UNIVERSITÀ DEGLI STUDI GUGLIELMO MARCONI

FACOLTÀ DI INGEGNERIA

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

« Analysis of Correlations between Environmental Data and Public Health: Development and Evaluation of Predictive Models. A Data Science Approach»

Relatrice:
Chiar.ma Prof.ssa Francesca Fallucchi

Candidato:
Ciro Tafuri
Matr. N°: 22579

ANNO ACCADEMICO

2023/2024

"Sometimes people think something is impossible until someone else proves otherwise."

Alan Turing

I would like to express my deepest gratitude to everyone who believed in me and my work, showing me that with dedication and support, nothing is impossible.

Index

CHAPTER 1 – Introduction -

Artificial intelligence technologies have been reshaping a multitude of domains at an unapprehended pace. With the capability to outdo human capabilities in virtually all scientific domains, machine learning is one of the most potent subfields of AI. For instance, these are cognitive functions, generation of language, and tasks that, up to this day, are only imagined to be exclusive functions of the human intellect. Machine learning is turning out to be a revolution in the scientific field, enabling great advances in, for example, drug discovery, pattern matching, and neuro-symbolic AI. An excellent example is how well IBM DeepMind has been able to generate languages, opening newer horizons that broaden the capability and opportunity of the scientific world. Automated inference techniques, in this process of scientific discovery, have turned into instruments for problem-solving that will allow researchers to determine which hypotheses to test, conduct experiments, and structure the principles that describe the observed phenomena. It is the availability of vast data volumes, or "big data," which facilitates the main factor for machine learning to expand its frontier to the natural and social sciences. Such data enables the singular opportunity where "the data speaks for itself," an explanation with an increment of the structure, relationships, and inherent links within the data. For example, in medicine, for the most part, the conventional approaches relying on data-based predictive modes have been superseded by a demand for more precise outputs, such as matching the most adequate preventive and therapeutic measures for the patient at the right time. Machine learning algorithms must be embedded with a clear understanding of the specific expertise in the healthcare sector so that, by further coupling with causal discovery, the qualitative change

in medicine occurs and there is better patient outcome. "Big data" is becoming more and more of a quantitative analytical tool in medicine but also in climate change research. Artificial intelligence and machine learning are also now revolutionizing the field of climate change research, in particular, the management of big data volumes and overcoming one of the main barriers to defining policies concerning climate change: a lack of concrete evidence.

In addition, innovative spatial and temporal econometric models added to our understanding of the impact that climate change has on human health and simultaneously set the way forward for defining global economic policies. This thesis will try to fill this gap of inconclusive evidence by trying to understand how machine learning can be applied to understanding the linkage between climate change and human health. Tools of machine learning for the analysis of data will project future trends in mortality data and explore any causal links that can be established by climate change. In this way, a larger knowledge base will be opened up for policy decisions, and it will also pave the way to apply the tools of causal discovery in a detailed mapping of causal relations.

1.1. Context and Significance of the Topic

In fact, with each passing day and the rapidly growing body of data, a consensus is being developed among professionals, researchers, and scientists that most of the solutions to a scientific concern lie in this very "big data." It is under these conditions of accelerated development in the field of data science, specifically its great volume in the last few years, that the development of research areas like "health data science" and "emerging machines for climate change" was suggested. This presents a new opportunity for data-based perspectives evaluation, since "big data" is now becoming a fact in the analysis of quantitative information and the development of comprehensive evidence. Much greater familiarity with machine learning and its processes is required for the interpretation of data and application of scientific discoveries in climate change sciences. For example, most of the traditional approaches in climate change sciences can be purely predictive in a mode based on data, hence unable to develop precise policies that can align economic goals with the most appropriate actions and measures within specific time constraints. Hence, the use of machine learning tools is important in the availability of big data, which is a key enabler in addressing the lack of evidence for efficient and effective decision-making. There is, therefore, an urgent need for research on the state of climate change, given that there is the present alarming situation in Europe concerning climate change for the mitigation of the adversities. Having this in view, the present thesis tries to clarify and investigate the issue of climate change in Europe, within the framework of machine learning, to analyze the relationship between climate change indicators and regional mortality. There is a gap in providing comprehensive and

conclusive evidence of the impact of climate change on human health through the use of machine learning tools, and it does exactly that. In other words, it will be possible to project trends that can be expected in the future with the most efficient models. Climate change and machine learning are two models that this paper will apply its tools to. In sum, the evident contribution of this study includes the methodologies employed as the innovative application of the ML framework, opening up a new and significant empirical insight into climate change research.

1.2. Objectives of the Thesis and Research problem

The most important challenge in defining a mitigation policy for climate change is the lack of solid evidence that is conclusive at a national scale but varies over both spatial and timelines. Following the devastating floods and shooting challenges in the global and national climate, there is a need to come up with an urgent intervention policy to mitigate the socio-economic crisis that comes as a direct result of this fundamental factor for economic growth. This study focuses mainly on forecasting the futuristic trend of mortality induced by climate changes over the European continent, providing conclusive evidence on how the new machine-learning techniques have been applied to learn the effect of climate change on human health. There is still a large gap in the exploration and identification of causal pathways through which climate change influences human mortality, with special mention of the numerous barriers about data availability and literature on climate change that are crucial in adaptation and mitigation. This paper sought to close the gaps by creating data-based evidence. It, therefore, deems it necessary to use big data and evaluate the best machine learning tools about climate change data and its projections to create comprehensive evidence. The main aim of the current paper is to project the futuristic trend of mortality related to climate change in Europe by gathering conclusive evidence on climate change and human health (mortality) using machine learning techniques. To apply machine learning tools in the emerging climate change and machine learning sector. To investigate interactions between climate change and mortality, with the aim of understanding and describing relations between climate change and mortality and illustrating them through spatial mapping. This will

therefore be used in the development of key research and knowledge gaps, the need for funding, and resources that would be used to support research in areas of climate change, health, and machine learning. It places all future research on climate change and machine learning on an even playing field in the pursuit of applying and assessing tools in machine learning. First, it is likely that the application of machine learning methods will result in the development of "living" evidence platforms. This may enable governments to prioritize and invest in actions that would be most likely to reduce and prevent the next phases of climate change's impact on human health. It also calls for the development of development economists, environmentalists, social scientists, and policy experts in their search for answers to real-world problems, not through conventional approaches, but via ML. To be precise, scientific evidence related to low-income countries, which are mostly affected by health impacts due to climate change, is deficient. There is a very small amount of research on how climate change affects physical and mental health. No or little evidence is provided for the promotion of health by climate change adaptation interventions. Though the literature review mentions some examples of CC adaptation activities that state their ability to improve health in general, they did not provide quality data.

CHAPTER 2 - Theoretical Foundations -

2.1. Machine learning overview

Forecasting of macroeconomic variables is essential for designing proper policy measures. In the absence of proper predictions, designing and implementing policy measures are seriously impaired. Real Gross Domestic Product (GDP) growth is an important variable in macroeconomic indicators, but this is difficult to predict due to delays in the publication of official data, normally released by quarter. This results in a situation where policy has to be formulated and implemented by the policymakers in the absence of proper data. Forecasting real GDP growth is a complex variable. Conventional economic forecasting models attempt to predict data on theoretical top-down approaches, linking causal associations between various variables. These models require forecasters to provide an economic sense to the data and methodology adopted, which sometimes may result in a forecast that is not valid.

In this context, the derived models by the machine learning technique are more flexible and less assumptive in forecasting compared to more traditional ones. A handful of studies attest to outperforming traditional econometric methods, such as those from Plakandaras et al. (2019) in real estate forecasts or from Medeiros and Gabriel Vasconcelos (2019) in low-frequency inflation forecasting. In the meantime, the DFM and, respectively, the DFM and its many extenuations are of great utility in forecasting macroeconomic time series in the last decades. Studies similar to Stock & Watson (2002) have shown the DFM to be superior in anticipation of the Federal Reserve Board's industrial production index. This is achieved by, first, extracting the factors using the PCA technique

and, then, regressing the target variable on the obtained factors. Although this methodology has shown great effectiveness, the LASSO has shown to be more effective, comparing their out-of-sample predictive performance with DFM on, for example, the study from Li & Chen (2014), which used 107 monthly/quarterly collected macroeconomic indicators. Another more common study was conducted by Bai & Ng (2008) which applied the LARS to select the predictors and achieved superior results compared to PCA in most data sub-samples.

These methodological breakthroughs certainly warrant further research, particularly for low-income economies where the effects of macroeconomic change on physical and mental health are the greatest. For governments, the availability of machine learning "live testing" platforms is further going to enable them to prioritize and implement interventions designed to mitigate the future effects of macroeconomic change. Finally, the scarcity of high-quality data with less adjustment calls for investment in research in the area of macroeconomic change and its socioeconomic implications.

CMIP5 is among the most spread models in the use of climate change research and adopted by NASA and multiple other organizations to track the variation in the climate. New techniques, the E2 Model for atmospheric modeling among others, have been used and influenced studies in the climate change field immensely. "Pizzulli et al., 2021". In this respect, long-term global warming prediction research and development are used in a multitude of areas, in areas such as climate research, agriculture, energy, and medicine. "Hema et al. 2019", a study measured by multiple machine learning algorithms, the annual rate of global warming in India. The main causes of global warming were estimated and then compared for the

validity of predictions between linear regression and Support Vector Regression SVR. It was concluded that CO2 was the main factor responsible for the increase in temperature, followed by CH4 and N2O. In terms of comparing the causes of mortality with the factors affecting climate change, machine learning models and statistical models have been used to track and analyze how multiple risk factors contribute to significant causes of death. Anthropogenic climate change has thus been detected as having a significant impact on the cause of human disease.

Berrang-Ford et al. used supervised machine learning with NLP to map the literature on climate change and health. They found all-cause mortality and infectious diseases to be the most studied health outcomes, but there were striking gaps regarding outcomes on mental health, malnutrition, and maternal and child health. Scheelbeek et al. observed that, although it is conceivable that some adaptation strategies could lead to collateral benefits to human health, in general, the lack of depth of data on the various plausible effects of these strategies represents a missed learning opportunity. It developed an AI-aided framework based on NASA's predictive data on climate change to infer the link between climate variability and major causes of deadly diseases worldwide—that is, predicting the future with the most effective correlation models. Automated machine learning is key to properly mapping the vast literature on climate and health. It is the key to finding causal correlations from observation data, which would be effective not only in understanding but also in possibly mitigating the effects of climate change on human health.

2.2. Correlation between climate change and health outcomes

These variations in the factors that cause diseases or that prevent treatment may well result in some seasonal variation in mortality over time. Such changes in mortality are important for any evaluation of the effectiveness of interventions to reduce seasonal mortality. Although it is an important factor in some of the detailed work on mortality impacts, the timing and magnitude of fluctuations in mortality rates have been little explored to date, especially in relation to the local environment, age group, gender, and medical causes.

Building on this literature, Parks et al. (2018) studied the seasonality of mortality rates across age groups and by sex in the USA and its sub-national climatic zones over the 1980-2016 period using geocoded mortality data. They analyzed seasonal trends of general and cause-specific mortality in the USA by the tools of Wavelet analysis which earlier had been applied to examination of the dynamics of meteorological phenomena and infectious diseases. The circular statistics and the techniques of analysis of the center of gravity were employed to address the timing of extremes in the range of mortality and revealed that the percent difference between mortality rates in the months with the highest and lowest mortality has changed. Among men and women aged in the late 40s, peak mortality was experienced from December to February and lowest from June through August, with primary causes being injuries and disorders of the cardiorespiratory system. From 1980 to 2016, the percent difference in peak-to-low monthly mortality rates did not differ across climate zones. In addition, it was shown that since the 1990s, seasonality in all-cause mortality for children under five years has declined sharply. Evidence from thermal biology suggests a decrease in the spread of vector-

borne diseases at thermal extremes and a peak at moderate temperatures. (Guo et al., 2016; Hansen et al., 2010; Liu et al., 2014; Mordecai EA et al., 2013; Paull et al., 2017). However, for most vector-borne diseases, their thermal optima and limits have not yet been described. To fill this causality literature gap between temperature variation and disease transmission, Shocket et al. (2018) developed a mechanistic model for the heat response of the Ross River virus, a major pathogen transmitted by mosquitoes in Australia and the Pacific Islands, using data from lab experiments designed to evaluate the performance of viruses and mosquitoes across a wide temperature range. They stated that "transmission peaks at moderate temperatures (26.4°C) and falls to zero at thermal extremes (17.0°C and 31.5°C)." The model, as it pointed out, "rightly predicted that transmission is ubiquitous year-round in tropical regions and seasonal in temperate regions, leading to a seasonal peak of human cases countrywide. Most Australians live in temperate regions, where the model predicts transmission to increase. In tropical regions, where the mean temperature is already close to the thermal optima, the effect will likely be to decrease the transmission. Possible policy implications of this model's findings for the Australian Government and mosquito control organizations in more effective long-term planning are also discussed.

Admittedly, the paper by Gaythorpe et al. (2020) was the first to analyze the potential impacts on disease burden arising from a change in climate, which is a critical determinant of the transmission of Yellow Fever in the Americas and Africa. They projected the level of transmission intensity in the African endemic region under four different climate change scenarios by expanding a pre-existing transmission model of Yellow Fever to account for rainfall and a temperature appropriateness index. Results

indicated that there was disproportionately varying disease incidence across the zone. The further analysis of the burden change in 2050 and 2070 took these transmission projections into account. The probability that annual deaths would increase in 2050 was 93.0 percent [95 percent CI (92.7, 93.2%)] which is indicative that the future programs in disease control would become tougher.

The study by Rerolle et al. (2021) elaborated on the linkage between forest fall and malaria transmission in GMS by using high-resolution forest coverage data by Hansen et al. (2010) and monthly malaria incidence data from 2013 to 2016. They further proved that the loss of rural forests has increased transmission during initial years of fall, but slows down after a few years of forest loss. The geographic location of deforested areas also was a concern. It did not have an impact on the rate of malaria within one to ten kilometers of the settlements. Thirty kilometers radius further away showed an effect in the transmission of malaria. The high forested intensity with twenty-five kilometers of deforestation in the radius could see the effect. Deforestation in high forested intensity areas seems to be the main influence of the results. These findings prove that activities in the forest affect malaria spread in the GMS. The study by Jaeggi et al. measured the relationship between social status with health. They tested the health of communities that had very weak social hierarchies in Bolivia. They compared the relative wealth and income of 870 households from 40 Tasmanian settlements with a variety of health outcomes including blood pressure, self-rated health, levels of stress hormones, depressive symptoms, and multiple diseases. We found that not all of the health indicators examined suffered negative impacts as a result of poverty and inequality, as could be seen to be occurring in industrialized cultures.

Nonetheless, overall, individuals of lower income and residing in more unequal neighborhoods had higher blood pressure levels.

Epidemiological study on the relationship between birth weight and PM2.5 at birth, due to fire in ambient air over time, from 2000-2014 in 54 lower-middle-income countries. J. Li et al. (2021) in a sibling-matched case-control study among 227,948 newborns. Fire-attributable PM2.5 was measured to estimate within-group differences in gestational exposure to landscape fire smoke (LFS) for each newborn. They used a fixed-effects regression model to examine the associations between within-matched-sibling-set differences in birth weight and within-group differences in LFS exposure. were noted. In general, the results indicated that prenatal exposure to LFS compromises neonatal health. The experiment by Kunze et al. (2022) was quite a well-designed one in showing that an increase in the mean temperature and wild fluctuation in the temperature regimes would impact host-pathogen interaction. They point to the necessity of processes that underlie species interaction in changing habitats and how the dynamics of disease would change because of climate change.

2.3. Factors illustrating the links between climate change and causes of death

Human activities that degenerate the environment have a long-term effect on the economies and well-being of people at local and global levels. Greenhouse gases, which are anthropogenic in nature, cause an abnormal upsurge in global warming that causes the melting of ice caps, polar caps, and glaciers, which together produce most of the freshwater of the world. There are a number of causes related to the lack of clean water in order to analyzeect to irrigation and drinking water consumption, such as catastrophic glacier advances, high rates of glacier melt, GLOF that results in the flooding of glacial lakes, massive recessions, and a negative glacier mass balance. In addition to the depletion of the reserves of global freshwater, there are also alarms raised in the context of the aging of deglaciation due to the fact that human settlement and physical infrastructure have also been severely endangered because of environmental degradation and hence the need for a great restoration. Similar socio-cultural effects of varying degrees are exhibited by climatic calamities, more so to peoples whose subsistence is wholly and directly hinged on freshwater resources that are becoming scant. In this regard, freshwater scarcity has set the stage for the local and global social fabric, terminating social characteristics such as relationships, trust, and networking to arrive at a common solution to problems. Instead of securing access to a minor freshwater source, it further reinforces the dependence of the communities and increases vulnerability to natural disasters, increasing the likelihood of social conflicts. It also speeds up the depletion of resources, thereby widening the scope of social ills.

The most frequently studied factors in climate health literature as categorized by (Berrang-Ford et al., 2021; Scheelbeek et al., 2021) are presented below:

Climate and Weather Events	Climate Change Forcing	Health Outcome
Extreme Temperatures	Temperature, CO_2 & CH_4 Concentration	All-cause mortality, Mental & Behavioral Disease
Frequent Heat waves	Temperature, CO_2 Concentration	All-cause mortality, Nervous, Mental & Behavioral Disease
Extreme Precipitation and Flooding	Rainfall, Temperature, CO_2 Concentration	Vector-Borne Diseases Digestive Diseases
Air Quality	CO_2 Concentration, CH_4, NO_2 CFCs	Respiratory Diseases, Infectious Diseases, Mortality, Nervous, Mental & Behavioral Disease
High Ocean Temperature & Acidification	Temperature, Anthropogenic forcing	Infectious Diseases,
Coastal Flooding	Seasonal Rainfall Temperature	Vector-Borne Diseases Digestive Diseases
Weather variability	Anthropogenic forcing, CO_2 Concentration	Vector-Borne Diseases Digestive Diseases, All-cause mortality, Mental & Behavioral Disease

Figure 2.1 : Indicators in climate change literature

McIver and Kim, 2011 in a literature review established that the areas that have received the least amount of attention in research on climate health include the issues of mental health, maternal, child health, and nutrition. According to the study by Berrang-Ford et al., 2021 in Asia and Europe, the impacts of particulates on the quality of air have been addressed thoroughly. In North America, hurricanes were the most common issues that were addressed; thus, extreme events were one of the three top risks.

In Europe and Oceania during that time, the major risks included the issues of heatwaves. However, in comparison to other places, the literature on the African and Latin regions identifies the variability of precipitation and weather as risks more common than the other risks.

Similarly, Berrang-Ford et al. 2021 reported a list of health issues about air quality, all-cause mortality, infectious diseases, and heat stress. A range of health outcomes is reported, with the greatest emphasis on respiratory outcomes, particularly about air pollution. Respiratory health is most prominent in Asia, while heat stress is one of the three most prominent health outcomes reported for Europe, North America, and Oceania. All-cause mortality is the most common health outcome examined in the literature across all regions. Other common disease-specific topics, besides cholera, dengue, influenza, leptospirosis, and malaria, include dengue, which is the primary health issue reported in Latin America and the second most common topic reported in Asia. Literature reports that evidence on climate health is majorly in some geographical regions of high-income countries, and from low-income countries severely affected by the health problems due to the consequences of climate, it is minimal. Berrang-Ford et al., 2021 reported that 79% of the 15,914 studies on climate and health are in high and upper-middle-income countries, and this is dominated by studies on China.

A major income gradient characterizes the published works on climate change and health. While the studies from upper-middle-income and high-income countries are equivalent, it is misleading; the high count of publications from China explains well the underrepresentation of research from Central Asia.

Studies from low-income areas are on infectious diseases, followed by food and nutrition and maternal and child health, remaining with significant attention (Checkley et al., 2004; Singh et al., 2001). As much as there is a gradient of greater emphasis put on infectious diseases, food and nutrition, water, sanitation, and hygiene services, as well as maternal and child health among the low-income status, this is well depicted by a gradient about chronic diseases, respiratory health, and demand towards health care systems. It was only recently discovered by Pizzulli et al., 2021 that there is a close correlation between climate change and human health globally. Many different prevalent diseases in the world have spread more rapidly and shown their worsening symptoms. Climate change could be linked to this recent development. The World Health Organization stated that environmental factors are responsible for 23% of global mortality (Confalonieri, 2007).

WHO04 is the only empirical analysis of diarrhea effects and probably one of the best health studies ever done (Kolstad and Johansson, 2011). WHO04 concluded a 1°C increase in temperature is associated with a 5% increase in diarrhea and states this is an approximate estimate based on empirical data from Fiji and Peru (Singh et al., 2001). These relationships and causal pathways are summarized in the following figure.

Climate change is the causative factor for the rise in incidence of diseases and mortality rates globally. Climate variability directly affects human life

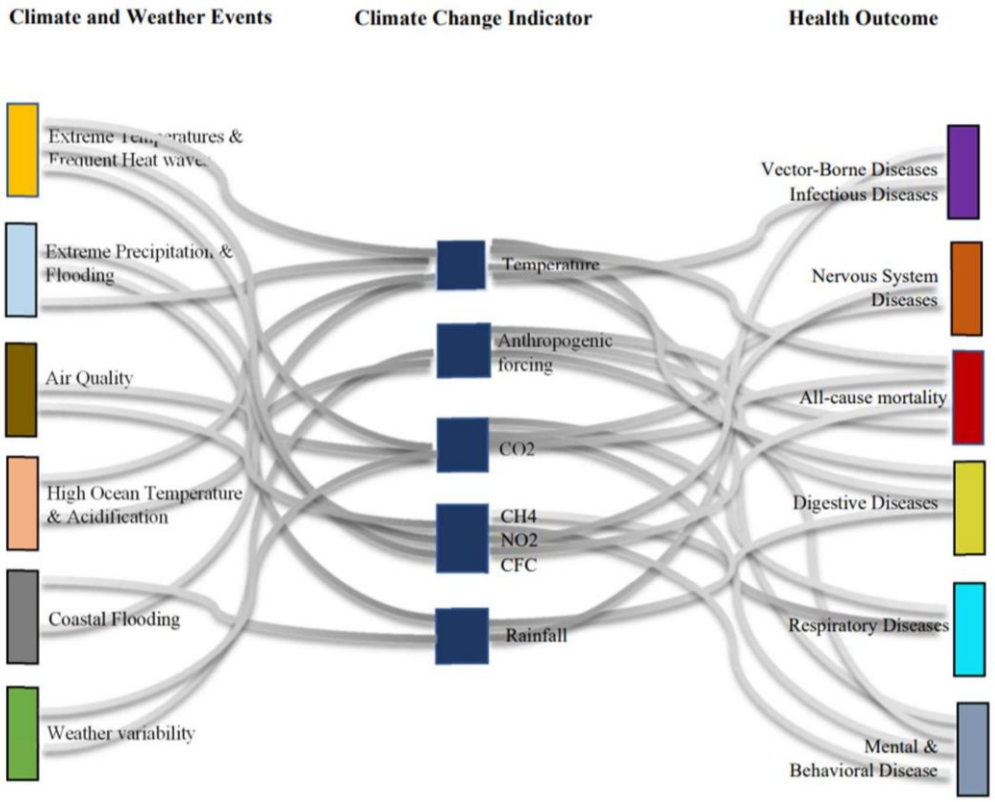

Figure 2.2: Causal linkages diagram

through the alteration of such patterns as extreme temperatures, precipitation, rising of the sea level, and more frequent extreme events, or indirectly through alterations in the quality of water, air, and food as well as ecosystems, agriculture, industry, settlements, and the economy. At present, the effects are small, but an increase expected to be slow will touch on the globe (Gallopín, 2006).

The impacts from changes in weather in human health take place from alerts to intense heat and violent storms to others probably much more distant. Higher concentrations of carbon dioxide in the upper stratosphere cause severe environmental damage due to the rise in temperature.

Climate and environment directly affect pathogenic insects' survival and behavior. Food and water quality will also be impacted due to climate change that, consequently, may affect human health (Checkley et al., 2004). In addition, the world faces huge risks linked to human exposure to the adverse effects of global warming on mental health and psychological well-being (Hayes et al., 2018).

2.4. Working Framework

The WHO states that close to 4 million people succumb every year as a result of causes related or directly linked to change in climate, representing about 23% of deaths all over the world. Climate change through both exacerbation and direct causation is responsible for the continued increase and spread of the disease symptoms. Climate change is a factor in the high global disease burden and premature mortality. The magnitude of climate change effects at this time is small, but the general consensus of science and research is that the rate of change and the magnitude in space and time will significantly increase and change in an unprecedented manner in the future. The impacts of climate change on human life and health, experienced directly through variations in the weather pattern, include changes in precipitation, temperature, sea level rise, and increased frequency of extreme events, while the indirect impacts include changes in ecosystems, such as changes in agriculture, industries, production, settlements, economy, quality of water, air, and food. Scientifically, the rationale that (Pizzulli et al., 2021) followed in examining mortality precipitated by climatic change is that in understanding the individual exposure pathways that are capable of causing human diseases, the approach would be efficient in learning the nature in which climatic change affects health. The mode of health effect is of an exposure pathway nature, in terms borrowed from the field of chemical risk assessment. Humans may react to the various exposure pathways differently depending on the environmental context and the duration of exposure. Single or multiple climatic changes, as well as geographical location, are potential human risk factors.

Climate change has a directly negative effect on human health through causal channels, such as meteorological variability, which deteriorates human health through mutations in biological and chemical pathways causing diseases, and indirect causal channels, such as food chain disruption, caused by changes in the ecosystem, leading to shifts in biodiversity of affected populations' biomes (Kunze et al., 2022). The spread of the vectors that carry dengue is a classic example of direct causality. Other indirect channels leading to the deterioration of human health are adverse income distribution, population explosion, and regional conflicts. A non-theoretical, evidence-based approach has thus, in effect, helped cull out the following conceptual framework, as can be elucidated below in Figure 2.2. Climate change indicators, such as the increasing temperature, increased levels of CH4 and CO2 concentrations, human-induced pollution, and anthropogenic forcing, pose a danger to human life, health, and the environment. Scientific research, as indicated by Berrang-Ford et al. 2021, Pizzulli et al. 2021, and Scheelbeek et al. 2021, shows the importance of choosing these climate change indicators in human health in terms of importance and health areas. The scientific reasoning for using mortality as an indicator of human health has been explained by (Berrang-Ford et al., 2021; Pizzulli et al., 2021) in the research studies. Mortality counts and rates are classical measures to evaluate the burden and the impact comparison of diseases in the domain of medical sciences, public health, health economics, and demography. Furthermore, in a couple of individual research studies by (Melillo et al., 2014; Schwartz et al., 2015) for the summer season from April to September, the scientists found that global warming will increase mortality. On the other hand, from October to March in the winter season, the studies predicted a decrease in deaths owing to global warming. With no such methodological

change to adapt in the coming future, these results maintain the human population levels at 2010. For this reason, studies forecast the twenty-first century, whereby the discovered links between the temperature with mortality produced in the last decade of evidence for 1997-2006 will not change (Gosling et al., 2009; Kalkstein & Greene, 1997). These show the evidence provided in these studies of mortality evoked by climate changes and further explain how mortality is an indicator of human health.

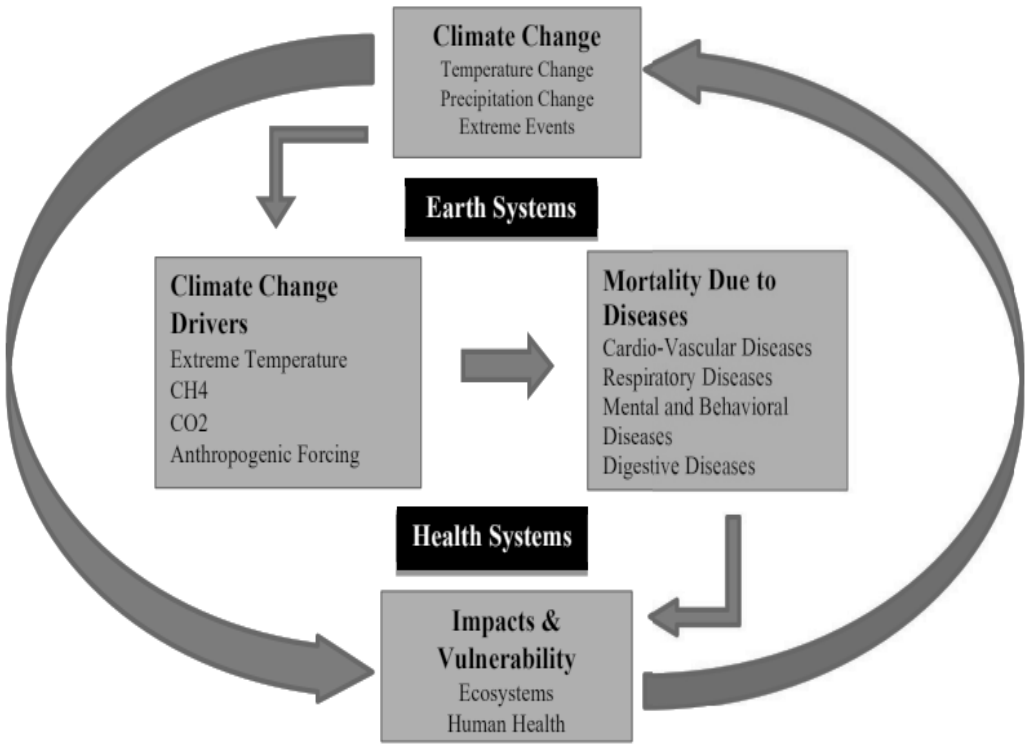

Figure 2.3: Framework of Climate Change Drivers and Their Impacts on Human Health

CHAPTER 3 - Data and Methodology -

Machine Learning is the integration of a set of algorithms utilized in modeling, interpreting, and understanding a large dataset under consideration. Machine learning is the process by which computers can be attuned to learn without being programmed explicitly. Generally, the ability of the computer to impersonate human intellect is referred to as machine learning. Machine learning adopts an automatic way in order to program computers through experience.

Machine learning is not something new; the first idea was formulated in the 1950s, and then multiple prototypes paved the way for early artificial intelligence. However, the hardware was not ready for it. A few decades later, with the concept of automation and processing of sophisticated mathematical operations on immense data, it picked up a lot of hype.

The ML algorithms learn from data and experiences, just as humans do. More precisely, the algorithms learn, grow, adapt, and evolve in view of new data. In other words, the goal of machine learning is for computers to be able to make inferences without being explicitly instructed to do so. Instead, the iterative learning algorithm is followed to attain this goal.

ML involves a rich variety of algorithms that could be grouped into two broad sets of algorithms: supervised learning and unsupervised learning. Supervised machine learning is a model trained on labeled input data to generate an input-output mapping that allows predicting the output values from the input instances. To date, it is the most widely applied machine learning approach. Unsupervised machine learning is the computer system trying to learn the patterns of input data. Unsupervised machine learning can find patterns or trends not actively looked for by

humans. Predominantly, supervised learning finds association or prediction in the dataset, while unsupervised learning explores data description, which might include dimensionality reductions and clustering, etc.

The main algorithms of supervised learning include:

Algorithms	Type	Usage	Advantages
Lasso	Supervised Learning	Association/ Prediction	Automatic covariate selection. Simple interpretability.
Neural Network		Association/ Prediction	Many predictors & non-linear relations can be accommodated. Better prediction performance.
Random Forests		Association/ Prediction	Identification of heterogeneous treatment & effects
Hierarchical Clustering / K-means	Unsupervised Learning	Dimensional reduction/ Clustering	Free of hypothesis. High-dimensional data could be mapped to lower dimensions.
Propensity Score Matching		Counterfactual/ Casual Inference	Simple Interpretability

Figure 3.1: Framework of Climate Change Drivers and Their Impacts on Human Health

Machine learning can start by feeding the selected algorithm with training data. When feeding training data into the selected algorithm, the system refers to the known or unknown information that is labeled or unlabelled data, needed by an algorithm.

After training an algorithm, new data can be fed to the same algorithm by using it as an input. The testing phase starts next. Forecasts are tested against results produced by the trained and tested phases. In cases of mismatched predictions and outcomes, the algorithm is retrained on various occasions until the required goal is achieved. In that way, machine learning systems keep learning and providing the perfect answer, improving with time.

3.1. Machine learning versus traditional methods in climate and health

Professionals have been dealing with an exponentially increased quantity of health data, comprising clinical, pharmacological, and genetic data. There is a growing belief that these statistics, which are too big, can resolve all issues in medical and scientific matters and, therefore, revolutionize healthcare into precision healthcare. However, statistics itself does not tell much of the story or the causes and issues lying behind it. The difference is made by the algorithms that encode domain knowledge—for instance, medical and biological—and causal reasoning. The advent of data science, and the scale it brought with it, especially the large amount of data that has given rise to health data science in the last years, offers the possibility to change this data-driven view.

Most up-to-date techniques applied in biological sciences and medicine, therefore, can mostly work in predictive modes based on data, hence they are not in the position to achieve goals based on precision, for instance, timely evaluation and matching of the patients with the most appropriate preventive and therapeutic measures. A deeper familiarity with machine

learning and its processes will be necessary for the correct data interpretation and implementation of innovations in health care.

Merging machine learning algorithms with domain expertise in the medical field and further attaching them to causal reasoning is paramount to lead to a transformation in qualitative medicine that will result in better patient outcomes, especially when "big data" is increasingly becoming a tool for the analysis of quantitative information. Anyway, there are several conventional means by which to relate climate change to human health. The results are, however, variable based on what may be used as the details to be analyzed and the amount of data being used. This thesis uses a machine learning-based framework to measure the relationship between selected indicators of climate change and national-scale mortality and project what might happen in the future, given the best-performing models.

3.2. Machine Learning and Causal Discovery

Machine learning uses multilayer neural networks to increase performance and validate results. Machine learning algorithms based on linear regression models are engaged to make up for the limitations of methods based on neural ones. The purpose of Causal Discovery is to infer the underlying causes from the observational evidence. Precisely, in this work, causal discovery techniques are used to investigate and explore the causal pathways through which climate change affects human health, as measured by the number of deaths due to the main diseases, as classified by the causes of death (frequency) in Europe. In most scientific fields, nowadays, a big stream of research is devoted to solving the problem of selecting the causal features and reconstructing the interaction networks from multivariate observational time series. In large part, two reasons cause this interest.

The first one is the huge amount of observational time series data that is nowadays available, in the era of big data. The second one is the research in scientific fields in which controlled experiments are impossible, unethical, or very costly, such as the climate, earth systems, or the human body. Studies based on correlation over pairwise association networks cannot be causally interpreted. More critically, the problem of causal network reconstruction goes beyond that of inferring association and directionality between two-time series.

The problem dealt with in causal discovery is that of distinguishing between direct dependencies and indirect ones and between common factors and time series. The drive for such reliable artificial intelligence systems has fueled causality techniques in machine learning studies. As pointed out by (Pearl, 2018), it is causal reasoning that gets around the

shackles that the current ML systems have. The backbone of the typical ML algorithms is the correlation between variables, which does not assure robust causal structures and, in consequence, can obtain inappropriate, biased, or even destructive inferences.

3.3. Methodological Framework

This methodological framework is adopted from the study by Pizzulli et al., 2021. It utilizes individual neural networks, machine learning tools, and causal discovery tools to explore, investigate, and understand the issue of climate change in the European context. Moreover, mapping the findings will assist in visualizing the impact evidence from observational data. For GIS software visualization, ArcGIS and Panoply will be used. The use of these tools adds to the existing methodological framework presented by Pizzulli et al., 2021.

3.3.1. Data Generation Process and Data Description

The following section explains the data generation process for the datasets.

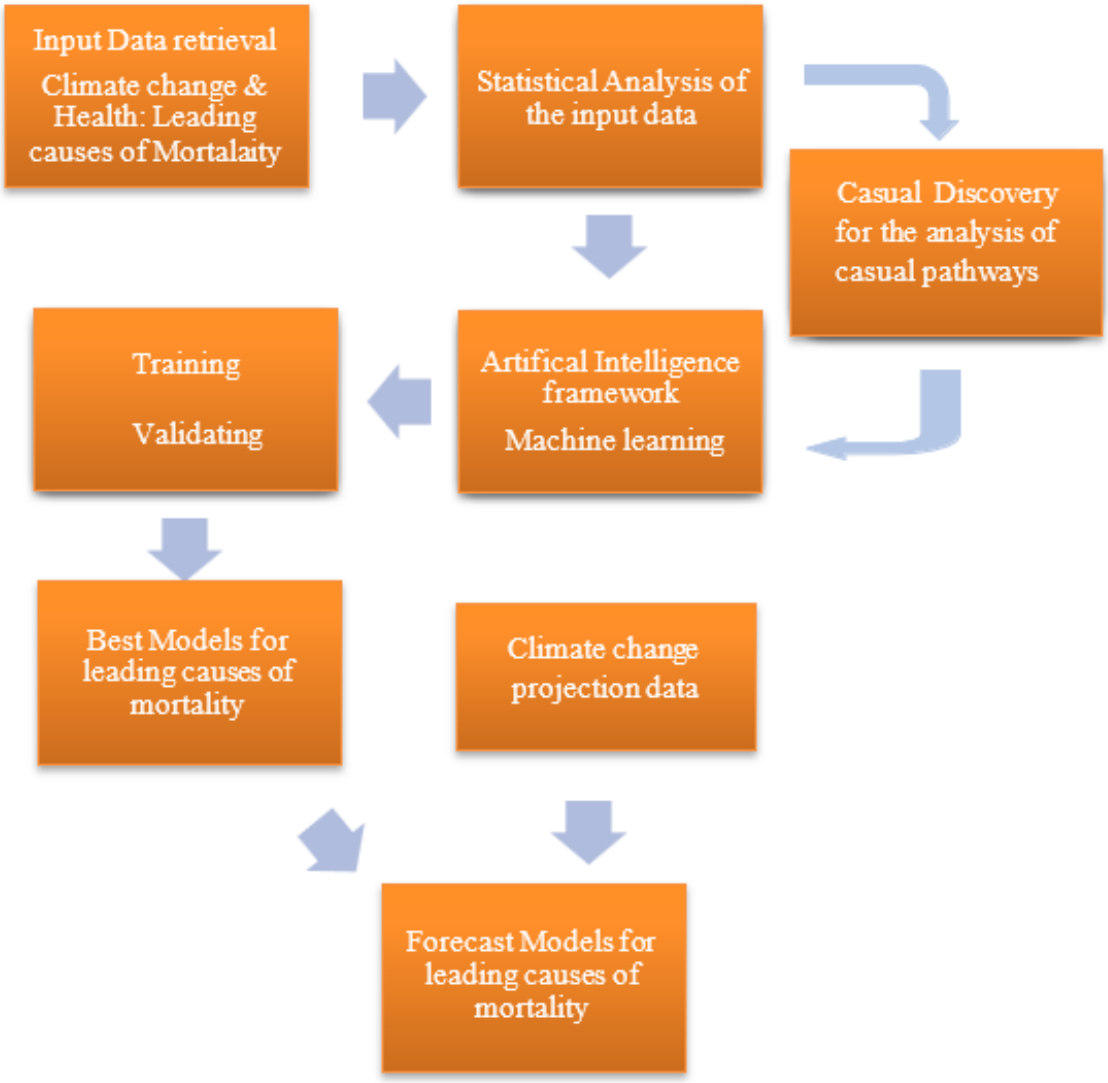

Figure 3.2: Data Generation Process

3.4. Data Generation Process: CMIP6 Model and NASA-GISS-E2-1-H

Climate models are very key to scientists in researching past and future climate changes. These are employed by conducting complex simulations of the Earth's atmosphere, land, and oceans with the world's biggest supercomputers, in turn generating climate projections. Climate models render in equations the physics, chemistry, and biology underlying the Earth's climate system.

As the climate system is complex and the computing power is limiting, it is required that the Earth is divided in a climate model into a grid of boxes or "grid cells." Consideration is given to temperature, air pressure, humidity, and wind speed at particular points within each grid cell in several layers over and through the atmosphere and oceans. Model spatial resolution is the size of grid cells used by the model in its computation. Global models use around 100 km cells in longitude and latitude at mid-latitude.

These climate models produce petabytes of information and cover many variables over space and time, such as temperature, clouds, and the salinity of the ocean. All these models have to incorporate outside, uncontrollable factors that alter the amount of solar energy absorbed by the Earth or trapped by the atmosphere. These are generally recognized under the name "forcings," which include changes in solar output, long-lived greenhouse gases such as carbon dioxide, methane, nitrous oxide, halocarbons, and aerosols released from the combustion of fossil fuels and wildfires, and volcanic eruptions. A further effect of aerosols, in addition to the effect on cloud formation, is reflected in their effect on incoming sunlight. Climate models can provide such a level of detail for the Earth's climate—a plurality of variables, covering a very broad range of time

intervals: from atmospheric temperatures and humidity profiles from surface to upper stratosphere, to oceanic parameters of temperature and salinity, to estimates of snowfall, precipitation, snow cover, and extent of glaciers and ice caps, to wind speed and direction and phenomena such as the jet stream and ocean currents that are representative of the climate.

General circulation models are also referred to as global climate models. These models apply to the basic physics of the climate system. The most recent models of this family, the Earth System Models (ESM), describe biogeochemical cycles and their interactions with the climate. ESMs consider the carbon and nitrogen cycles, atmospheric chemistry, ocean ecology, vegetation dynamics, and land-use changes, thereby including all processes in which the climate responds to anthropogenic greenhouse gas emissions. The climate response of vegetation to temperature and precipitation, with feedback on the exchange of carbon and other greenhouse gases with the atmosphere, is a part of an ESM.

3.4.1. Coupled Model Inter-comparison Project (CMIP6)

CMIP stands for Coupled Model Intercomparison Project; this gives a protocol that is designed by scientists for the comparison of climate models. It is run in a way that the analysis, validation, and improvement of GCMs are done systematically. It treats the Earth's climate system as a real dynamical system. CMIP6 is the scope and the sixth phase of it, making it global. For purposes of carrying out an up-to-date comprehensive analysis of the climate system, CMIP has better models, with much-improved computing capabilities, and a variety of scenarios. The following note provides an overview of CMIP6, including its purpose, salient features, and contribution to science (Studies (NASA/GISS), 2018).

CMIP6 is home to numerous climate models from various research institutes worldwide. In addition, the CMIP6 archive encompasses 35 diverse models from different modeling centers. This ensemble of models covers a broad gamut of approaches, resolutions, and parameterizations in support of the global coordinated effort to simulate and understand the Earth's climate system.

Experiments: In CMIP6, experiments represent a suite of coordinated simulations directed at addressing specific science questions regarding exploring various aspects of climate variability and change. The total number of experiments conducted in CMIP6 is not pinned down and depends a lot on the scientific objectives and interests of the participating modeling groups. Experiments span a broad range of climate-related topics including historical simulations to reproduce past climate conditions, future projections under different emission scenarios, and

specialized experiments targeted at specific components of the climate system.

The main goal of CMIP6 is to enhance our knowledge of the climate system of the Earth and the interactions that take place within the system, in addition to the influence of the interactions on regional and global climate patterns. Since the CMIP6 ensemble includes many models, its projections have robust data, and those data are very important in informing policymakers and researchers on the ground who will come up with strategies for intervention and adaptation to climate change. These projected future scenarios based on the expected concentrations of greenhouse gases, aerosols, and other climate forcings help evaluate the potential future climate conditions. In addition, CMIP6 is a necessary framework to conduct and coordinate climate model experiments, allowing the scientific community to explore the climate system of Earth. Improved models, different experiments, and updated projections from CMIP6 improve knowledge of climate variability and change and support informed decision-making when the world faces environmental challenges.

3.4.2. NASA-GISS-E2-1-H Model NASA Goddard Institute for Space Studies E2.1

This is one of the widely recognized models of CMIP6. The horizontal resolution of the model is 2.25o x 2.8o, briefly named as NASA-GISS-E2.1-H. It's a model of Earth's system and includes all important subsystems and physical processes of interaction between the atmosphere, oceans, land surface, and cryosphere. The model is very spatially detailed and includes the best available parameterizations and observational data measurements that are necessary for the improvement of the representation of the Earth's climate system. The model allows us to generate climate projections and analyze emission scenarios, using the observational data for the fine-tuning representation of the current climate state.

Such models, like NASA-GISS-E2-1-H, make outstanding contributions to the development of the science of climate and knowledge about regional climate patterns, climate sensitivity, and the consequences of greenhouse gas emissions for global and regional climates, which are used by policymakers and scientists in developing adequate strategies to mitigate and adapt to climate change.

3.4.3.Data Generating Process: WHO Causes of Death dataset

The World Health Organization's dataset on causes of death is one of the most important and largest sets of data on the details of diseases of the world, including causes of death. The process of generating it goes through a series of elaborate steps, which result in the data's accuracy, uniformity, and representativeness.

Data harmonization by WHO ensures comparability and consistency across different sources and countries. This means harmonizing discrepancies in coding systems, disease classifications, and data formats using standard methods and definitions. This way, WHO develops a harmonized and standardized data set that allows making meaningful comparisons across countries and sources.

Wherever some degree of uncertainty is always found, WHO, moving toward an all-embracing approach, takes this in the context of mortality data. It applies statistical methods for the quantification of uncertainty intervals and validation techniques representing the precision and reliability of the estimates lying within the dataset. In this manner, a transparent approach is laid down where the users interpret the data given uncertainties.

The WHO Mortality Database holds the data reported by Member States for deaths and causes of death. However, the quality of the data reported to the WHO Mortality Database varies greatly across different countries and over time. The quality of mortality data has two major dimensions: completeness and quality of cause-of-death information. Completeness is

the percentage of deaths for which registration includes cause-of-death information.

Even when deaths are registered, information about their cause may be missing incomplete, ill-defined, or erroneous. The proportion of deaths assigned to one of a shortlist of leading garbage codes—a cause that cannot be a valid underlying cause of death or is ill-defined—is used as an indicator of the quality of cause-of-death information. Together, completeness and ill-defined percentage of deaths can be used to determine the quality of the data. An indicator called usability, which combines both dimensions of quality, gives the percentage of deaths that are registered with meaningful information on the cause of death.

3.4.4. Geographic Area of Interest and Time Interval

Regarding the geographic area, a zone encompassing the following extremes was selected: A(52.3730; -9.1365), B(52.3730; 28.9758), C(36.1285; 28.9758), D(36.1285; -9.1365). This area covers approximately 93% of the European surface, and for meteorological purposes, annual averages of the zone using the enclosed cells have been calculated.

Figure 3.3: Geographic area

Statistics related to deaths have been extracted from 24 countries (Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, France, Germany, Greece, Hungary, Italy, Luxembourg, Malta, Netherlands, North Macedonia, Poland, Portugal, Republic of Moldova, Romania, Serbia, Slovakia, Slovenia, Spain, Switzerland) enclosed in this area. Some countries were not considered because they did not provide data continuously.

For the temporal interval, a period of 30 years from 1990 to 2019 was considered, as this allows for continuity of both meteorological and public health data.

3.5. Data Source

A single source of data on climate change is selected, namely NASA GISS E2-1-H (Studies (NASA/GISS), 2018), and another on causes of death from WHO. The dataset is selected based on the reliability and veracity of this source. The data on climate change are selected by directly extracting the data from the Historical experiment from the CMIP 6 database for data up to 2014 and SSP2-4.5, data from 2015, and associated future projections (See Appendix A for a detailed discussion on the CMIP project) on the basis of the selection of the factors based on a literature review that caused the most significant impact on the human physical health globally and in Europe. The NASA CMIP 6 database is the database of the collection of various kinds of climate data. For this study, the data set chosen is in the form of a forced model of worldwide climate change.

3.5.1. Climate Change Factors

Here is a list of climate change drivers/factors:

- **Surface Temperature (Tas):**

Temperature of Near-Surface Air (Tas) is a main variable in CMIP6 model experiments for the estimation of climate change. It represents an average of the Earth's surface temperature about 2 m above the Earth's surface. Climate models estimate Tas through physical processes such as radiative transfer, energy balance, and heat exchange between the atmosphere and the Earth's surface. Projections of Tas under different scenarios of emission, for both global and regional temperature patterns, play an important role in order to evaluate which could be the level at which the climatic warming will be, with which type of distribution. Later also referred to as temperature anomaly

- **Carbon Dioxide (CO2):**

The initial condition of the simulation of climate change is the concentration of carbon dioxide in the atmosphere. Carbon dioxide is important as it represents the single most important variable in CMIP6: the radiative effects of enhanced greenhouse gas concentrations by human activities. Historical and future scenarios of CO2 emissions are adopted by climate models as a basis for the study of its effects on the Earth climate system. Models used involve interactions between the atmosphere, Earth's surface, and oceans, including CO2 emissions from fossil fuel combustion, land-use changes, and natural processes. Projections of climate change require proper representation of CO2.

- **Atmospheric Methane (CH4)**

Another important greenhouse gas is atmospheric methane (CH4). Concentrations of CH4 are considered and simulated for CMIP6 radiative forcing and climate variability experiments. CH4 is emitted both naturally, for example, via microbes in wetlands, and anthropogenically, for example, via agriculture and fossil fuel use. In CMIP6, the climate models consider the complex interaction between the atmosphere and the biosphere and anthropogenic sources in order to take into account the temporal and spatial variability of CH4 concentrations. The current climate model requires consideration of CH4 dynamics.

- **Anthropogenic Forcings:**

Anthropogenic forcings refer to the entire array of human-induced factors modulating the climate system, such as greenhouse gas and aerosol emissions, land use changes, and the more general alterations of Earth's atmospheric energy balance made by humans. In the CMIP6 experiments, the anthropogenic forcings are implemented with models to derive the estimates of their contribution to climate change. That is, the quantification of its historical and future scenarios in driving climate change in terms of the surface temperature and precipitation pattern, among other climatic variables, will be better estimated. An accurate representation of anthropogenic forcings is imperative for the understanding of the driving processes of climate change and the development of effective mitigation strategies.

3.5.2 Motivation for Variable Selection

Selection of the factors leading to climate change is premised on a literature review of the changes that have been experienced in the climate. The setting of the IPCC AR-6 is laid on factors that will lead to existing and future climate changes. Temperature, CH4, CO2, and anthropogenic forcing (ensemble for all the forcings) are the leading determinants.

CHAPTER 4 - Used Model data Analysis -

In this part of the thesis, the model for analysing the influence of anthropogenic and climatic variables on human health was developed, following the study conducted by Pizzuli. The four input variables of the model, chosen to represent the anthropogenic and climatic parameters, are 'Anthropogenic Forcing,' ' Anomaly Temperature ', 'CO2,' and 'CH4.' A selection of these variables was done because they play a very significant role in the context of climate change and greenhouse gas emissions.

There are five output variables, and these are chosen out of five categories of patients and diseases- 'Victims Without External Causes,' 'Respiratory Diseases,' 'Digestive Diseases,' 'Mental and Behavioral Disorder,' and 'Nervous System Diseases.' The architecture selected for the neural network was; input layer consisting of four neurons, hidden layer with ten neurons, while 5 neurons are further included in the output layer.

The model has been trained through 200 epochs over a dataset that is divided into 60% for training, 20% for validation, and the remaining 20% for testing. The model has been partitioned in this ratio to have a proper balance between the learning phase and its power to generalize on unseen data. Data, taken for the purpose of training or testing, have been preprocessed and normalized so that the network can 'feel' every input equally.

The simulation process was run, which showed different iterations for the improvement of the model. The count of neurons in the hidden layer is increased to 40, and a value of selected learning rate parameter is 0.008 is to optimize the learning of the model. Other than that, overfitting prevention techniques are used, mentioned further in the proceeding

sections. This adopted method of approach, therefore, seeks to build a robust and accurate model for independent validation that will help in providing a detailed understanding of the interactions between climatic variables and human health, thus making substantial contributions to the field of research in data science applied to environmental and health issues.

4.1. Neural network using

The implementation, or creation, of a neural network in Python will be based on some powerful and versatile libraries with specific functionalities for handling, processing, analyzing, and visualizing data. Among the listed libraries are Pandas, NumPy, Scikit-learn, PyTorch, Xarray, NetCDF4, Matplotlib, and Seaborn. Down here are detailed the sections of such libraries in the development process.

Pandas is a central data manipulation and analysis library. It is used to read data from a variety of sources, such as CSV, Excel, SQL, and many others, and to create powerful data structures like DataFrame and Series, within which data can be cleaned, filtered, or transformed. For instance, missing values should be handled, the data should be converted into normalization, and only then should it be fed into the neural network.

NumPy is a numerical base package for Python that supports multidimensional arrays and complex high-level mathematical functions. It is most used for performing linear algebra operations, mathematical transformations, and any other numeric manipulations needed during the preparation of the data before feeding it into a neural network. When there is a need to compute fast, perform heavy computations on large datasets, and do that within a very short period of time, NumPy will easily come to mind.

Scikit-learn (sklearn) is a machine learning library with some data modeling capabilities. It ships with functions to split the data into the training, validation, and test sets; standardize the data; and automatically select the most important features. Scikit-learn also gives access to many machine learning algorithms, techniques for cross-validations, and tools to

construct predictive models that will be helpful in estimating model performance.

PyTorch is unarguably the most popular framework when it comes to developing deep neural networks. It provides a flexible and dynamic interface in building, training, and evaluating deep learning models. PyTorch supports tensor computation just like NumPy arrays but with the benefit of acceleration on hardware like GPUs. It also has multiple modules for defining neural network layers, loss functions, and optimizers, tools for training, which are going to make the process of training faster, and sharply improve the efficiency and intuitiveness of model development.

Xarray is a library for operating on multi-dimensional data with high-level indexing and computation tools. These features offer meteorologists and climatologists a very effective way to analyze data. More importantly, Xarray builds on top of both NumPy and Pandas and allows for real multi-dimensional operations on data labeled with multiple dimensions. It supports more complex analysis and transformation operations than data already classified as multi-dimensional. It is beneficial in the analysis of data from multidimensional scientific data files in NetCDF format.

NetCDF4 is a read-and-write library to files formatted in NetCDF, which is a common format in storing multidimensional scientific data. This library will avail data in such files for post-processing and assimilation in climatic and environmental studies.

The libraries prove to be really useful in data handling and visualization. **Matplotlib** is a plotting library that contains a wide set of features for plotting, including line, histogram, and scatter plotting.

Seaborn builds on this and provides a high-level interface for implementing attractive and informative statistical graphics. These libraries allow one to carry out exploratory data analyses on variable distributions and model performance trends during training, thus making the interpretation of results easy.

Development Process

The data will first be collected and prepared with the help of a neural network developed from scratch using Pandas and NumPy. In the next process, the data gets split into their respective parts: the training set, validation set, and test set with the help of Scikit-learn. Afterward, a deep feedforward neural network architecture with input, hidden, and output layers, along with the activation functions, loss functions, and optimizer, have been defined using PyTorch. The climatic and environmental data can be acquired through Xarray and NetCDF4, after which it can be manipulated within the process of model training. Finally, model performance and data distributions are visualized for useful insights into model improvement and evaluation with Matplotlib and Seaborn.

All source codes, extracted data, this thesis, and any attached materials are available online in a Git repository accessible at the following link **https://bit.ly/TESI0001** or via the QR code included in the appendix.

4.2. Data extraction

Monitoring and studying meteorological and death observations are important variables that impact human health. The present work uses two very big datasets: the meteorological dataset from CMIP6 in NetCDF format and the death data set from WHO in CSV format. For processing of each dataset, Python is used to get aggregated data by year and by geographic coordinates, following strictly methodological specifications.

The CMIP6 dataset is among the exhaustive datasets for climate data. It is in NetCDF format, self-descriptive, machine-independent, and very suitable for a large volume of meteorological multidimensional data. The WHO cause of death dataset is in the CSV file format, so commonly used in the management of tabular data, that shows in detail deaths from different causes, broken down by year and country.

Data Extraction and Aggregation

The meteorological data extraction from the CMIP6 dataset is carried out through Python libraries such as netCDF4 and xarray, the ones that allow for reading and working with NetCDF files. The NetCDF files passed their initial loading, after which the variables of interest—temperature, precipitation, among other related climatic parameters—were to be selected. Death data from the WHO dataset in CSV format was extracted using the pandas library, which provides tools for efficient operations on tabular data. It was loaded, and columns of interest were extracted for analysis. If obtained successfully, the next step was to group by year and geographic coordinates. In the case of meteorological data, data grouping for the computation of annual averages for each variable was to be conducted using the geographic coordinates (latitude and longitude).

In the case of death data, it was to be done year-wise and by respective countries. For meteorological data, the data aggregation was implemented with the xarray group by and resample methods to be able to compute annual averages for each combination of latitude and longitude. For all death data, the pandas group method was selected to group data by years and countries to compute the annual total number of deaths for each of the cause.

Year	*Anthropogenic Forcing*	*Anomaly Temperature*	*Co2*	*CH4*
mean	1,790	276,40	372,91	1.744,53
dev	0,648	55,13	72,54	342,35
min	-0,691	0,57	17,46	44,89
25%	1,658	286,81	376,14	1.782,96
50%	1,820	287,03	384,68	1.797,69
75%	2,119	287,21	396,57	1.826,28
Max	2,574	287,86	411,43	1.898,53

Figure 4.1: Statistical indicators for input data (1990–2019) from NASA and WHO (victim numbers are based on yearly global deaths).

Year	*Victims Without External Causes*	*Respiratory Diseases*	*Digestive Diseases*	*Mental and Behavioral Disorder*	*Nervous System Diseases*
mean	161.579	186.498	183.816	229.210	125.500
dev	42.279	38.919	36.076	84.273	39.542
min	27.355	20.967	4.844	86.144	42.099
25%	148.207	174.121	187.336	171.382	103.923
50%	157.337	187.925	190.593	212.043	122.663
75%	174.335	202.078	194.320	282.152	151.096
Max	252.967	228.324	197.135	371.505	185.521

Figure 4.2: Statistical indicators for input data (1990–2019) from WHO (victim numbers are based on yearly global deaths).

4.3. General Model Implementation

For model training, advanced backpropagation with Bayesian regularization was used, which provides most of the properties of Bayesian statistics and the principles of a gradient-based optimization approach to enhance the generalization ability of the model. The introduction of a regularization parameter in Bayesian learning introduces a penalty on the complexity of the model, affecting simple solutions and the mitigation of the risk of overfitting. This is especially useful in complex contexts like ours, where the input and the output variables can have non-linear and intricate relationships. For such reasons, we first used a search algorithm that combines binary search and grid search for the purpose of tuning the parameters and thereby enhancing the performances. Grid search is a systematic exploration of an enormous number of values for each parameter in order to find possibly optimal combinations. However, it can be very demanding in terms of computation. To make up for that, a binary search has been implemented for the fast narrowing down of the most promising value intervals. In this way, the hybrid approach explored the parameter space in a highly efficient manner and identified optimal combinations for both regularization and learning rate quite fast. Subsequent to this, the generalization capability of the trained model on unseen test data was drastically enhanced, while at the same time the risk of overfitting was significantly decreased. This resulted in the development of a robust predictive model that should be able to provide reliable and accurate forecasts about the effect of the climatic and anthropogenic variables on human health.

Backpropagation Algorithm with Bayesian Regularization

The backpropagation algorithm with Bayesian regularization is an optimization technique that combines machine learning methods with Bayesian statistical principles. This technique aims to improve the model's generalization capability, reducing the risk of overfitting by keeping the network weights within plausible limits based on prior knowledge.

Principles of Bayesian Regularization

Bayes' Theorem: Bayes' theorem allows updating the beliefs about the model parameters (the weights) given the observed data. The formula is:

$$P(\theta \mid D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

where:

- $P(\theta|D)$ is the posterior probability of the parameters given the data.
- $P(D|\theta)$ is the likelihood of the data given the parameters.
- $P(\theta)$ is the prior probability of the parameters.
- $P(D)$ is the marginal probability of the data.

Prior and Posterior Distributions:

Prior Distribution ($P(\theta)$): This represents the initial knowledge about the weights before observing the data. It is often assumed that the weights follow a normal distribution with zero mean and small variance.

Posterior Distribution ($P(\theta|D)$): This is the updated distribution of the weights after observing the data. It integrates the prior distribution with the information provided by the data. The backpropagation algorithm with Bayesian regularization combines the standard loss function with a regularization term based on the prior distribution of the weights. The

regularization term can be seen as an additional penalty on large weights, pushing the model towards simpler and more plausible solutions.

Calculating the Loss Function: The total loss function combines the standard loss (e.g., mean squared error) with a Bayesian regularization term:

$$\text{Loss} = \text{Loss}_{\text{standard}} + \lambda \cdot \text{Loss}_{\text{regularization}}$$

where λ is a scaling factor that controls the importance of regularization.

Regularization Term: The Bayesian regularization term can be approximated using a logarithmic function that penalizes large weights:

$$\text{Loss}_{\text{regularization}} = \sum_i \log(1 + \beta \cdot \theta_i^2)$$

where β is a parameter that controls the sensitivity of the regularization and θi are the model weights.

Backpropagation: The backpropagation algorithm is used to calculate the gradients of the total loss function with respect to the model weights. These gradients are then used to update the model weights, minimizing the loss function. In Python, Bayesian regularization can be simulated by adding a regularization term to the loss function during training with backpropagation. Using PyTorch, we can define a feedforward neural network model with a hidden layer and a dropout layer to prevent overfitting. The loss function is MSELoss (Mean Squared Error Loss), and the chosen optimizer is Adam.

4.4. Preventing Overfitting

One of the major problems experienced in the course of the development of the neural network model was overfitting. It usually took place when a model fit the training set too well and lost the power of generalizing for new, unseen data. Some implementation strategies that were adopted for overfitting were early stopping and Bayesian regularization.

Early stopping means monitoring the performance of the model on a validation set during training and stopping the process if performance worsens. It generally prevents the wastage of additional training cycles and reduces the chance of overfitting.

The second method was done by running the backpropagation with an additional Bayesian regularization algorithm.

The current implementation, therefore, marries the Bayesian approach to the treatment of stochastic inputs with gradient-driven optimization in the process of backpropagation. Therefore, such incorporation would also have made the generalization capacity of the model very strong. The Bayesian approach to regularization was hence implemented under the Bayesian statistical sitting to impose the penalty on model complexity begotten from the posterior probability of the model weights and hence encourage simpler solutions while reducing the risk of overfitting. It prevents the model from being too complex, hence making a very good balance between generalization and accuracy.

Moreover, the drop-out technique has been used in the neural network hidden layers, randomly omitting some neurons at some moment of training so that the model doesn't rely too much on a few neurons and can generalize better.

The Bayesian regularization algorithm can be fine-tuned using such a hybrid way of binary and grid search of parameters for optimization. For the regularization algorithm with the incorporation of Bayesian priors, the developed hybrid of binary and grid search would enhance speed in locating an optimal parameter space point obtained by combinations that would be optimal for improved model performance. Lastly, the use of the cross-validation technique enabled the verification of model performance on independent data partitions, thus stopping the process of over-optimization on one data partition. These combined techniques aid in building the robustness and reliability of the model in generalizing upon unseen test data to, therefore, provide accurate predictions in practical scenarios.

4.5. Setting Alpha and Beta parameters for the model

The algorithm searched for the best values of the Alpha and Beta parameters, and the best available values are Alpha = 0.072 and Beta = 0.11. This led to a validation loss of 0.304570. I left some data on several Alpha and Beta values here. The heat map is representing the best value with purple. Simulation was made several times and optimal values do not change, so it is possible to confirm robustness of the algorithm in finding the optimal parameters for the model.

Alpha	Beta	Mean Validation Loss
0,0718	0,1090	0,33023
0,0718	0,1100	0,32242
0,0718	0,1110	0,31719
0,0719	0,1090	0,32220
0,0719	0,1100	0,32486
0,0719	0,1110	0,30665
0,0720	0,1090	0,32402
0,0720	**0,1100**	**0,30457**
0,0720	0,1110	0,32036
0,0721	0,1090	0,31367
0,0721	0,1100	0,32010
0,0721	0,1110	0,33944

Figure 4.2: Grid Search Result

The heat maps describe any direct or indirect percentage correlation between the variables, and make it easy to read

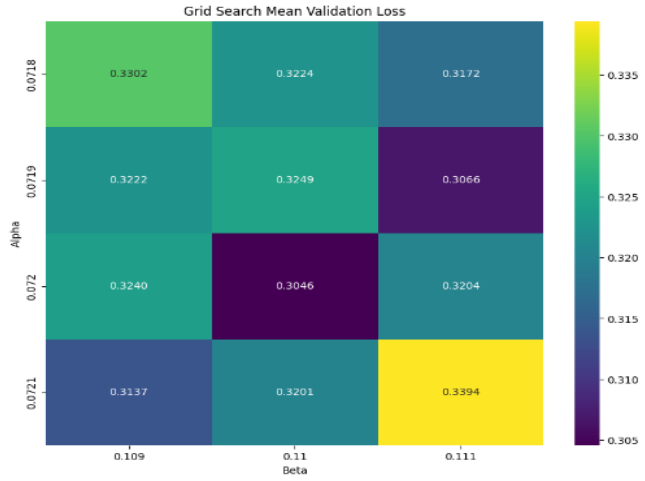

Figure 4.3: Heatmap

By adding the model parameters, we obtain the following average values of the indicators. We make initial assessments to see whether the choice of input and output variables, selected on the basis of previous global studies, can also serve as a reference for Europe or whether adjustments are necessary.

Output Variable	*Train Variance*	*Validation Variance*	*Test Variance*	*Correlation*
Digestive Diseases	0,82134	0,95552	0,89338	-0,15532
Mental and Behavioural Disorder	0,08908	0,03856	0,10365	0,93927
Nervous System Diseases	0,06803	0,04467	0,07285	0,97757
Respiratory Diseases	0,15709	0,11568	0,17058	0,88596
Victims Without External Causes	0,26004	0,24771	0,27726	0,75220

Figure 4.4: Variance & Correlation

and the associated variance and correlation graphs:

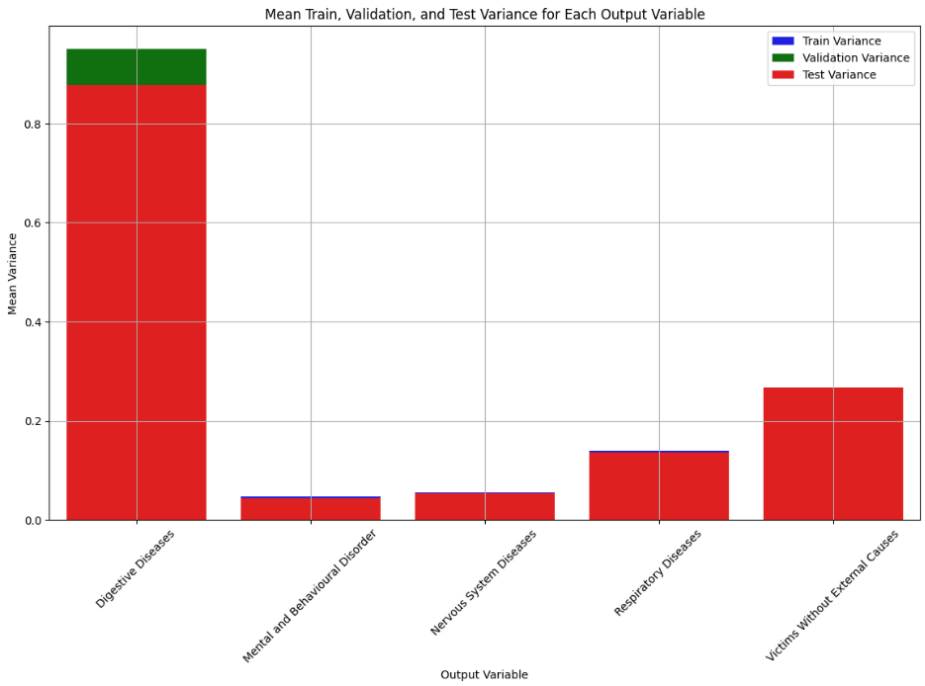

Figure 4.5: Variance

The previous bar chart shows the mean variances of training, validation, and test errors for each output variable. The variance of errors indicates the dispersion of errors around the mean.

High variance indicates that errors are widely dispersed, suggesting the model is not very precise. Low variance indicates that errors are more concentrated around the mean, suggesting greater model precision.

Next chart shows the average correlation between the model's predictions and the actual values for each output variable.

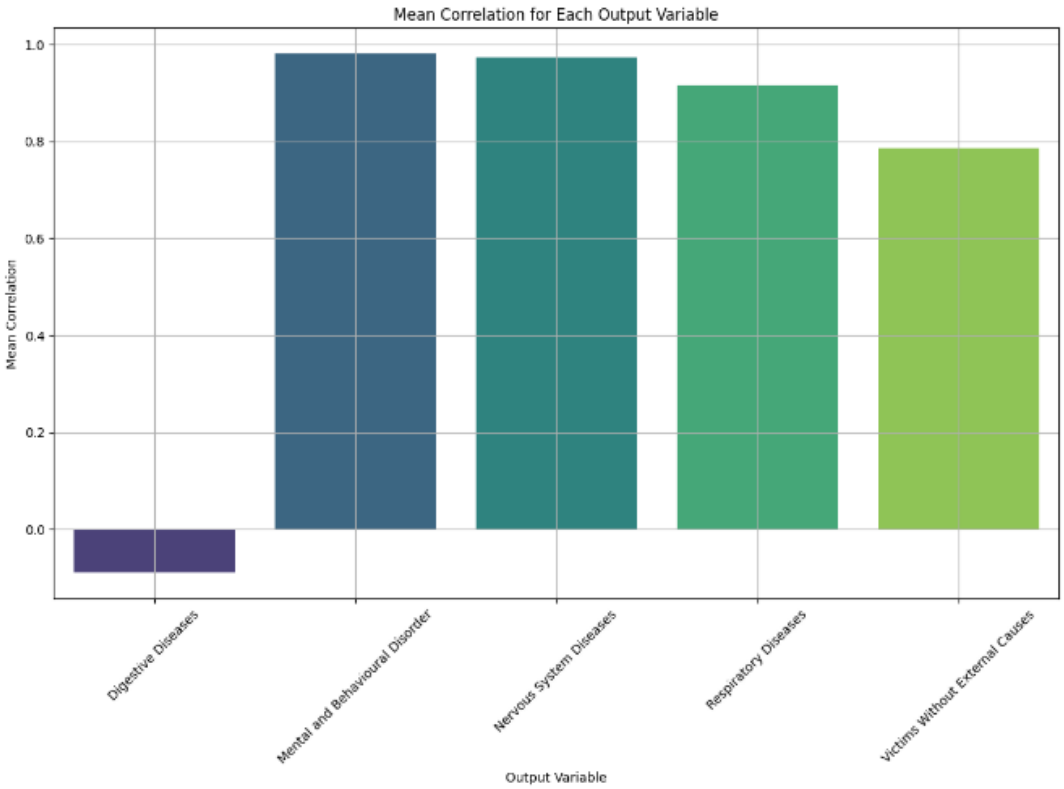

Figure 4.6: Correlation

Correlation measures the strength and direction of the linear relationship between two variables. A correlation close to 1 indicates a strong positive linear relationship, suggesting that the model accurately predicts the actual values. A correlation close to -1 indicates a strong negative linear relationship. A correlation close to 0 indicates little or no linear relationship between the model's predictions and the actual values.

4.5.1 Analysis of Each Output Variable

Digestive Diseases:

The model shows a similar average loss between training, validation, and testing, suggesting good generalization. However, the negative correlation indicates that the model's predictions do not align with the actual values.

Mental and Behavioral Disorder:

The average loss is consistent across training, validation, and testing, but a very high average correlation might indicate over-optimization or a non-linear relationship that the model has captured well.

Nervous System Diseases:

The model shows good generalization with similar average loss across training, validation, and testing, and a good positive correlation, suggesting accurate predictions.

Respiratory Diseases:

The average loss is consistent, and the high positive correlation indicates that the model accurately predicts the actual values.

Victims Without External Causes:

The model shows good generalization with similar average loss and decent positive correlation, suggesting relatively accurate predictions.

The results show that the model has good generalization for most output variables, with consistent average losses between training, validation, and testing. However, some variables show unusual correlations, which may require further analysis or model improvements. To do this, we will create

a correlation matrix in Python to verify the correspondence between each input and output and plot the graph.

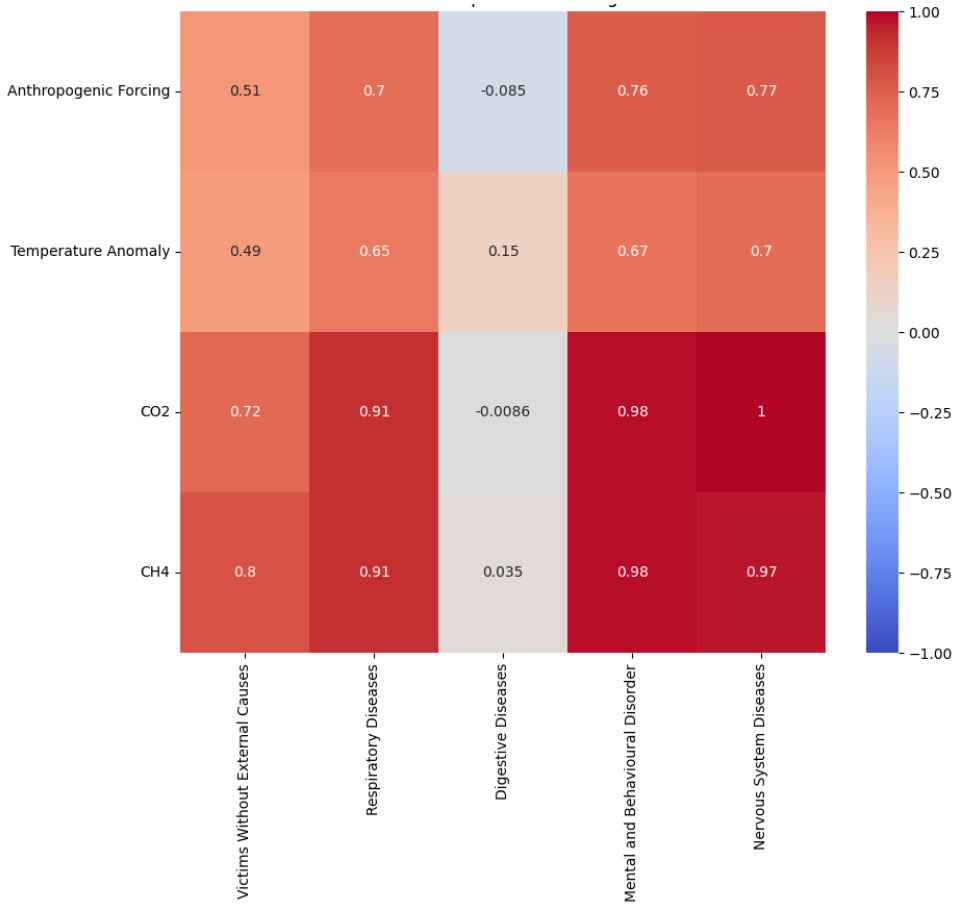

Figure 4.7: Heatmap Input / Output

As shown in the graph, the variable "Digestive Diseases" has no correlation with any input variable, so we can exclude it. We will create an accurate model for each remaining variable, excluding those with a correlation lower than 0.7. Thus, we will have four models with the following inputs:

Output Variable	***Input Variable***
Victims Without External Causes	CO2, CH4
Respiratory Diseases	Anthropogenic Forcing, CO2, CH4
Mental and Behavioral Disorder	Anthropogenic Forcing, CO2, CH4
Nervous System Diseases	Anthropogenic Forcing, CO2, CH4, Anomaly Temperature

4.6. Implementing and Analysing Neural Network Predictions

Neural networks are powerful tools for modelling complex data due to their ability to learn non-linear representations. However, once a neural network model is trained, it is crucial to analyse its predictions to understand how accurately the model performs and to identify any biases or systematic errors. In this chapter, we will explore how linear regression, the exponential kernel, and the squared exponential kernel can be used to analyse neural network predictions. Additionally, we will discuss key metrics for evaluating these methods.

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In the context of analysing neural network predictions, linear regression can be used to compare the neural network predictions $(\hat{y}_{\mathrm{NN}})$ with the actual observed values (y). This comparison can reveal how well the neural network is performing and whether there is any systematic bias in its predictions. Suppose we have a neural network model making predictions on a dataset. After obtaining the predictions, we can perform a linear regression between the neural network predictions and the observed values (y):

$$y = \beta_0 + \beta_1 \hat{y}_{\mathrm{NN}} + \epsilon$$

Where:

- $\beta 0$ is the intercept.
- $\beta 1$ is the regression coefficient.
- ϵ represents the error term.

If the neural network model is accurate, we expect $\beta 0$ to be close to 0 and $\beta 1$to be close to 1.

Metrics for evaluating linear regression include:

- Mean Squared Error (MSE): The mean of the squared errors, calculated as:

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

  Where y_i are the observed values and $\hat{y}_i$ are the predictions.

- R-squared (R^2): The proportion of variance in the observed data explained by the model, calculated as:

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

  A value closer to 1 indicates a better fit of the model.

- Mean Absolute Error (MAE): The mean of the absolute errors, calculated as:

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

A detailed analysis of the residuals (ϵ) can reveal any biases or patterns not captured by the neural network model. If the residuals show a random distribution around zero, the model is considered adequate. Conversely, if the residuals show a pattern, it may indicate that the model needs improvement.

The exponential kernel, also known as the Laplace kernel, is a kernel function used in Gaussian processes to model similarities between data points. It is often used when the functions modelling the data are expected to be continuous but not necessarily differentiable. In the context of analysing neural network predictions, the exponential kernel can be used to model the uncertainty of the predictions and evaluate their performance.

The exponential kernel is defined as:

$$k(x, x') = \exp\left(-\frac{|x-x'|}{2\ell^2}\right)$$

Where ℓ is the length scale parameter that controls the scale over which similarities decrease. By using this kernel in a Gaussian process, we can obtain probabilistic predictions that include an estimate of uncertainty.

Metrics for evaluating a model based on the exponential kernel include:

- Mean Squared Error (MSE): The mean of the squared errors.
- Root Mean Squared Error (RMSE): The square root of the MSE, providing a measure of the average error in the original units.

  $$\text{RMSE} = \sqrt{\text{MSE}}$$

- Mean Absolute Error (MAE): The mean of the absolute errors.
- Negative Log Predictive Density (NLPD): A measure of the quality of probabilistic predictions, where lower values indicate better quality.

Using a Gaussian process with an exponential kernel, we can visualize the predictions along with uncertainty bands. This is useful for identifying areas where the model is less certain about its predictions, providing insights into where the model might benefit from more data or improvements.

The squared exponential kernel, also known as the RBF (Radial Basis Function) kernel or Gaussian kernel, is another type of kernel function used in Gaussian processes. This kernel is suitable for smooth and differentiable functions and is often used to model continuous data.

The squared exponential kernel is defined as:

$$k(x, x') = \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

Where ℓ is the length scale parameter. By using this kernel, we can obtain smooth and differentiable predictions. Metrics for evaluating a model based on the squared exponential kernel include the same metrics used for the exponential kernel, such as MSE, RMSE, MAE, and NLPD. Additionally, we can evaluate the smoothness of the predictions by observing the continuity of the uncertainty bands.

Predictions from a Gaussian process with a squared exponential kernel tend to be smoother compared to those obtained with the exponential kernel. This can be advantageous for data that are expected to be generated by smooth and continuous functions. Analysing the uncertainty bands can provide further insights into the robustness of the predictions.

Linear Regression vs Exponential Kernel vs Squared Exponential Kernel

Linear regression is simple to implement and interpret, but it may not capture non-linear relationships in the data.

The exponential and squared exponential kernels are more complex and can model non-linear relationships and prediction uncertainty.

Ability to Model Non-Linearity:

Linear regression is limited to linear relationships. Both kernels can model non-linear relationships, with the squared exponential kernel offering smoother predictions.

Uncertainty Estimation and Evaluation Metrics:

Linear regression does not provide an estimate of prediction uncertainty. Gaussian processes with exponential and squared exponential kernels provide useful uncertainty bands for evaluating prediction robustness. MSE, RMSE, MAE, and R^2 are common to all techniques. NLPD is specific to probabilistic models like Gaussian processes.

When to Use Each Technique

Linear Regression: Use it if your neural network model is relatively simple and you want a quick and interpretable evaluation of its performance.

Exponential Kernel: Use it if your data are continuous but not necessarily smooth, and you want to model prediction uncertainty.

Squared Exponential Kernel: Use it if your data are continuous and smooth, and you want smooth predictions with uncertainty estimates.

CHAPTER 5 - Model Setup -

For each dependent variable, we have developed a model from the selected input variables chosen from the correlation factors calculated earlier. In the proceeding steps, we will analyze the reliability of the model based on the present values using the regression matrix. For reliable models, considerations and projections have been made for the future decades based on the values provided by the GMIP models.

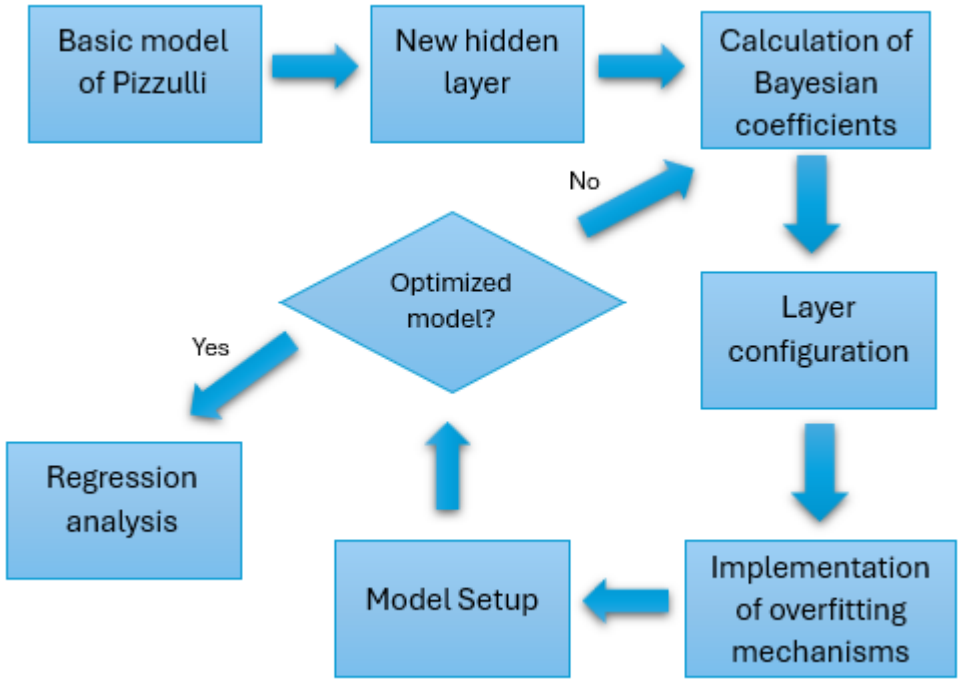

Figure 5.1: Graph of the model optimization process

5.1. Setup of the Nervous System Diseases Model

The first model created represents diseases of the nervous system, it showed correlation with all four variables although for anthropogenic forcing and temperature it does not have a very high correlation factor but allows the use of modelling techniques to create a reliable model. The first step is the creation of the neural network based on Pizzulli's study, a network with an input layer with four neurons, a hidden layer of 20 and an output layer with only one neuron. We thus proceeded with the calculation of the losses for the training, validation and testing phases with visualization of the gradient and error graphs. Based on the values and trends of the graphs, techniques were implemented to avoid overfitting such as drop out and early stopping. Yes, a further hidden layer was added, and it was evident the need to use a number of neurons in the hundreds to optimize the model also because the data was not free of noise. The algorithm uses Bayesian tuning with optimized alpha and beta coefficients.

Input variables:

Anthropogenic Forcing, CO2, CH4, Anomaly Temperature

Alfa: 0.012 **Beta:** 0.016 **Learning Rate:** 0.0001 **epoch:** 215 (early stopping)

Train R^2: 0.9843095149844885 - **Variances**: [0.01427742]

Validation R^2: 0.9531616866588593 - **Variances**: [0.03147064]

Test R^2: 0.9803068693727255 - **Variances**: [0.0239818]

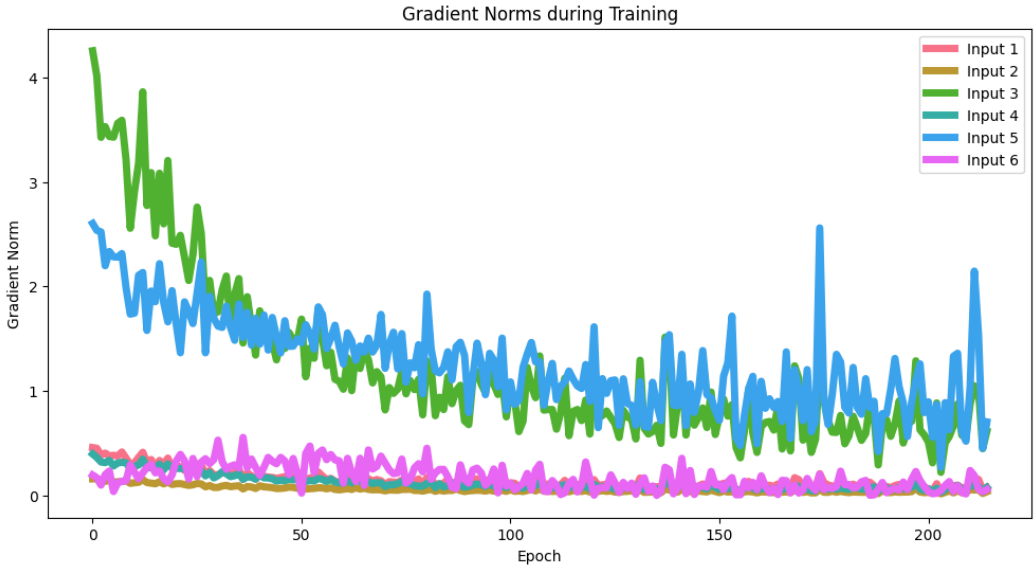

Figure 5.2

Dropout mechanism can be seen in this gradient graph, where randomly disabled neurons act on the gradients to produce jagged lines, although the overall trend is preserved.

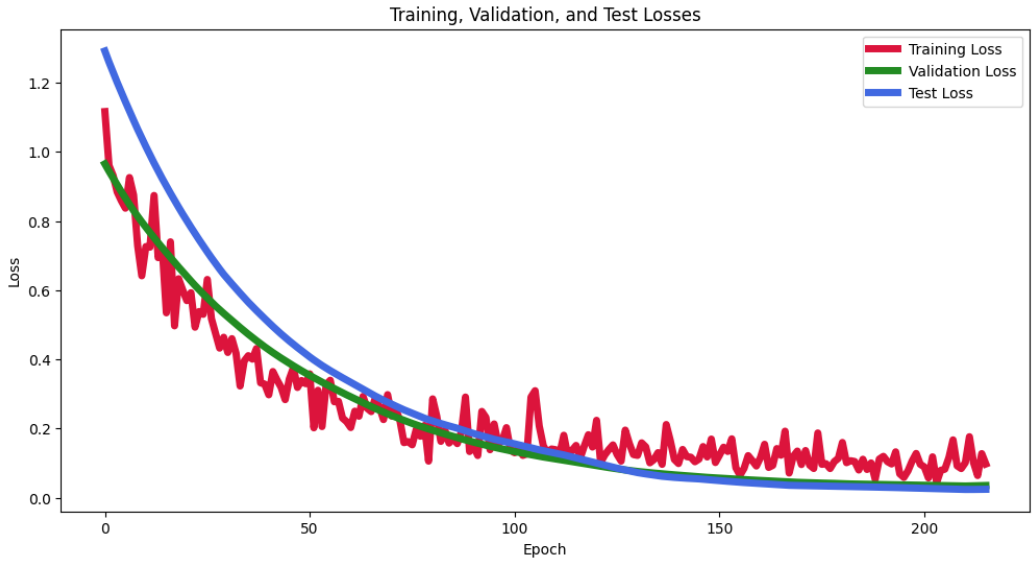

Figure 5.3

The loss chart captures in another way the nonlinearity that the dropout mechanism imposes into the model during training. This is also manifested in the trend of zeroing validation and test losses while the training loss is a bit higher to ensure the model does not overfit.

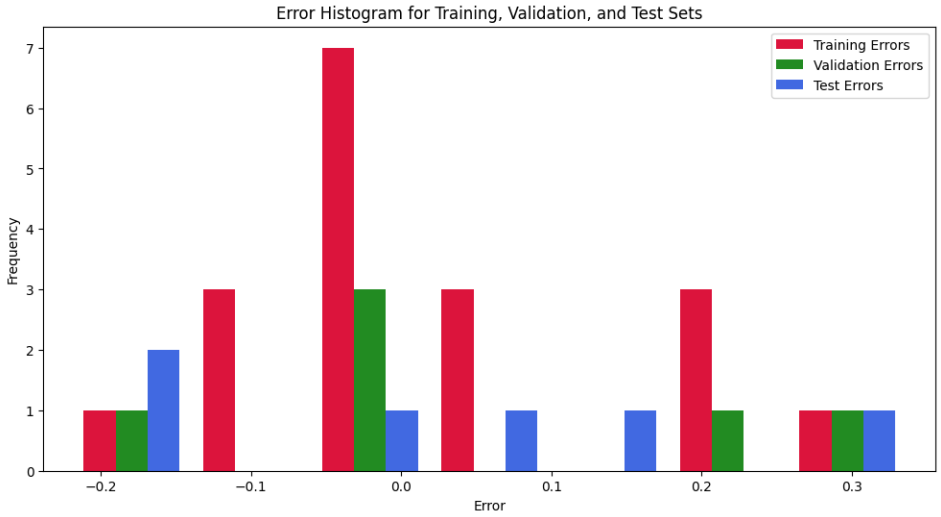

Figure 5.4

The real vs. predicted graph in itself will give a representation of how good the model is to reality.

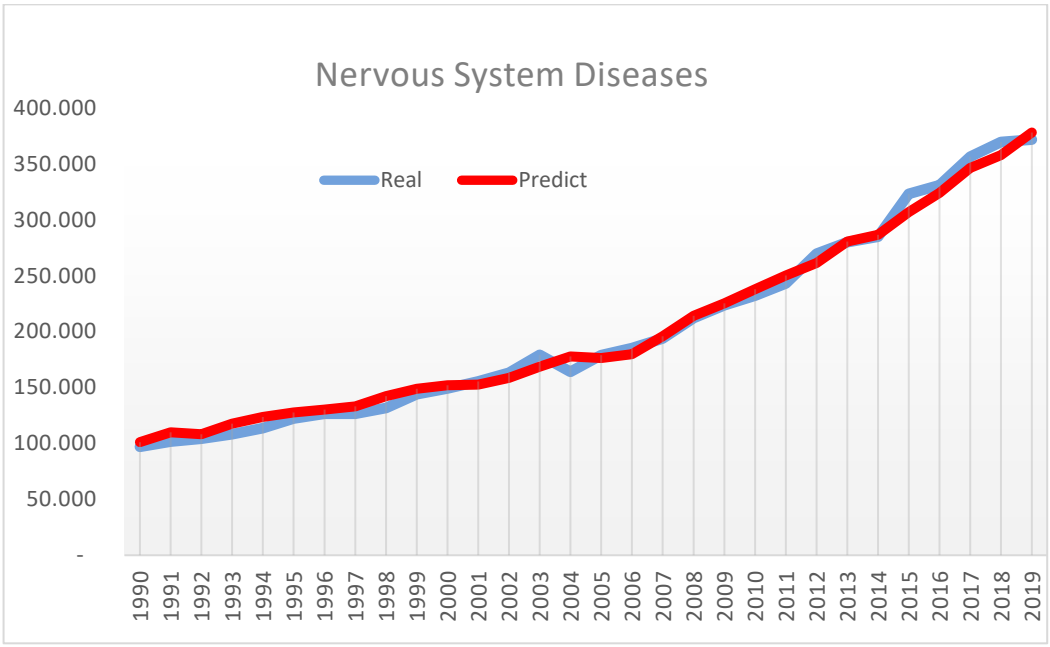

Figure 5.5

Analysis of Neural Network Predictions

Esponenzial Kernel:

MSE: 264677739.67723167 **RMSE:** 16268.919437910794

MAE: 10572.15153685522 **R²:** 0.9631029998895217

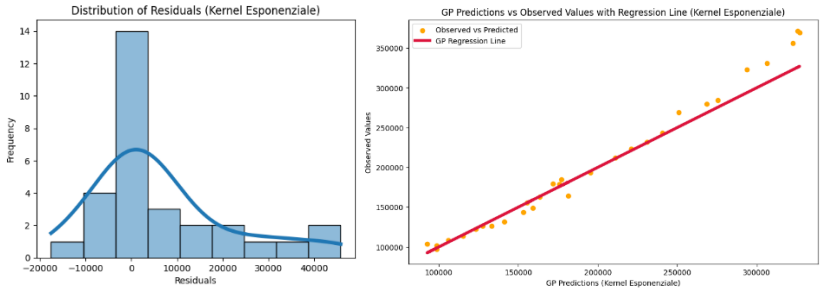

Figure 5.6

Quadratic Exponential Kernel:

MSE: 514963477.24759585 **RMSE:** 22692.80672917292

MAE: 16197.10128289085 **R²:** 0.9282122950722353

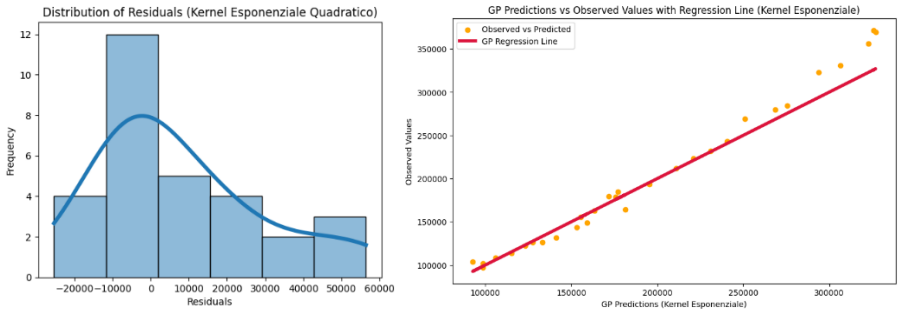

Figure 5.7

Linear Regression:

Coefficients model: [1.03699725] **Intercept model:** -4229.22810641682

MSE: 51448713.10173786 **RMSE:** 7172.77582960306

MAE: 5803.309877625103 **R²:** 0.992827869939823

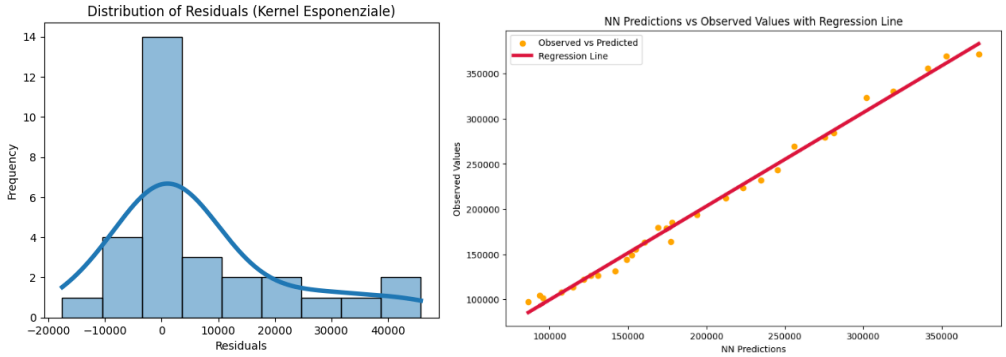

Figure 5.8

From the charts, one can clearly see that the decrease in loss has already been steady and significant when applying the set of training, which gives proof to the efficiency of the model regarding its functionality. Ideally, the good behavior of the gradients suggests that there exists no gradient explosion and vanishing. The errors made on predictions are contained, and the residuals are scattered at random across zero, which means that the model is free from any error patterns. Besides, a good R^2 value, being 0.99, explains 99% of the variance in the output data. In the conclusion, this neural network model seems to be quite effective and highly reliable for the task it was designed for.

5.2. Setup of the Mental and Behavioural Disorder Model

This model only considers three independent variables: Anthropogenic, CO2, and CH4; it does not take temperature as a variable because it is not otherwise very tightly related to this phenomenon. And the methodology evolved is the same in all the models. First, we reached the ideal quantity of neurons for hidden layers, adjusting the training rate and correctly configuring the Alpha and Beta values. The techniques implemented to avoid overfitting have been preventive in fact; to be concrete, the training of the model lasted one-fifth of the initial setup.

Input variables: Anthropogenic Forcing, CO2, CH4

Alfa 0.59 Beta 0.0125 LR 0.00023 epoch: 52 (early stopping)

Train R2: 0.9896580101922154 - Variances: [0.00872088]

Validation R^2: 0.9877709969878197 - Variances: [0.00904628]

Test R^2: 0.992531846743077 - Variances: [0.00663876]

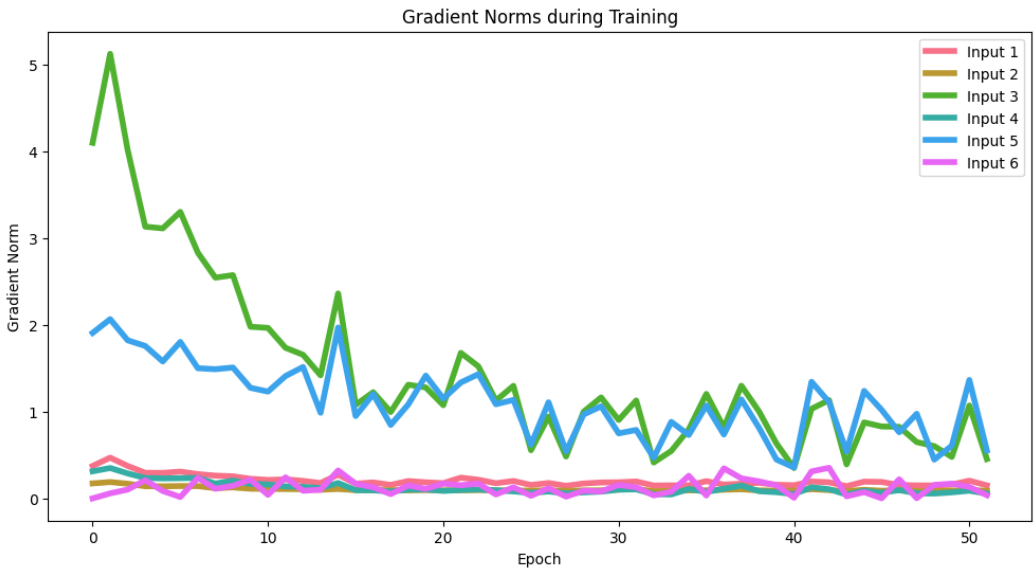

Figure 5.9

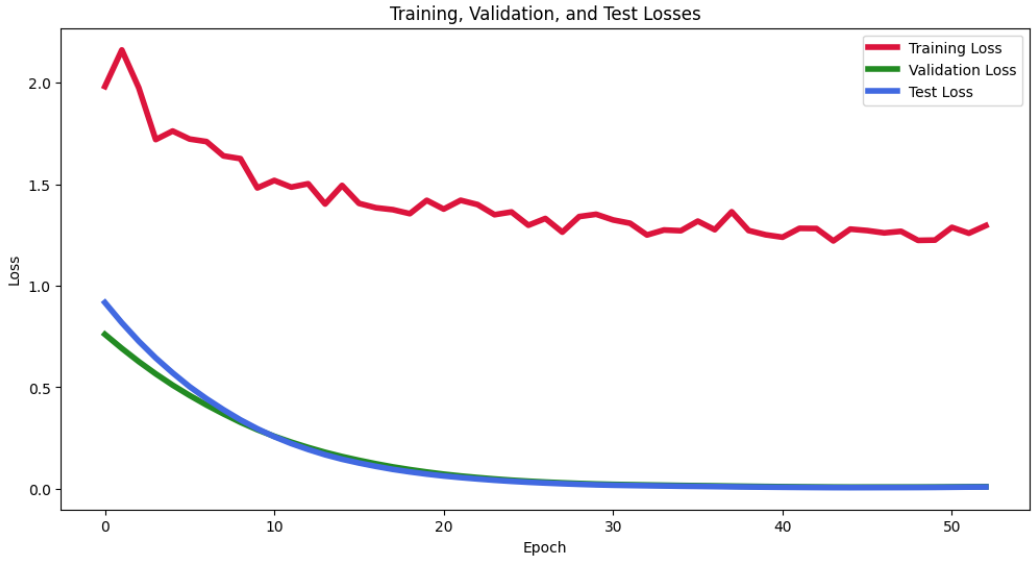

Figure 5.10

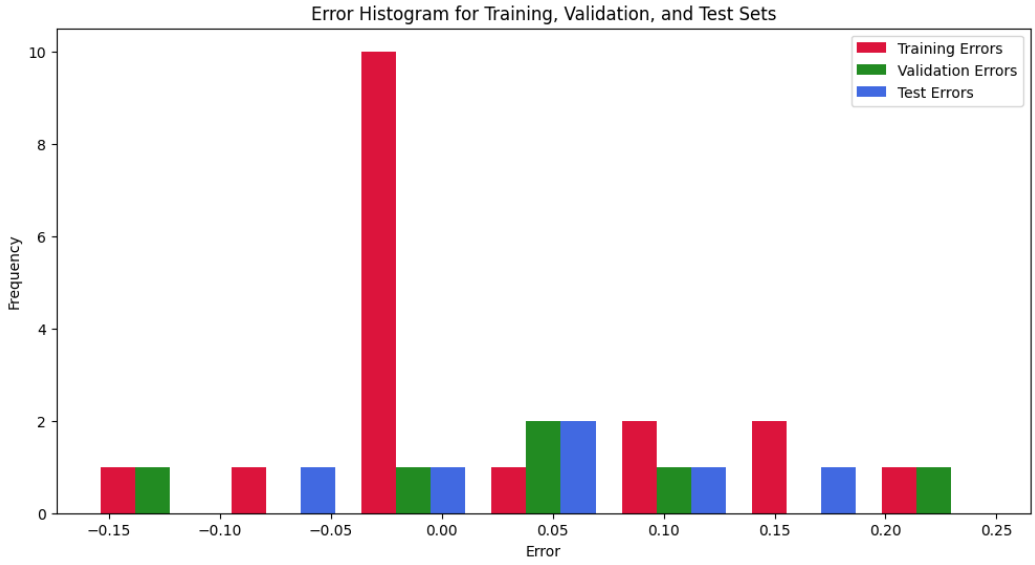

Figure 5.11

These graphs show clearly the effect of dropout, leading to a ridge-like shape, more clearly with these graphs(loss and particularly the gradient). As for the errors, they will naturally be higher for the training phase because it is a larger set, but on average, they take a very low value at the validation and test phases, giving an R^2 of 0.99.

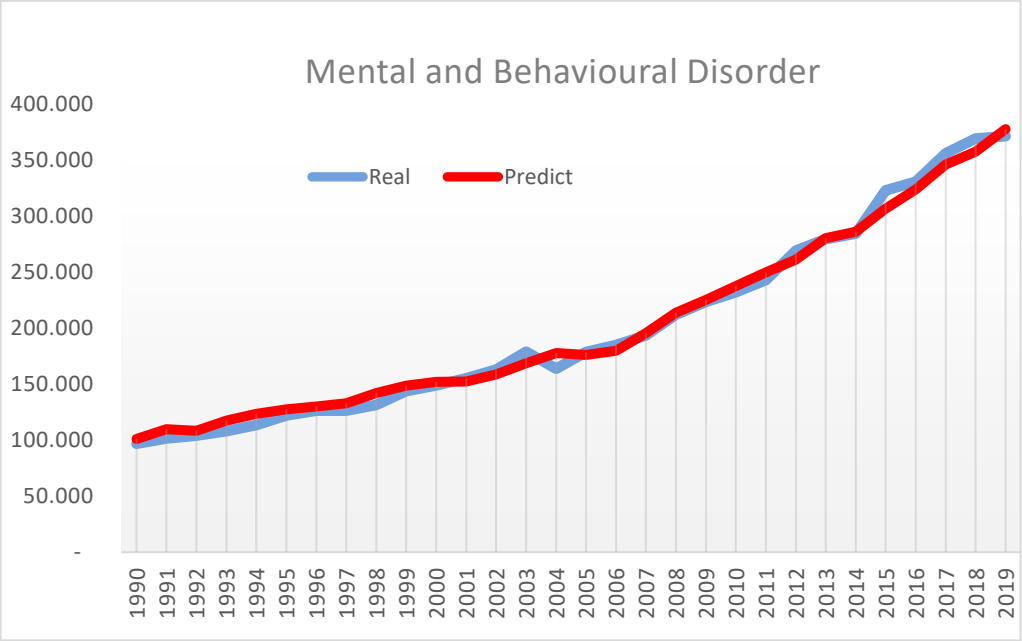

Figure 5.12

Although with this model we have an almost perfect overlap of the true data with the forecast data. Let us now illustrate the metrics of the three regressions.

Esponenzial Kernel:
MSE: 264677739.67723167 **RMSE:** 16268.919437910794

MAE: 10572.15153685522 **R^2:** 0.9631029998895217

Quadratic Exponential Kernel:

MSE: 514963477.24759585 **RMSE:** 22692.80672917292

MAE: 16197.10128289085 **R^2:** 0.9282122950722353

Linear Regression:

Coefficients model: [1.03699725] **Intercept model:** -4229.22810641682

MSE: 51448713.10173786 **RMSE:** 7172.77582960306

MAE: 5803.309877625103 **R^2:** 0.992827869939823

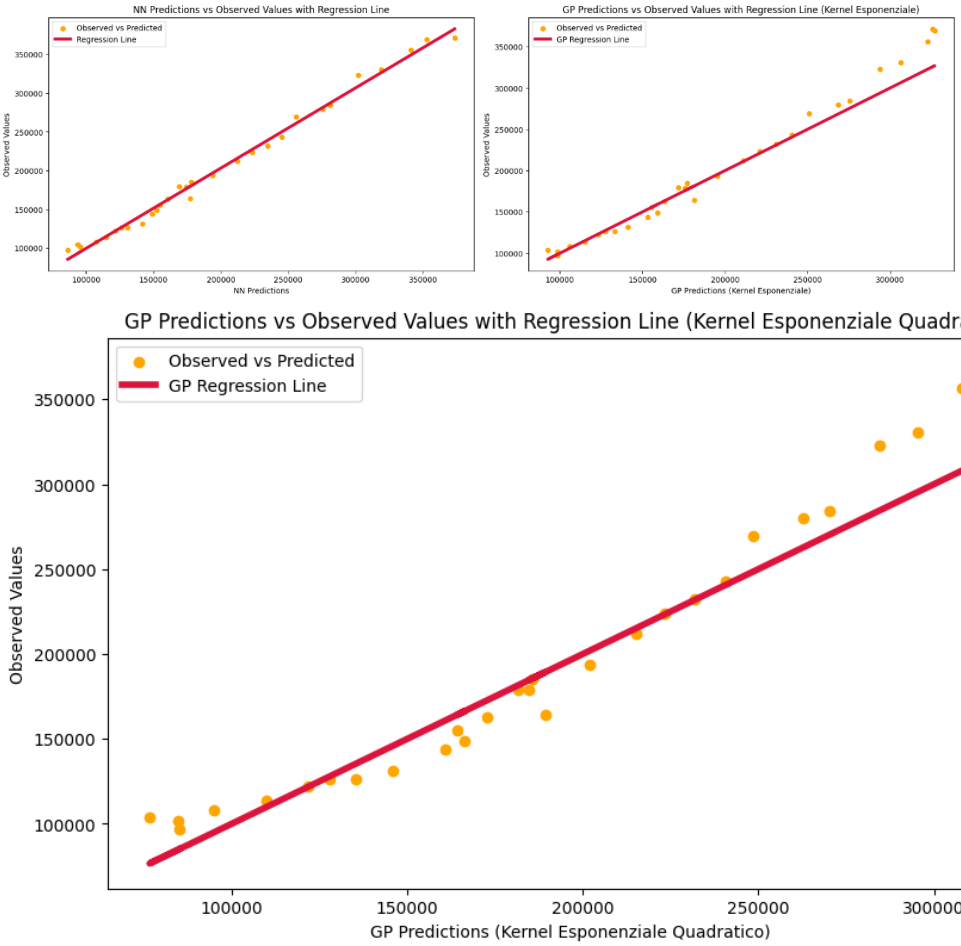

Figure 5.13

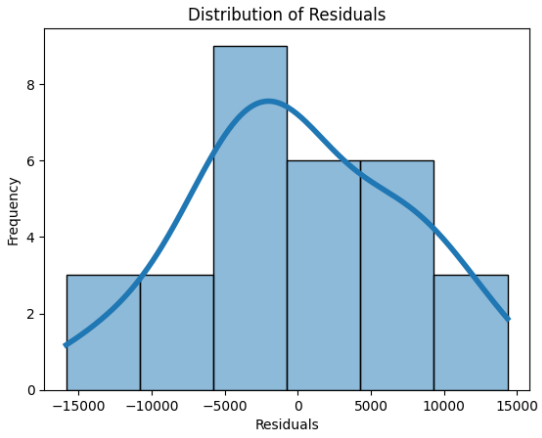

Figure 5.14

As it was in the previous case, the metric that best analyzed the model's predictions is the linear regression. As can be seen, obviously a linear regression line fits the values better since it presents a score of 0.99

5.3. Setup of the Respiratory Diseases

This model considers three independent variables: Anthropogenic, CO2, and CH4; it does not take temperature as a variable because it is not strictly correlated with this phenomenon, for reasons that we will analyze later. As the reference metrics suggest, we are far from having a model that represents reality, but it can still be useful for perceiving a trend.

Input variables: Anthropogenic Forcing, CO2, CH4

Alfa 0.01 Beta 0.05 LR 0.0001 epoch: 183 (early stopping)

Train R^2: 0.8875791728496552 **- Variances:** [0.1050207]

Validation R^2 :0.8816852420568466 **- Variances:** [0.11016586]

Test R^2 :0.8656796514987946 **- Variances:** [0.12991709]

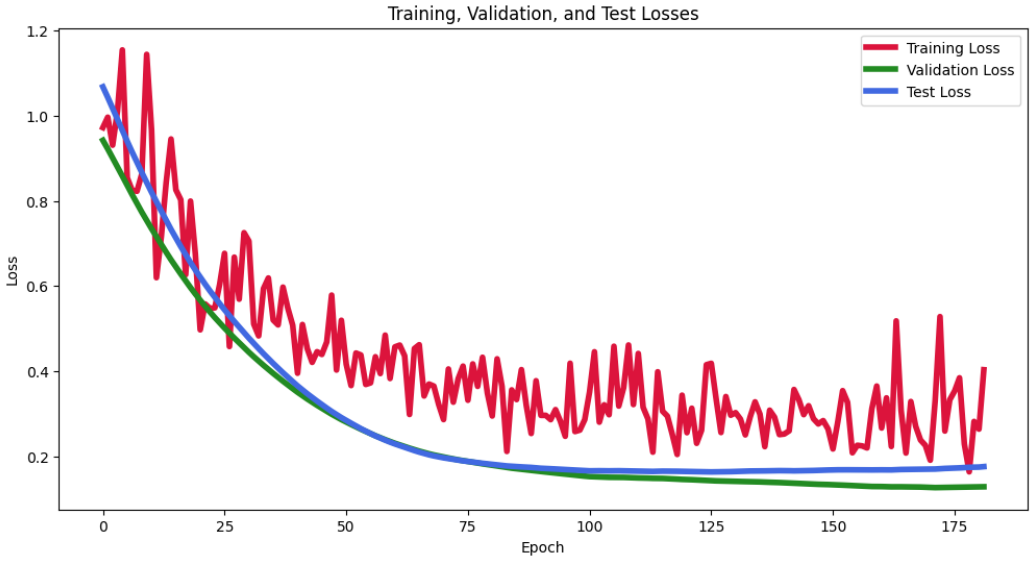

Figure 5.15

From this graph, we can appreciate how the validation and test losses change direction, a clear sign of the onset of overfitting. In this case, the control mechanisms have stopped the training.

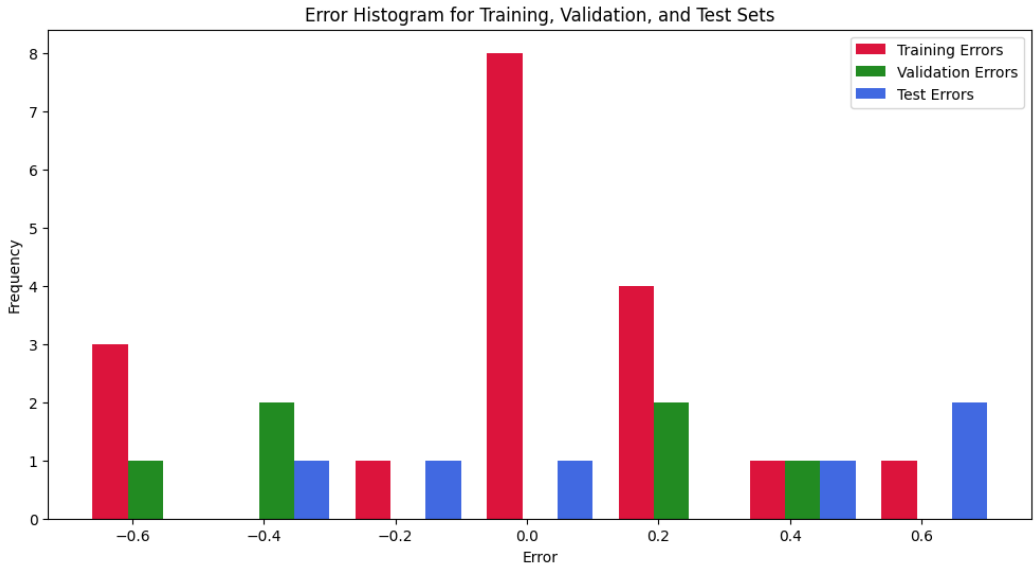

Figure 5.16

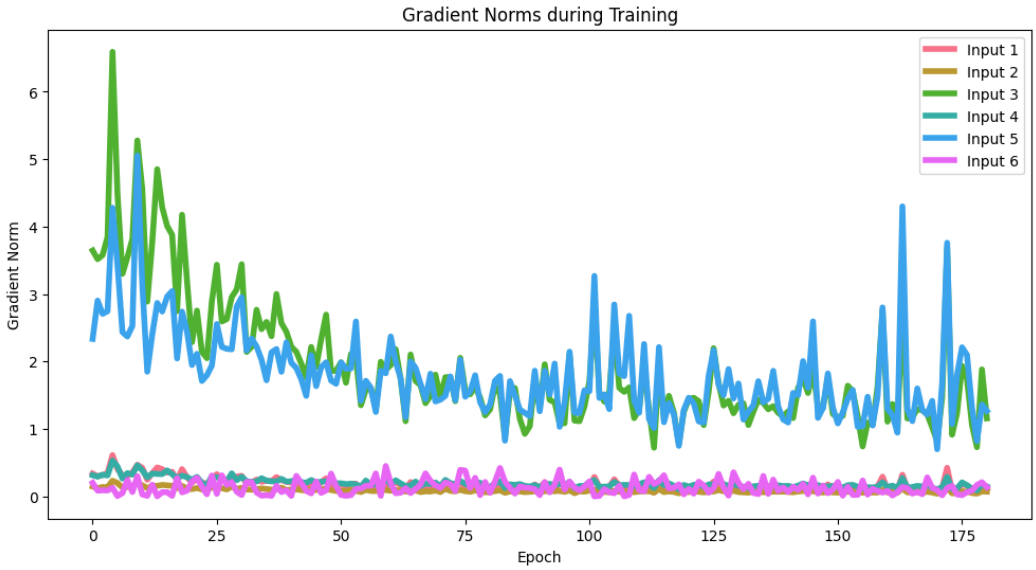

Figure 5.17

From the error graph, it is clear that their distribution is constant in all directions, indicating errors spread across all phases. The gradient graph,

even if "disturbed" by dropout, is not descending due to insufficiently representative training data, given the data variability.

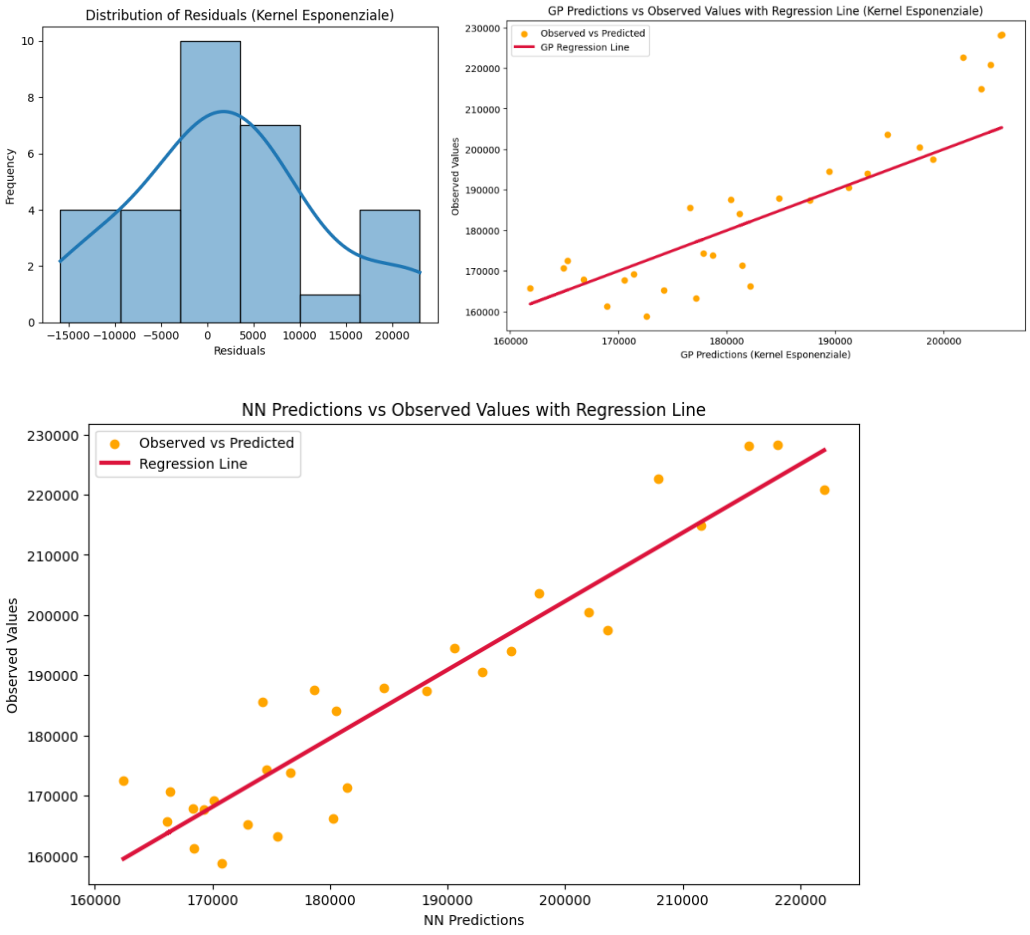

Figure 5.18

Esponenzial Kernel:
MSE: 106156979.8509096
RMSE: 10303.25093603517
MAE: 7950.4452027522
R_2: 0.7502077399111962

Quadratic Exponential Kernel:
MSE: 435171449.4753905
RMSE: 20860.763396275564
MAE: 17161.984802652194
R^2: -0.0239784519420565

Linear Regression:
Coefficients model: [1.1385739]
Intercept model: -25401.83770736071
MSE: 49597678.759491935
RMSE: 7042.5619457333805
MAE: 5750.599775266366
R^2: 0.8832943788539234

The model in question has an R^2 value of 0.88, indicating a good capacity to explain the variability of the response data. However, the analysis of the non-descending gradients highlights a significant problem in the model's optimization. The gradients remain stable or even increase during the iterations, suggesting that the model might be stuck in a local minimum or that the learning rate is inadequate. Despite various simulations and parameter optimizations, it was not possible to obtain results that deviate from these.

Linear regression still captures the trend of the phenomenon, suggesting a direction for the future. The conclusion is that these phenomena are not directly correlated, or at least the training data do not allow an adequate model to be generate.

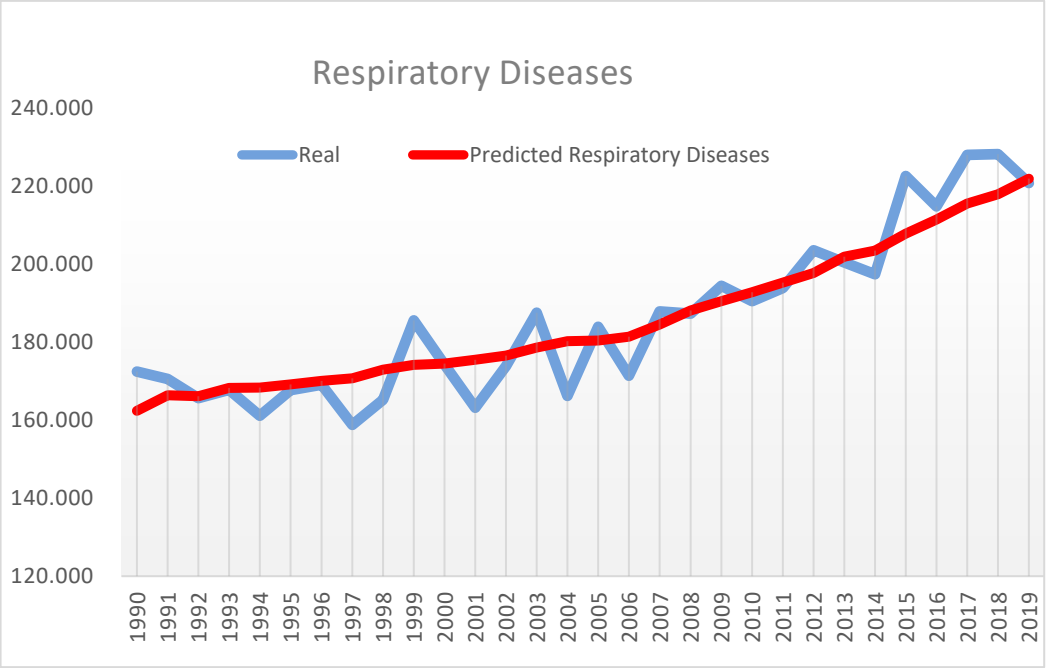

Figure 5.19

5.4. Setup of the Victims Without External Causes

This model has been a real challenge. Even by adjusting all possible parameters, attempting to tune the learning rate, and differently modulating the number of neurons, it was difficult to reach an acceptable conclusion without disabling the overfitting mechanisms. These mechanisms are crucial to ensure that the model represents reality and potentially the future. Various combinations of input variables were tested, including those different from the ones suggested by the correlation matrix, but no result was validated.

Input variables: CO2, CH4

Alfa 0.05 Beta 0.1 LR 0.0001 epoch: 250

Train R^2: 0.7922091782093048 **- Variances:** [0.2064738]

Validation R^2: 0.7781321108341217 **- Variances:** [0.13580649]

Test R^2: 0.9086213633418083 **- Variances:** [0.03525508]

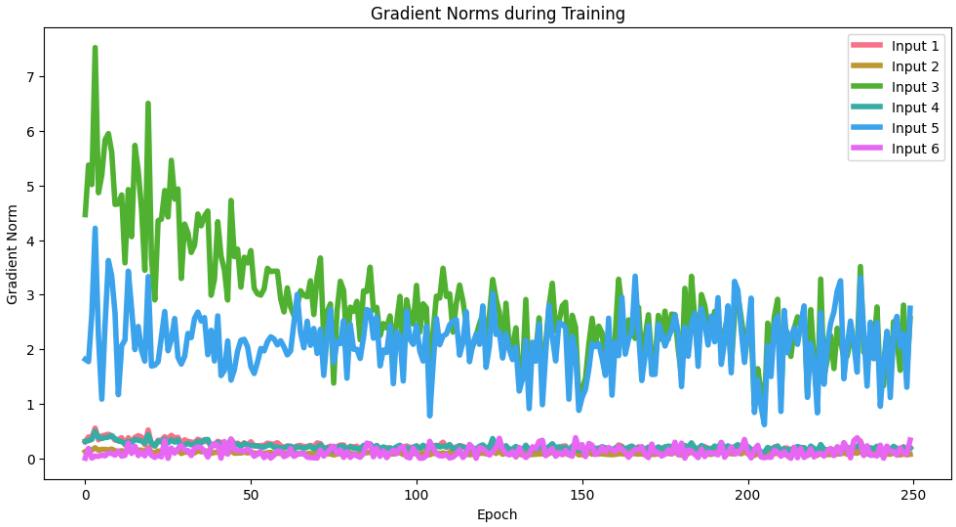

Figure 5.20

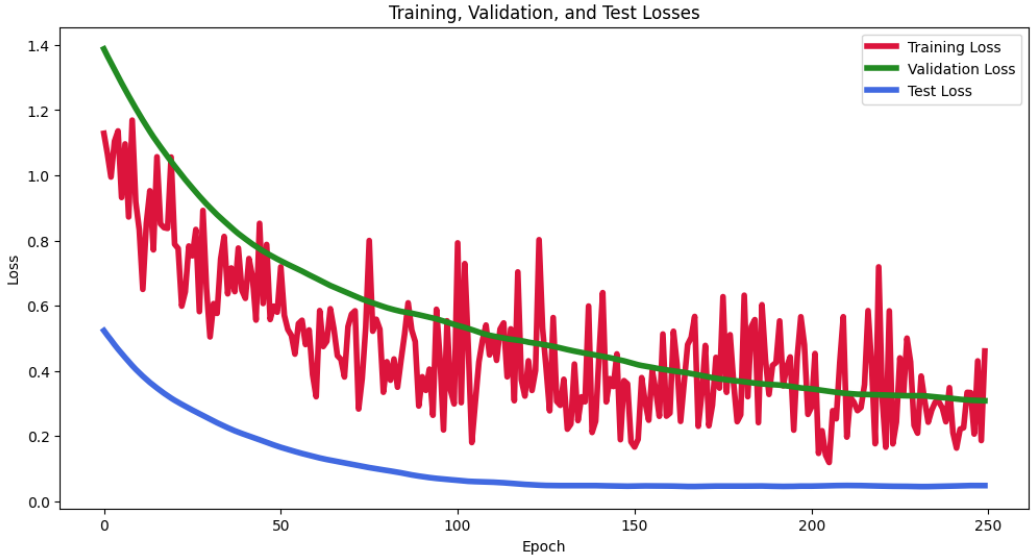

Figure 5.21

Even though the Loss graph seems to suggest that the model has achieved a good level of learning thanks to the training data, a closer look at the gradient graph reveals that many gradients are flat, indicating that the model is not optimized.

In addition, the mechanisms for avoiding overfitting were reactivated, as can be seen from the fluctuating gradients and formation loss.

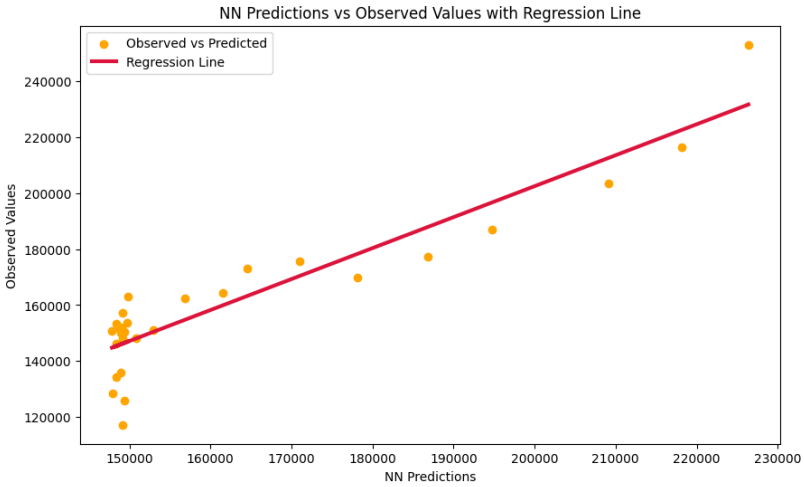

Figure 5.22

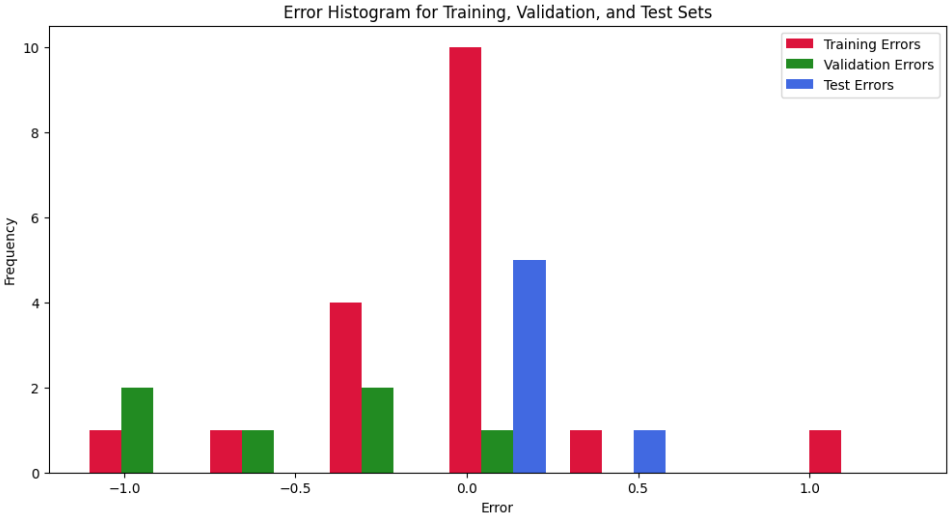

Figure 5.23

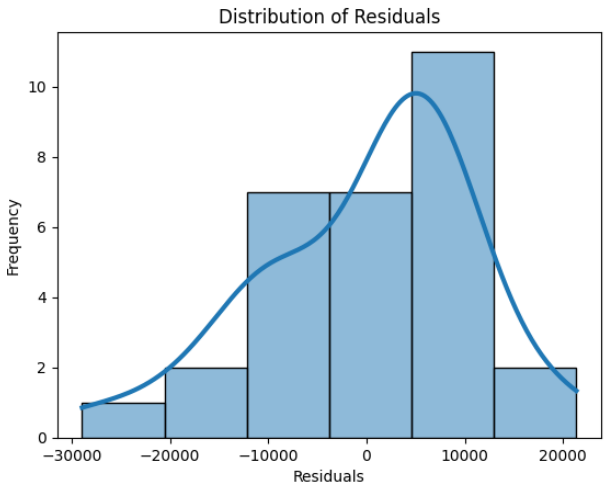

Figure 5.24

Esponenzial Kernel:
MSE: 304225965.9919625
RMSE: 17442.07458968005
MAE: 11441.149867657976
R²: 0.5794313798789432

Quadratic Exponential Kernel:
MSE: 736834733.1361947
RMSE: 27144.69990875189
MAE: 18007.784513067963
R²:: -0.018616428620506387

Quadratic Exponential Kernel:
Coefficients model: [1.10559487]
Intercept model: -18674.440056013846
MSE: 117037686.81685509
RMSE: 10818.395759855297
MAE: 8747.947968889528
R²: 0.8382045454725401

Even from the prediction graph, it is evident that the model fails to capture the values, unlike the previous model which could at least be used to obtain a possible trend. In this case, we can completely discard it.

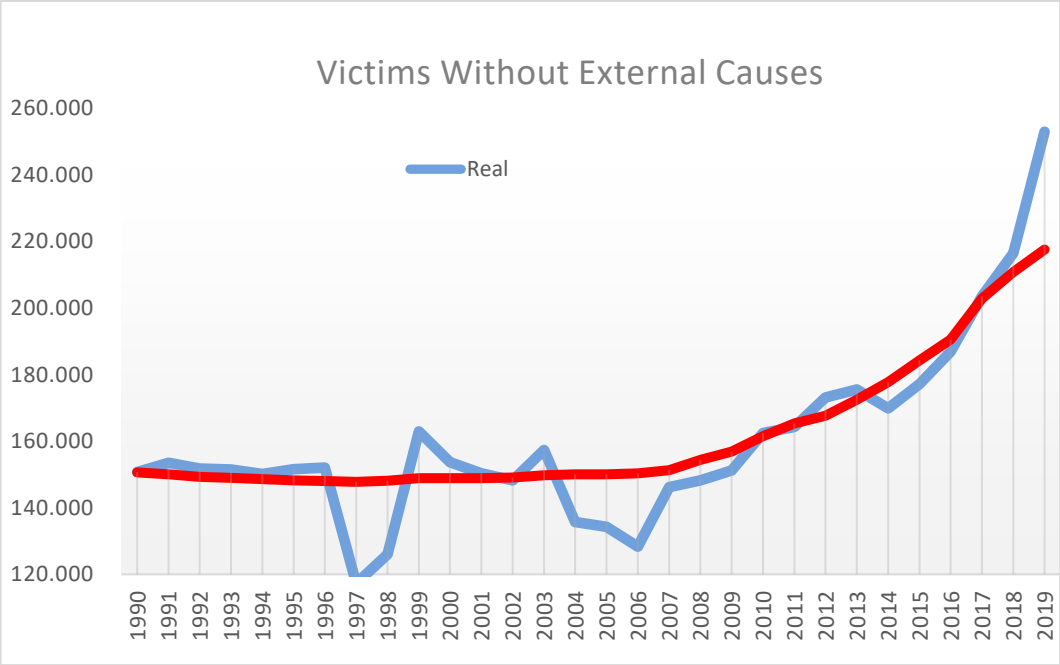

Figure 5.25

CHAPTER 6 - Analysis of model predictions -

Impacts of development, climate and environmental scenarios on population health and future health trends are relevant for health planning processes because the health of populations is an essential element of adaptive capacity. Then the models were validated and analyzed through the metrics of regression; so, then, they were used by entering new input data, referring to future years. The future forecast data for 2020 onward up to the year 2050 is given by NASA; this is extracted from the SSP2 4-5 model, which forecasts the variation in the concentration of some gases in the air, and this is of particular concern to climate change. The input data is reported until 2019 since that is the year that the model generates. The year 2019 is the base year for which the four input data are given. The predictions are made considering regulations currently in force relating to the reduction in polluting emissions and an increase in the use of energy from renewable sources. All the future trends for the four variables predict an increase in the climate change variables. For this reason, it is reasonable to deduce that an increase will occur in future scenarios. The new data was extracted using Python from NETCDF files, filtering them by period and coordinates. Of the four models derived using the data, only two will simulate a future trend (mental and behavioral diseases and nervous system diseases), with the lowest (R2) being those related to the victims of nervous system diseases. With a horizon until the year 2050, it only considers the forecast, but since it is a mathematical model, it is not trustworthy, so we will analyze it until the year 2030. In the second step, he will analyze why these diseases have not shown a correlation in Europe as in other countries after previous studies.

6.1. Nervous system diseases

Let's build a chart by adding the forecast data to the real data:

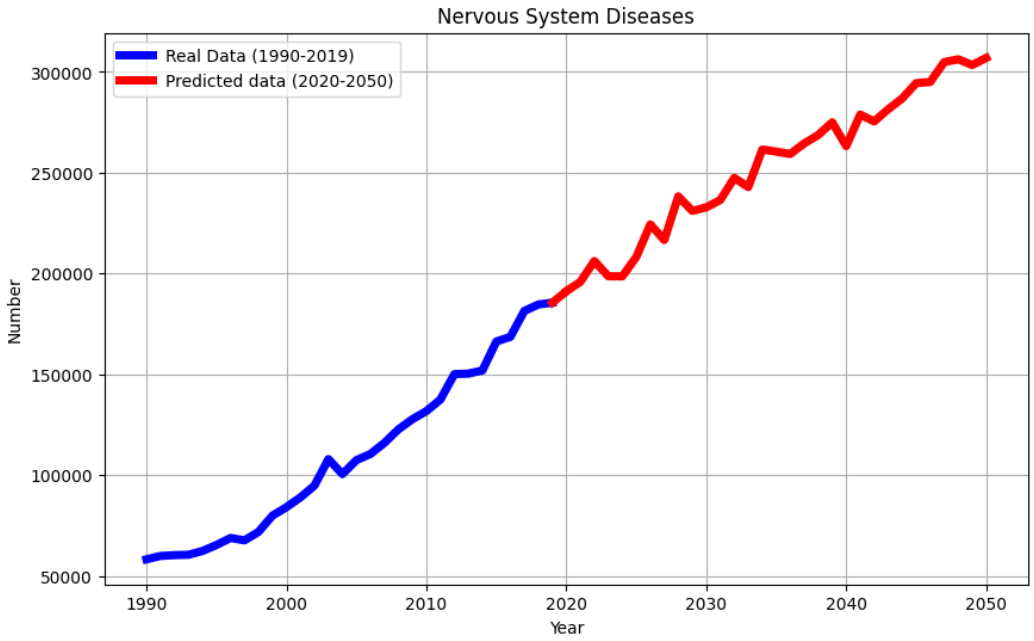

Figure 6.1

The trend is clearly increasing, and even though the forecasts show some oscillations, which are surely due to the characteristics of the training data, the direction of the line is evident. For this type of pathology, it is therefore predictable with a relatively low degree of uncertainty that the number of deaths in the coming years will increase.

Year	*Number*	*%*
2024	198.561	
2025	208.242	5
2026	224.359	8
2027	216.598	-3
2028	238.278	10
2029	231.061	-3
2030	232.809	1
2031	236.448	2
2032	247.449	5
2033	242.844	-2
2034	261.499	8

This is the trend for the next 10 years, and the increase at the end of the period is 32%, which is certainly unsustainable from a healthcare and economic perspective.

Figure 6.2

This model uses the four variables Anthropogenic Forcing, CO2, CH4, and Temperature Anomaly as inputs and has proven to be extremely effective. It has therefore confirmed various studies highlighting how high temperatures, air pollution, and other effects of climate change can worsen existing neurological conditions and increase the incidence of new pathologies not only globally but also in the European context.

In summary, by analyzing and contextualizing the work done, we can affirm that:

During heatwaves, an increase in deaths related to nervous system diseases has been recorded. High temperatures can cause heat stress, worsening pre-existing neurological conditions such as multiple sclerosis and Parkinson's disease. The increase in air pollution levels is closely linked to climate change and represents a risk factor for various neurological diseases. Air pollution has been associated with an increase in symptoms in people with diseases like Alzheimer's and other forms of dementia. Climate change also facilitates the spread of infectious diseases that can affect the central nervous system. For example, rising temperatures and changes in precipitation patterns can favor the proliferation of vectors such as mosquitoes, leading to a higher incidence of diseases like viral meningitis. Besides physical impacts, climate change can have significant psychological effects, contributing to increased stress, anxiety, and depression, which in turn can negatively affect neurological health.

6.2 Mental and Behavioural Disorder

The progression of this pathology is very similar to the previous case even if the scale of magnitude is different:

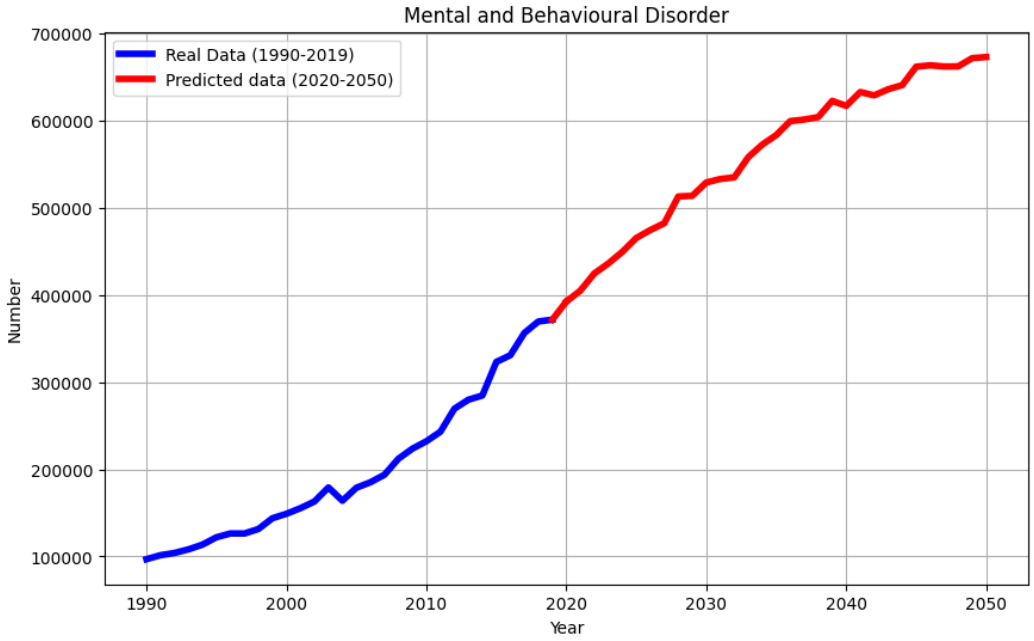

Figure 6.3

Year	*Number*	*%*
2024	*449.209*	*3,5%*
2025	*465.130*	*2,0%*
2026	*474.203*	*1,7%*
2027	*482.066*	*6,4%*
2028	*512.697*	*0,1%*
2029	*513.451*	*3,0%*
2030	*528.725*	*0,8%*
2031	*532.761*	*0,3%*
2032	*534.614*	*4,3%*
2033	*557.741*	*2,6%*
2034	*572.128*	*3,5%*

This is the trend for the next 10 years, and the increase at the end of the period is 28%.

Figure 6.4

The data science study demonstrated a significant correlation between mental and behavioral disorders and climate change in Europe, with a forecast of a 28% increase in the prevalence of these disorders in the next decade. Compared to previous models, the new data suggests that while rising temperatures continue to play an important role, their impact is not as significant as other factors related to climate change. Among these, air pollution and climate variability emerge as the main contributors to the increase in mental and behavioral disorders.

Air pollution is closely linked to the rise in respiratory and cardiovascular diseases, which in turn can exacerbate pre-existing mental conditions such as anxiety and depression. Additionally, climate instability, characterized by extreme weather events like floods and droughts, generates psychological stress and trauma that profoundly affect the mental health of affected populations.

The study's results underline the need for greater attention and targeted policies to mitigate the impact of climate change on mental health. It is essential to promote adaptation measures that include reducing air pollution and implementing psychological support strategies for communities most vulnerable to extreme climate events. Furthermore, additional research is needed to deepen the understanding of the interactions between climate change and mental health, to develop more effective interventions.

These findings indicate that to effectively address the anticipated increase in mental and behavioral disorders, it will be crucial to adopt a holistic approach that considers a range of environmental and socio-economic factors influenced by climate change.

6.3. Analysis of non-performing models

It is equally interesting, not only for academic purposes, to analyze underperforming models, particularly to understand which phenomena can refute a global study when applied to a smaller set. The first cause to evaluate is the global climatic differences compared to those in Europe.

Let us now examine the variables to consider and their differences at a European level.

6.3.1. Climate Difference

Central Europe is characterized by a temperate climate, with cold but not extreme winters and warm but not scorching summers. This region benefits from the Gulf Stream, which moderates winter temperatures and makes the climate more stable compared to other parts of the world. The Gulf Stream, a warm ocean current that flows from the tropical Atlantic towards Europe, plays a fundamental role in maintaining the mild climate of Central Europe. Thanks to this current, winter temperatures are less severe, and summers are more moderate compared to regions at the same latitude in North America or Asia (European Environment Agency's home page). Central Europe experiences moderate seasonal variations, with average temperatures rarely dropping below -10°C in winter and exceeding 35°C in summer. This temperate climate promotes stability, reducing thermal stress on the population and the environment. Climatic moderation also contributes to less precipitation variability, reducing the risk of extreme events such as prolonged droughts or devastating floods (ScienceDaily). Anthropogenic forces, such as greenhouse gas emissions and air pollution, have a less dramatic impact in Central Europe compared to other regions. This is partly due to the higher adaptability of European infrastructure and greater environmental awareness. The environmental

policies of the European Union, such as the European Green Deal, aim to reduce emissions and promote sustainability, further mitigating the effects of climate change (European Environment Agency's home page).

Based on these considerations, we can understand how digestive diseases evolve differently compared to hotter climates, where they can spread more quickly. Another parameter to consider is the higher population density in equatorial countries, particularly India, which accounts for about 18% of the world's population. This, combined with the humid climate, results in the mortality rate for these diseases being double that of Europe.

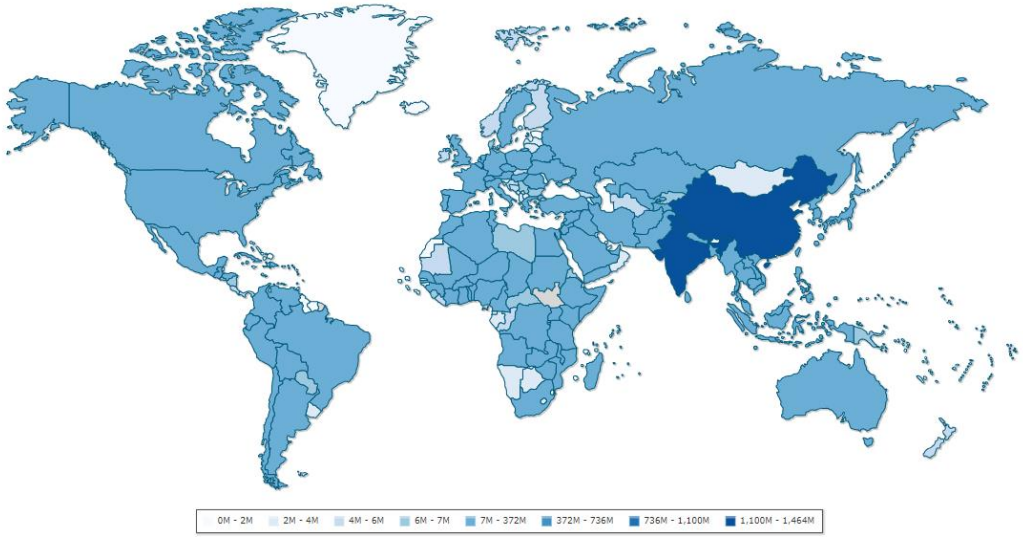

Figure 6.5

We can therefore conclude that the main components of climate change affecting the number of deaths in Europe, among those considered, are the percentages of CH4 and CO2, but in a general sense, the increase in greenhouse gases. As for the temperature increase, it will become much more impactful in the coming decades when it reaches significant levels even in the old continent.

6.3.2 Respiratory Diseases

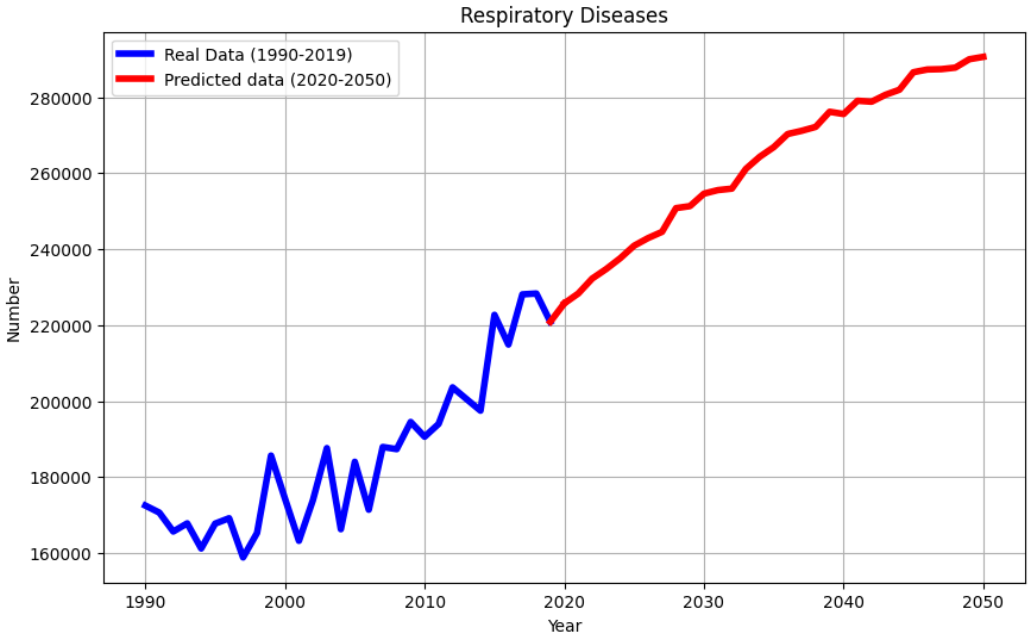

Figure 6.7

Further, a critical analysis of the performance of the developed neural network for verifying the correspondence of the climate change variables with respiratory diseases revealed that the model failed to verify. Inadequate prediction, generalization, and poor interpretability are, therefore, an indication of the inadequacy of the approach to represent such complex interactions of climate with respiratory health. Performance can be enhanced further with more advanced and complex models, more relevant variables, more data that should be of better quality and quantity, and more advanced preprocessing techniques. Most importantly, the models must be more interpretable to provide useful information to policymakers and health providers. Research in this area is critical to better understand how climate change interacts with human health and in turn to the development of strategies for mitigation and adaptation. Despite that, the model can suggest the average trend of the phenomenon according to the previous studies.

6.3.3 Victims Without External Causes

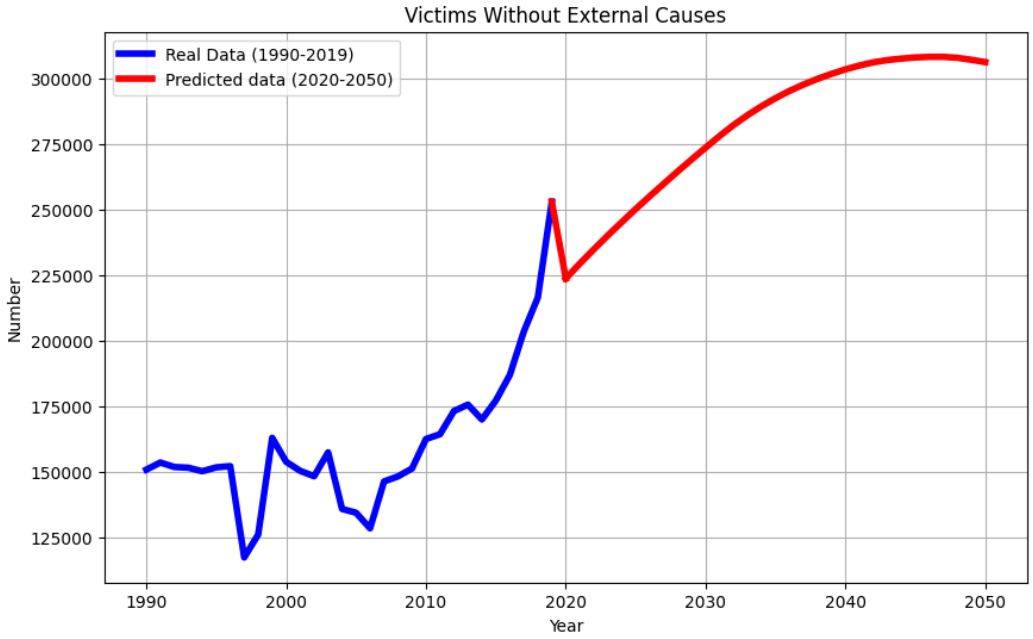

Figure 6.8

Scientific interest in the consequences of climate change often extends to public health. One of the poorest investigated aspects is the relationship between climate variables and deaths from no external cause.

In this chapter, we review the development and application of an ANN model designed for verification of the correspondence of climate change variables and the number of deaths from no external causes, and analyze the performance and reliability of the model. The result from the neural network was awful and revealed many problems that impede making the model usable.

The model presented a very high MSE on the test, which speaks of very inaccurate predictions. In this way, it can be inferred that the neural network is a failure in capturing climate-variable–number-of-deaths relationships from no-external causes. That the many times distant

predictions are with respect to the observed values certainly makes this model unreliable.

Primary among the reasons for the model's failure is that of the training data, in which there is a high level of noise. Data on mortality are embedded with other nonclimatic factors which influence them—access to health care services, socio-economic conditions, chronic diseases, and health policies. This very high noise compromised the capacity of the neural network to learn meaningful patterns. The low performance of the model do not allow the use of the neural network even to find long-term trends: the predictions are so unreliable that any extrapolation of future patterns would be meaningless.

6.4. Temperature Increase

Using the simulation models created by NASA, selecting the most balanced scenario SSP2-4.5 (the CMIP model and related scenarios are detailed in the appendix), we observe how the temperature in Europe will change in the coming decades.

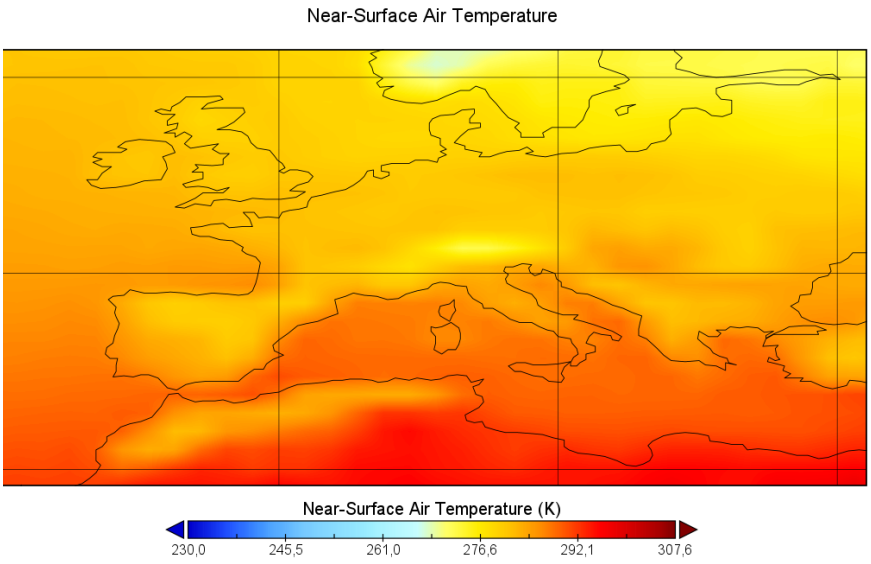

Figure 6.9 Average temperature in May 2024

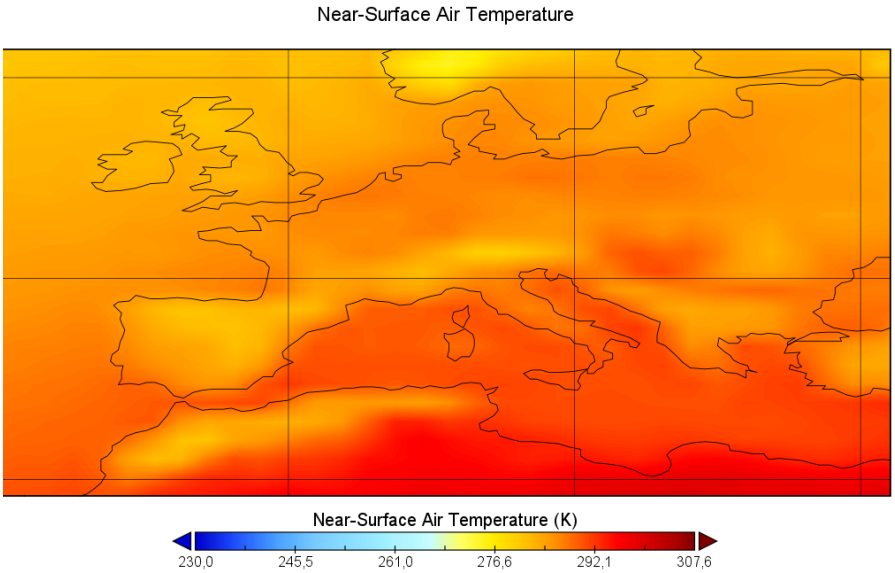

Figure 6.10 Average temperature in May 2044

As these maps show, the yellow areas (colder regions) will almost completely vanish over a span of 20 years. In these maps, we have considered an intermediate month, namely May. Obviously, these changes become more pronounced during months with extreme temperatures, resulting in much hotter summers and milder winters.

These variations, although relatively significant, will not represent a decisive factor for the selected pathologies. They will certainly have an impact, but not on the models considered.

6.5. Change in the concentration of CO2 and CH4 in the atmosphere in the coming decades

In order to better understand these aspects, we will use the data provided by the CMIP model for carbon dioxide and methane concentrations. we will analyse the two graphs to assess how the concentration of these gases will increase over the reference period and how this increase might produce the results produced by our models, considering that we have used an intermediate scenario as the source of our data, which is often preferred for studies from various scientific fields.

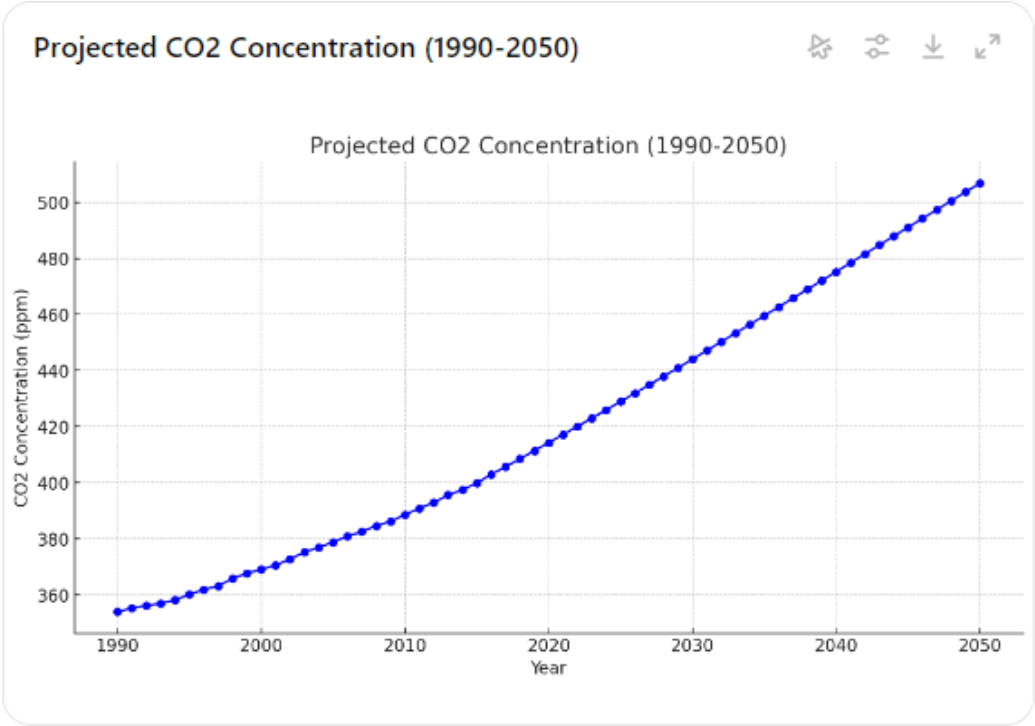

Figure 6.11

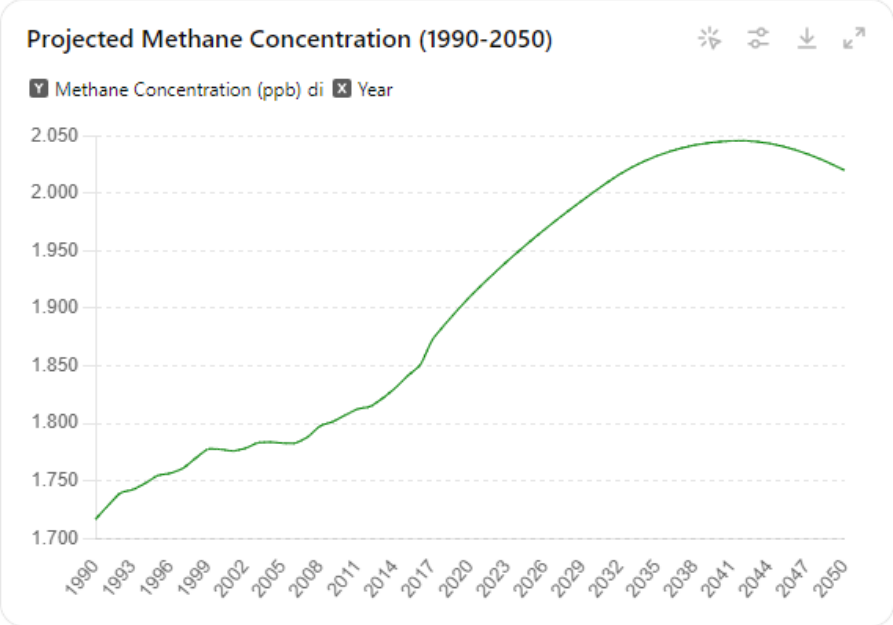

Figure 6.12

The set of charts below shows how the concentrations of CO2 and CH4 gases have been changing from 1990 until 2050. This makes it possible to notice the patterns of changing emissions over time:

CO2 graph displays a smooth, average increase of CO2 concentration advancing in the atmosphere.

This only proves that increased CO2 emissions indeed are due to continued human activities, which are connected with the burning of fossil fuels, as well as with deforestation and other industrial practices.

The methane diagram indicates that the concentration peaks at a significant high until the 2040s before it flattens, and even in one scenario, it reduces slightly in concentration by 2050.

This might indicate that the effect of methane emission mitigation has taken place or that the sources of methane emissions are changing with time.

Consequences on Global Warming

The increased levels of CO2 and CH4 can be directly correlated with global warming because both are greenhouse gases and retain heat on Earth. The steady rise of CO2 is happening, in particular, as this gas has a long lifetime in the atmosphere and is an important greenhouse gas.

Need for Mitigation Actions

These graphs illustrate the urgent necessity of stronger actions toward reducing greenhouse gas emissions. Other important policies include the adoption of renewables, forest protection, energy efficiency, our forests' preservation, and reduction of methane from sources like agriculture and the natural gas industry, to name a few.

CHAPTER 7 - Results and Conclusion -

The results of the neural networks make it possible to state that there is a relationship between climate change and human health at an international level, confirming the results of past literature. The greater depth of analysis of the machine learning techniques once again confirmed the connection, and also provided details on the environmental risk factors most closely related to the diseases studied, while deleting some that were not strictly related to climate change in the area under consideration. Internationally, climate change is strongly correlated with mental and behavioural diseases and diseases of the nervous system, as it has one of the lowest relative errors for all the diseases analysed.

In the above-mentioned global studies for respiratory diseases, the relative error amounts to 7.5%; mathematically, the relationship of climate change with respiratory diseases is therefore weaker. Note that the relative error is high with or without temperature as an input variable.

Another unassociated lack of correlation is attributed to the geographical area in question and deals with digestive diseases, which remain unaffected by climate change.

A new output was elicited in the inquiry into the relationship between climate change and human health by the use of artificial intelligence. Two forms and combinations of two types of artificial intelligence are used to enhance the better analysis of the relationship.

Such neural network-detected studies provide excellent results as they plunge further into understanding which of the input variables are most influential for this analysis performed. The machine learning technique

reconfirmed a previous correlation, increasing the reliability of two of the four pathologies studied.

The predictions obtained have good reliability, and comparing the scenarios deduced with those of other scientific articles, it can be said that they are very similar and that the error is mainly due to the limited amount of data and the large number of variables needed to describe such a complex phenomenon. Of course, focusing on data extracted only from the European territory complicated the data gathering and aggregation phase, but naturally, it made possible a more precise analysis of the phenomena, since some kinds of mortality are predominant in the selected countries. The results could be improved with further refinement by selecting one country and increasing the input time interval. The complexity and difficulty of collecting information on human health remains a major obstacle to the study of correlations between human health and climate change.

7.1 Conclusions

The monitoring data obtained were analysed on a European basis to identify scientific correlations between climate change and human health, using machine learning techniques. The correlations identified so far express, even on a continental scale, the link between climate change and human health, indicating that mortality for two types of diseases, respiratory and nervous system diseases, has a very high level of correlation. Two different types of artificial intelligence were implemented to analyse the best correlation. Our study within the neural network gave very good results for a better understanding of the input variables with a significant influence on the analysed diseases. The use of machine learning techniques confirmed the correlation previously obtained, turning it into an increase in reliability for three of the four pathologies studied.

These analyses assisted by artificial intelligence were significantly more successful than purely statistical analyses. In particular, the application of machine learning techniques provided the model with the lowest absolute errors.

It was found that the occurrence of climate change has little or no correlation with digestive casualties. The mortality variable needs to be studied with other climate change causation variables that are associated with water and food quality and their scarcity rather than the main cause of climate change.

The curve showed near-linearity and a predictable trend in the association between climate change and fatalities in mental and nervous system disorders, suggesting a high correlation factor.

The correlation between climate change and victims of respiratory diseases partly confirms previous studies, although with a lower value for Europe.

In conclusion, there is a link between climate change and human health, in particular between changes in greenhouse gases and victims of mental and nervous system disorders.

Appendix A

The Coupled Model Intercomparison Project Phase 6 (CMIP6) represents the sixth phase of an international climate modeling initiative coordinated by the World Climate Research Programme (WCRP). This initiative aims to improve the understanding of climate variability and change and to predict future global climate conditions. CMIP6 includes a wide range of climate models developed by research institutes around the world, using socio-economic and climatic scenarios to simulate possible future climates. In this chapter, we will explore in detail the scenarios used, the main institutes involved, the key variables considered in the models, and the models themselves.

Shared Socioeconomic Pathways (SSPs)

SSPs describe different trajectories of global socio-economic development, which influence greenhouse gas emissions, climate policies, and the ability to adapt to climate change. The main SSPs used in CMIP6 are five, each with an identifying number and a narrative description.

SSP1: Sustainability (Green Road)

Description: This scenario imagines a world that undertakes sustainable development, with a strong commitment to environmental protection, reducing inequalities, and international cooperation.

Key Features: Low greenhouse gas emissions, high energy efficiency, widespread clean technologies, poverty reduction.

Climate Implications: Low levels of global warming, potential to limit temperature rise to 1.5-2°C above pre-industrial levels.

SSP2: Middle of the Road

Description: This scenario represents a middle path, where global socio-economic trends follow an intermediate course without radical changes compared to current policies.

Key Features: Moderate greenhouse gas emissions, balanced economic growth, moderate technological and social progress.

Climate Implications: Moderate warming, with a temperature rise of around 2-3°C above pre-industrial levels.

SSP3: Regional Rivalry (A Rocky Road)

Description: In this scenario, the world is fragmented into regional blocks pursuing protectionist economic policies and facing significant socio-economic challenges.

Key Features: High greenhouse gas emissions, low international cooperation, slow technological development, marked economic and social inequalities.

Climate Implications: High levels of global warming, with temperatures potentially rising by over 3°C above pre-industrial levels.

SSP4: Inequality (A Road Divided)

Description: This scenario sees a world divided between rich countries, which adopt advanced climate and technological policies, and poor countries, which remain vulnerable to climate change and have limited access to technology.

Key Features: Variable greenhouse gas emissions, high social and economic inequality, uneven technological progress.

Climate Implications: Variable global warming, with significant regional impacts due to inequalities.

SSP5: Fossil-fueled Development (Taking the Highway)

Description: In this scenario, the world follows a path of rapid economic development fueled by high fossil fuel consumption, with an emphasis on **economic growth rather than environmental sustainability.**

Key Features: High greenhouse gas emissions, rapid economic growth, low commitment to environmental sustainability, high dependence on fossil fuels.

Climate Implications: Extremely high levels of global warming, with temperatures potentially rising by over 4°C above pre-industrial levels.

Representative Concentration Pathways (RCPs)

RCPs describe different trajectories of greenhouse gas concentrations in the atmosphere and are used in combination with SSPs to create complete climate scenarios. The RCPs are four and vary based on the amount of radiative forcing (measured in W/m^2) they represent for the year 2100.

RCP2.6

Description: This pathway represents an ambitious mitigation scenario, with strong policies to reduce greenhouse gas emissions and a resulting radiative forcing of about 2.6 W/m^2 by 2100.

Climate Implications: Limits global warming to about 1.5-2°C above pre-industrial levels.

RCP4.5

Description: This pathway represents an intermediate scenario where mitigation policies manage to stabilize greenhouse gas emissions, with a radiative forcing of about 4.5 W/m² by 2100.

Climate Implications: Moderate warming, with a temperature rise of around 2.5-3°C above pre-industrial levels.

RCP6.0

Description: This pathway represents a scenario where mitigation policies are less effective, with a radiative forcing of about 6.0 W/m² by 2100.

Climate Implications: Higher warming, with a temperature rise of around 3-3.5°C above pre-industrial levels.

RCP8.5

Description: This pathway represents a high emissions scenario, with weak or nonexistent policies for mitigating greenhouse gases, and a radiative forcing of about 8.5 W/m² by 2100.

Climate Implications: Extremely high global warming, with a temperature rise of over 4°C above pre-industrial levels.

SSP-RCP Combinations

In CMIP6, SSPs are combined with RCPs to create specific climate scenarios, known as SSP-RCP, which offer a comprehensive range of possible future climates. Some of the key combinations include:

SSP1-1.9: Represents a sustainable world with ambitious climate policies, combining SSP1 with a very low concentration pathway, RCP1.9. This **scenario aims to limit the temperature increase to well below 2°C.**

SSP1-2.6: Another sustainable scenario, but less ambitious than SSP1-1.9, combining SSP1 with RCP2.6, limiting the temperature rise to about 2°C.

SSP2-4.5: Represents an intermediate scenario, combining SSP2 with RCP4.5, with moderate mitigation policies and moderate global warming.

SSP3-7.0: Combines SSP3, characterized by regional rivalries and high emissions, with RCP7.0, representing a future with high emissions and significant climate impacts.

SSP5-8.5: An extreme scenario that combines SSP5, based on high fossil fuel use, with RCP8.5, predicting a drastic increase in global temperatures.

Scenario Implications

The CMIP6 scenarios provide a wide range of possible future climates that help researchers explore the consequences of different policies and socio-economic development paths. The implications of these scenarios are crucial for:

Climate Policy Planning: Scenarios help policymakers understand the potential consequences of their choices and develop effective mitigation and adaptation strategies.

Climate Research: Scenarios provide a basis for scientific research, allowing researchers to explore climate dynamics under different future conditions.

Natural Resource Management: Understanding possible future climates is essential for the sustainable management of natural resources, such as water, agriculture, and biodiversity.

Preparation and Adaptation: Climate projections help communities and industries prepare for the impacts of climate change and adopt adaptation measures to reduce vulnerability.

Main Participating Institutes

CMIP6 involves numerous research institutes and universities from around the world, contributing their climate models and data. Some of the main participating institutes include:

NASA Goddard Institute for Space Studies (GISS): Provides models like the GISS ModelE.

Met Office Hadley Centre (MOHC): Contributes models like HadGEM3.

Max Planck Institute for Meteorology (MPI-M): Provides models like MPI-ESM.

National Center for Atmospheric Research (NCAR): Contributes the CESM (Community Earth System Model).

Institute Pierre Simon Laplace (IPSL): Provides the IPSL-CM model.

China Meteorological Administration (CMA): Contributes models like BCC-CSM.

These institutes collaborate to ensure that CMIP6 models are comprehensive and accurate, facilitating a deeper understanding of climate change and its implications.

Key Variables Considered in Models

CMIP6 models consider a wide range of climatic and environmental variables. The main variables considered include:

Surface Temperature (land and ocean): Critical measure for assessing global warming.

Precipitation: Important for understanding changes in rainfall patterns and water availability.

Snow and Ice Cover: Key indicator of changes in glaciers and sea ice.

Greenhouse Gas Concentrations (CO2, CH4, N2O): Measurement of atmospheric concentrations of gases that contribute to the greenhouse effect.

Sea Level: Measure of sea level rises caused by glacier melt and thermal expansion of the oceans.

Soil Moisture: Important for agriculture and water resource management.

Wind: Important for understanding changes in atmospheric circulation patterns.

Solar Radiation: Measure of the amount of solar energy reaching the Earth's surface.

These variables are essential for accurately simulating past, present, and future climates, and for understanding the potential impacts of climate change on global and regional scales.

Models Used

CMIP6 comprises a variety of climate models developed by different institutes. Each model has its own characteristics and specifications, but all aim to simulate the Earth's climate accurately. Some of the main models used in CMIP6 include:

GISS ModelE: Developed by NASA GISS, it is a global climate model that simulates the interaction between the atmosphere, ocean, cryosphere, and biosphere.

HadGEM3: Developed by the Met Office Hadley Centre, it is an advanced climate model that includes detailed atmospheric and oceanic physics schemes.

MPI-ESM: Developed by the Max Planck Institute for Meteorology, it is a climate model that integrates atmospheric, oceanic, and terrestrial biosphere dynamics.

CESM: The Community Earth System Model, developed by NCAR, is a modular climate model that allows studying different components of the Earth system separately or together.

IPSL-CM: Developed by the Institute Pierre Simon Laplace, it is a complex climate model that includes various modules to simulate the atmosphere, ocean, and cryosphere.

BCC-CSM: Developed by the China Meteorological Administration, it is a climate model that simulates interactions between the atmosphere and ocean, with a particular focus on the Asian climate.

These models are used to perform century-scale simulations, providing climate projections that help policymakers, researchers, and international organizations prepare for and adapt to climate change.

Technical Details of the Models

Each CMIP6 model has technical specifications that make it unique. For example:

Spatial and Temporal Resolution: Models vary in resolution, with some offering high-resolution details (up to 10 km) while others use coarser resolutions (up to 200 km). Temporal resolution can also vary, with simulations using hourly, daily, monthly, or annual intervals.

Parameterization of Physical Processes: Each model uses different schemes to parameterize physical processes such as convection, cloud formation, heat transport in the ocean, and land-atmosphere interactions.

Climate Feedbacks: CMIP6 models consider various climate feedbacks, such as the ice-albedo effect, carbon cycle, and aerosol-cloud interactions, which significantly influence climate projections.

Initialization and Boundary Conditions: Each model starts its simulations from specific initial conditions, which can be based on observational data or previous simulations. Boundary conditions include external factors such as solar radiation, greenhouse gas emissions, and aerosols.

Appendix B

Support Files and Thesis

Through the link pointed to by the following QR code, it is possible to access a GitHub repository containing the Python files used in the thesis, Excel sheets, CSV files containing the data, and all the files used for this thesis.

In the README.md file, there is a detailed list of the files contained.

Bibliography

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F. D., Haddaway, N. R., Haines, A., & Dangour, A. D. (2021). Systematic mapping of global research on climate and health: A machine learning review. *The Lancet Planetary Health*, *5*(8), e514– e525. https://doi.org/10.1016/S2542-5196(21)00179-0

Confalonieri, U. ; M., B. ;. Akhtar, R. ;. Ebi, K. L. ;. Hauengue, M. ;. Kovats, R. S. ;. Revich, B. ;. Woodward, A. (2007). *Climate Change: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* Cambridge University Press: https://doi.org/10.2134/jeq2008.0015br

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J. M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J. I., Stine, R. A., Turlach, B. A., Weisberg, S., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Gosling, S. N., Lowe, J. A., McGregor, G. R., Pelling, M., & Malamud, B. D. (2009). Associations between elevated atmospheric temperature and human mortality: A critical review of the literature. *Climatic Change*, *92*(3–4), 299–341. https://doi.org/10.1007/s10584-008-9441-x

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4). https://doi.org/10.1029/2010RG000345

Hayes, K., Blashki, G., Wiseman, J., Burke, S., & Reifels, L. (2018). Climate change and mental health: Risks, impacts

and priority actions. *International Journal of Mental Health Systems*, *12*(1). https://doi.org/10.1186/s13033-018-0210-6

Melillo, J. M., Richmond, T. C., Yohe, G. W., & Assessment, U. N. C. (2014). Climate change impacts in the United States: The third national climate assessment. In *US Global change research program* (Vol. 841, p. 841). U.S. Global Change Research Program: Washington.

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Tausnev, N., Bauer, S. E., Delgenio, A. D., Kelley, M., Lo, K. K., Ruedy, R., Shindell, D. T., Aleinov, I., Bauer, M., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T. L., Faluvegi, G., Hansen, J. E., … Zhang, J. (2014). CMIP5 historical simulations (1850-2012) with GISS ModelE2. *Journal of Advances in Modeling Earth Systems*, *6*(2), 441–477. https://doi.org/10.1002/2013MS000266

Parks, R. M., Bennett, J. E., Foreman, K. J., Toumi, R., & Ezzati, M. (2018). National and regional seasonal dynamics of all-cause and cause-specific mortality in the USA from 1980 to 2016. *ELife*, *7*. https://doi.org/10.7554/eLife.35500

Pearl, J. (2018). *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*. 3–3. https://doi.org/10.1145/3159652.3176182

Pizzulli, V. A., Telesca, V., & Covatariu, G. (2021). Analysis of correlation between climate change and human health based on a machine learning approach. *Healthcare (Switzerland)*, *9*(1). https://doi.org/10.3390/healthcare9010086

Raita, Y., Camargo, C. A., Liang, L., & Hasegawa, K. (2021). Big Data, Data Science, and Causal Inference: A Primer for Clinicians. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.678047

Scheelbeek, P. F. D., Dangour, A. D., Jarmul, S., Turner, G., Sietsma, A. J., Minx, J. C., Callaghan, M., Ajibade, I., Austin, S. E., Biesbroek, R.,

Bowen, K. J., Chen, T., Davis, K., Ensor, T., Ford, J. D., Galappaththi, E. K., Joe, E. T., Musah-Surugu, I. J., Alverio, G. N., … Berrang-Ford, L. (2021). The effects on public health of climate change adaptation responses: A systematic review of evidence from low- And middle-income countries. *Environmental Research Letters*, *16*(7). https://doi.org/10.1088/1748-9326/ac092c

Solomon, S., Intergovernmental Panel on Climate Change, & Intergovernmental Panel on Climate Change (Eds.). (2007). *Climate change 2007: The physical science basis: contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Song, X., Wang, S., Hu, Y., Yue, M., Zhang, T., Liu, Y., Tian, J., & Shang, K. (2017). Impact of ambient temperature on morbidity and mortality: An overview of reviews. *Science of The Total Environment*, *586*, 241–254. https://doi.org/10.1016/j.scitotenv.2017.01.212

Studies (NASA/GISS), N. G. I. for S. (2018). *NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.1421

Data Availability Statement:

Climate data (NASA database):

Available online: https://data.giss.nasa.gov/modelE/

Mortality data (WHO database):

Available online: https://platform.who.int/mortality