Bachelor Thesis

Explaining Weather Data Predictions Using SHAP

Sára Jurovatá

Subject Area: Data Science

Student Number: 11919776

Supervisor: Prof. Sabrina Kirrane

Date of Submission: 28. September 2022

Institute for Information Systems and New Media
Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Vienna, Austria

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I have to thank Prof. Sabrina Kirrane for not only supervising my bachelor thesis but also for helping me to come up with this interesting topic. I would like to thank her for her patience, support, and all the useful feedback and advice she gave me during the last few months. As an active researcher, she gave me a great insight into the research process as such and taught me how to effectively conduct research on my own. Also, I would like to thank all the professors, lecturers, and teaching assistants from the Data Science specialization who provided me with knowledge without which I would not be able to complete this thesis. Last but not least, I am grateful to my family and friends for their active encouragement throughout this whole process.

Abstract

With the ongoing global climate crisis, exploring and understanding weather data is germane. To highlight the importance of the topic of explainable artificial intelligence (xAI), this research focuses on using SHAP approach to evaluate machine learning predictions on chosen weather data. The first chapters of this thesis summarize the existing literature on the topic of machine learning used for weather predictions as well as weather explanations in general. Later, we describe the machine learning approaches and the datasets used during this research project. In this thesis, we examine weather data collected in three different locations in Switzerland. After a brief introduction of the data and its variables, we use four machine learning models - Linear Regression, Decision Tree, Random Forest, and LightGBM - to make predictions. The performance of these algorithms is then compared using respective performance metrics. On top of this, SHAP comes into play to help understand the models. By using SHAP, one can easily comprehend what features and how these are significant when making predictions. Not only is this beneficial when interpreting the results, but it can also come in useful when improving the model performance. Afterward, corresponding limitations and future work is discussed in appropriate detail. At the end of the thesis, we answer the main research question, *"How effective is SHAP for providing explanations for existing weather prediction models?"*.

1 Introduction and Thesis Structure

This thesis focuses on analyzing and explaining machine learning models forecasting the weather with the help of SHAP (SHAPley Additive exPlanations),[1] a game-theoretic approach. The main motivation of this research lies in the explainability of machine learning models [1].

This thesis, however, also focuses on highlighting environmental changes using real-world data. By applying basic machine learning techniques together with the SHAP method, it aims to explain the possible meteorological and air quality predictions. Thus, this research tries to underline the severity of such environmental issues justified by widely used data science methods.

When talking about the relevance of this research, both societal and economic perspectives need to be mentioned. These predictions, since impacting human lives, need to be accurate. This is essential not only from the societal but also from the economic perspective. As stated by The Organization for Economic Cooperation and Development [2] during their Global Forum on Environment in 2016, *"the links between the economy and the environment are manifold"*. For instance, economic growth causes pollution, especially in poor parts of the world, which then lowers the quality and quantity of resources, thus harming economic growth again.

Moreover, the fact that weather is closely connected to climate change is undoubtful. The United States Environmental Protection Agency (EPA) classifies the changes in temperature as one of the indicators of the extreme weather conditions that we are facing more and more often [3]. Therefore, in this thesis, we would like to point out these issues by using relevant data together with different machine learning methods and explaining the results with the help of SHAP visualizations.

[1] `https://shap.readthedocs.io/en/latest/index.html`

1.1 Research Questions

For this research project, we developed four research questions – one main research question and three subquestions.

Main research question:

How effective is SHAP for providing explanations for existing weather prediction models?

Subquestions:

1. Subquestion: *How effective are existing machine learning approaches in forecasting weather?*

2. Subquestion: *What datasets and how do they need to be adjusted in order to use them for machine learning?*

3. Subquestion: *How effective is SHAP when compared to other approaches to xAI?*

1.2 Thesis Structure

The rest of the thesis is organized in the following way:

In **Chapter 2**, three types of proposed research methods are defined and described with regard to this thesis.

The main purpose of **Chapter 3** is to give an overview of existing literature on both machine learning used for weather predictions and weather explanations in general.

Chapter 4 introduces the datasets and machine learning models used in the thesis. Data used for the predictions is described in sufficient detail with the help of various figures. A brief explanation of the chosen machine learning algorithms is given.

The focus of **Chapter 5** lies in the evaluation of the machine learning models. In this section of our thesis, we make use of respective performance metrics to compare the models and discuss the model performance.

Chapter 6 details the usage of SHAP for this research. First, an introduction to explainability in machine learning in general is provided, which is then followed by concrete examples of SHAP used for weather predictions. Also, the SHAP visualizations are available in this section together with a discussion on the respective application of SHAP and a short reflection on the usage of SHAP in this thesis. Last but not least, with the help of SHAP visualizations, we perform a simple feature selection of our best-performing model.

Chapter 7 covers the research limitations. In this part of the thesis, three different weaknesses of our models are explained.

The last chapter, **Chapter 8**, is divided into two sections. In the first part of this chapter, the main findings of the thesis are summarized, and the research questions are revisited and answered. In the second section, we state the key ideas for possible future work.

2 Research Methodology

To describe the methodology of our research, we focus on three different categories – sources of information, type of data used, and purpose. What types of research and why these are appropriate for our research is outlined below.

2.1 Sources of Information: Secondary Research

When it comes to information sources, opting for secondary instead of primary data is both convenient and suitable. As opposed to primary data analysis, in secondary data analysis, individuals who analyze the data did not collect it [4]. Hence, data of adequate size and quality is needed for our research.

According to [5], the process of secondary analysis consists of three main steps – developing the research question, identifying the dataset, and evaluating it. All of these are described below in the context of our research.

2.1.1 Develop the Research Questions

As the literature suggests, *"the key to secondary data analysis is to apply theoretical knowledge and conceptual skills to utilize existing data to address the research questions"* [5]. Hence, as the very first step of our research, we developed one main research question and three subquestions stated in *Chapter* 1.

2.1.2 Identify the Dataset

The next step of our secondary research was the identification of data we will work with. We chose two datasets – **Daily updated air quality measurements, since 1983**[2] and **Daily**

[2] `https://data.europa.eu/data/datasets/6db44316-9717-4a98-8a83-577d4cb25afc-stadt-zurich?locale=en`

updated Meteodata, since 1992.[3] We agreed that we are able to use these datasets for the purpose of this thesis since their size as well as contents are suitable for the purpose of this thesis, which is discussed in more detail in *Chapter* 4. All in all, after creating research questions and doing the first literature review, we came to the conclusion that the dataset found is suitable for our research.

2.1.3 Evaluate the Dataset

It is our responsibility to work with data that comes from a reliable source. Thus, we made sure that this is the case by choosing datasets from the *data.europa.eu* website published by *opendata.swiss*. Since *data.europa.eu* is the official portal for European data and is funded by the European Union and managed by the Publications Office of the European Union,[4] we believe such source may be considered as a reliable one.

2.2 Type of Data Used: Quantitative Research

By definition, *"quantitative research encompasses a range of methods concerned with the systematic investigation of social phenomena, using statistical or numerical data"* [6]. The involvement of measurement, which is essential for quantitative research, is imperative in our research too. The process of deduction is continued by analyzing these measurements and finalized by drawing respective conclusions [6].

There exist two categories of quantitative research, namely experimental and survey designs [6]. In this thesis, experimental design is envisaged. According to Watson [6], *"an experiment is a study where the researcher can manipulate one variable, the in-*

[3] `https://data.europa.eu/data/datasets/0ece9cfa-49ad-4aef-b923-3a0ec2520736-stadt-zurich?locale=en`

[4] `https://data.europa.eu/en`

dependent variable, and study its effect on a dependent variable". This is exactly what we are doing with our dataset since we make predictions of our dependent variable (`temperature`) while making use of our independent variables. We not only study the overall performance of our machine learning models, but we also examine which specific features and how much they effect our dependent variable by using SHAP.

2.3 Purpose: Applied Research

The main purpose of this thesis is to apply practical knowledge. Compared with basic research, applied research focuses more on understanding and addressing problems rather than developing universal knowledge. Furthermore, as is typical for applied research, in our thesis, we aim to answer multiple questions and make use of multiple methods [7]. In particular, we pre-process the datasets found online, build four machine learning models using this data, apply SHAP approach to explain the results of these algorithms, and in the end, we answer four research questions in our thesis. On top of this, we discuss related literature together with our conclusions and observations from doing this research.

3 State of the Art

The first part of this section focuses on machine learning for weather predictions. This is continued with an introduction to weather explanations where a few interesting papers are mentioned. Furthermore, a brief explanation of SHAP as well as of our choice to work with this approach in particular is explained.

3.1 Machine Learning for Weather Predictions

When analyzing data on weather, making use of machine learning algorithms is undoubtedly the right approach. This claim is supported by the amount of literature available online. For example, an article written by Bochenek and Ustrnul [8], demonstrates extensive research of *"the 500 most relevant scientific articles published since 2018, concerning machine learning methods in the field of climate and numerical weather prediction"*. In this paper, multiple machine learning models are built, of which the Random Forest method is used in our research too. The reason for choosing Random Forest and the three other machine learning models are further explained in *Chapter* 4.

Another paper where the goal is to anticipate weather predictions uses four different algorithms – Gradient Boosting Decision Tree, Random Forest, Naive Bayes Bernoulli, and KNN Algorithm [9]. For their predictions they make use of various features such as `rainfall`, `wind direction`, and `cloud`. The best-performing model, however, is the ensemble-based model[5] with an overall accuracy of 0.957. In our thesis, we also make use of the Decision Tree and Random Forest algorithms, and their performance is shown in *Chapter* 5.

Since the choice of machine learning models is a crucial one, we apply the knowledge from a 'Machine learning algorithms: Popu-

[5] `https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f`

lar algorithms for Data Science and Machine Learning' book [10] to make suitable decisions. This book not only explains the theory behind the machine learning models, but also shows how to use these techniques in practice with the help of Python.

3.2 Weather Explanations

To get an insight into the weather explanations, we explore a paper written by Dieber and Kirrane [11], 'A Novel Model Usability Evaluation Framework (MUsE) for Explainable Artificial Intelligence'. During their research, they worked with different machine learning models and used model agnostic explanations to better understand their results. Compared with our approach, however, they used the LIME model,[6] which differs from SHAP in various aspects. In contrast to SHAP, LIME does not use Shapley values to compute the feature importance but it trains an interpretable model by creating a new dataset consisting of some of the original variables.[7]. Also, LIME works by explaining single predictions of a model, while SHAP is both locally and globally interpretable.[8] Nevertheless, the data used in their research is on the weather too, and the models they chose to implement are similar to our preferences.

To analyze our machine learning models, we decided to use the SHAP explanations that are *"a popular feature-attribution mechanism for explainable AI"* [12]. The reason for our choice, however, stems from SHAP's universal usability. As indicated in the 'Explainable AI: A Review of Machine Learning Interpretability Methods', *"SHAP is the most complete method, providing explanations for any model and any type of data, doing so at both a*

[6] https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

[7] https://christophm.github.io/interpretable-ml-book/lime.html

[8] https://ernesto.net/lime-vs-shap-which-is-better-for-explaining-machine-learning-models/

global and local scope" [1]. Here, SHAP was compared to multiple approaches to xAI of which some are LIME, InterpretML,[9], or AIX360.[10] Additionally, SHAP together with LIME are considered the most comprehensive and dominant approaches to explaining feature importance [1].

Apart from gathering examples on how to use SHAP to interpret various machine learning models, the official SHAP documentation[11] explains the SHAP approach using this simple figure (*Figure* 1). SHAP helps to understand the "black box" by visualizing the output of machine learning models. In other words, it explains the results of machine learning models by visualizing the effect the independent variables (features) have on respective predictions of a dependent variable. For this, SHAP uses *"Shapley values from game theory and their related extensions"*.[12]

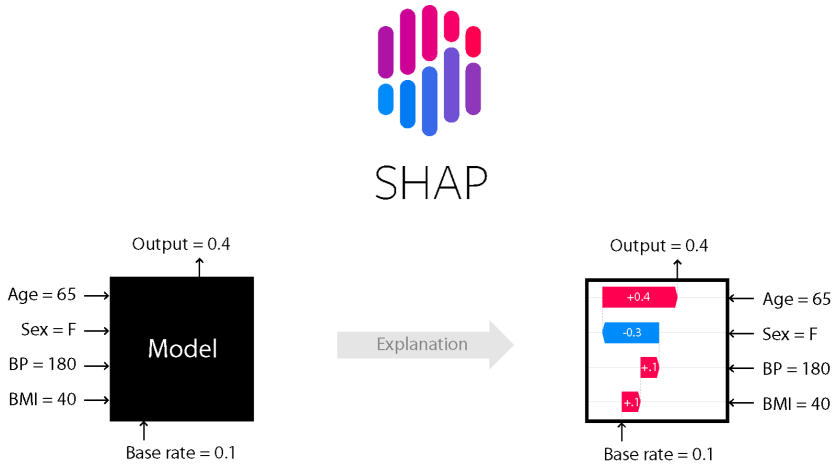


Figure 1: SHAP model example[13]

[9] https://interpret.ml/
[10] https://aix360.readthedocs.io/en/latest/
[11] https://shap.readthedocs.io/en/latest/index.html
[12] https://shap.readthedocs.io/en/latest/index.html

A paper published by Lubo et al. [13] shows how the SHAP approach works in practice. There, SHAP values are used to interpret data on seismic facies. More precisely, they implement SHAP's `TreeExplainer` and analyze both global and local interpretations of their predictions. In contrast to our research, in [13] only one machine learning model, Random Forest, is built, and the data used in this paper is a 3D seismic survey[14] which is very different from our data.

SHAP is also used to show the relationship between climatology and other fields. In [14], SHAP explains the heatstroke predictions while looking at multiple weather information. With the help of SHAP, Ogata et al.[14] are able to identify which predictors work for predicting a higher as well as lower number of heatstrokes of deaths and hospital admissions. Other examples where SHAP is used interdisciplinary while still focusing on weather are outlined in *Chapter* 6.

[14] `https://www.youtube.com/watch?v=hxJa7EvYoFI&ab_channel=ge0physicsrocks`

4 Data and Machine Learning

In *Chapter* 4, we provide a short yet informative data introduction and describe the process of its pre-processing. Moreover, we make use of multiple tables and figures to communicate our findings more effectively. In the second part of this chapter, the machine learning models used in our thesis are mentioned with the support of relevant literature. Ultimately, we explain how we implement machine learning using Python.

4.1 Data

The datasets used in this thesis are published by *opendata.swiss* on the *data.europa.eu* website. Two types of datasets are used, namely **Daily updated air quality measurements, since 1983** [15] and **Daily updated Meteodata, since 1992**.[16] These two datasets are, as their names suggest, updated daily, and they contain measurements from three locations in Switzerland. For this reason, the datasets are in the German language.

All in all, these dataset contain one datetime variable each (`Date`), one integer variable (`Value`) each, and five string variables (`Location`, `Parameter`, `Interval`, `Unit`, and `Status`) each. There are five different parameters available in the meteo dataset (`global_radiation`, `temperature`, `air_pressure`, `T_max_h1`, and `precipitation_duration`) and eleven parameters available in the air quality dataset (`O3`, `O3_max_h1`, `O3_nb_h1>120`, `CO`, `NO2`, `NO`, `NOx`, `SO2`, `PM10`, `PM2.5`, and `PN`). Respective units for these parameters are outlined in *Table* 2. The `Status` variable indicates whether the specific measurement is revised or provisional. Finally, there is the `Interval` variable that has only one unique

[15] `https://data.europa.eu/data/datasets/6db44316-9717-4a98-8a83-577d4cb25afc-stadt-zurich?locale=en`

[16] `https://data.europa.eu/data/datasets/0ece9cfa-49ad-4aef-b923-3a0ec2520736-stadt-zurich?locale=en`

value since the datasets are updated every day.

4.1.1 Data Pre-processing

Before training the data for the purpose of machine learning described in the second part of this chapter, we had to pre-process the data. This consisted of several steps that are explained below.

The very first step was to put all the meteo and air quality .csv files together. To be able to create one dataset out of these two .csv files, we merged them based on `Date`, `Location`, `Interval`, and `Status` columns. In the merged dataset, there were found no duplicates, but we found a few missing values. However, at this stage, we did not want to drop these since we were planning to make considerable changes to the dataset, and we did not want to lose any data.

Next, the variable names, as well as some of the values, were translated from German to English for the sake of consistency of this research. Afterward, we changed the variable types where needed, for instance, for the `Date` variable. *Table* 1 and *Table* 2 show the variable overview we got after these amendments. *Table* 1 shows the datetime and integer variables and the number of unique occurrences, their mean, minimum, and maximum values. *Table* 2, on the other hand, displays the string variables, the number of unique occurrences, and the list of unique values for each of them.

Table 1: Overview of date, value_meteo, and value_air_quality variables

Variable Name	**#Unique**	**Mean**	**Min**	**Max**
`Date`	10,847	2011-03-13	1992-07-01	2022-03-15
`Value_Meteo`	20,266	256.32	-10.93	1440
`Value_Air_Quality`	19,072	525.82	-0.02	78,863.57

Table 2: Overview of meteo and air quality string variables

Variable Name	**#Unique**	**Unique Values**
Location	3	Zch_Stampfenbachstrasse Zch_Schimmelstrasse Zch_Rosengartenstrasse
Parameter_Meteo	5	global_radiation temperature air_pressure T_max_h1 precipitation_duration
Interval	1	d1
Unit_Meteo	4	W/m2 °C hPa min
Status	2	revised provisional
Parameter_Air_Quality	11	O3 O3_max_h1 O3_nb_h1>120 CO NO2 NO NOx SO2 PM10 PM2.5 PN
Unit_Air_Quality	5	µg/m3 1 mg/m3 ppb 1/cm3

The original variables, however, had to be changed for the purpose of this thesis. To train the machine learning model, we needed to have all the values of `Parameter_Meteo` and `Parameter_Air_Quality` in separate columns. This, together with SHAP, was needed to be able to see what features and how much they influence our predictions. Also, we omitted the `Interval` variable since it had only one unique value, which would have not helped our predictions. Columns `Unit_Meteo` and `Unit_Air_Quality` were dropped too for a similar reason – there was only one unique value for each new variable created out of the `Parameter_Meteo` and `Parameter_Air_Quality` columns.

Lastly, since new variables were created out of `Parameter_Air_Quality` and `Parameter_Meteo` columns, new missing values were created too. This occurred because not all of these values were measured every day. Due to the fact that deleting such observations would have extensively reduced the data size, we decided to replace our missing values with the corresponding mean values.

4.1.2 Data Visualization

To get a better understanding of the data, we examine a few visualizations. We are especially interested in inspecting the `temperature` variable since it will be later defined as a dependent variable for our machine learning models.

Figure 2 depicts the correlation coefficients for our meteo and air quality variables. By looking at the `temperature` row/column, we may see a rather strong correlation with some other variables. The highest coefficient can be seen where the relationship with `T_max_h1` is examined, however, this is basically the same variable and should not be taken into account when making predictions later. The second strongest correlation is visible with the ozone variables, especially with `O3_max_h1` and `O3`. Furthermore,

there is some nonnegligible relationship between `temperature` and `global_radiation`. If these variables, however, could actually help with our predictions will be seen in the *Chapter* 7.

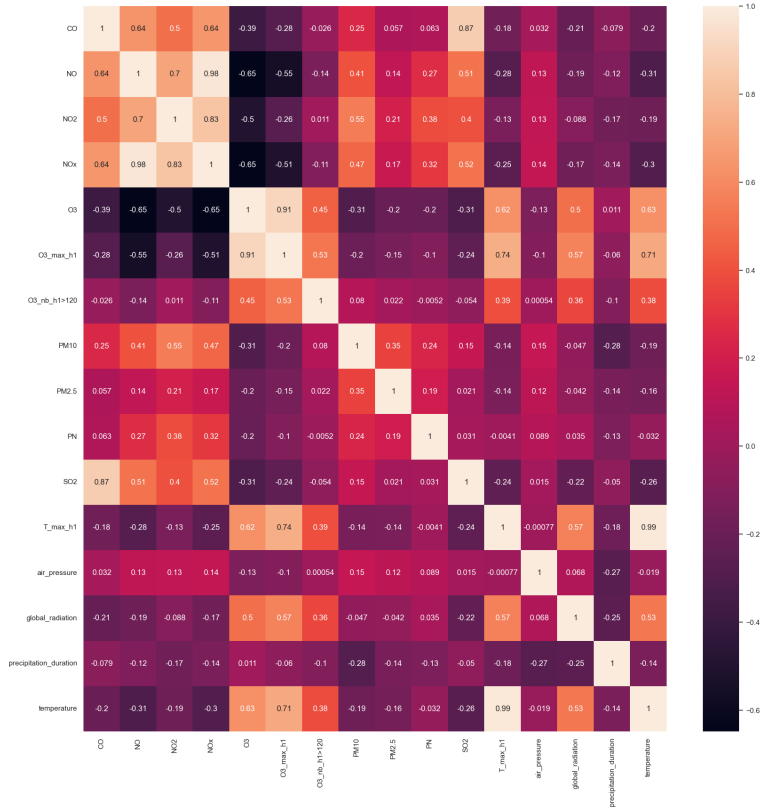


Figure 2: Correlation matrix of meteo and air quality variables

To observe how the levels of `temperature`, `O3`, and `global_radiation`, are developing over time, we look at these three graphs below. Because of the high correlation between `temperature` and `T_max_h1` as well as `O3` and `O3_max_h1`, we decide not to examine the development of `T_max_h1` and `O3_max_h1` over time.

In the first line plot, the temperature levels measured in °C can be seen over years 1992 to 2022. Even though the measurements over the years seem to be rather constant in general, we could make some more assumptions. For instance, the levels in recent years appear to be more extreme than the ones from thirty years ago.

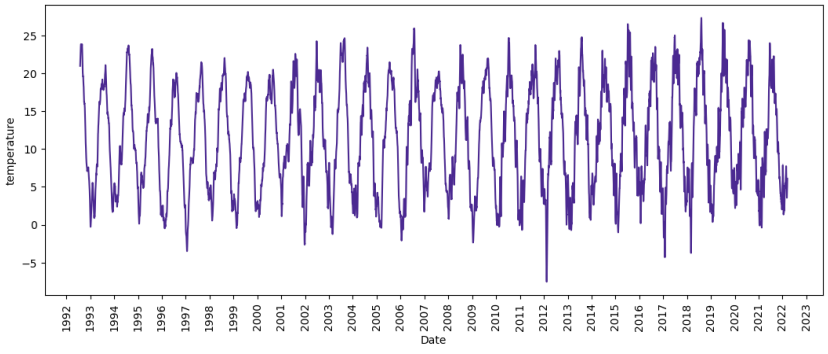


Figure 3: Temperature levels [^{o}C] over years 1992-2022

Figure 4 shows the levels of ozone measured in µg/m3 over the same time period. Similar to the chart before, the differences between the minimum and maximum values seem more substantial in recent years.

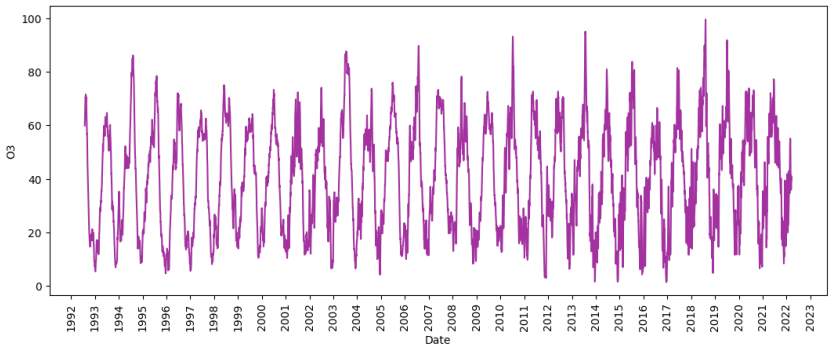


Figure 4: O3 levels $[\mu g/m^3]$ over years 1992-2022

The third graph of this type, *Figure* 5, shows the development of global radiation values measured in `W/m2`. Here, the change over years is most visible, and it is almost the opposite of the other two charts. The straight line between the years 2008 and 2010 indicates that there were no measurements taken during that time period.

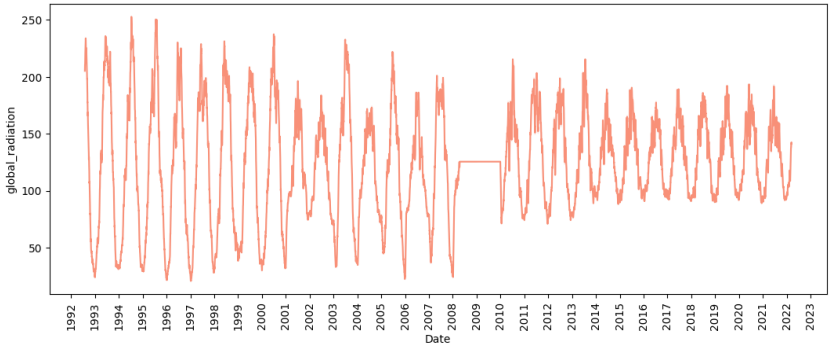


Figure 5: Global radiation levels $[W/m^2]$ over years 1992-2022

The last two graphs portray the relationships `O3` and `global_radiation` have with `temperature`. These regression plots show that with higher temperatures, both ozone and global radiation levels rise.

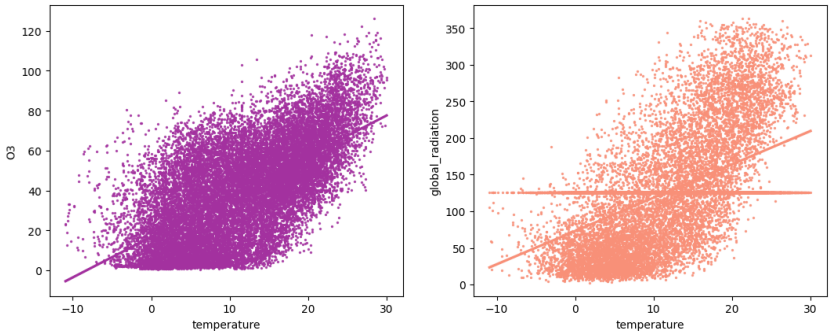


Figure 6: Regression plots: temperature vs. O3 and temperature vs. global radiation

4.2 Machine Learning for Weather Predictions

Below, the reasons for our model choice are briefly explained for each algorithm separately by summarizing relevant literature.

4.2.1 Machine Learning Models

Linear Regression The first and simplest algorithm performed on our data is Linear Regression. In spite of being one the most basic machine learning models, Linear Regression is oftentimes used for making predictions. For example, Holmstrom et al. [15] use Linear Regression to forecast weather in Stanford, CA. The dataset they use in this research comes from Weather Underground[17] and was collected for years 2011-2015. They use nine

[17] `https://www.wunderground.com/`

different variables for their machine learning models, which of some are `maximum temperature` and `precipitation`. Although professional weather forecasting outperforms Linear Regression in their case, the advantages of this machine learning model are indubitable. That is to say, Linear Regression could potentially outperform professional models over longer time periods [15].

In [16], the authors use a stepwise Linear Regression model to predict daily maximum stream temperatures. They use data for the Truckee River in California and Nevada and some of the variables available in their dataset are `hourly stream temperature` and `average daily flow`. The results of the stepwise procedure show that `daily maximum air temperature` and `average daily flow` features are able to predict maximum daily stream temperature at Reno, Nevada. In the next chapter, we will see whether Linear Regression is able to make satisfactory predictions for our research too.

Decision Tree The inspiration for selecting a Decision Tree as our next model stems from various sources. One of them is a paper written specifically on Decision Trees in weather predictions [17]. Here, they aim to forecast temperature in Hong Kong using parameters such as `relative humidity` or `average temperature`. In our thesis, however, we make use of various machine learning models and work with different software for our calculations (in their case Weka(Waikato Environment for Knowledge Analysis)[18] is used).

Another example of weather prediction using the Decision Tree model is outlined in [18]. In this article, Pekel tries to estimate soil moisture with the help of various parameters, such as `time` and `soil temperature`. In contrast to our research, Pekel concludes that the performance of the Decision Tree algorithm is satisfactory with R Squared value of 0.842 [18].

[18] `https://www.weka.io/`

Random Forest Making use of a Random Forest model after implementing a Decision Tree algorithm is a natural choice. The Random Forest model is based on a set of Decision Trees while often performing better than the simpler algorithm. *"Instead of looking for the best choice, a random subset of features (for each tree) is used, trying to find the threshold that best separates the data"*, explains Bonaccorso in [10]. This technique is implemented inter alia by Hill et al. [19] to forecast severe weather. They use various dynamical model fields (for example, `mean sea level pressure` and `relative humidity two meters above ground`) and make use of nine years of such data. Their predictions are rather accurate which confirms our initial thought that this is the right model to use with our type of data.

Random Forest algorithm is also used in [11] as one of the four machine learning algorithms used for rain prediction. In this research, they work with the 'Rain in Australia' dataset from Kaggle[19]. In their case, Random Forest performs similarly to Logistic Regression whilst outperforming Decision Tree and falling behind the performance of an XGBoost model.[20] Our results, as explained later, are slightly different, however, similar to this research, Random Forest is neither the best nor the worst-performing model in our thesis too.

LightGBM The last model used in our thesis is the LightGBM algorithm. This model is a type of gradient boosting Decision Tree which is *"a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability"* [20]. According to [21], this algorithm outperforms other gradient boosting methods in terms of accuracy and computational speed. This is also the case for weather predictions which is explained by Liu in [22]. According to Liu, the LightGBM model is able to quickly and

[19] `https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package`
[20] `https://xgboost.readthedocs.io/en/stable/`

automatically recognize three main types of severe weather whilst outperforming the other models when looking at accuracy and false alarm. For these reasons, we came to the conclusion that the LightGBM could perform pretty accurately which is affirmed in the Evaluation section.

4.2.2 Machine Learning Implementation

Almost all of our machine learning models are implemented using Python's `scikit-learn` library.[21] This is a *"simple and efficient tool for predictive data analysis"* [23] which is open source too. The only machine learning algorithm for which we have to use a different package, the `lightgbm` package [24], is the LightGBM framework.

Before running the models themselves, we had to do some more data pre-processing for the purpose of machine learning. First, we had to change the type of `Date` column to integer so that it can be used in the models. Then, we created dummy variables out of `Location` and `Status` columns to be able to use them as features in machine learning. Therefore, instead of the `Location` variable, we ended up with columns `ML_Zch_Schimmelstrasse` and `ML_Zch_Stampfenbachstrasse`, and instead of `Status` variable, the `ML_revised` column was created. The last step before the dataset was ready was dropping the `T_max_h1` column which would certainly bias our predictions. The final list of variables, their mean, minimum, and maximum values can be found in *Appendix*.

Next, as already mentioned before, we defined `temperature` as our dependent variable. The last step that had to be taken before running the models was dividing the dataset into training (80 percent of the dataset) and testing (20 percent of the dataset) samples.

[21] `https://scikit-learn.org/stable/modules/classes.htmachinelearning`

For all our models, we first fitted the respective model to the training data (*Figure* 7). Then, we used the `predict` function to make the predictions which is depicted in *Figure* 8.

```
# Fit the model
lr = LinearRegression()
lr = lr.fit(X_train, Y_train)
```

Figure 7: Fitting the linear regression model to the training dataset

```
# Make predictions
Y_pred_lr = lr.predict(X_test)
```

Figure 8: Making predictions for the linear regression model

5 Evaluation

To evaluate and compare the different machine learning models used, we implemented suitable performance metrics for regression algorithms. Below, we mention five popular performance metrics (R Squared, Adjusted R Squared, Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE)[22]) together with the reasons for their choice and their performance on our models (*Table* 3).

5.1 Performance Metrics

R Squared and Adjusted R Squared The first evaluation metrics used are R Squared and Adjusted R Squared. These metrics are, according to Chicco et al. [25], highly informative because they are bounded and cannot go to infinity. In this sense, R Squared alone would be more explanatory than the other metrics used in this research, however, more metrics give insights into the performance from various angles. R Squared was used, for instance, by Stern in [26] to evaluate the accuracy of the weather forecast in Melbourne, Australia. We decided to include Adjusted R Squared too since this metric, in contrast to R Squared, penalizes for adding independent variables which do not improve the model.[23]

Mean Square Error and Root Mean Square Error Other popular metrics that are, as reported by Botchkarev [27], one of the most frequently used performance metrics in research studies are MSE and RMSE. Both of these metrics are *"sensitive to large errors, to large variance of errors, and to errors due to outliers"* [28].

[22] https://medium.com/analytics-vidhya/performance-metrics-regression-model-69f68a18504f

[23] investopedia, https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp

For example, in a paper written by Pandey et al. in [29], they use Adaptive Neural Fuzzy Inference Systems[24] and fuzzy logic[25] methods on weather data while using MSE to compute the predictive power of these algorithms. RMSE is also used used when forecasting weather to assess the predictive ability of machine learning models (Linear Regression, Multiple Linear Regression,[26] Support Vector Regression,[27] and Auto Regressive Integrated Moving Average[28]) in [30].

Mean Absolute Error The last metric used to evaluate the model performance is the MAE measure. Similar to MSE and RMSE, this metric is also *"sensitive to outlier errors"* [28]. In [31], the author uses, in addition to RMSE and other metrics, the MAE measure to calculate the performance of various regression models, such as Linear Regression and Regression Tree. In this case, Jahnavi [31] is also interested in the analysis of weather data by looking at different primary atmospheric parameters of which temperature is the main feature in our research too.

In *Table* 3, all five metrics are compared for each machine learning model. The models are ordered according to their performance, and the best-performing model is highlighted in orange.

[24] https://www.youtube.com/watch?v=HPaqPHT08vY&ab_channel=AmitMishra

[25] https://www.techtarget.com/searchenterpriseai/definition/fuzzy-logic

[26] https://corporatefinanceinstitute.com/resources/knowledge/other/multiple-linear-regression/

[27] https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[28] https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Table 3: Machine learning models performance comparison

	$R^{2\,a}$	${R^2}_{adj}{}^{b}$	MSE	RMSE	MAE
Linear Regression	0.404	0.403	34.321	5.858	4.727
Decision Tree	0.558	0.557	25.447	5.044	3.686
Random Forest	0.737	0.736	15.157	3.893	2.988
LightGBM	0.828	0.828	9.875	3.142	2.463

[a] R Squared
[b] Adjusted R Squared

Looking at *Table* 3, we may notice that even if reviewing each performance metric separately, we would conclude for all of them that the best-performing model is LightGBM. It explains more than twice the variance explained by Linear Regression when looking both at R Squared and Adjusted R Squared. Also, the value of MSE for LightGBM is almost a fourth of the value of MSE for Linear Regression. All in all, the performance differences between the models are clearly visible when looking at all the evaluation measures used. The biggest improvement in the model performance can be seen for the Random Forest algorithm (almost 18 percent difference in explained variance compared to Decision Tree). On the contrary, the smallest but still considerable improvement may be observed for LightGBM (roughly 9 percent difference in explained variance compared to Random Forest).

6 SHAP

In this chapter, we first give a short introduction to xAI in machine learning in general. Next, we look at a few weather prediction models where SHAP was used by other researchers to explain and comprehend the respective results. Afterward, we describe how SHAP is used in our thesis, and compare the SHAP visualizations of our four machine learning models. Here, we also briefly summarize why using SHAP was beneficial for the purpose of this thesis. Last, we discuss a simple feature selection that we performed thanks to the SHAP plots.

6.1 Explainability in Machine Learning

The opaqueness of machine learning models for humans is just one of the reasons why xAI is crucial to understand the results of the respective algorithms [32]. In cases like ours, when the models are too complex, relevant techniques need to be employed to understand the algorithms and their results [33]. In this thesis, Shapley values are used to explain our machine learning models, however, other techniques, such as layer-wise relevance propagation,[29] are frequently applied too [33].

The relevance and different approaches to explainability in machine learning are further outlined in [32]. There, Burkart and Huber highlight the importance of understanding the decision making in delicate matters like health. Even though the accuracy of temperature predictions is not life-threatening, as discussed at the beginning of this thesis, our topic is directly associated with climate change which is indeed a pressing issue. Therefore, explainability is an important aspect of our work since underlining environmental problems is one of our main motivations.

[29] https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea

As Gilpin et al. add in [34], xAI does not serve merely to explain the model results but is also *"important to ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected"*. Altogether it is recommended in both [34] and [35] to use diverse metrics that conform to the intent and thoroughness of the respective explanation.

6.2 Weather Prediction Model

Because of SHAP's popularity, it is used by researchers to forecast weather too. For instance, Straaten et al. [36] applies SHAP to *"discover subseasonal drivers of high summer temperatures in western and central Europe"*. Another research performed by Beucler et al. [37] makes use of the SHAP approach to understand climate invariance.

Other researchers make use of SHAP to explain the relationships between weather and various other fields, such as healthcare [38], food industry [39], or aviation [40]. In [38], the authors use SHAP to explain the performance of a *"random forest-based model for estimating the occurrence of heat-related mortality in a detailed spatial unit within a city"* [38]. The SHAP approach helped them to identify the most important sectors when estimating heat-related mortality. Next, Zhu et al. [39] perform a SHAP analysis to find out what extreme weather predictors and how they contribute to yield shock events. Last, SHAP is also used together with the XGBoost models to give trustworthy explanations that can help to make aviation operations more economical and safer [40].

6.3 The Application of SHAP

The next step of our research is the implementation of SHAP. To interpret our machine learning models using this game theoretical approach, we decided to compare the respective summary plots. First, however, we had to calculate Shapley values using `shap.Explainer`.[30] For each algorithm, we specified the `shap.Explainer` so that it fits the model. Thus, for Linear Regression, we used the `shap.LinearExplainer` (*Figure* 9), and for the other three models (Decision Tree, Random Forest, and LighGBM), we switched to the `shap.TreeExplainer`. In addition, when calculating Shapley values for our Linear Regression model, we had to specify a so-called *masker*[31] (*Figure* 10). Here, masker provides background data for our `shap.LinearExplainer` to work properly. After calculating respective Shapley values, we created summary plots for each of our four models (*Figure* 11).

```
# Create a masker
masker = shap.maskers.Independent(X_train)
```

Figure 9: Specifying masker when calculating Shapley values for the linear regression model

```
# Calculate shapley values
explainer_lr = shap.LinearExplainer(lr, masker=masker)
shap_values_lr = explainer_lr.shap_values(X_train)
```

Figure 10: Calculating Shapley values for the linear regression model

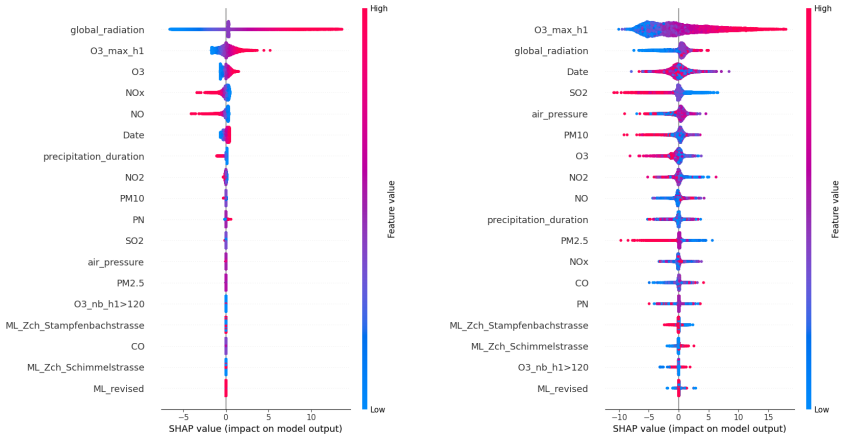
[30] https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html?highlight=explainer

[31] https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html?highlight=explainer

```
# SHAP summary plot
shap.summary_plot(shap_values_lr, X_train)
```

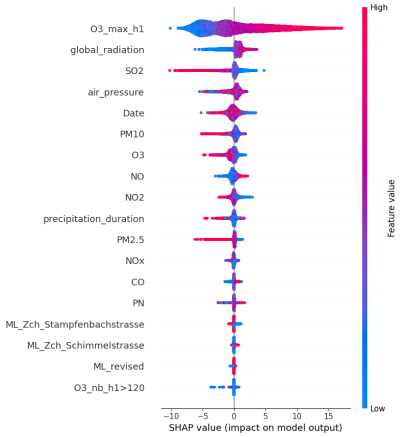
Figure 11: Creating a SHAP summary plot for the linear regression model

To compare our models, we may look at *Figure* 12 displaying the four summary plots in which we can see what features and how strongly they effect our predictions. For instance, `O3_max_h1` seems to have the strongest impact on our predictions when looking at the results of our tree models. This is, however, not true for the Linear Regression model, where `global_radiation` has the strongest effect. Another apparent difference among the graphs is that for the first graph, a lot of the features seem to have almost no effect on our predictions. But when looking at charts (b), (c), and (d) here, the effect of the variables at the bottom of the list is more visible.

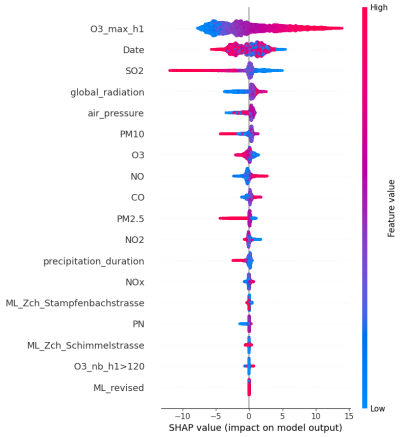


(a) Linear Regression

(b) Decision Tree



(c) Random Forest



(d) LightGBM

Figure 12: SHAP summary plots comparison

All in all, we may conclude that exploring SHAP visualizations is definitely useful. Without them, we would not be able to see what features and how much they impact our predictions for each model. When looking at the SHAP visualization of our best-performing model, we can conclude that the most important features when predicting `temperature` are `O3_max_h1`, `Date`, and `SO2`. Nonetheless, there exist various limitations of SHAP that one needs to be aware of when working with it, and these are briefly mentioned in *Chapter* 7.

6.4 Feature Selection

To get the best performance possible with our models, we decided to perform feature selection on our best-performing model, LightGBM. Even though the difference between LightGBM and LightGBM with feature selection is not drastic (*Table* 3), there is an improvement. By omitting the least useful variables for predictions, namely `NOx`, `ML_revised`, `ML_Zch_Schimmelstrasse`, `ML_Zch_Stampfenbachstrasse`, `PN`, `O3_nb_h1>120`, `NO_2`, and `precipitation_duration`, we manage to slightly improve our results. Without SHAP, this would be, however, not possible. Therefore, we may conclude that SHAP not only improves the model understanding but also plays an important role in refining the model itself.

Table 4: Comparison of the performance of LightGBM and LightGBM with feature selection

	R^2	R^2_{adj}	MSE	RMSE	MAE
LightGBM	0.828	0.828	9.875	3.142	2.463
LightGBM with Feature Selection	0.833	0.833	9.585	3.096	2.437

7 Limitations

Despite training several models, analyzing various performance metrics, and implementing feature selection, there are still some limitations of our machine learning algorithms as well as SHAP. Below, three such weaknesses are introduced.

7.1 Data Bias

Data bias occurs when *"the available data is not representative of the population or phenomenon of study"*.[32] For our research, only one data source, *data.europa.eu*, is used. Being an official open data portal of the European Union makes it a reliable source of information, however, it can still be a potential subject to bias in our data. Gathering data from multiple sources could reduce the chance of such an error. Nevertheless, data bias occurring in our research could have negative effects in terms of weather prediction. As already stated, the change in weather is one of the indicators of climate change [3]. Hence, since impacting human lives, the issue of incorrect weather predictions is far more severe than it can seem at first sight.

7.2 Missing Values

After completing the data pre-processing, quite some missing values were created. Losing all these observations just because of a few missing values in one row would considerably decrease the size of the dataset. This would result in a smaller training and testing dataset, and the predictions would most probably not be as accurate which would become a limitation itself. Hence, we decided to replace all the missing values with respective mean values. However, since such values are not the real ones, this may result in imprecise predictions, thus erroneous results of our models. Still,

[32] `https://towardsdatascience.com/survey-d4f168791e57`

we believe replacing missing values instead of deleting them was a better solution because of the reasons explained above.

7.3 SHAP

Even though SHAP is a powerful tool, it has its limitations. For instance, SHAP relies solely on mathematics, however, oftentimes the common sense of a human mind can result in better assessments [41]. In addition, SHAP explains the correlations only, and it is also limited to the features of a model which can sometimes not reflect reality correctly.[33] Another limitation that comes with using SHAP is the issue of multicollinearity. It can happen that if there are variables with high multicollinearity, one of them could be predicted as more important than in reality whilst the contrary would happen for the other variable.[34] Regardless of SHAP's limitations, we gain more insight into our data than if we would have not used it.

[33] https://towardsdatascience.com/using-shap-for-explainability-understand-these-limitations-first-1bed91c9d21

[34] https://towardsdatascience.com/using-shap-for-explainability-understand-these-limitations-first-1bed91c9d21

8 Conclusion and Future Work

8.1 Conclusion

To conclude our work, first, we would like to summarize our key findings. By making use of respective literature, we were able to explain the results of four machine learning models. We carefully chose the method for explaining our models as well as the models themselves. Additionally, we pre-processed two raw datasets so that they could be used for machine learning. All in all, we performed end-to-end research that not only leads to accurate results but also builds a foundation for future analysis.

8.1.1 Research Questions

Here, we will revisit and answer our four research questions in the following order – main research questions, subquestion regarding the state of art, subquestion about data and machine learning, and subquestion regarding SHAP.

How effective is SHAP for providing explanations for existing weather prediction models?

The effectiveness of SHAP for providing explanations for weather prediction models is not only documented in previous research, but is also demonstrated in our own work. The added value of SHAP in our research lies in enhancing the understanding of our results, as well as being able to further improve our models.

How effective are existing machine learning approaches in forecasting weather?

As mentioned before, using machine learning models to forecast weather is rather common. When opting for the right models, these predictions can be accurate and informative.

What datasets and how do they need to be adjusted in order to use them for machine learning?

The whole process of choosing the right datasets as well as preprocessing them is explained in our thesis. We had to make sure that our data comes from a reliable source and is suitable for our research. Still, we had to make quite a few changes to our data so that we would be able to use it for machine learning. These, among other things, included handling missing values, pivoting our dataset, or dropping irrelevant variables. These modifications can often be very tedious and require a lot of time, however, they are an integral part of our work. That is, without correct data there would be no correct results.

How effective is SHAP when compared to other approaches to xAI?

As previously noted, SHAP is a very popular game theoretic approach for xAI. It is, therefore, no surprise that SHAP is, as explained before, one of the most complete xAI methods. Its universal usability is the main reason why we chose this approach in particular.

8.2 Future Work

For future work, various points could be mentioned. We decided to briefly discuss the possibility of using more data for the predictions, upgrading our machine learning algorithms, as well as adding more techniques to explain and understand our machine learning models.

8.2.1 More Data

Using more data for our models could potentially improve our predictions. Not only adding more historical data but also including more measurements for the time period observed would give us more information. This could then be used by the algorithms to train themselves and perform more accurately. Additionally, new data could be added by including more variables that would help with our predictions. None of these was, however, an option for our datasets since we made complete use of them.

Nonetheless, we managed to find similar datasets[35,36] that could help with our results. These were, however, measured on an hourly basis so we would not be able to combine them with our datasets. Still, we could apply the same machine learning models and SHAP approach to this data to see whether any new information would have been gained since not all the variables are the same as in our datasets.

8.2.2 Improvement of Existing Models

Even though we have already done some feature selection of our best-performing model, there is more that could be done to refine a machine learning model. For example, one could make use of the `sklearn.feature_selection` library to apply wrapper methods. As defined by Karagiannopoulos et al. [42], *"wrapper methods wrap the feature selection around the induction algorithm to be used, using cross validation to predict the benefits of adding or removing a feature from the feature subset used"*. These methods, even though quite computationally heavy, often perform better than so-called filter methods [42]. Ideally, both methods would be used and compared to achieve the best possible results for our

[35] `https://data.stadt-zuerich.ch/dataset/ugz_luftschadstoffmessung_stundenwerte`

[36] `https://data.stadt-zuerich.ch/dataset/ugz_meteodaten_stundenmittelwerte`

predictions. There even exists a method that combines these two approaches – it uses *"a low-cost filter method to rank features and a costly wrapper method to further eliminate irrelevant variables"* [28].

Possibly, new models could be added to our machine learning portfolio. We, however, chose to use the four models for the reasons explained in *Chapter* 4.

8.2.3 More Approaches to xAI

The knowledge gained by using the SHAP approach confirms that our choice to use this method was the right decision. Nevertheless, including more approaches, such as LIME, could lead to new findings and improve the overall understanding of our models. Thus, for potential future work, we recommend comparing the results of a few different techniques for xAI.

References

[1] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[2] OECD. Global forum on environment and economic growth, 2016.

[3] Environmental Protection Agency. Climate change indicators: Weather and climate, May 2021.

[4] Russell M. Church. The effective use of secondary data. *Learning and Motivation*, 33(1):32–45, 2002.

[5] Melissa Johnston. Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, 3(3):619–626, 2017.

[6] Roger Watson. Quantitative research. *Nursing Standard (2014+)*, 29(31):44, 2015.

[7] Terry Elizabeth Hedrick, Leonard Bickman, and Debra J Rog. *Applied research design: A practical guide*. Sage Publications, 1993.

[8] Bogdan Bochenek and Zbigniew Ustrnul. Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, 13(2), 2022.

[9] Hadil Shaiba, Radwa Marzouk, Mohamed K Nour, Noha Negm, Anwer Mustafa Hilal, Abdullah Mohamed, Abdelwahed Motwakel, Ishfaq Yaseen, Abu Sarwar Zamani, and Mohammed Rizwanullah. Weather forecasting prediction using ensemble machine learning for big data applications. *CMC-COMPUTERS MATERIALS & CONTINUA*, 73(2):3367–3382, 2022.

[10] Giuseppe Bonaccorso. *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd, 2018.

[11] Jürgen Dieber and Sabrina Kirrane. A novel model usability evaluation framework (muse) for explainable artificial intelligence. *Information Fusion*, 81:143–153, 2022.

[12] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of shap explanations. *arXiv preprint arXiv:2009.08634*, 2020.

[13] David Lubo-Robles, Deepak Devegowda, Vikram Jayaram, Heather Bedle, Kurt J Marfurt, and Matthew J Pranter. Machine learning model interpretability using shap values: Application to a seismic facies classification task. In *SEG International Exposition and Annual Meeting*. OnePetro, 2020.

[14] Soshiro Ogata, Misa Takegami, Taira Ozaki, Takahiro Nakashima, Daisuke Onozuka, Shunsuke Murata, Yuriko Nakaoku, Koyu Suzuki, Akihito Hagihara, Teruo Noguchi, Koji Iihara, Keiichi Kitazume, Tohru Morioka, Shin Yamazaki, Takahiro Yoshida, Yoshiki Yamagata, and Kunihiro Nishimura. Heatstroke predictions by machine learning, weather information, and an all-population registry for 12-hour heatstroke alerts. *Nature Communications*, 12:4575, 07 2021.

[15] Mark Holmstrom, Dylan Liu, and Christopher Vo. Machine learning applied to weather forecasting. *Meteorol. Appl*, 10:1–5, 2016.

[16] David W Neumann, Balaji Rajagopalan, and Edith A Zagona. Regression model for daily maximum stream temperature. *Journal of Environmental Engineering*, 129(7):667–674, 2003.

[17] Elia Georgiana Petre. A decision tree for weather prediction. *Universitatea Petrol-Gaze din Ploiesti*, 61(1):77–82, 2009.

[18] Engin Pekel. Estimation of soil moisture using decision tree regression. *Theoretical and Applied Climatology*, 139(3):1111–1119, 2020.

[19] Aaron J. Hill, Gregory R. Herman, and Russ S. Schumacher. Forecasting severe weather with random forests. *Monthly Weather Review*, 148(5):2135 – 2161, 2020.

[20] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[21] Essam Al Daoud. Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6 – 10, 2019.

[22] Xinwei Liu, Haixia Duan, Wubin Huang, Runxia Guo, and Bolong Duan. Classified early warning and forecast of severe convective weather based on lightgbm algorithm. *Atmospheric and Climate Sciences*, 11:284–301, 01 2021.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Microsoft Corporation. Python-package introduction, 2022.

[25] Jurman G Chicco D, Warrens MJ. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 08 2021.

[26] Harvey Stern. The accuracy of weather forecasts for melbourne, australia. *Meteorological Applications*, 15(1):65–71, 2008.

[27] Alexei Botchkarev. A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:45–76, 2019.

[28] Ioannis Kyriakidis, Jaakko Kukkonen, Kostas Karatzas, Giorgos Papadourakis, and J. Ware. New statistical indices for evaluating model forecasting performance. 08 2015.

[29] A. K. Pandey, C. P. Agrawal, and Meena Agrawal. A hadoop based weather prediction model for classification of weather data. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–5, 2017.

[30] Gaurav Chavan and Bashirahamad Momin. An integrated approach for weather forecasting over internet of things: A brief review. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 83–88, 2017.

[31] Yeturu Jahnavi. Analysis of weather data using various regression algorithms. *International Journal of Data Science*, 4(2):117–141, 2019.

[32] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.*, 70:245–317, 2021.

[33] Michal Choras, Marek Pawlicki, Damian Puchalski, and Rafal Kozik. Machine learning - the results are not the only thing that matters! what about security, explainability and fairness? In *Computational Science - ICCS 2020*, pages 615–628. Springer International Publishing, 2020.

[34] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[35] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019.

[36] Chiem van Straaten, Kirien Whan, Dim Coumou, Bart van den Hurk, and Maurice Schmeits. Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe. *Monthly Weather Review*, 150(5):1115 – 1134, 2022.

[37] Tom Beucler, Michael Pritchard, Janni Yuval, Ankitesh Gupta, Liran Peng, Stephan Rasp, Fiaz Ahmed, Paul A O'Gorman, J David Neelin, Nicholas J Lutsko, et al. "using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe". *arXiv preprint arXiv:2112.08440*, 2021.

[38] Yesuel Kim and Youngchul Kim. Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. *Sustainable Cities and Society*, 79:103677, 2022.

[39] Peng Zhu, Rose Abramoff, David Makowski, and Philippe Ciais. Uncovering the past and future climate drivers of wheat

yield shocks in europe with machine learning. *Earth's Future*, 9(5):e2020EF001815, 2021. e2020EF001815 2020EF001815.

[40] Alise Danielle Midtfjord, Riccardo De Bin, and Arne Bang Huseby. A machine learning approach to safer airplane landings: Predicting runway conditions using weather and flight data. *arXiv preprint arXiv:2107.04010*, 2021.

[41] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[42] M Karagiannopoulos, D Anyfantis, SB Kotsiantis, and PE Pintelas. Feature selection for regression problems. *Educational Software Development Laboratory, Department of Mathematics, University of Patras, Greece*, 2004.

A Appendix I

The following Table shows the overview of the variables present in the pre-processed dataset used for machine learning. For each feature, its name, mean, minimum, and maximum values are shown.

Table 5: Overview of the final dataset variables

Variable Name	Mean Value	Min Value	Max Value
Date	1,283,018,710.262	709,945,200	1,647,298,800
CO	0.548	0.070	3.980
NO	29.841	0.440	337.760
NO2	39.637	4.500	112.040
NOx	44.651	2.820	307.610
O3	40.097	0.510	126.210
O3_max_h1	68.554	0.990	266.390
O3_nb_h1>120	0.365	0	16
PM10	20.821	0.930	163.450
PM2.5	11.026	1.960	60.660
PN	14508.928	0	78,863.570
SO2	5.336	-0.020	99.510
air_pressure	966.931	931.500	992.940
global_radiation	125.492	2.320	363.270
precipitation_duration	131.204	0	1,440
temperature	11.525	-10.930	30.040
ML_revised	0.881	0	1
ML_Zch_Schimmelstrasse	0.310	0	1
ML_Zch_Stampfenbachstrasse	0.537	0	1

Note: Since the type of the Date variables was changed to integer for the purpose of machine learning, the values now represent the number of seconds since 1970-01-01 - Unix epoch time.
(https://unixtime.org/)