reported in 2022 [2].

According to data from the Cancer Observatory of the Spanish Association Against Cancer [8], 30,948 new cases of lung cancer were diagnosed in Spain in 2022, making it the fourth most common cancer in the country. As a result, lung cancer has become a major public health concern in Spain. It is the leading cause of cancer-related death among men and the third among women. In total, Spain reports around 240,000 new cancer cases and 100,000 cancer-related deaths annually, with lung cancer contributing significantly to these figures.

Additionally, with 23,129 deaths due to this cause in 2023, lung cancer is the leading cause of cancer death in Spain, being a very frequent disease, with tobacco consumption being its main risk factor, affecting approximately 80% of lung cancer [9].

Lung cancer typically begins in a way that's like many other types of cancer: when a normal cell transforms into a tumor cell. This transformation typically occurs in the epithelium lining the entire respiratory tree, from the trachea to the thinnest terminal bronchioles, as well as in the cells found in the pulmonary alveoli (Fig. 2).

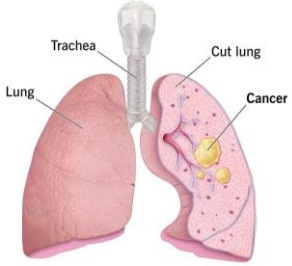


Figure 2. Lung Cancer.
Source: Cleveland Clinic, "Small Cell Lung Cancer," 2024.

Lung cancer is classified into two main types [10] non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC).

- **NSCLC:** accounts for approximately 85 % of all lung cancer cases and includes three primary subtypes, adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Adenocarcinoma often occurs in the outer regions of the lung and is the most common subtype, especially among non-smokers. Squamous cell carcinoma typically develops in the central airways and has a strong association with smoking. Large cell carcinoma, a less common form, arises in various lung locations and tends to grow and spread more rapidly than the other NSCLC subtypes.

- **SCLC:** represents about 10–15 % of lung cancers. It is characterized by small, round cells with a high mitotic rate, aggressive growth, and early metastasis. SCLC is strongly linked to tobacco exposure and, although it often responds initially to chemotherapy and radiotherapy, it has a high recurrence rate and poor long-term prognosis.

Given the significant influence of various external factors in the development of cancer, prevention strategies play a crucial role. Public health initiatives focused on reducing tobacco use, raising awareness, and promoting early detection could significantly decrease both incidence and mortality rates.

Moreover, due to its prevalence and aggressiveness, continued research is vital. Advancing our understanding of lung cancer's risk factors will be key to developing more effective diagnostic tools, treatments, and preventive measures.

Carcer prevention involves two main strategies: primary prevention, which aims to reduce the incidence by avoiding exposure to risk factors, and secondary prevention, which focuses on early detection in healthy individuals so that, through appropriate intervention at this early stage, the natural history of the disease can be modified.

According to the World Cancer Research Fund's Third Expert Report published in 2018 [11], which references the World Health Organization (WHO), between 30% and 50% of cancer cases are preventable through healthy lifestyle choices and by avoiding exposure to occupational carcinogens, environmental pollution, and certain chronic infections. Combining these measures with science-based preventive strategies could significantly improve cancer control.

Currently, about one third of cancer deaths are due to the five main behavioral and dietary risk factors: tobacco and alcohol consumption, high body mass index, low fruit and vegetable intake and lack of physical activity. The European Code Against Cancer contains 12 recommendations in this regard [12].

2.2 Environmental Pollutants and Lung Cancer Incidence

It is widely recognized that tobacco use remains the primary risk factor for lung cancer, being responsible for approximately 80% of cases. However, other contributors such as environmental pollution, indoor air contamination, occupational exposures, and diet, account for a significant proportion of lung

cancer-related deaths (Fig. 3). These factors are estimated to result in about 908,000 deaths annually [1].

Additionally, the rising incidence of lung cancer among non-smokers highlights the need for a closer look at the impact air quality plays on its development. Environmental pollution alone is implicated in about 36% of lung cancer deaths, resulting in around 265,000 annual fatalities [13, 14].

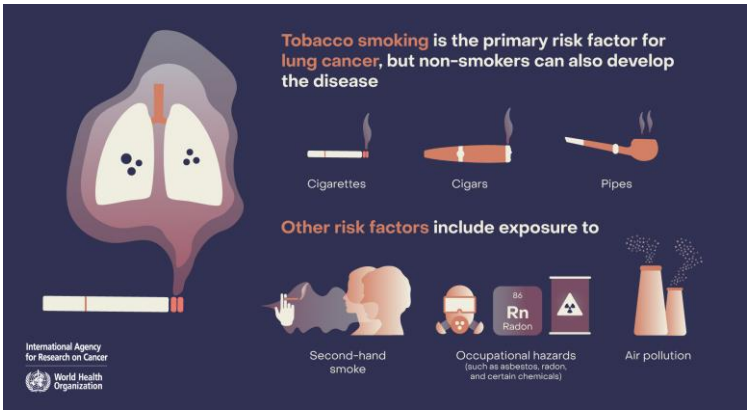


Figure 3. Lung Cancer Factors.

Source: International Agency for Research on Cancer, "Lung Cancer Awareness Month", 2024.

The air we breathe is heavily polluted with numerous carcinogens substances. In 2013, the WHO and IARC classified outdoor air pollution as a Group 1 human carcinogen [15]. Among the various pollutants, particulate matter (PM) is considered the most harmful. It is a mix of nitrates, sulfates, carbon, organic chemicals, and metals, released from natural sources like wildfires and volcanoes, and human activities such as traffic, industry, farming, and biomass burning [16].

Particulate matter is categorized by size: PM10 refers to particles with a diameter smaller than 10 µm, while PM2.5 includes particles smaller than 2.5 µm. Their ability to penetrate the respiratory system depends on their size, although the precise contribution of particle size to disease development remains being studied.

Extensive epidemiological research supports the association between air pollution and lung cancer risk. The ESCAPE study, which combined data from 17 European cohort studies, showed that every 5 µg/m³ increase in PM2.5 raised lung cancer risk by 18%, and every 10 µg/m³ increase in PM10 raised it by 22% [17]. Similarly, a meta-analysis by Hamra et al. (2014), covering 18 cohort studies from Asia, North America, and Europe, reported comparable risks [18]. They also found that higher environmental concentrations of nitrogen dioxide (NO_2), an indicator of traffic-related pollution, significantly increased the risk [19].

The consistent results across multiple populations provide strong evidence linking air pollution to lung cancer. Recognizing pollution as a contributing factor is crucial for developing prevention strategies and improving early detection efforts.

2.3 Key Pollutants and Their Geographical Distribution

In terms of its constituents, air pollution shows wide geographical variation and consists of many different substances. The proportion and concentration of these pollutants can also differ. However, the information available to characterize the full pollution mix remains limited. Some pollutants, such as particulate matter (PM) are routinely measured worldwide, while others are less frequently monitored, although some data exists. In addition, air pollution may contain harmful substances that have yet to be identified.

Recent scientific evidence, primarily based on studies in Europe and North America, consistently suggests that urban air pollution causes adverse health impacts [20]. According to estimates from the World Health Organization (WHO) Global Burden of Disease project, approximately 5% of all cancer-related mortality involving the trachea, bronchus, and lung can be attributed to urban air pollution, measured through particulate matter (PM) concentrations. Although the highest absolute number of deaths occurs in developing countries, proportionally, some of the most affected areas are located within Europe [21].

The air quality index (AQI) provides an accessible and real-time indication of air quality at national monitoring network stations. In addition to reporting current air pollution levels, the AQI includes health recommendations tailored both to the general population and to sensitive groups and allows the public and authorities to track trends in air quality over recent months [22].

It defines six air quality categories: good, reasonably good, fair, unfavorable, very unfavorable, and extremely unfavorable (Fig. 4). Each station is assigned the worst air quality category for any of the pollutants considered for its estimation, whether measured data or derived from the Copernicus Atmosphere Monitoring Service (CAMS).

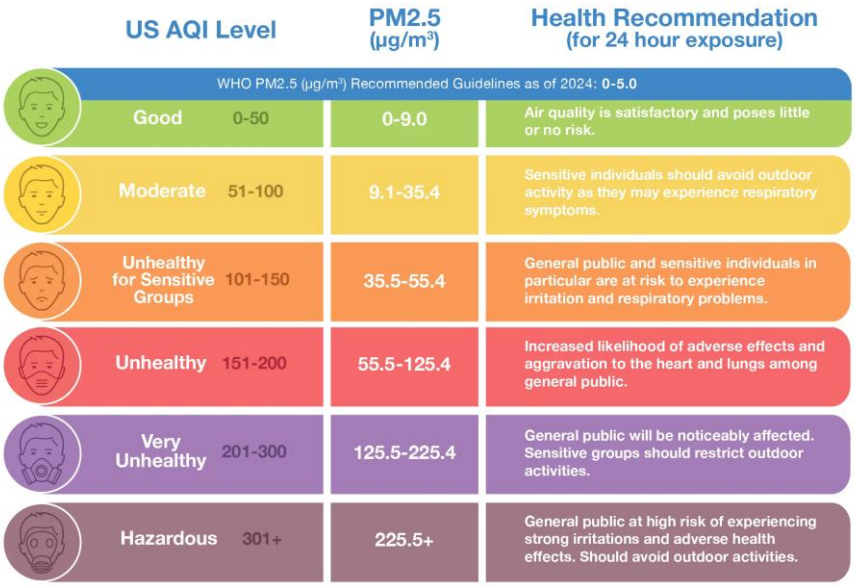


Figure 4. AQI Levels.

Source: IQAir, "What is the difference between the US AQI and WHO air quality guidelines?", 2016.

Air quality assessment involves methods to measure, calculate, predict, or estimate pollutant concentrations in ambient air or their deposition onto surfaces. For this purpose, authorities divide the territory into zones and agglomerations with similar air quality. Two locations are considered equivalent if their pollutant levels fall within the same legislative range, either above or below the established limit and target values on an annual basis [23].

In Spain, air quality evaluation and reporting are managed by the Deputy Directorate-General for Pollution Prevention of the Ministry for the Ecological Transition and the Demographic Challenge. The objectives are to provide comparable information across the territory, assess the effectiveness of measures, and inform both the public and the European Commission.

In this context, identifying the key pollutants that significantly impact human health is crucial for understanding the broader effects of air pollution.

A comprehensive review published in Environmental Health Perspectives outlines the wide range of health effects associated with long-term arsenic exposure [24]. The authors present strong evidence linking chronic arsenic exposure to several types of cancer, particularly skin and lung cancers. The review also highlights epidemiological studies from Taiwan that demonstrate a clear dose-response relationship between elevated arsenic levels in drinking water and increased mortality from liver cancer. It further reports associations with kidney and bladder cancers, as well as neurological disorders and other systemic health conditions.

A study published in Environmental Pollution estimated current BaP concentration levels, population exposure, and potential health impacts in Europe. The results showed significant exceedances of the European target value for BaP, especially in Central and Eastern Europe. Approximately 20% of the European population is exposed to BaP concentrations above the EU target value, leading to an estimated 370 lung cancer incidences per year among the 60% of the European population included in the estimation. The study emphasizes the need for more BaP measurements in areas with low monitoring density and highlights the health risks posed by polycyclic aromatic hydrocarbon exposure [25].

A 2023 review on cadmium-induced lung carcinogenesis highlights that cadmium inhalation leads to oxidative stress, DNA damage, and chronic inflammation. These effects are driven by increased reactive oxygen species (ROS), which damage cellular components and impair DNA repair mechanisms. Cadmium also activates inflammatory pathways, such as the cGAS-STING and NLRP3 inflammasome, contributing to cell death and promoting tumor development in lung tissue [26].

A 2023 study published in the International Journal of Cancer investigated the relationship between occupational exposure to nickel and the risk of lung cancer. The findings indicated that even relatively low cumulative exposure levels were significantly associated with an elevated risk of lung cancer, with the effect being more pronounced in male workers. These results suggest that occupational exposure to nickel, even at lower intensities, may pose a substantial carcinogenic risk to respiratory health [27].

A cohort study by Anttila et al. (2022) investigated 20,729 Finnish workers monitored for blood lead levels between 1973–1983 and followed for lung cancer incidence until 2014. The study found a clear dose–response relationship: workers with mean blood lead levels of 1.0–1.9 μmol/L had a 72% increased risk of lung cancer (RR = 1.72), and those with ≥2.0 μmol/L had a 163% increase (RR = 2.63), compared to those with <0.5 μmol/L. These associations remained significant after adjusting for smoking and co-exposures. This supports lead as a probable human lung carcinogen and highlights that even relatively low occupational exposures to airborne lead particles can significantly elevate lung cancer risk [28].

According to the World Health Organization (WHO), exposure to fine particulate matter (PM2.5) is linked to cardiovascular and respiratory diseases, as well as lung cancer. Low- and middle-income countries bear the greatest burden, with about 89% of the 4.2 million annual premature deaths from outdoor air pollution occurring in these regions [29].

The U.S. Environmental Protection Agency (EPA) states that particle size is directly linked to the potential to cause health problems. Particles less than 10 micrometers in diameter (PM10) can penetrate deep into the lungs, and some may even enter the bloodstream, affecting both the lungs and the heart. Among these, fine particles (PM2.5) pose the greatest risk to health [30].

2.4 Artificial Intelligence

Artificial Intelligence can be defined as the field of study focused on creating intelligent agents or systems capable of perceiving their environment and making decisions that aim to maximize their chances of achieving specific objectives. As explained by Russell and Norvig, the core idea behind AI is not merely to mimic human behavior, but to build rational agents that select optimal actions based on available information to achieve their goals efficiently [31].

Using artificial intelligence, knowledge can be acquired from data obtained through various sources, enabling the automation and acceleration of tasks that were traditionally performed exclusively by humans. Moreover, AI systems have the capability to adapt their responses by analyzing the impact of their actions on the environment in which they operate.

Artificial intelligence encompasses various techniques aimed at enabling machines to perform complex and adaptive tasks. Among these, machine learning has emerged as a particularly influential approach, contributing significantly to the advancement of AI by utilizing data to enhance system performance and adaptability. Understanding the role of machine learning is essential to grasp the current developments within the field.

2.5 Machine Learning models and techniques used for data processing

Machine learning is a subset of artificial intelligence that involves the development of algorithms capable of learning from data and improving their

performance as they are exposed to more information. These algorithms enable machines to make predictions or decisions without requiring explicit programming for each specific task, thus minimizing the need for human intervention [32].

Rather than receiving direct instructions, machines that employ machine learning are provided with data and patterns, which allow them to independently derive conclusions and make decisions. These algorithms can be used for a wide range of tasks, including classifying data, predicting future outcomes, grouping similar data, or even deciding on the best actions to take in a situation.

According to Mitchell [32], machine learning algorithms can be broadly categorized based on the type of supervision they receive during training (Fig 5). These categories include:

- **Supervised Learning:** involves learning a function that maps an input to an output based on example input-output pairs. The algorithm is trained on a labeled dataset, meaning each training example is paired with an output label. Common supervised learning tasks include classification and regression. *Examples:* Decision Trees, Support Vector Machines, Neural Networks, Linear Regression.

- **Unsupervised Learning:** the algorithm is provided with input data without labeled responses. The objective is to infer the natural structure or distribution in the data. This category often includes clustering and dimensionality reduction. *Examples:* K-Means Clustering, Principal Component Analysis (PCA), Autoencoders.

- **Reinforcement Learning:** is a type of machine learning where an agent learns to make decisions by performing actions in an environment to maximize some notion of cumulative reward. The feedback is in the form of rewards or penalties, rather than explicit labels. *Examples:* Q-Learning, Deep Q-Networks (DQN), Policy Gradient Methods.

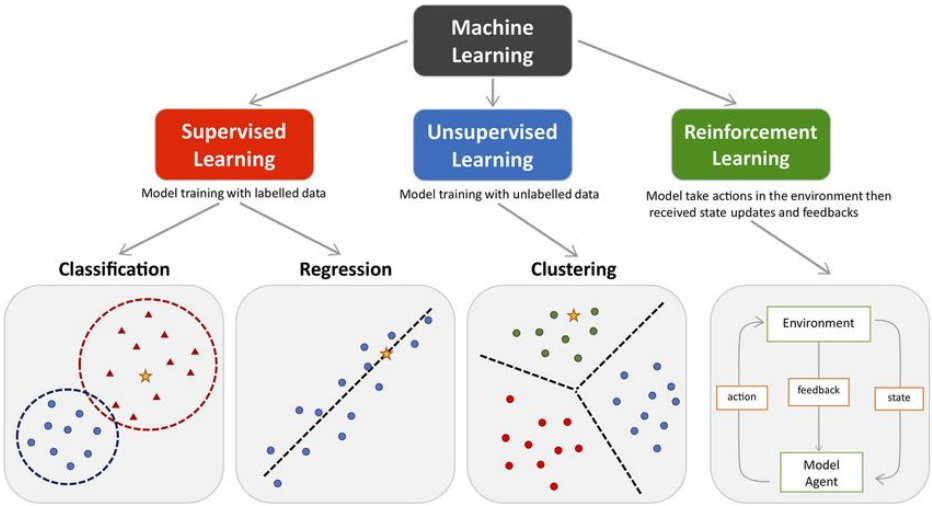


Figure 5. Machine Learning Algorithms Classification.

Source: World Journal of Advanced Research and Reviews, "Comprehensive review of machine learning models for SQL injection detection in e-commerce" 2024

To develop the system that has been proposed in this TFM, three widely known supervised machine learning models (Linear Regression, Random Forest and XG Boost) are selected to train, test and validate the data to predict the incidence of lung cancer in Spain.

2.5.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables through a linear function. As described by Montgomery et al. in Introduction to Linear Regression Analysis, the model assumes that the observed response can be expressed as a combination of a deterministic component and a random error term [33]. The simple linear regression model is given by the following equation (Eq. 1):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Equation 1. Linear Regression.

where y is the response variable, x is the predictor, β_0 is the intercept, β_1 is the slope, and ε represents random error. Parameter estimation is typically performed using the least squares method, minimizing the sum of squared deviations between observed and predicted values. The model generalizes to multiple regression when several predictors are involved.

One of the principal advantages of linear regression lies in its interpretability. The model provides explicit coefficients that quantify the effect of each independent variable on the dependent variable, thereby offering valuable insights into the relationships among variables. This transparency makes linear regression not only easy to understand but also straightforward to implement, rendering it an essential tool in both statistical analysis and machine learning [34].

Furthermore, its simplicity facilitates its use as a foundational framework for more complex modeling techniques. The primary objective in employing linear regression is to identify the best-fitting line that minimizes the discrepancy between the observed and predicted values, typically achieved through the minimization of the residual sum of squares [34].

2.5.2 Random Forest

Random forest is an ensemble learning method primarily used for classification and regression problems. It works by constructing multiple decision trees during the training phase, where each tree is trained on a different bootstrap sample of the data and considers a random subset of features when making splits (Fig. 6). This randomness helps to create diverse trees, which reduces the risk of overfitting that can occur with individual decision trees. [35]

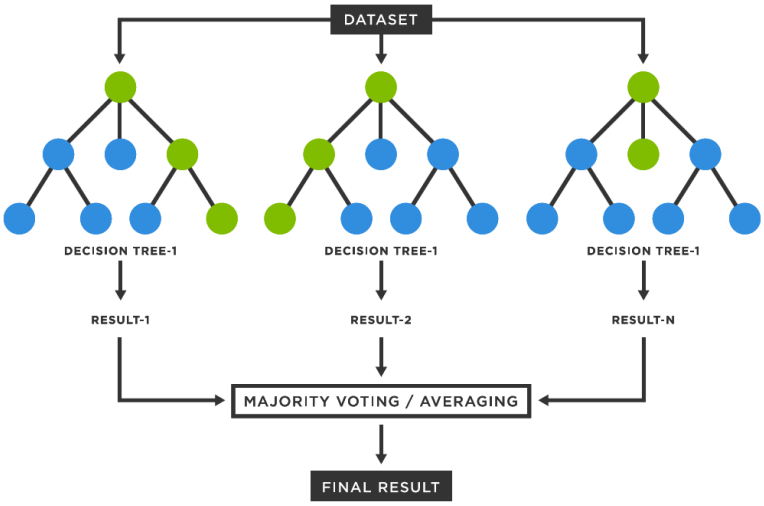


Figure 6. Random Forest.

Source: Applied Sciences, "Example of a decision tree and random forest", 2025.

The final prediction of the random forest is obtained by aggregating the predictions from all individual trees using majority voting for classification tasks or averaging for regression tasks. According to Géron [35], this approach enhances the model's robustness and accuracy by combining the strengths of many weak learners into a stronger overall predictor.

Random forests offer several key advantages, including a reduced risk of overfitting compared to individual decision trees. By aggregating predictions from multiple uncorrelated trees, they lower variance and improve predictive accuracy. They are flexible, effectively handling both classification and regression tasks, and can maintain performance even when some data values are missing. Random forests also provide useful measures of feature importance, aiding in model interpretation [36].

However, these benefits come with challenges, random forests can be computationally intensive and require more memory due to the large number of trees involved. Additionally, while single decision trees are easy to interpret, the ensemble nature of random forests makes their predictions less transparent [37].

2.5.3 XGBoost (Extreme Gradient Boosting)

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting algorithms designed for scalable and efficient supervised learning. It builds an ensemble of decision trees sequentially, where each new tree aims to correct the errors made by the previous ensemble by optimizing a differentiable loss function using gradient descent techniques (Fig. 7). XGBoost introduces system optimizations such as parallel processing, tree pruning, and regularization to prevent overfitting, resulting in faster training times and improved predictive performance compared to traditional gradient boosting methods [38].

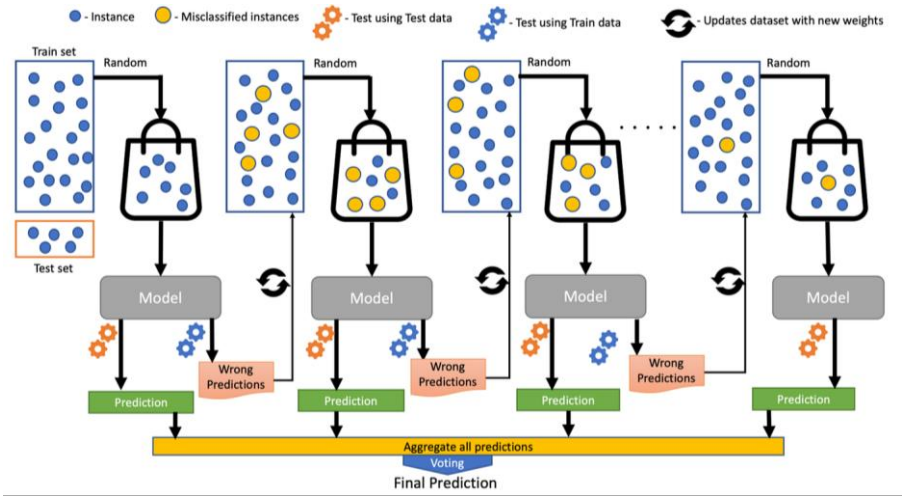


Figure 7. XGBoost.

Source: Medium, "Introducción a los Métodos de Ensamble y al Algoritmo de XGBoost: Caso Práctico," 2022.

This algorithm uses L1 and L2 regularization to prevent overfitting. It handles sparse and weighted data through a sparsity-aware method and weighted quantile sketch. Its block structure allows fast parallel training by optimizing memory and CPU usage. Additionally, it supports out-of-core computing to process datasets larger than memory. These features contribute to high accuracy and speed in tasks like classification and regression [38].

2.6 Data Preparation Techniques

2.6.1 Cross-validation

Cross-validation is a widely used statistical resampling method for evaluating the predictive performance and generalization ability of machine learning models. It works by partitioning the available dataset into a series of complementary subsets: the model is trained on one subset (the training set) and validated on another (the validation set), rotating through multiple combinations to ensure robust evaluation (Fig. 8). One of the most common forms is k-fold cross-validation, where the data is divided into k equal parts, and the model is trained and validated k times, each time leaving out a different fold for validation [39].

This method reduces the variance associated with a single train-test split and is especially effective when data availability is limited. By ensuring that every data point is used for both training and validation, k-fold cross-validation provides a more stable and comprehensive evaluation metric [40]. Common choices for k, such as 5 or 10, represent a trade-off between computational efficiency and the bias-variance characteristics of the performance estimate.

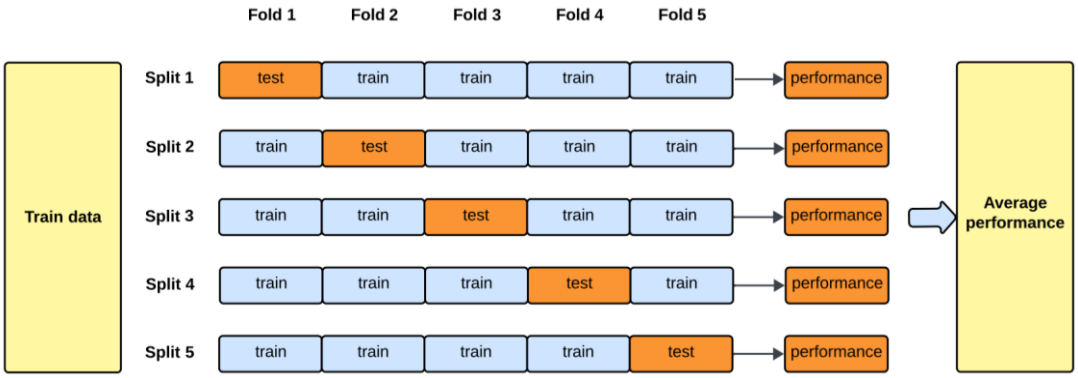


Figure 8. K-fold Validation Process

Source: Medium, "How-To: Cross Validation with Time Series Data", 2023

2.6.2 K-NN Imputation

K-Nearest Neighbors (KNN) imputation is a non-parametric method used to estimate missing values in datasets by leveraging the similarity between observations. For each missing value, the algorithm identifies the k most similar instances (neighbors) based on a defined distance metric, commonly Euclidean distance, using the available (non-missing) features.

The missing value is then imputed by aggregating the values of these neighbors, often through a mean (for numerical data) or mode (for categorical data). This approach preserves local data patterns and is particularly effective when the dataset has underlying structure or correlation between features. However, KNN imputation can be computationally intensive for large datasets and may be sensitive to the choice of k and distance metric [41, 42].

The KNN method offers several advantages [43, 44]:

- **Preserves multivariate structure:** Missing values are estimated based on similar instances, helping to retain the relationships among variables in the dataset.
- **Captures local patterns:** Unlike global methods (e.g., mean imputation), KNN uses neighborhood information, making it more sensitive to local data behavior.
- **Non-parametric:** It does not assume any specific data distribution, which makes it broadly applicable across different domains and dataset types.
- **Handles mixed-type data:** KNN can be adapted to work with both numerical and categorical variables using appropriate distance metrics, such as Gower's distance.
- **Robust to outliers:** Imputations are based on neighboring instances, which reduces the influence of extreme or anomalous values.

- **Simple and interpretable:** The method is easy to implement and understand, making it accessible for practical use in various research settings.

2.6.3 One Hot Encoder

One-hot encoding is a preprocessing technique used to convert categorical variables into a binary vector representation, enabling machine learning algorithms to process non-numeric data. For a categorical feature with n distinct categories, one-hot encoding creates n new binary features, each representing the presence (1) or absence (0) of one category (Fig. 9).

This approach avoids imposing any ordinal relationship among categories, preserving the nominal nature of the data. While one-hot encoding increases the dimensionality of the dataset, it is widely used due to its simplicity and effectiveness in representing categorical variables for algorithms that require numerical input [40, 45].

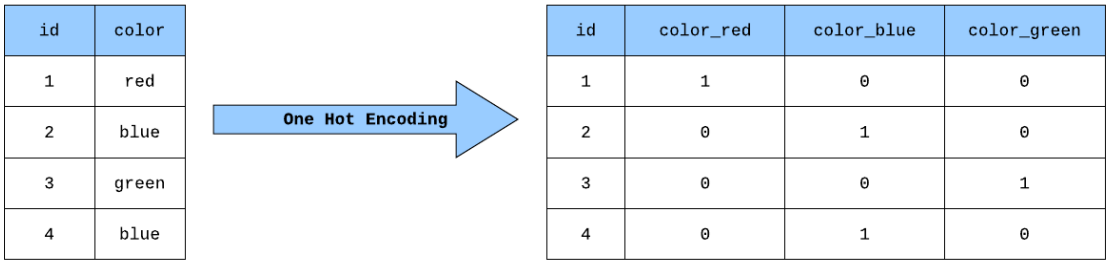


Figure 9. One Hot Encoding.

Source: Medium, "Building a One Hot Encoding Layer with TensorFlow", 2020.

2.6.4 Standard Scaling

Standard scaling, also known as z-score normalization, is a feature scaling technique that transforms data to have a mean of zero and a standard deviation of one. For each feature, the scaling is performed by subtracting the mean and dividing by the standard deviation of that feature, according to the next formula (Eq. 2):

$$z = \frac{(x - \mu)}{\sigma}$$

Equation 2. Standard Scaling.

where x is the original value, μ is the mean of the feature, and σ is the standard deviation. This transformation ensures that features are centered and scaled, which can improve the convergence and performance of many machine learning algorithms, especially those sensitive to the scale of input data such as k-nearest neighbors, support vector machines, and gradient-based methods [40, 44].

2.6.5 Winsorization

Winsorization is a statistical technique used to reduce the influence of extreme values in a dataset by limiting (or capping) the values at specified percentiles. Rather than removing outliers, it replaces values above a certain upper percentile and below a lower percentile with the respective boundary values. It is particularly useful in situations where outliers may distort parameter estimates, such as the mean or standard deviation, while preserving the sample size [46]

In a 5% Winsorization procedure, all data values below the 5th percentile are replaced with the value at the 5th percentile, and all values above the 95th percentile are replaced with the value at the 95th percentile. This method reduces the influence of extreme outliers while preserving the overall structure and size of the dataset.

2.7 Evaluation metrics

2.7.1 R-squared (R^2)

R-squared, also known as the coefficient of determination, is a statistical measure that indicates the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It provides an assessment of the model's goodness of fit, with values ranging from 0 (no explanatory power) to 1 (perfect explanatory power). The formula is given by the following (Eq. 3):

$$R^2 = 1 - \frac{\sum_{\{i=1\}}^{n}(y_i - \hat{y}_i)^2}{\sum_{\{i=1\}}^{n}(y_i - \bar{y})^2}$$

Equation 3. R-squared.

where y_i is the observed value, y_i is the predicted value, and $\bar{y}$ is the mean of the observed values [32].

Despite its usefulness, R-squared should be interpreted with caution, especially in models with many predictors or in non-linear contexts, where complementary metrics such as adjusted R-squared, RMSE, or MAE may offer a more complete evaluation of model performance.

2.7.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a commonly used loss function and evaluation metric in regression tasks. It measures the average magnitude of the errors between predicted and actual values without considering their direction, with this formula (Eq. 4):

$$MAE = \frac{1}{n}\sum_{\{i=1\}}^{n}|y_i - \hat{y}_i|$$

Equation 4. Mean Absolute Error.

Where y_i are the true values, $\hat{y}_i$ are the predicted values, and n is the number of observations. MAE is easy to interpret and less sensitive to outliers than metrics like the Mean Squared Error (MSE), making it suitable for many real-world prediction tasks. [40]

2.8 Machine Learning in Environmental and Health Data

Machine learning (ML) is becoming more common in healthcare, changing important tasks like diagnosing with images, predicting disease outcomes, classifying patients, and monitoring treatments. With the increasing availability of digital health data, especially from electronic health records, there is a growing need for advanced tools to analyze this information. These algorithms can find complex patterns and connections in medical data that are difficult to detect manually. [47]

As these data-driven methods become more widely used in healthcare, providers are better able to use predictive approaches in precision medicine. This shift helps create a more efficient healthcare system, leading to better care, improved outcomes for patients, and smarter decision-making. Oncology is a key area where ML is helping improve early detection, personalizing treatments, and predicting patient outcomes.

Several recent studies have used machine learning (ML) to predict lung cancer risk. In [48], Pathan et al. tested four ML models: Support Vector Machine (SVM), K Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF), to classify lung cancer risk levels (low, medium, high) based on clinical and environmental data. One important part of this study was the use of explainable artificial intelligence (XAI) methods, such as LIME and decision boundary plots, to help understand how the models made decisions. This helped make the models more trustworthy for doctors and patients. After tuning the model parameters, the authors achieved very high accuracy, close to 100 percent. They also showed which input features were most important, such as coughing up blood, being a passive smoker, and feeling tired.

Another study by Wang et al. [48] looked at lung cancer from a population-level view. They used ML to predict lung cancer incidence rates in Taiwan by combining health, environmental, and social data. Five algorithms were tested: Linear Regression, Support Vector Regression (SVR), Random Forest, KNN, and Cubist Model Tree.

To improve the models, they selected the most relevant features and removed others that caused multicollinearity. The Cubist model had the best performance, with an R^2 of 0.96 and a low error rate. Their results showed that, besides smoking, air pollution (especially NO_2), number of factories, and employment levels were key factors related to lung cancer. This study shows that using different types of data, not only clinical, can help understand and predict cancer trends in a country [49].

3 Project Development

The present research aims to develop a machine learning model that can identify correlations between geographic distribution, chemical pollutants, and the incidence of lung cancer.

The study is organized into four main phases: Data Understanding, Data Preparation, Modeling and Evaluation, structured adapting the CRISP-DM framework to the specific needs of the project, with each phase addressing a distinct area of focus as described in the following sections (Fig. 10). All the code is available on github[1].

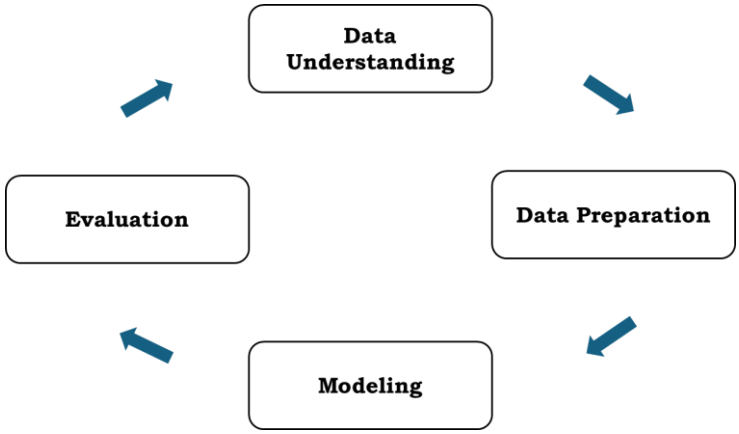


Figure 10. Phases of the study.

3.1 Data Understanding

3.1.1 Data Collection

The data collection corresponds to a complex procedure due to the diversity, the variables involved and the large amount of data that can be found in the spectrum referred to Cancer Risk Factors.

The selected data source for the carcer information was the webpage of the Spanish Association Against Cancer [2] (Fig 11). It has information about Dimensions of cancer, risk factors, psychosocial problems and healthcare, being the first two the most relevant to the study.

[1] TFM: Lung Cancer and Environmental Risk Factors, (2025). Available: https://github.com/laurandara/TFM

[2] Asociación Española Contra el Cáncer (AECC), Observatorio del Cáncer, [Online]. Available: https://observatorio.contraelcancer.es/.

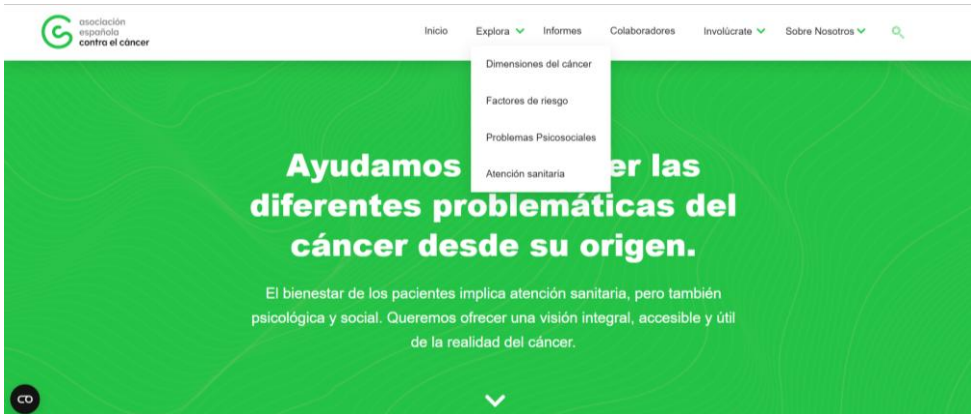


Figure 11. Spanish Association Against Cancer Home Page.

The page was selected for its credible sources, including the International Agency for Research on Cancer for medical data and the National Statistics Institute for population data. In the "Dimensions of Cancer" section, three key characteristics are presented: incidence, prevalence, and mortality. For the purposes of this study, incidence was chosen as the primary characteristic for evaluation and data collection. In the "Risk Factors" section, various categories are examined like tobacco, alcohol, obesity, physical activity, and consumption among young people. From these, tobacco was selected as the main risk factor for analysis.

Another key variable in this research is air pollution data. Initially, the approach focused on identifying chemical substances present in the air by scraping various sources, such as chemical databases. However, this method provided only general information about the pollutants, without associated concentration data, which had to be searched for separately.

Obtaining reliable concentration data for the list of scraped chemicals proved particularly challenging, as many of these pollutants are not routinely measured or consistently reported. Also, numerous parameters beyond chemical names and basic classifications were either inconsistent or irrelevant to the objectives of the study, complicating data integration.

Ultimately, the European Environment Agency (EEA) was identified as a reliable and relevant source for the study, providing standardized pollutant concentration data through its official website via the Download Service[3] (Fig. 12).

The Download Service provides access to air quality measurements time series. Three sets of time series are available for download:

- Historical Airbase data delivered between 2002 and 2012 before Air Quality Directive 2008/50/EC entered into force.

[3] European Environment Agency, Air quality in Europe – data dashboard, [Online]. Available: https://www.eea.europa.eu/en/datahub/datahubitem-view/778ef9f5-6293-4846-badd-56a29c70880d

- Verified data (E1a) from 2013 to 2023 reported by countries by 30 September each year for the previous year.

- Unverified data transmitted continuously (Up To Date/UTD/E2a) data from the beginning of 2024.

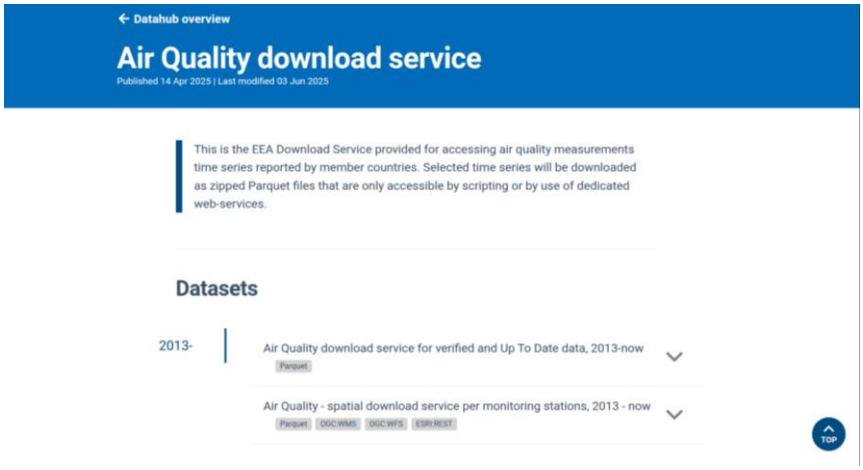


Figure 12. Air Quality Download Service Home Page.

Initially, it was planned to limit the data to a five-year period, from 2018 to 2022, for processing purposes. However, this was later reduced to four years after it was determined that only the 2019–2022 period was common to both the cancer incidence and risk factor datasets required for the study. As a result, the final analysis was conducted using data from the years 2019 to 2022, inclusive.

The objective is to gather three types of information (cancer incidence, risk factors, and air pollutants) for the specified time and merge them into a single final dataset in the following data preparation steps. Cancer incidence and risk factor data are linked by province, while the resulting dataset is then combined with air pollution data by aggregating and matching records by year.

3.1.1.1 Incidence Dataset

The dimension webpage (Fig. 13) provided the option to download the dataset in CSV format, with the ability to select specific years. Data from 2019 to 2022 (inclusive) was chosen. The dataset included columns for province, general incidence rate, and dimension, as well as the number of cases grouped by sex, age, and type of cancer. An example of the generated csv can be seen in Fig 14.

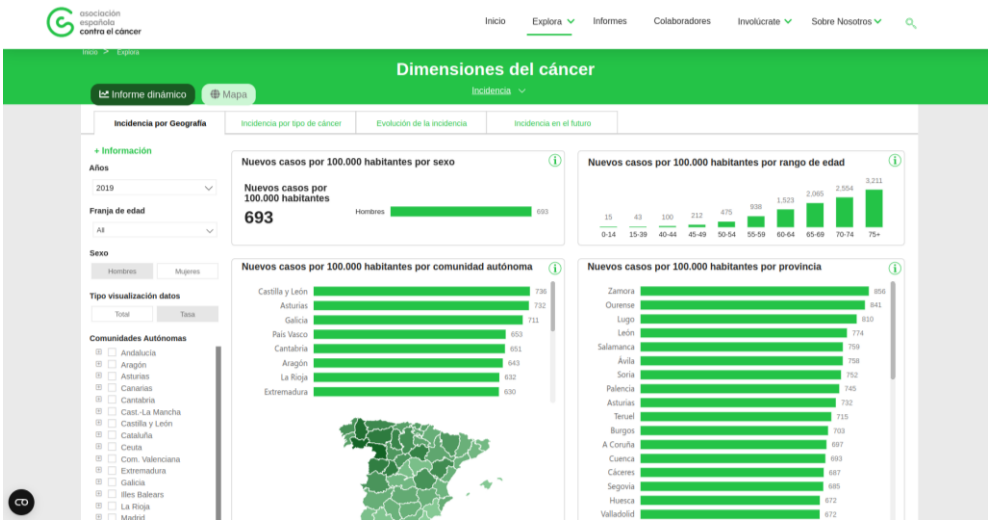
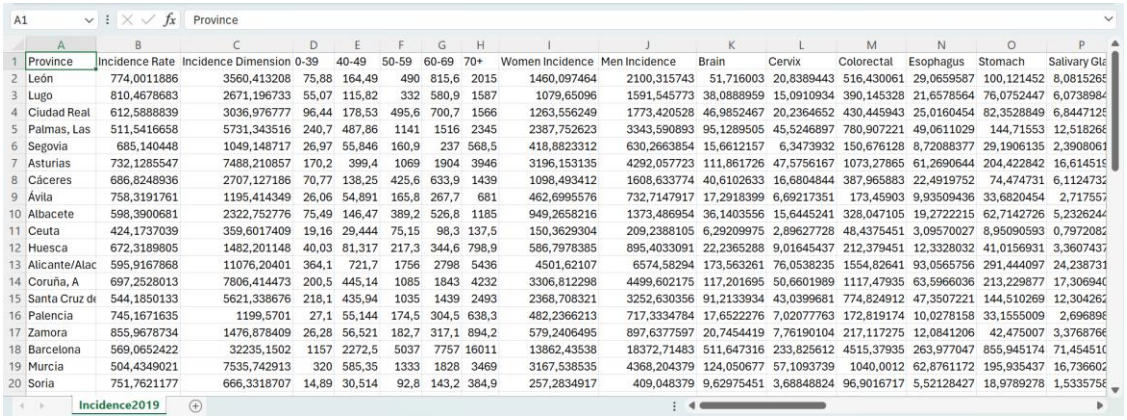


Figure 13. Cancer Incidence page.



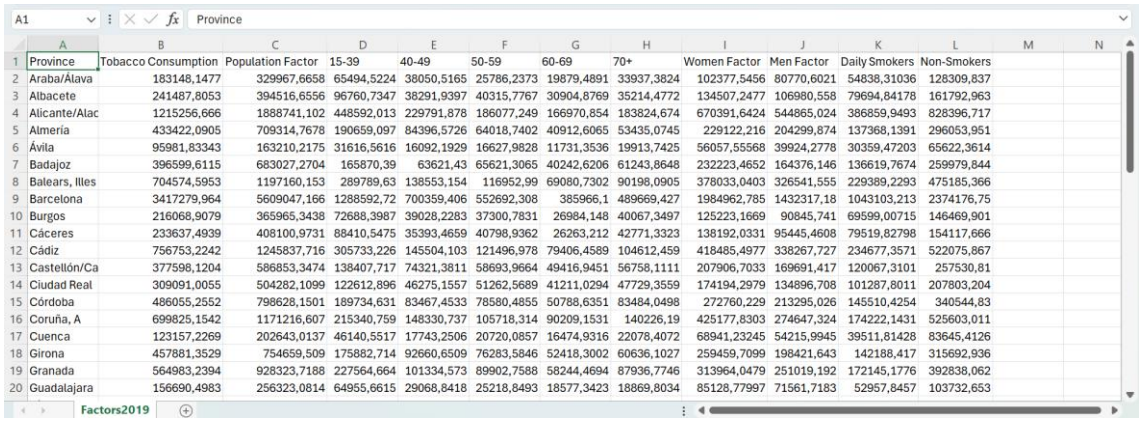
	Province	Incidence Rate	Incidence Dimension	0-39	40-49	50-59	60-69	70+	Women Incidence	Men Incidence	Brain	Cervix	Colorectal	Esophagus	Stomach	Salivary Gla
2	León	774,0011886	3560,413208	75,88	164,49	490	815,6	2015	1460,097464	2100,315743	51,716003	20,8389443	516,430061	29,0659587	100,121452	8,0815265
3	Lugo	810,4678683	2671,196733	55,07	115,82	332	580,9	1587	1079,65096	1591,545773	38,0888959	15,0910934	390,145328	21,6578564	76,0752447	6,0738984
4	Ciudad Real	612,5888839	3036,976777	96,44	178,53	495,6	700,7	1566	1263,556249	1773,420528	46,9852467	20,2364652	430,445943	25,0160454	82,3528849	6,8447125
5	Palmas, Las	511,5416658	5731,343516	240,7	487,86	1141	1516	2345	2387,752623	3343,590893	95,1289505	45,5246897	780,907221	49,0611029	144,71553	12,518268
6	Segovia	685,140448	1049,148717	26,97	55,846	160,9	237	568,5	418,8823312	630,2663854	15,6612157	6,3473932	150,676128	8,72088377	29,1906135	2,3908061
7	Asturias	732,1285547	7488,210857	170,2	399,4	1069	1904	3946	3196,153135	4292,057723	111,861726	47,5756167	1073,27865	61,2690644	204,422842	16,614519
8	Cáceres	686,8248936	2707,127186	70,77	138,25	425,6	633,9	1439	1098,493412	1608,633774	40,6102633	16,6804844	387,965883	22,4919752	74,474731	6,1124732
9	Ávila	758,3191761	1195,414349	26,06	54,891	165,8	267,7	681	462,6995576	732,7147917	17,2918399	6,69217351	173,45903	9,93509436	33,6820454	2,717557
10	Albacete	598,3900681	2322,752776	75,49	146,47	389,2	526,8	1185	949,2658216	1373,486954	36,1403556	15,6445241	328,047105	19,2722215	62,7142726	5,2326244
11	Ceuta	424,1737039	359,6017409	19,16	29,444	75,15	98,3	137,5	150,3629304	209,2388105	6,29209975	2,89627728	48,4375451	3,09570027	8,95090593	0,7972082
12	Huesca	672,3189805	1482,201148	40,03	81,317	217,3	344,6	798,9	586,7978385	895,4033091	22,2365288	9,01645437	212,379451	12,3328032	41,0156931	3,3607437
13	Alicante/Alac	595,9167868	11076,20401	364,1	721,7	1756	2798	5436	4501,62107	6574,58294	173,563261	76,0538235	1554,82641	93,0565756	291,444097	24,238731
14	Coruña, A	697,2528013	7806,414473	200,5	445,14	1085	1843	4232	3306,812298	4499,602175	117,201695	50,6601989	1117,47935	63,5966036	213,229877	17,306940
15	Santa Cruz de	544,1850133	5621,338676	218,1	435,94	1035	1439	2493	2368,708321	3252,630356	91,2133934	43,0399681	774,824912	47,3507221	144,510269	12,304262
16	Palencia	745,1671635	1199,5701	27,1	55,144	174,5	304,5	638,3	482,2366213	717,3334784	17,6522276	7,02077763	172,819174	10,0278158	33,1555009	2,696898
17	Zamora	855,9678734	1476,878409	26,28	56,521	182,7	317,1	894,2	579,2406495	897,6377597	20,7454419	7,76190104	217,117275	12,0841206	42,475007	3,3768766
18	Barcelona	569,0652422	32235,1502	1157	2272,5	5037	7757	16011	13862,43538	18372,71483	511,647316	233,825612	4515,37935	263,977047	855,945174	71,454510
19	Murcia	504,4349021	7535,742913	320	585,35	1333	1828	3469	3167,538535	4368,204379	124,050677	57,1093739	1040,0012	62,8761172	195,935437	16,736602
20	Soria	751,7621177	666,3318707	14,89	30,514	92,8	143,2	384,9	257,2834917	409,048379	9,62975451	3,68848824	96,9016717	5,52128427	18,9789278	1,5335758

Figure 14. Example of Downloaded Incidence Dataset.

3.1.1.2 Risk Factors (Tobacco) Dataset

The Risk Factors webpage (Fig. 15) also provided the option to download the dataset in csv format specifying the years. Data from 2019 to 2022 (inclusive) was chosen. The dataset included columns like province, tobacco consumption by territorial unit (which will also serve as a check on the validity of the models), population factor and number of people by age and sex. An example of the generated csv can be seen in Fig 16.

Figure 15. Risk Factors (Tobacco) page.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Province	Tobacco Consumption	Population Factor	15-39	40-49	50-59	60-69	70+	Women Factor	Men Factor	Daily Smokers	Non-Smokers		
2	Araba/Álava	183148,1477	329967,6658	65494,5224	38050,5165	25786,2373	19879,4891	33937,3824	102377,5456	80770,6021	54838,31036	128309,837		
3	Albacete	241487,8053	394516,6556	96760,7347	38291,9397	40315,7767	30904,8769	35214,4772	134507,2477	106980,558	79694,84178	161792,963		
4	Alicante/Alac	1215256,666	1888741,102	448592,013	229791,878	186077,249	166970,854	183824,674	670391,6424	544865,024	386859,9493	828396,717		
5	Almería	433422,0905	709314,7678	190659,097	84396,5726	64018,7402	40912,6065	53435,0745	229122,216	204299,874	137368,1391	296053,951		
6	Ávila	95981,83343	163210,2175	31616,5616	16092,1929	16627,9828	11731,3536	19913,7425	56057,55568	39924,2778	30359,47203	65622,3614		
7	Badajoz	396599,6115	683027,2704	165870,39	63621,43	65621,3065	40242,6206	61243,8648	232223,4652	164376,146	136619,7674	259979,844		
8	Balears, Illes	704574,5953	1197160,153	289789,63	138553,154	116952,99	69080,7302	90198,0905	378033,0403	326541,555	229389,2293	475185,366		
9	Barcelona	3417279,964	5609047,166	1288592,72	700359,406	552692,308	385966,1	489669,427	1984962,785	1432317,18	1043103,213	2374176,75		
10	Burgos	216068,9079	365965,3438	72688,3987	39028,2283	37300,7831	26984,148	40067,3497	125223,1669	90845,741	69599,00715	146469,901		
11	Cáceres	233637,4939	408100,9731	88410,5475	35393,4659	40798,9362	26263,212	42771,3323	138192,0331	95445,4608	79519,82798	154117,666		
12	Cádiz	756753,2242	1245837,716	305733,226	145504,103	121496,978	79406,4589	104612,459	418485,4977	338267,727	234677,3571	522075,867		
13	Castellón/Ca	377598,1204	586853,3474	138407,717	74321,3811	58693,9664	49416,9451	56758,1111	207906,7033	169691,417	120067,3101	257530,81		
14	Ciudad Real	309091,0055	504282,1099	122612,896	46275,1557	51262,5689	41211,0294	47729,3559	174194,2979	134896,708	101287,8011	207803,204		
15	Córdoba	486055,2552	798628,1501	189734,631	83467,4533	78580,4855	50788,6351	83484,0498	272760,229	213295,026	145510,4254	340544,83		
16	Coruña, A	699825,1542	1171216,607	215340,759	148330,737	105718,314	90209,1531	140226,19	425177,8303	274647,324	174222,1431	525603,011		
17	Cuenca	123157,2269	202643,0137	46140,5517	17743,2506	20720,0857	16474,9316	22078,4072	68941,23245	54215,9945	39511,81428	83645,4126		
18	Girona	457881,3529	754659,509	175882,714	92660,6509	76283,5846	52418,3002	60636,1027	259459,7099	198421,643	142188,417	315692,936		
19	Granada	564983,2394	928323,7188	227564,664	101334,573	89902,7588	58244,4694	87936,7746	313964,0479	251019,192	172145,1776	392838,062		
20	Guadalajara	156690,4983	256323,0814	64955,6615	29068,8418	25218,8493	18577,3423	18869,8034	85128,77997	71561,7183	52957,8457	103732,653		

Factors2019

Figure 16. Example of Downloaded Risk Factors (Tobacco) Dataset.

3.1.1.3 Pollutants Dataset

To tailor the data to specific research needs, several filters are available in the download service, including country, pollutant type, dataset category (Historical, E1a, or E2a), data resolution (hourly or daily), and the desired temporal coverage (Fig. 17).

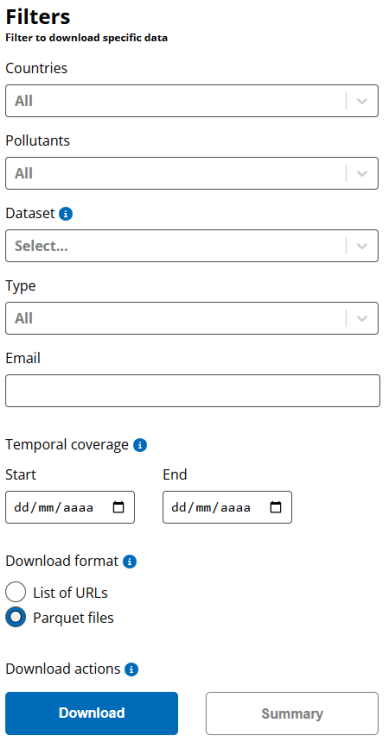


Figure 17. EEA Air Quality Download Service.

The selected time series according to the filters are downloaded as zipped Parquet files, with each file containing the complete time series for a specific monitoring location (Sampling Point). The filename corresponds to the unique identifier (localId) of the respective Sampling Point. Since large date ranges were selected, the download did not include the data itself but instead provided links to the corresponding Parquet files. These files were then accessed and extracted using Python.

From the page, data was extracted for seven pollutants: Arsenic (As), Benzo[a]pyrene (BaP), Cadmium (Cd), Nickel (Ni), Lead (Pb), Particulate Matter ≤10 micrometers in diameter (PM10), and Particulate Matter ≤2.5 micrometers in diameter (PM2.5), as they were considered representative indicators related to lung cancer incidence according to the research and findings from the state of art.

For each pollutant, the dataset included information on the province, municipality, monitoring station, pollutant type (magnitude), sampling point, year, month, and daily measurements from the 1st to the 31st.

3.2 Data Preparation

3.2.1 Initial Data Cleaning and Integration

After collection, the data was divided into two separate datasets to facilitate processing. One dataset includes risk factors and cancer incidence by province, while the other contains air quality data for six selected pollutants, also organized by province. Both datasets cover the period from 2019 to 2022, inclusive. Data cleaning was first performed on each individual dataset. After

merging them, additional preprocessing techniques were applied to obtain the final consolidated dataset.

3.2.1.1 Cancer Dataset

First, it was verified that all columns included in the datasets were relevant to the objectives of the project. Any columns that did not contribute meaningful information to the analysis were excluded. Additionally, to maintain data integrity and avoid multicollinearity, variables that were calculated as combinations or sums of other features were removed, ensuring that each input variable remained independent and did not distort the results.

From the cancer incidence dataset, the selected columns were: Province (necessary for integration with the risk factors dataset), Women Incidence, Men Incidence, and Lung Cancer cases. In the case of the risk factors dataset, the chosen columns included: Province (for alignment with the incidence data), Tobacco Consumption, Women Factor, and Men Factor.

Additionally, data types were reviewed. Several numerical variables had been incorrectly imported as strings, which caused issues during data processing and analysis. To correct this, formatting inconsistencies were addressed, including replacing commas with periods where appropriate, to ensure proper numeric interpretation. Finally, the resulting dataset was organized in a consistent manner by sorting them alphabetically by province and chronologically by year and month, facilitating easier integration and analysis in later stages of the project (Fig. 18).

	Province	Women Incidence	Men Incidence	Lung	Tobacco Consumption	Women Factor	Men Factor	Year
112	Albacete	949.265822	1373.486954	237.500169	2.414878e+05	134507.247666	106980.557603	2019
60	Albacete	960.537154	1400.983145	244.105062	2.561948e+05	139756.552750	116438.265834	2020
8	Albacete	962.087540	1408.153048	246.264484	2.552340e+05	139241.613489	115992.348748	2021
164	Albacete	970.517172	1426.669210	250.324926	7.323479e+04	29427.222817	43807.565616	2022
115	Alicante/Alacant	4501.621070	6574.582940	1155.149254	1.215257e+06	670391.642429	544865.024003	2019
...	...	...	...	...	...	...	...	...
200	Zaragoza	2610.456745	3638.566971	645.334446	1.696831e+05	62929.955533	106753.110326	2022
111	Ávila	462.699558	732.714792	123.152846	9.598183e+04	56057.555676	39924.277754	2019
59	Ávila	464.835164	742.824954	124.809384	1.013412e+05	53426.308738	47914.872427	2020
7	Ávila	466.278103	749.440601	126.516346	1.017439e+05	53610.057865	48133.806024	2021
163	Ávila	470.945118	759.731008	128.542231	2.831690e+04	11450.726615	16866.173776	2022

208 rows × 8 columns

Figure 18. Final Cancer Dataset.

3.2.1.2 Pollutants Dataset

From the air pollutants dataset, only the records corresponding to province, year, month, and daily concentration values (from the 1st to the 31st) were selected for the years 2019 to 2022, inclusive, to align with the previously collected cancer incidence data.

Columns that did not contribute meaningful information to the analysis were removed from the dataset and variables that were calculated from or closely related to other variables were excluded.

In the original dataset, provinces were represented by numerical codes. To make the data interpretable and compatible with other datasets, these codes were mapped to their corresponding province names using the official reference provided by the National Statistics Institute.

The dataset also contained missing values in the pollutant concentration measurements, with several rows showing incomplete data. Since the cause of these gaps was unknown and to avoid reducing the dataset size, row deletion was ruled out. Instead, missing values were imputed using the K-Nearest Neighbors (KNN) imputation method from the scikit-learn library.

This technique estimates missing values based on the similarity to other samples and was selected because it preserves the integrity of the dataset without negatively impacting the performance of subsequent machine learning models.

The resulting Dataset can be seen in Fig. 19

	Province	Year	PM2.5	As	Cd	PM10	BaP	Ni	Pb
0	A Coruña/A	2019	14.942448	1.610701	1.241326	7.041560	0.255765	7.532996	0.051873
1	A Coruña/A	2020	12.296623	1.216923	0.936570	7.327442	0.148533	5.728319	0.038758
2	A Coruña/A	2021	12.058553	0.689311	0.482194	9.098441	0.111049	3.174727	0.035912
3	A Coruña/A	2022	8.951011	0.388326	0.333821	14.832757	0.114946	1.991365	0.001899
4	Albacete	2019	11.019928	0.532751	0.047784	20.212178	0.060483	0.919113	0.001535
...	...	...	...	...	...	...	...	...	...
203	Zaragoza	2022	11.192133	0.951410	0.922580	20.117208	0.160270	6.272242	0.004818
204	Ávila	2019	11.019928	0.612154	0.246606	20.212178	0.152557	2.704280	0.010244
205	Ávila	2020	9.994890	0.546008	0.455209	18.475109	0.133354	1.999153	0.008230
206	Ávila	2021	9.792402	0.309084	0.122868	13.422774	0.133444	1.783297	0.002234
207	Ávila	2022	10.184902	0.485884	0.178930	14.979894	0.160270	2.213033	0.006811

208 rows × 9 columns

Figure 19. Final Pollutants Dataset.

3.2.1.3 Final Dataset Integration

To merge the cancer dataset with the pollutants dataset, the common fields used were province and year. However, inconsistencies in province names were identified, caused by typographical errors, variations in spelling, and the presence of accents or special characters. These differences initially prevented a seamless match between the two datasets.

To resolve this, province names were carefully reviewed and standardized to ensure consistency across both sources. Once aligned, the datasets were successfully merged into a single, unified dataset. After merging, the resulting data was reviewed to verify that there were no duplicate or redundant rows, confirming the accuracy and integrity of each record obtaining the dataset in Fig. 20.

	Province	Women Incidence	Men Incidence	Lung	Tobacco Consumption	Women Factor	Men Factor	Year	PM2.5	As	Cd	PM10	BaP	Ni	Pb
0	Albacete	949.265822	1373.486954	237.500169	2.414878e+05	134507.247666	106980.557603	2019	11.019928	0.532751	0.047784	20.212178	0.060483	0.919113	0.001535
1	Albacete	960.537154	1400.983145	244.105062	2.561948e+05	139756.552750	116438.265834	2020	9.994890	0.419448	0.035514	18.475109	0.036403	0.423344	0.000862
2	Albacete	962.087540	1408.153048	246.264484	2.552340e+05	139241.613489	115992.348748	2021	9.792402	0.420495	0.049551	18.686462	0.015470	0.726687	0.001912
3	Albacete	970.517172	1426.669210	250.324926	7.323479e+04	29427.222817	43807.565616	2022	9.826559	0.143493	0.061479	30.560495	0.101108	0.697396	0.000988
4	Alicante/Alacant	4501.621070	6574.582940	1155.149254	1.215257e+06	670391.642429	544865.024003	2019	11.703945	0.306421	0.078084	18.629939	0.099793	2.481382	0.011714
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
203	Zaragoza	2610.456745	3638.566971	645.334446	1.696831e+05	62929.955533	106753.110326	2022	11.192133	0.951410	0.922580	20.117208	0.160270	6.272242	0.004818
204	Ávila	462.699558	732.714792	123.152846	9.598183e+04	56057.555676	39924.277754	2019	11.019928	0.612154	0.246606	20.212178	0.152557	2.704280	0.010244
205	Ávila	464.835164	742.824954	124.809384	1.013412e+05	53426.308738	47914.872427	2020	9.994890	0.546008	0.455209	18.475109	0.133354	1.999153	0.008230
206	Ávila	466.278103	749.440601	126.516346	1.017439e+05	53610.057865	48133.806024	2021	9.792402	0.309084	0.122868	13.422774	0.133444	1.783297	0.002234
207	Ávila	470.945118	759.731008	128.542231	2.831690e+04	11450.726615	16866.173776	2022	10.184902	0.485884	0.178930	14.979894	0.160270	2.213033	0.006811

208 rows × 15 columns

Figure 20. Cancer-Pollutants Dataset.

3.2.2 Data Exploration

Data exploration, also known as exploratory data analysis (EDA), involves examining and understanding a dataset using a variety of techniques and tools. The main objective is to uncover patterns, relationships, anomalies, and to generate hypotheses that guide further analysis.

After obtaining the merged dataset, it was essential to examine the data. This involved reviewing the structure and relationships between variables to gain a solid understanding of the dataset's behavior. Identifying underlying trends and associations at this stage was crucial for guiding the selection of features and refining the modeling strategy.

In this study, data exploration was carried out through the sequential application of several key methods: Descriptive Statistics, Data Visualization, Correlation Analysis and Contingency Tables.

3.2.2.1 Descriptive Statistics

Descriptive statistics were used to summarize the key characteristics of the dataset by calculating measures of central tendency and dispersion like mean, median, mode, standard deviation and extreme values (Fig. 21). The analysis showed that lung cancer incidence tends to be higher in men than in women, with significant variation observed across provinces. Several variables, including tobacco consumption and population-related factors, exhibited right-skewed distributions, suggesting the presence of high outlier values.

Among the environmental pollutants, PM2.5 and PM10 demonstrated moderate variability, whereas elements like Pb and Cd showed relatively little fluctuation. Notably, Ni and PM10 presented higher standard deviations, indicating they may hold stronger predictive potential. These observations highlight the importance of applying appropriate scaling or transformation techniques to normalize skewed variables before proceeding with modeling.

	mean	median	mode	std_dev	min	max
Women Incidence	2289.602610	1562.500749	140.985501	2914.207875	140.985501	1.684317e+04
Men Incidence	3165.295663	2186.664083	197.073746	3782.383357	197.073746	2.151322e+04
Lung	563.636046	388.768952	35.138556	683.871892	35.138556	3.925086e+03
Tobacco Consumption	471098.846138	269061.432628	10414.041020	659743.547208	10414.041020	4.152105e+06
Women Factor	256755.297694	142646.516666	2989.953060	372457.421873	2989.953060	2.352812e+06
Men Factor	214343.548444	129028.959841	7146.008609	288292.121703	7146.008609	1.799293e+06
Year	2020.500000	2020.500000	2019.000000	1.120731	2019.000000	2.022000e+03
PM2.5	9.652318	9.989194	9.792402	1.895164	3.964909	1.494245e+01
As	0.497239	0.460465	0.612154	0.303980	0.130000	1.943280e+00
Cd	0.262615	0.178930	0.246606	0.280520	0.035514	1.427312e+00
PM10	18.318589	18.658200	20.212178	4.708995	7.041560	3.952833e+01
BaP	0.137582	0.133354	0.152557	0.079578	0.007135	5.909731e-01
Ni	2.334576	2.031290	2.704280	1.378533	0.423344	8.353978e+00
Pb	0.008805	0.005574	0.010244	0.012634	0.000862	9.491311e-02

Figure 21. Descriptive Statistics of the Dataset.

3.2.2.2 Data Visualization

Visual representations of the data were created to facilitate the identification of patterns, trends, and relationships, underlying structures and potential anomalies.

A scatter plot was generated to explore the relationship between air pollutant levels and lung cancer incidence. Figure 22 illustrates these associations across various pollutants. A moderate positive trend is noticeable for PM2.5 and PM10, where higher concentrations generally align with increased cancer cases.

In contrast, the patterns for arsenic (As), cadmium (Cd), benzo[a]pyrene (BaP), nickel (Ni), and lead (Pb) appear more dispersed, showing weaker but potentially meaningful relationships.

Overall, Figure 22 shows that PM2.5 and PM10 exhibit a more direct correlation with cancer incidence, which aligns with previous research findings [17]. In contrast, the other pollutants require more in-depth statistical analysis to determine their potential impact

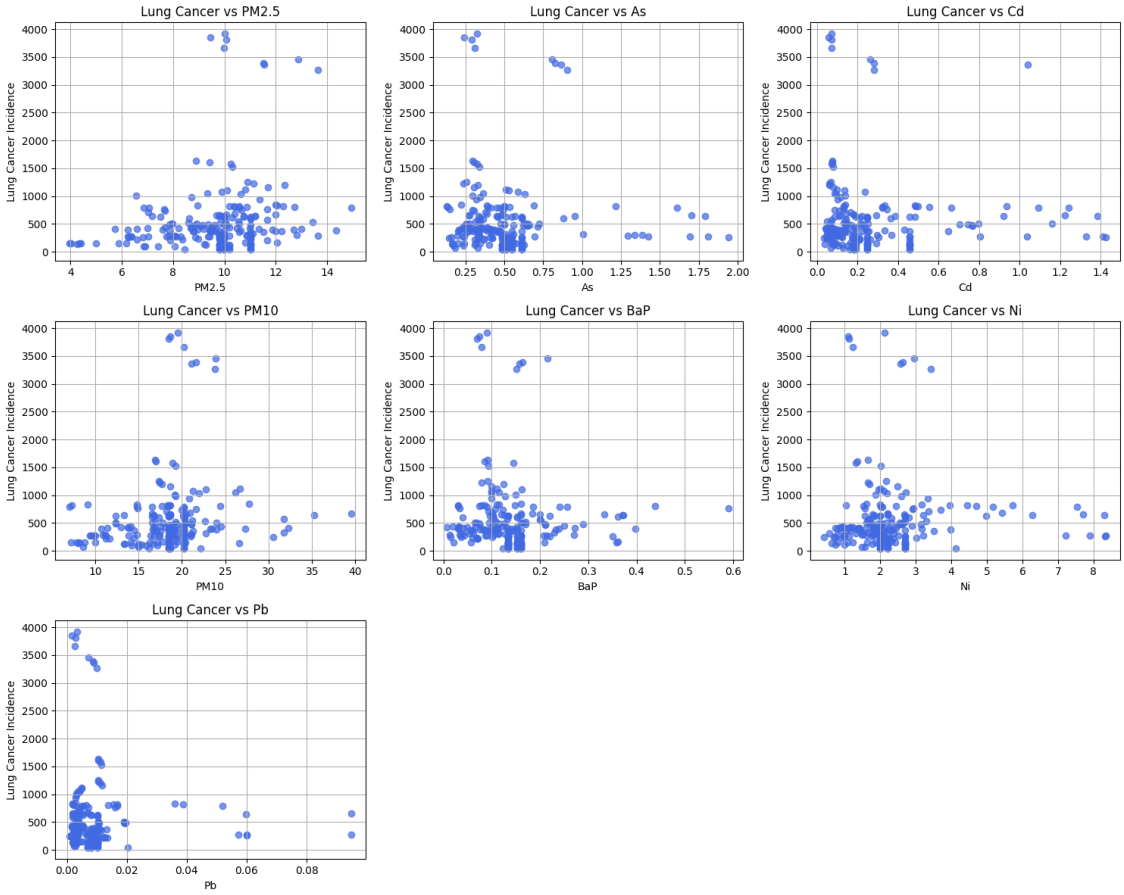


Figure 22. Scatterplot of pollutants vs Lung Cancer

The histograms in Fig. 23 reveal that most variables in the dataset, particularly those related to cancer incidence and heavy metal concentrations, exhibit right-skewed distributions with visible outliers. PM2.5 and PM10, however, show more symmetric distributions, which may facilitate their interpretation in linear models.

Given the presence of extreme values, it is advisable to apply techniques that limit their influence in order to enhance the reliability of statistical analysis and model performance.

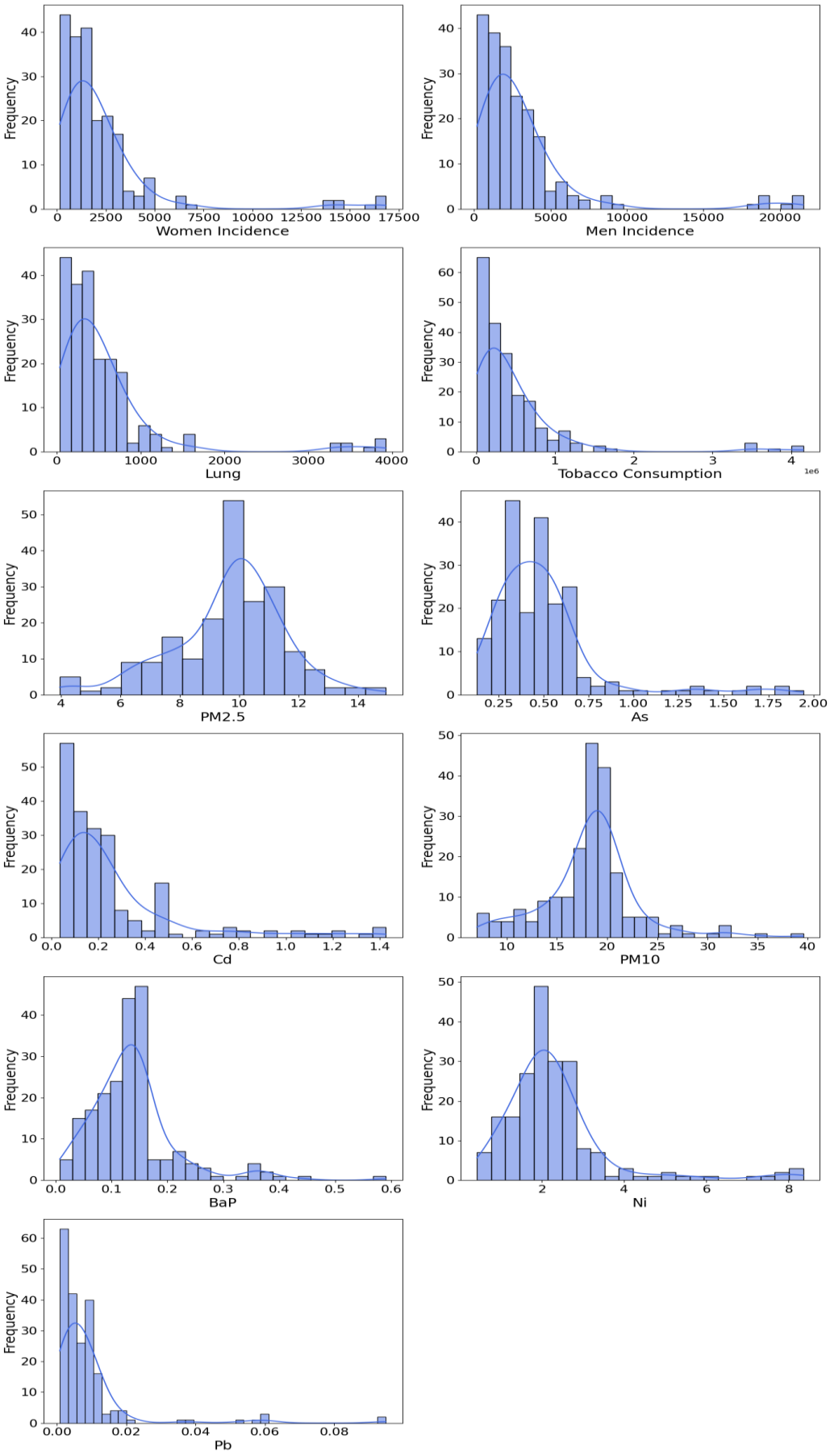
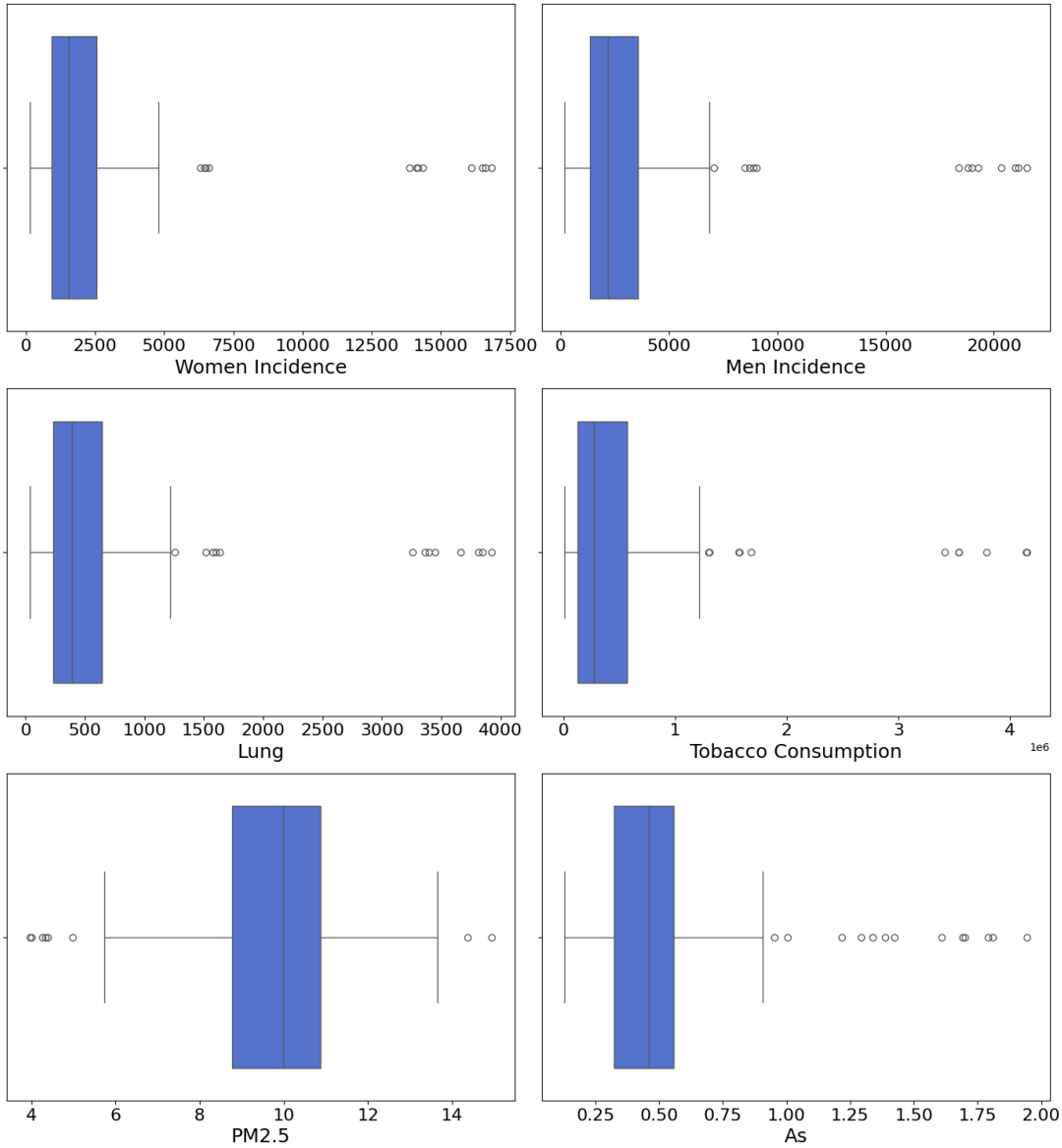


Figure 23. Histograms of the Parameters.

The boxplots in Fig. 24 confirm the presence of significant outliers across most variables, particularly in cancer incidence rates and heavy metals such as cadmium (Cd) and lead (Pb). While variables like PM2.5 display more compact and balanced distributions, others like cadmium (Cd), benzo[a]pyrene (BaP) and nickel (Ni), exhibit pronounced right-skewness and high-leverage points.

These findings support the need for applying appropriate outlier-handling techniques to minimize their impact and improve the robustness of subsequent statistical modeling.



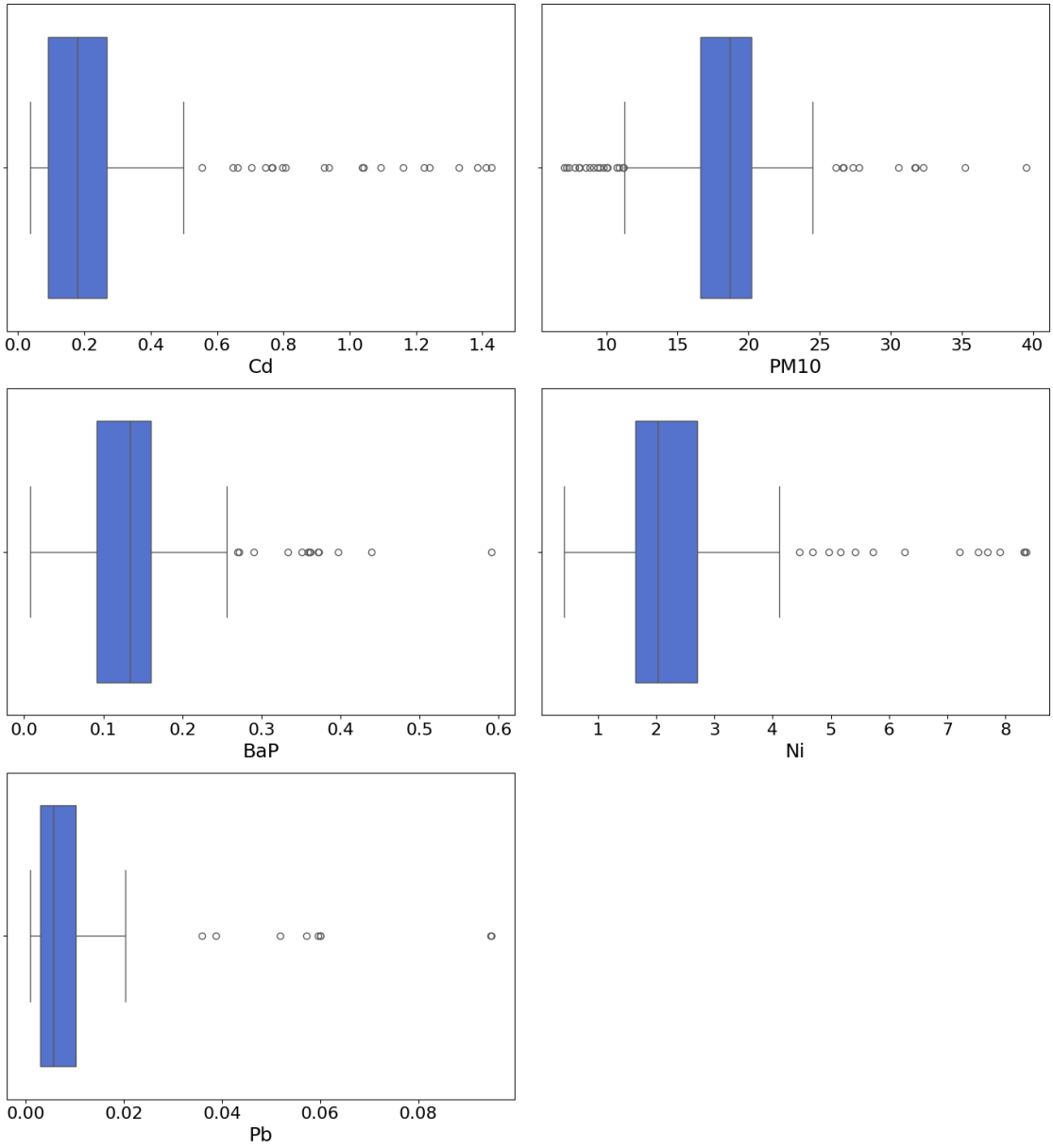


Figure 24. Boxplots of the Parameters.

Fig. 25 displays a bar chart of the average lung cancer incidence by province, based on data collected over multiple years. The chart clearly illustrates regional disparities in incidence rates. Provinces such as Barcelona, Madrid, and Valencia/València report the highest average numbers, each exceeding 1,200 cases, while provinces like Melilla, Ceuta, and Soria show much lower averages, often below 100 cases.

These differences may reflect variations in environmental exposures, population density, smoking prevalence, and access to healthcare services, including screening and early detection programs. The visual representation emphasizes geographic variability and suggests that public health interventions should be adapted to the specific needs of each region.

The higher incidence observed in more urbanized and industrial provinces could indicate a potential association with air pollution, consistent with findings from previous epidemiological studies.

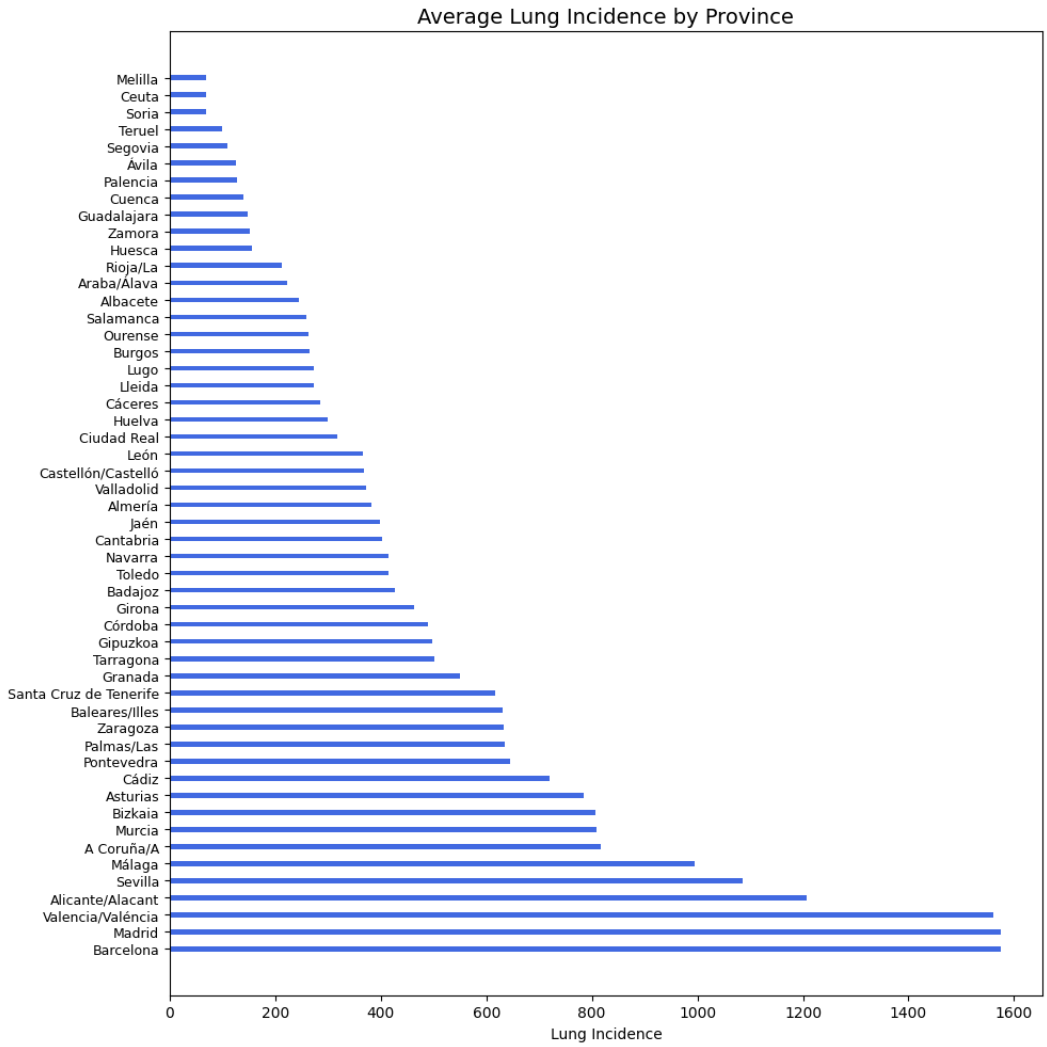


Figure 25. Incidence by Province.

3.2.2.3 Correlation analysis

To examine potential correlations between variables, the Spearman correlation method was used. It was chosen over the Pearson correlation because most of the variables in the dataset are not normally distributed and do not necessarily have linear relationships. Spearman correlation is more appropriate in these cases, as it measures the strength and direction of monotonic relationships without assuming normality.

The Spearman correlation matrix was computed using air pollutants, tobacco consumption, and province-level binary indicators. Given the presence of 52 provinces, displaying all of them in a heatmap would reduce interpretability. Therefore, only the top five provinces with the highest lung cancer incidence identified in the previous bar chart (Fig. 25) were selected to provide a clearer

view of the correlations (Fig. 26). As expected, tobacco consumption shows a strong positive correlation with lung cancer incidence, consistent with extensive research that identifies smoking as the primary risk factor for this disease.

Among pollutants, PM2.5 and PM10 show weak-to-moderate positive correlations with lung cancer, which is consistent with literature linking fine particulate matter to respiratory and carcinogenic effects [15]. Other pollutants such as Nickel, Lead, and Cadmium also show some correlation, though at lower levels, possibly reflecting impact may be more indirect or context dependent.

The province-level binary indicators were included to look for geographic variation. These variables show low-to-moderate correlations with lung incidence, likely capturing localized differences not fully explained by pollutants or smoking rates alone but also related to other factors such as industrial activity, urban density, or healthcare access.

It's important to note that while the overall correlation between pollutants and lung incidence remains modest, this does not necessarily indicate that environmental factors lack predictive value. Lung cancer is multifactorial, and pollution-related cases may be diluted when averaged across entire populations where smoking remains the dominant risk factor. Moreover, only a limited subset of pollutants was analyzed, whereas real-world exposure involves complex mixtures.

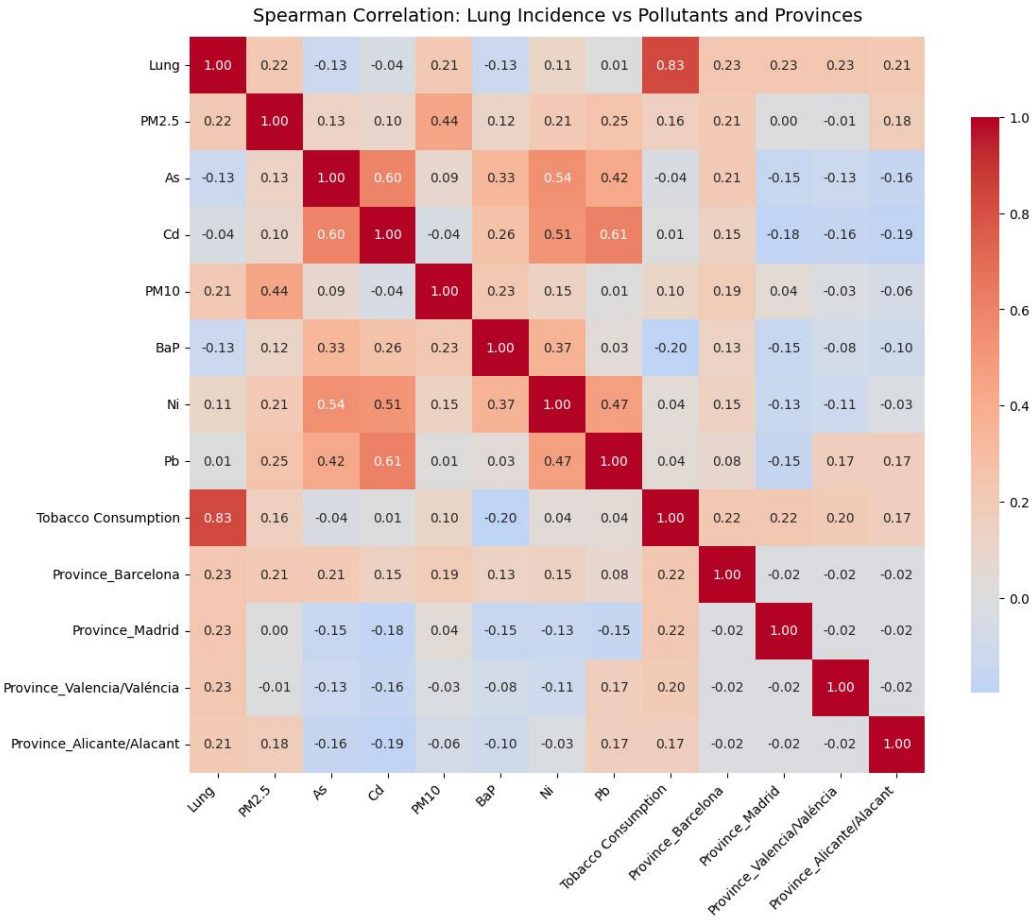


Figure 26. Spearman Matrix of Correlation.

3.2.3 Data Refinement and Transformation

The exploratory data analysis (EDA) revealed significant differences in the scale of various features. These variations in magnitude could cause some machine learning algorithms to give undue weight to certain variables, even if they are not inherently more important. To mitigate this issue and ensure that all features contribute equally during model training, the dataset was standardized using the StandardScaler from the scikit-learn library. This transformation normalized the data, bringing all numerical variables to a comparable scale and improving the reliability of the models.

Additionally, data visualization highlighted the need to address the presence of outliers. To manage them effectively, three main strategies were considered:

- **Remove the affected data:** This would result in a reduction of the overall dataset, potentially weakening the analysis. Therefore, this option was ruled out.

- **Keep the dataset unchanged and proceed with modeling:** While this approach preserves all data, it risks compromising model performance, as some algorithms may be sensitive to outliers. For this reason, it was also discarded.

- **Apply outlier treatment methods:** This option was selected to address the issue without losing valuable data, ensuring more reliable results from the machine learning models.

For this study, the Winsorization method was applied to handle outliers. This statistical technique reduces the influence of extreme values by replacing them with the nearest values within a specified percentile range, from Fig. 27 to Fig. 30 the comparison between the values before and after Windsorization can be seen with more detail for some of the parameters. By applying Winsorization, the impact of outliers on the analysis is minimized, resulting in more stable and reliable outcomes in the statistical and machine learning models.

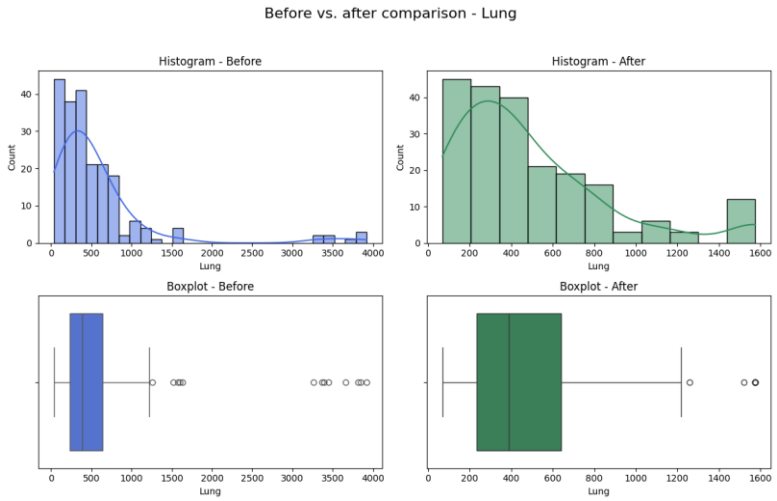


Figure 27. Comparison before/after winsorization Lung incidence.

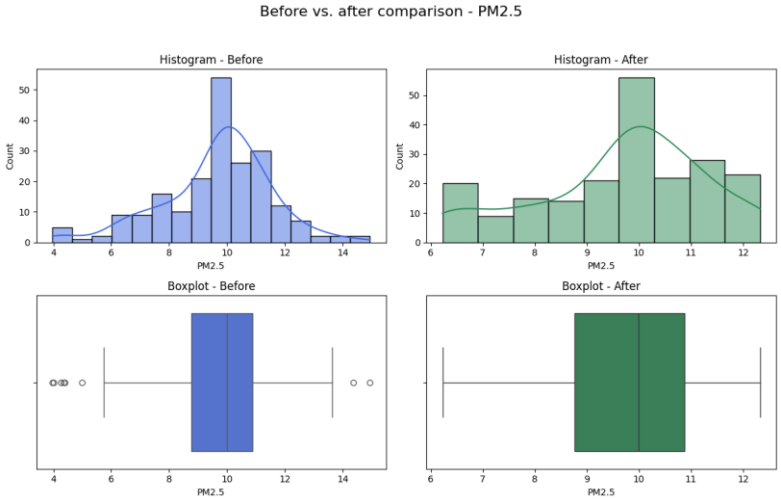


Figure 28. Comparison before/after winsorization PM2.5

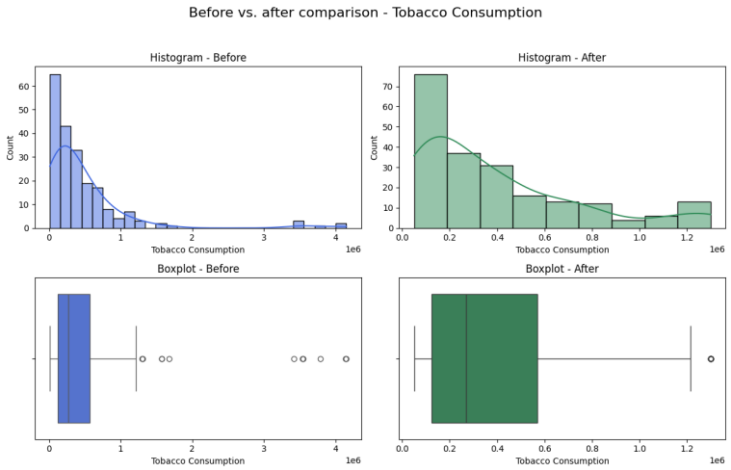


Figure 29. Comparison before/after winsorization Tobacco consumption.

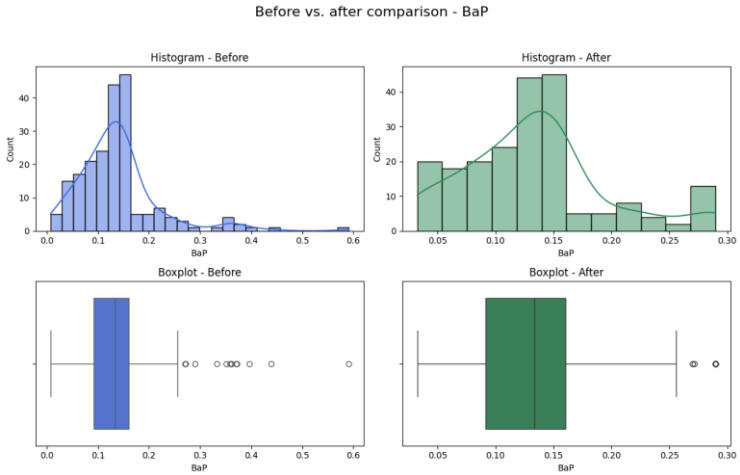


Figure 30. Comparison before/after winsorization BaP.

Additionally, one-hot encoding was applied to the categorical variable Province to convert it into a numerical format suitable for machine learning models. This step was necessary because the selected algorithms require numerical input and cannot directly process categorical data. To avoid multicollinearity, one of

the resulting dummy variables was dropped; specifically, Melilla was excluded due to its notably low cancer incidence data, minimizing redundancy while preserving model integrity.

The final dataset obtained after all preprocessing steps, including cleaning, merging, and outlier treatment, is presented below in Fig. 31:

	Women Incidence	Men Incidence	Lung	Tobacco Consumption	Women Factor	Men Factor	Year	PM2.5	As	Cd	...	Province_Sevilla	Province_Soria	Province_Tarragona	Province_Teruel	Province_Toledo	Province_Valencia/València
0	-0.648727	-0.656584	-0.661179	-0.443951	-0.329012	-0.373309	-1.341641	0.827748	0.221921	-0.854648	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
1	-0.641430	-0.643382	-0.643669	-0.401894	-0.314884	-0.340424	-0.447214	0.204084	-0.262366	-0.854648	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
2	-0.640426	-0.639940	-0.637945	-0.404642	-0.316270	-0.341974	0.447214	0.080885	-0.257891	-0.854648	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
3	-0.634968	-0.631050	-0.627181	-0.925092	-0.611819	-0.592966	1.341641	0.101667	-1.196169	-0.809467	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
4	1.651204	1.840501	1.771484	2.340664	1.113239	1.149247	-1.341641	1.243924	-0.745473	-0.737621	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
203	0.426792	0.430898	0.419979	-0.649286	-0.521651	-0.374100	1.341641	0.932522	2.011385	2.916212	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
204	-0.963749	-0.964223	-0.964310	-0.860044	-0.540147	-0.606468	-1.341641	0.827748	0.561313	-0.008486	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
205	-0.962366	-0.959369	-0.959919	-0.844718	-0.547229	-0.578685	-0.447214	0.204084	0.278588	0.894063	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
206	-0.961432	-0.956193	-0.955394	-0.843566	-0.546735	-0.577923	0.447214	0.080885	-0.734089	-0.543857	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028
207	-0.958410	-0.951252	-0.950023	-0.992316	-0.660200	-0.686643	1.341641	0.319693	0.021602	-0.301298	...	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028	-0.140028

208 rows × 65 columns

Figure 31. Final Dataset.

3.2.4 Feature Selection

The independent variables or predictors are a key component in constructing a machine learning model. In this study, multiple independent variables were used to develop a predictive model that explores the correlations between environmental factors, chemical exposures, and the incidence rate of lung cancer, which serves as the dependent (Y) variable, across Spanish provinces.

Previous research has indicated that air pollution may contribute to a slight increase in the risk of lung cancer [15, 16]. Based on this, seven air pollutants were selected as independent variables for the analysis, along with tobacco use, a well-established risk factor.

Finally, following the data cleaning and preparation process, the dataset included a total of 64 independent variables. These consisted of 7 industrial air pollutants, 3 tobacco-related factors by province, 2 cancer related incidence by sex, the year and the 51 Spanish provinces represented as columns through one-hot encoding (minus one to avoid multicollinearity). These independent variables were used to build predictive models of lung cancer incidence rate (the dependent variable), aiming to uncover potential correlations among environmental factors, chemical exposures, and the incidence of lung cancer across provinces in Spain.

Table 1 shows the amounts corresponding to each variable.

INDEPENDENT VARIABLE	**QUANTITY**
Industrial air pollutants	7
Factors / habits	3
Cancer	2
Provinces	51
Year	1
Total Independent Variables	***64***

Table 1. Quantity of Independent Variables.

The following table (Table 2) provides a more detailed breakdown of the variables used in the study.

FACTOR	VARIABLE	DESCRIPCION
Air Pollutants	1. Arsenic (As)	Average concentration of As µg/m³
	2. Benzo (a) Pyrene (BaP)	Average concentration of BaP µg/m³
	3. Cadmium (Cd)	Average concentration of Cd µg/m³
	4. Nickel (Ni)	Average concentration of Ni µg/m³
	5. Lead (Pb)	Average concentration of Pb µg/m³
	6. Particulate Matter PM10 (PM10)	Average concentration of PM10 µg/m³
	7. Particulate Matter PM25(PM25)	Average concentration of PM25 µg/m³
Tobacco Use	8. Tobacco consumption per region	Consumption of tobacco per region
	9. Tobacco consumption Men	Consumption of tobacco per sex (Men)
	10. Tobacco consumption Women	Consumption of tobacco per sex (Women)
Risk Factors Group	11. Cancer Men	General cancer Incidence per sex (Men)
	12. Cancer Women	General cancer Incidence per sex (Women)
51 Region of Spain: From ARABA/ALABA TO CEUTA	13. Araba/Alava Province	Lung cancer incidence in Araba/Alava province of Spain per year
	to	to
	63. Ceuta Province	Lung cancer incidence in Ceuta province of Spain per year
Time	64. Year	Year of the cancer and pollutants information
Dependent Variable	65. Incidence	Lung cancer incidence in provinces of Spain

Table 2. Description of the model variables.

3.3 Modeling

After completing the stages of data collection, feature selection, cleaning, and preparation, which included handling missing values through imputation or removal, correcting outliers, addressing inconsistencies and errors, creating and deleting features, scaling, normalization, and encoding categorical variables, the dataset was fully assembled and ready for analysis. At this point, the machine learning models to be used and compared were defined based on the problem type, which is regression, and the characteristics of the data.

For this study, three well-established machine learning algorithms were selected: Linear Regression, Random Forest, and XGBoost. These models were

trained, tested, and validated to predict lung cancer incidence rates across provinces in Spain.

The first model selected was Linear Regression, due to its simplicity and interpretability. Its coefficients represent the expected change in the dependent variable for a one-unit change in each independent variable, offering valuable insights into the relationships between predictors and lung cancer incidence.

Additionally, the Random Forest model was selected. This algorithm is known for offering a strong balance between accuracy, robustness, and ease of use, making it a reliable choice for various machine learning tasks. Its ability to handle complex interactions and reduce overfitting contributes to its effectiveness in predictive modeling.

The third model chosen was XGBoost. This algorithm is widely recognized for its high performance, combining accuracy, speed, and robustness. Due to its efficient handling of large datasets and strong predictive power, XGBoost is considered a powerful tool for a broad range of machine learning applications.

The next step in the process is data splitting, which involves dividing the dataset into training and test subsets. The training set is used to teach the model by allowing it to learn patterns and relationships within the data, while the test set is reserved for evaluating the model's performance on unseen data. In this case, the dataset was split with 80 percent allocated to training and 20 percent to testing.

To further ensure a reliable assessment of model performance, five-fold cross validation was applied. This technique divides the training data into five parts, using four for training and one for validation in each iteration, cycling through all combinations. This helps provide a more robust and unbiased estimate of how the model will generalize to new, unseen data.

With the data prepared, the selected models are then trained using the training set. Initially, all three models, Linear Regression, Random Forest, and XGBoost, are trained using their default hyperparameters to establish a baseline performance before any tuning is performed.

3.3.1 Baseline Models

The baseline linear regression model was trained using the following hyperparameters:

- **fit_intercept:** True (the model estimates the intercept)
- **copy_X:** True (the input data matrix is copied before fitting)
- **n_jobs:** None (no parallel computation)
- **positive:** False (coefficients were not constrained to be positive)

On the test set, the model achieved a mean absolute error (MAE) of 0.0067, indicating very precise predictions on unseen data. The coefficient of determination (R^2) was 0.9998, meaning the model explained almost all the variance in the test outcomes.

Cross-validation with multiple folds resulted in a higher MAE of 0.3091 and an R^2 of 0.7853, reflecting the model's performance when evaluated on different splits of the data. This suggests some reduction in predictive accuracy when generalizing, which is common in real-world datasets.

Sample predictions demonstrate that the predicted values closely approximate the actual values, reinforcing the model's strong predictive capabilities.

```
BASELINE LINEAR REGRESSION RESULTS
----------------------------------
Hyperparameters used:
  fit_intercept: True
  copy_X: True
  n_jobs: None
  positive: False

MAE (test set): 0.0067
R²  (test set): 0.9998
MAE (cross-validation): 0.3091
R²  (cross-validation): 0.7853

Example predictions:
  Example 1: Predicted = -0.62, Actual = -0.6162264143148274
  Example 2: Predicted = 0.78, Actual = 0.784991179344639
  Example 3: Predicted = -0.45, Actual = -0.4447073755281683
  Example 4: Predicted = 1.94, Actual = 1.9401736186594385
  Example 5: Predicted = -0.29, Actual = -0.286792538830543
```

Figure 32. Linear Regression baseline.

The baseline Random Forest model was trained using default hyperparameters, which included:

- **bootstrap:** True (samples are drawn with replacement for each tree)
- **ccp_alpha:** 0.0 (no complexity pruning)
- **criterion:** squared_error (used to measure the quality of splits)
- **max_depth:** None (trees grow until all leaves are pure or contain fewer samples than min_samples_split)
- **max_features:** 1.0 (consider all features when looking for the best split)
- **min_samples_leaf:** 1 (minimum samples required to be at a leaf node)
- **min_samples_split:** 2 (minimum samples required to split an internal node)
- **n_estimators:** 100 (number of trees in the forest)
- **random_state:** 21 (seed for reproducibility)

On the test set, the model achieved a mean absolute error (MAE) of 0.0156 and an R^2 of 0.9993, indicating very accurate predictions and strong explanatory power on unseen data.

Cross-validation showed a MAE of 0.0443 and an R^2 of 0.9889, confirming consistent and robust model performance across different subsets of the data.

Sample predictions illustrate that the model's predictions are very close to actual values, demonstrating its effectiveness for the regression task.

```
RANDOM FOREST BASELINE RESULTS
--------------------------------
DEFAULT HYPERPARAMETERS:
bootstrap: True
ccp_alpha: 0.0
criterion: squared_error
max_depth: None
max_features: 1.0
max_leaf_nodes: None
max_samples: None
min_impurity_decrease: 0.0
min_samples_leaf: 1
min_samples_split: 2
min_weight_fraction_leaf: 0.0
monotonic_cst: None
n_estimators: 100
n_jobs: None
oob_score: False
random_state: 21
verbose: 0
warm_start: False
MAE (test set): 0.0156
R2  (test set): 0.9993
MAE (cross-validation): 0.0443
R2  (cross-validation): 0.9889

Example predictions:
Example 1: Predicted = -0.60, Real = -0.6162264143148274
Example 2: Predicted = 0.78, Real = 0.784991179344639
Example 3: Predicted = -0.46, Real = -0.4447073755281683
Example 4: Predicted = 1.88, Real = 1.9401736186594385
Example 5: Predicted = -0.30, Real = -0.286792538830543
```

Figure 33. Random Forest baseline.

The baseline XGBoost regression model was configured with the following hyperparameters:

- **max_depth:** 6 (maximum depth of each tree)
- **learning_rate:** 0.3 (step size shrinkage to prevent overfitting)
- **n_estimators:** 100 (number of boosting rounds)
- **subsample:** 1 (using all samples for each tree)
- **colsample_bytree:** 1 (using all features for each tree)
- **gamma:** 0 (no minimum loss reduction required to make a further partition)
- **reg_alpha:** 0 (no L1 regularization)
- **reg_lambda:** 1 (L2 regularization term)
- **objective:** 'reg:squarederror' (squared error for regression)

On the test set, the model achieved a mean absolute error (MAE) of 0.0184 and a coefficient of determination (R^2) of 0.9994, indicating excellent predictive accuracy and strong fit to the data.

Cross-validation results demonstrated a MAE of 0.0547 and an R^2 of 0.9846, confirming the model's robustness and generalizability across different data splits.

Example predictions closely match the actual values, highlighting the model's reliability for lung cancer incidence estimation based on the given features.

```
XGBOOST BASELINE RESULTS
------------------------------
Used Hyperparameters:
max_depth=6, learning_rate=0.3, n_estimators=100, subsample=1, colsample_bytree=1, gamma=0
reg_alpha=0, reg_lambda=1, objective='reg:squarederror'

MAE (test set): 0.0184
R2  (test set): 0.9994
MAE (cross-validation): 0.0547
R2  (cross-validation): 0.9846

Example predictions:
Example 1: Predicted = -0.61, Real = -0.6162264143148274
Example 2: Predicted = 0.75, Real = 0.784991179344639
Example 3: Predicted = -0.50, Real = -0.4447073755281683
Example 4: Predicted = 1.88, Real = 1.9401736186594385
Example 5: Predicted = -0.29, Real = -0.286792538830543
```

Figure 34. XGBoost baseline.

As a general view, this are the hyperparameter used and the metrics obtained (Table 3):

MODEL	DEFAULT HYPERPARAMETERS	MAE TEST	R^2 TEST	MAE CV	R^2 CV
Linear Regression	fit_intercept: True copy_X: True n_jobs: None positive: False	0.0067	0.9998	0.3091	0.7853
Random Forest	bootstrap: True ccp_alpha: 0.0 criterion: 'squared_error' max_depth: None max_features: 1.0 max_leaf_nodes: None max_samples: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 2 min_weight_fraction_leaf: 0.0 monotonic_cst: None n_estimators: 100 n_jobs: None oob_score: False random_state: 21 verbose: 0	0.0156	0.9993	0.0443	0.9889

	warm_start: False				
XGBoost	max_depth: 6 learning_rate: 0.3 n_estimators: 100 subsample: 1 colsample_bytree: 1 gamma: 0 reg_alpha: 0 reg_lambda: 1 objective: 'reg:squarederror'	0.0184	0.9994	0.0547	0.9846

Table 3. Baseline hyperparameters and metrics

3.3.2 Hyperparameter Tuning

To optimize model performance, hyperparameter tuning was conducted using tools from the scikit-learn library. For models with relatively limited hyperparameter spaces, such as Linear Regression and Random Forest, GridSearchCV was applied. This method exhaustively evaluates all possible combinations of specified parameters using k-fold cross-validation to ensure robust model selection.

In contrast, the XGBoost model involved a significantly larger and more complex set of hyperparameters. To manage computational efficiency while still ensuring effective tuning, RandomizedSearchCV was used instead. This approach randomly samples a fixed number of parameter combinations from defined distributions and evaluates them using cross-validation. Both methods incorporate cross-validation to prevent overfitting and ensure generalization of the models.

For Linear Regression the grid search explored different values (true or false) for two key hyperparameters:

- **fit_intercept**: which determines whether to calculate the intercept term in the regression model.
- **positive**: which constrains the regression coefficients to be non-negative if set to True.

Using 5-fold cross-validation, each combination of these hyperparameters was evaluated based on the negative mean absolute error. The combination that resulted in the lowest average error across the validation folds was selected as the best configuration.

After hyperparameter tuning, the linear regression model was updated with these parameters:

- **it_intercept:** False (do not fit the intercept term assumes data is centered)
- **positive:** True (force coefficients to be non-negative for easier interpretation and stability)

Performance metrics on the test set were excellent, with a very low Mean Absolute Error (MAE) of 0.0064 and an almost perfect coefficient of

determination R2 of 0.9999, indicating the model fits the test data almost perfectly.

However, the cross-validation results reveal a contrasting story: the MAE increases to 0.3958, and the R2 becomes negative (-0.4058), suggesting the model generalizes poorly on unseen data when evaluated via cross-validation. A negative R2 means the model's predictions are worse than simply predicting the mean value of the target variable.

This discrepancy could indicate overfitting to the training set or that the model assumptions with these hyperparameters do not hold well across different folds of the data. Example predictions on the test set show close alignment between predicted and actual values, reinforcing the strong fit on this particular split.

```
Best Hyperparameters:
{'fit_intercept': False, 'positive': True}

TUNED LINEAR REGRESSION RESULTS
--------------------------------
MAE (test set): 0.0064
R²  (test set): 0.9999
MAE (cross-validation): 0.3958
R²  (cross-validation): -0.4058

Example predictions with tuned model:
Example 1: Predicted = -0.62, Real = -0.6162264143148274
Example 2: Predicted = 0.78, Real = 0.784991179344639
Example 3: Predicted = -0.45, Real = -0.4447073755281683
Example 4: Predicted = 1.93, Real = 1.9401736186594385
Example 5: Predicted = -0.29, Real = -0.286792538830543
```

Figure 35. Linear Regression hyperparameter tuning.

For the Random Forest model the grid search evaluated multiple hyperparameters, including:

- **n_estimators:** Number of trees in the forest (tested values: 100, 200).
- **max_depth:** Maximum depth of each tree (tested values: None, 10, 20).
- **min_samples_split:** Minimum number of samples required to split an internal node (tested values: 2, 5).
- **min_samples_leaf**: Minimum number of samples required to be at a leaf node (tested values: 1, 2).
- **bootstrap:** Whether bootstrap samples are used when building trees (tested values: True, False).

For each combination of these hyperparameters, the model was trained and validated on different subsets of the training data. The model's performance was evaluated using the negative mean absolute error (neg_MAE**)** as the scoring metric. The hyperparameter combination that resulted in the lowest average error across the folds was selected.

The Random Forest model underwent hyperparameter tuning using 5-fold cross-validation over 48 candidate configurations, resulting in a total of 240 model fits. The best-performing hyperparameters found were:

- **bootstrap**: True
- **max_depth**: None (trees grow until pure or minimum samples reached)

- **min_samples_leaf**: 2 (minimum samples required at a leaf node)
- **min_samples_split**: 2 (minimum samples required to split a node)
- **n_estimators**: 200 (number of trees in the forest)

The tuned Random Forest model achieved strong predictive performance. On the test set, it reached a Mean Absolute Error (MAE) of 0.0172 and an R^2 of 0.9993, indicating an excellent fit and minimal prediction error. During cross-validation, it maintained solid generalization capability, with an MAE of 0.0500 and an R^2 of 0.9858, suggesting consistent performance across different data splits.

The example predictions further demonstrate that the model's outputs closely align with the actual values, reinforcing its reliability. Overall, hyperparameter tuning significantly improved the model's robustness and accuracy compared to the baseline configuration, positioning this Random Forest setup as a strong candidate for predicting lung cancer incidence based on the available features.

```
Fitting 5 folds for each of 48 candidates, totalling 240 fits

TUNED RANDOM FOREST RESULTS
------------------------------
Best Hyperparameters:
bootstrap: True
max_depth: None
min_samples_leaf: 2
min_samples_split: 2
n_estimators: 200

MAE (test set): 0.0172
R²  (test set): 0.9993
MAE (cross-validation): 0.0500
R²  (cross-validation): 0.9858

Example predictions with tuned model:
Example 1: Predicted = -0.60, Real = -0.6162264143148274
Example 2: Predicted = 0.77, Real = 0.784991179344639
Example 3: Predicted = -0.46, Real = -0.4447073755281683
Example 4: Predicted = 1.88, Real = 1.9401736186594385
Example 5: Predicted = -0.30, Real = -0.286792538830543
```

Figure 36. Random Forest hyperparameter tuning.

For the XGBoost regression model the hyperparameters tuned included:

- **n_estimators:** Number of boosting rounds (100 to 1000),
- **max_depth:** Maximum depth of each tree (3 to 8),
- **learning_rate:** Step size shrinkage to prevent overfitting (0.01 to 0.3),
- **subsample:** Fraction of samples used per tree (0.6 to 1.0),
- **colsample_bytree:** Fraction of features used per tree (0.6 to 1.0),
- **gamma:** Minimum loss reduction required to make a further partition (0 to 1),
- **reg_alpha:** L1 regularization term on weights (0 to 1),
- **reg_lambda:** L2 regularization term on weights (1 to 3).

The XGBoost model was optimized through hyperparameter tuning using 5-fold cross-validation over 50 randomly sampled candidate configurations, totaling 250 model fits. The best hyperparameters found were:

- **subsample**: 0.7 (fraction of samples used per tree to introduce randomness)
- **reg_lambda**: 3 (L2 regularization term to reduce overfitting)
- **reg_alpha**: 0 (no L1 regularization)
- **n_estimators**: 300 (number of boosting rounds)
- **max_depth**: 7 (maximum depth of each tree)
- **learning_rate**: 0.05 (step size shrinkage to prevent overfitting)
- **gamma**: 0 (minimum loss reduction for further splitting)
- **colsample_bytree**: 1.0 (use all features for each tree)

Performance metrics demonstrate that the tuned model achieved excellent predictive accuracy. On the test set, it reached a Mean Absolute Error (MAE) of 0.0168 and an R^2 of 0.9995, indicating highly accurate predictions with minimal error. Cross-validation results also showed strong generalization capabilities, with an MAE of 0.0449 and an R^2 of 0.9913, confirming the model's robustness across different data splits.

Example predictions reveal that the model's output closely matches the actual values, further supporting its reliability. Overall, the tuning process significantly improved the model's ability to capture complex patterns in the data, resulting in strong predictive performance for lung cancer incidence.

```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
Best hyperparameters found:
{'subsample': 0.7, 'reg_lambda': 3, 'reg_alpha': 0, 'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.05, 'gamma': 0, 'colsample_bytree': 1.0}

TUNED XGBOOST RESULTS
------------------------------
MAE (test set): 0.0168
R2  (test set): 0.9995
MAE (cross-validation): 0.0449
R2  (cross-validation): 0.9913

Example predictions:
Example 1: Predicted = -0.61, Real = -0.6162264143148274
Example 2: Predicted = 0.74, Real = 0.784991179344639
Example 3: Predicted = -0.48, Real = -0.4447073755281683
Example 4: Predicted = 1.89, Real = 1.9401736186594385
Example 5: Predicted = -0.28, Real = -0.286792538830543
```

Figure 37. XGBoost hyperparameter tuning.

As a general view, Table 4 shows the best parameters and results of each model:

MODEL	BEST HYPERPARAMETERS	MAE TEST	R^2 TEST	MAE CV	R^2 CV
Linear Regression	fit_intercept=False positive=True	0.0064	0.9999	0.3958	-0.4058
Random Forest	bootstrap=True max_depth=None min_samples_leaf=2 min_samples_split=2 n_estimators=200	0.0172	0.9993	0.0500	0.9858
XGBoost	subsample=0.7 reg_lambda=3 reg_alpha=0 n_estimators=300 max_depth=7 learning_rate=0.05 gamma=0 colsample_bytree=1.0	0.0168	0.9995	0.0449	0.9913

Table 4. Best hyperparameters and metrics.

3.4 Model Evaluation and Comparison

The performance of each regression model was assessed by evaluating both baseline and tuned versions using Mean Absolute Error (MAE) and R^2 metrics. These evaluations were performed on the test set as well as through 5-fold cross-validation. The results, presented in Table 5 and Figure x, reveal differences in accuracy and generalization among the models.

The comparison between the baseline and tuned linear regression models reveals a trade-off between test performance and generalization. The tuned model, configured with fit_intercept=False and positive=True, slightly improved the performance on the test set—achieving a lower mean absolute error (MAE of 0.0064 vs. 0.0067) and a marginally higher R^2 score (0.9999 vs. 0.9998).

However, its performance in cross-validation deteriorated significantly, with a higher MAE (0.3958 vs. 0.3091) and a negative R^2 value (-0.4058), suggesting that the model struggled to generalize across different data splits. This drop is likely due to the constraints imposed during tuning, particularly forcing all coefficients to be non-negative and removing the intercept, which may have limited the model's flexibility. In contrast, the baseline model, although slightly less accurate on the test set, showed more stable and reliable cross-validation results, making it the more robust choice overall.

The comparison between the baseline and tuned Random Forest models reveals that tuning led to only marginal differences in performance. The baseline model, using default hyperparameters, achieved a test set MAE of 0.0156 and an R^2 of 0.9993, with cross-validation scores of MAE = 0.0443 and R^2 = 0.9889. After tuning the model maintained the same R^2 on the test set (0.9993) but showed a slight increase in test MAE to 0.0172 and a minor decrease in cross-validation performance (MAE = 0.0500, R^2 = 0.9858).

These results suggest that the default configuration was already near-optimal, and while tuning slightly altered the bias-variance tradeoff, it did not significantly improve the model's overall predictive power. The minimal gain may also indicate that the model's performance plateaued, and further tuning would yield diminishing returns.

The comparison between the baseline and tuned versions of the XGBoost model shows that hyperparameter tuning led to meaningful improvements in performance, especially in cross-validation. The baseline model, using default settings like max_depth=6, learning_rate=0.3, and n_estimators=100, achieved an MAE of 0.0184 and an R^2 of 0.9994 on the test set, with cross-validation scores of MAE = 0.0547 and R^2 = 0.9846.

After tuning, the model was adjusted to a more regularized and conservative configuration: max_depth=7, learning_rate=0.05, n_estimators=300, subsample=0.7, and reg_lambda=3. This resulted in a slightly better test set MAE (0.0168) and R^2 (0.9995), and a more notable improvement in cross-validation performance (MAE = 0.0449, R^2 = 0.9913). These gains suggest that tuning helped reduce overfitting and improved the model's generalization. The lower learning rate and higher number of trees allowed for more gradual learning, while increased regularization enhanced stability, making the tuned model more robust across data splits.

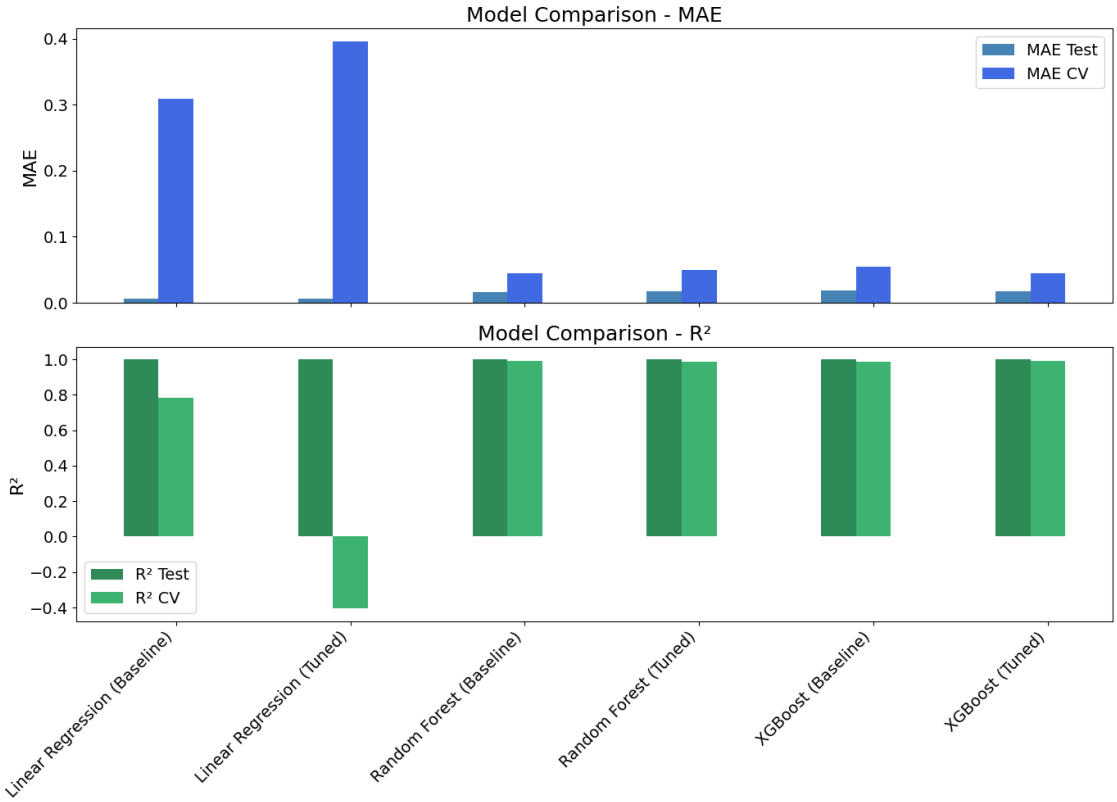


Figure 38. Model Performance Comparisons.

MODEL	VERSION	MAE TEST	R^2 TEST	MAE CV	R^2 CV
Linear Regression	Baseline	0.0067	0.9998	0.3091	0.7853
Linear Regression	Tuned	0.0064	0.9999	0.3958	-0.4058
Random Forest	Baseline	0.0156	0.9993	0.0443	0.9889
Random Forest	Tuned	0.0172	0.9993	0.0500	0.9858
XGBoost	Baseline	0.0184	0.9994	0.0547	0.9846
XGBoost	Tuned	0.0168	0.9995	0.0449	0.9913

Table 5. Model Performance Comparison.

Among all the models evaluated, the tuned XGBoost model demonstrated the best overall performance. It achieved the highest R^2 on the test set (0.9995) and the highest R^2 during cross-validation (0.9913), indicating excellent predictive power and strong generalization to unseen data. Additionally, it had one of the lowest mean absolute errors (MAE), with 0.0168 on the test set and 0.0449 in cross-validation. Compared to its baseline version, the tuned XGBoost model showed consistent improvements across all metrics. While other models like Random Forest and Linear Regression also performed well especially on the test set, XGBoost stood out due to its balance between accuracy and robustness, making it the best-performing model in this analysis.

Regarding interpretability, each model provides insights into the relative importance of features, highlighting which variables contribute most to lung cancer prediction. These feature importance rankings offer valuable information about the potential influence of different factors on lung cancer incidence.

Every model calculates feature importance differently based on the underlying algorithm. For the linear regression model, feature importance is derived from the model coefficients model.coef_, which reflect both the strength and direction of each feature's effect on the prediction. Features with larger absolute coefficients have a greater influence on the output.

For the random forest model, importance is obtained using the feature_importances_ attribute, which measures the average decrease in impurity (such as Gini impurity) caused by each feature across all trees. This provides a relative ranking of features based on how often and how effectively they split the data.

In the case of XGBoost, feature importance is calculated using the booster's get_score() function with importance_type='gain', which captures the average improvement in model performance from splits involving each feature. While random forest and XGBoost provide insight into which features contribute most to the model's predictive accuracy, they do not indicate whether the influence is positive or negative unlike linear regression.

Linear Regression

Feature	Coefficient
Men Incidence	0.958320
Province_Madrid	0.023697
Province_Barcelona	0.023057
Province_Sevilla	0.021560
Province_Valencia/Valéncia	0.020425
Province_Málaga	0.015071
Province_Cádiz	0.014002
Province_Palmas/Las	0.011625
Province_Bizkaia	0.011538
Province_Alicante/Alacant	0.011245

Random Forest

Feature	Importance
Women Incidence	0.500324
Men Incidence	0.488916
Men Factor	0.006872
Women Factor	0.000942
Pb	0.000874
Tobacco Consumption	0.000519
Cd	0.000297
Province_Alicante/Alacant	0.000247
PM10	0.000176
As	0.000145

XGBoost

Feature	Importance
Women Incidence	0.690079
Men Incidence	0.207332
Men Factor	0.019236
Province_Valencia/Valéncia	0.000689
Province_Baleares/Illes	0.000411
Province_Zaragoza	0.000284
Tobacco Consumption	0.000177
Province_Burgos	0.000172
PM10	0.000153
Women Factor	0.000134

Figure 39. Feature importance of the models.

Linear regression is the most interpretable model due to its simplicity and direct coefficient output. In the results, Men Incidence has the largest coefficient, indicating it is the most influential predictor for the target variable. Interestingly, geographical features (provinces) dominate the next top coefficients, such as Madrid, Barcelona, Sevilla, and Valencia, suggesting strong regional patterns in lung cancer cases. This directly supports the goal of exploring geographical disparities. However, pollutants are not among the top features, likely because their effect is overshadowed by incidence and location. Still, linear regression gives a clear picture of relative influence and aligns well with the geographic focus.

Random Forest provides feature importances that show both Women Incidence and Men Incidence are the primary drivers. Among the remaining variables, a few pollutants appear: Pb, Cd, PM10, and As, though with small importance scores. Some province indicators (e.g., Alicante) are also present but minimally weighted. This suggests that while demographic cancer incidence dominates, environmental pollutants have a detectable though small signal. Random Forest thus provides more nuanced interpretability around pollution's role.

XGBoost, also provides important scores. In the model, Women Incidence is the dominant feature, followed by Men Incidence. Beyond this, a few geographical and pollutant features appear, such as Province_Valencia, Province_Baleares, and PM10, with modest importances. This suggests that incidence variables are the main drivers, but the model is still detecting mild patterns in environmental and regional variables. XGBoost thus offers a compromise: slightly better performance than the others, while still allowing insight into the contributing variables.

Feature importance analysis revealed variation in how sex-related variables contributed to model predictions. In the linear regression model, Men Incidence had the highest coefficient, while in both Random Forest and XGBoost, Women Incidence was the most influential feature. This shift may reflect the ability of non-linear models to capture more complex interactions between sex and other risk factors.

Previous studies have shown that women may be more susceptible to lung cancer despite similar levels of exposure to carcinogens, potentially due to hormonal factors, differences in DNA repair mechanisms, and a higher prevalence of actionable genetic mutations such as epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) rearrangements [50-52].

These biological differences may contribute to the increased predictive importance of female incidence in ensemble-based models.

In summary, for explainability, particularly when examining geographic patterns, Linear Regression offers the greatest transparency and ease of interpretation. If the goal is to emphasize the influence of environmental pollutants, Random Forest provides slightly more informative insights due to its ability to capture non-linear relationships and feature importance. However, when aiming for high predictive performance while maintaining a balance between interpretability and accuracy, XGBoost stands out as the most effective option.

4 Results and Discussion

Predictive models were developed using supervised learning algorithms such as Linear Regression, Random Forest, and XGBoost. The machine learning models were initially trained and evaluated using their default hyperparameters, resulting in reasonable predictive performance. To obtain more reliable estimates of model generalization, cross-validation was employed. Subsequently, hyperparameter tuning was applied to improve accuracy. While XGBoost achieved the highest performance, Random Forest proved more effective at identifying relationships between pollutants and lung cancer incidence.

This study also identified associations between environmental pollutants (As, BaP, Cd, Ni, Pb, PM10, and PM2.5), lifestyle risk factors like tobacco use, and lung cancer incidence across regions in Spain. The analysis showed higher cancer incidence in specific provinces, consistent with known patterns.

These results align with the multifactorial nature of lung cancer. External risk factors like smoking habits, occupational exposure, and genetic predisposition, many of which were not included in the dataset, likely contributed to the modest predictive strength of the models. Nonetheless, the models ability to detect meaningful patterns suggests that even limited datasets can offer valuable insights when processed carefully.

The strong influence of incidence-related variables suggests potential feature redundancy, which may hinder the detection of subtler predictors. Future studies should consider reducing multicollinearity and dependence on direct incidence data to better isolate environmental influences.

Compared with existing literature, this study contributes a novel perspective by integrating regional environmental data into predictive modeling. In contrast to previous studies that primarily emphasize clinical and environmental factors in isolation, this geographically informed approach provides a more comprehensive perspective on how lung cancer risk varies across regions.

5 Conclusions

This study explored the relationship between environmental pollutants, lifestyle risk factors, and lung cancer incidence in Spain, using machine learning techniques to assess predictive potential. The process of data extraction, preprocessing, and model development provided valuable insights into the strengths and limitations of applying supervised learning methods such as Linear Regression, Random Forest, and XGBoost to public health data.

Although the predictive performance of the models was modest, the results are consistent with the multifactorial nature of lung cancer, which is influenced by complex interactions among environmental, behavioral, occupational, and genetic factors. The modest contribution of environmental variables in the models points to the limitations of the dataset and highlights the importance of integrating richer and more granular data in future research.

One of the key contributions of this work is the incorporation of geographic variability into the analysis. By analyzing data across Spanish provinces, the study offered a localized view of lung cancer risk, revealing spatial patterns consistent with known incidence rates. This approach provides a valuable perspective that is often missing in studies focused solely on clinical or national-level data.

Moreover, the use of explainable machine learning models allowed for the identification of relevant variables and interactions, even in the presence of limited data. Random Forest, in particular, proved useful for capturing non-linear relationships, suggesting its suitability for epidemiological modeling in complex settings.

However, the dominance of direct incidence-related features in model performance suggests a challenge: these variables may mask the effects of more subtle predictors. To enhance the explanatory power of future models, efforts should be made to reduce multicollinearity, eliminate redundant features, and incorporate proxy variables that better represent environmental exposure and behavioral patterns.

In conclusion, this work demonstrates that machine learning can serve as a valuable tool for environmental health research, especially when combined with thoughtful data curation and interpretation. While the current models are exploratory in nature, they provide a solid starting point for future studies that incorporate more detailed datasets, refined exposure measurements, and specialized modeling approaches. These advancements could enhance predictive accuracy and offer deeper insights into the regional dynamics of lung cancer risk driven by environmental and lifestyle factors.

6 Bibliography

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, pp. 229-263, 2024. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834.

[2] Ferlay, J., Ervik, M., Lam, F., Laversanne, M., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., Bray, F. (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available: https://gco.iarc.who.int/today.

[3] American Cancer Society. (2024). What Causes Lung Cancer?. Available: https://www.cancer.org/cancer/types/lung-cancer/causes-risks-prevention/what-causes.html.

[4] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, 8: Article 14. doi: 10.3389/fpubh.2020.00014. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2020.00014/full.

[5] Cohen, A.J. and Pope, C.A. III (1995). Lung Cancer and Air Pollution. Environmental Health Perspectives, 103(Suppl 8):219-224. The Health Effects Institute, Cambridge, Massachusetts; Brigham Young University, Provo, Utah.

[6] Kulhánová, I., Morelli, X., Le Tertre, A., Loomis, D., Charbotel, B., Medina, S.,Ormsby, J. N., Lepeule, J., Slama, R., Soerjomataram, I. (2018). The fraction of lung cancer incidence attributable to fine particulate air pollution in France: Impact of spatial resolution of air pollution models. Environment International, 121, 1079–1086. doi: 10.1016/j.envint.2018.09.055.

[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., 2000. [Online]. Available: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf.

[8] Observatorio Contra el Cáncer, El cáncer de un vistazo, Asociación Española Contra el Cáncer (AECC). [Online]. Available: https://observatorio.contraelcancer.es/informes/el-cancer-de-un-vistazo

[9] Asociación Española Contra el Cáncer, "Cáncer de pulmón", AECC, [Online]. Available: https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-pulmon.

[10] American Lung Association, “Types of Lung Cancer,” [Online]. Available: https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/basics/lung-cancer-types. [Accessed: Jul. 17, 2025].

[11] World Cancer Research Fund/American Institute for Cancer Research, Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. The Third Expert Report, 2018. [Online]. Available: https://www.wcrf.org/wp-content/uploads/2024/11/Summary-of-Third-Expert-Report-2018.pdf

[12] Sociedad Española de Oncología Médica (SEOM), Prevención del cáncer, [Online]. Available: https://www.seom.org/informacion-sobre-el-cancer/prevencion-cancer.

[13] D. Clofent, M. Culebras, K. Loor, and M. J. Cruz, "Contaminación ambiental y cáncer de pulmón: el poder carcinogénico del aire que respiramos," *Arch Bronconeumol.*, vol. 57, pp. 317–318, 2021.

[14] World Health Organization (WHO), "Air pollution and cancer," International Agency for Research on Cancer, IARC Scientific Publication No. 161, 2013.

[15] IARC Working Group, "Outdoor Air Pollution," *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, vol. 109, 2015.

[16] U.S. Environmental Protection Agency (EPA), "Particulate Matter (PM) Basics," 2023.

[17] R. Raaschou-Nielsen *et al.*, "Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE)," *Lancet Oncol.*, vol. 14, no. 9, pp. 813–822, 2013.

[18] G. B. Hamra *et al.*, "Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis," *Environ Health Perspect.*, vol. 122, no. 9, pp. 906–911, 2014.

[19] G. B. Hamra *et al.*, "NO_2 exposure and lung cancer: results from a meta-analysis," *Environ Health Perspect.*, vol. 123, no. 11, pp. 1107–1112, 2015.

[20] World Health Organization (WHO), *Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide: Report on a WHO Working Group*, Bonn, Germany, 13–15 January 2003.

[21] A. J. Cohen, H. R. Anderson, B. Ostro, K. D. Pandey, M. Krzyzanowski, N. Künzli, K. Gutschmidt, A. Pope, I. Romieu, J. M. Samet, and K. Smith, "Urban air pollution," in *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attribution to Selected Major Risk Factors*, vol. 2, M. Ezzati, A. D. Lopez, A. Rodgers, and C. J. L. Murray, Eds. Geneva: World Health Organization, 2004, pp. 1353–1433.

[22] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), "Índice de Calidad del Aire", [Online]. Available: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/visualizacion-datos-calidad-del-aire/ica.html.

[23] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), Informe de evaluación de la calidad del aire en España 2023, [Online].Available: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/evaluacion-y-datos-de-calidad-del-aire.html.

[24] M. F. Naujokas, B. Anderson, H. Ahsan, H. V. Aposhian, J. H. Graziano, C. Thompson y W. A. Suk, "The broad scope of health effects from chronic arsenic exposure: update on a worldwide public health problem", Environmental Health

Perspectives, vol. 121, n.º 3, pp. 295–302, mar. 2013, doi: 10.1289/ehp.1205875.

[25] C. B. B. Guerreiro, J. Horálek, F. de Leeuw y F. Couvidat, "Benzo(a)pyrene in Europe: Ambient air concentrations, population exposure and health effects," Environmental Pollution, vol. 214, pp. 657–667, jul. 2016, doi: 10.1016/j.envpol.2016.04.081.

[26] J. Lee, D. Y. Shin, Y. J. Jang, J. P. Han, E.-M. Cho y Y. R. Seo, "Cadmium-induced Carcinogenesis in Respiratory Organs and the Prostate: Insights from Three Perspectives on Toxicogenomic Approach," Journal of Cancer Prevention, vol. 28, n.º 4, pp. 150–159, dic. 2023, doi: 10.15430/JCP.2023.28.4.150.

[27] T. Behrens et al., "Occupational exposure to nickel and hexavalent chromium and the risk of lung cancer in a pooled analysis of case-control studies (SYNERGY)," Int. J. Cancer, vol. 152, no. 4, pp. 645–660, Feb. 2023, doi: 10.1002/ijc.34272.

[28] A. Anttila, S. Uuksulainen, M. Rantanen, and M. Sallmén, "Lung cancer incidence among workers biologically monitored for occupational exposure to lead: a cohort study," Scandinavian Journal of Work, Environment & Health, vol. 48, no. 6, pp. 457–465, 2022, doi: 10.5271/sjweh.4046.

[29] World Health Organization, "Ambient (outdoor) air quality and health," WHO, Oct. 24, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[30] U.S. Environmental Protection Agency (EPA), Health and Environmental Effects of Particulate Matter (PM) [Online]. Available: https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm.

[31] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.

[32] T. M. Mitchell, Machine learning, McGraw-Hill, 1997, New York, NY, USA.

[33] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[34] R. Kumar and S. Sharma, "A novel approach for solving nonlinear differential equations using homotopy perturbation transform method," International Journal of Mathematical Sciences, vol. 8, no. 6B, pp. 1463-1476, 2023, doi: 10.22271/maths.2023.v8.i6b.1463.

[35] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[36] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[37] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Burlington, MA, USA: Morgan Kaufmann, 2016.

[38] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," arXiv preprint arXiv:1603.02754, 2016.

[39] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, Canada, 1995, pp. 1137–1143.

[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[41] H. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[42] G. Batista and M. Monard, “An analysis of four missing data treatment methods for supervised learning,” Applied Artificial Intelligence, vol. 17, no. 5–6, pp. 519–533, 2003.

[43] M. C. Batista and M. C. Monard, “A study of K-Nearest Neighbour as an imputation method,” in Proc. 2nd Int. Conf. Hybrid Intelligent Systems (HIS), Santiago, Chile, 2002, pp. 251–260.

[44] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, 2nd ed. Hoboken, NJ, USA: Wiley, 2002.

[45] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*, Springer, 2006.

[46] G. W. Tukey and D. A. Mosteller, “Data analysis and regression: A second course in statistics,” Addison-Wesley, Reading, MA, USA, 1977.

[47] S. Barth, "Machine Learning in Healthcare: Guide to Applications & Benefits", ForeSee Medical, 3 de febrero de 2025. [Online]. Available: https://www.foreseemed.com/blog/machine-learning-in-healthcare.

[48] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd, “The efficacy of machine learning models in lung cancer risk prediction with explainability,” PLoS ONE, vol. 19, no. 6, e0305035, Jun. 2024. doi: 10.1371/journal.pone.0305035

[49] K.-M. Wang, K.-H. Chen, C. A. Hernanda, S.-H. Tseng, and K.-J. Wang, “How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking,” Int. J. Environ. Res. Public Health, vol. 19, no. 14, p. 8445, Jul. 2022. doi: 10.3390/ijerph19148445

[50] J. Subramanian and R. Govindan, "Lung cancer in 'never-smokers': a unique entity," The Oncologist, vol. 12, no. 5, pp. 494–500, 2007, doi: 10.1634/theoncologist.12-5-494.

[51] Y. Liu and K. He, "Sex differences in lung cancer susceptibility: genetics and molecular biology," Cancers (Basel), vol. 15, no. 7, p. 1902, 2023, doi: 10.3390/cancers15071902.

[52] A. G. Schwartz and M. L. Cote, "Epidemiology of lung cancer," in Advances in Experimental Medicine and Biology, vol. 893, pp. 21–41, 2016, doi: 10.1007/978-3-319-24223-1_2.