report_julian

January 9, 2024

course: Data Analytics with Statistics | lecturer: Prof. Dr. Jan Kirenz | Date: 29.12.2023 | Name: Julian Erath, Furkan Saygin, Sofie Pischl | Group: Group B

Weather Data Analysis: A Regression and Classification Approach on the ERA5 Dataset

1 Introduction and data

Scope of the Project:

The project aims to analyze and derive insights from the ERA5 (ECMWF Reanalysis version 5) weather dataset. It involves both regression and classification analyses to model and predict various weather parameters as well as gain insights on correlations, causations and patterns involving the parameters [^1]. The ERA5 dataset, provided by the European Centre for Medium-Range Weather Forecasts (ECMWF), is a high-quality global atmospheric reanalysis dataset covering multiple decades [^2]. The analysis will focus on the region of Bancroft in Ontario, Canada for the years 2015 to 2022.

1.1 Introduction to Subject - Motivation

Exploring the Atmosphere Through Data
Weather, a phenomenon as old as Earth itself, has long been a subject of human fascination. The complex interplay of temperature, wind, and precipitation shapes our environment, affects our lives, and challenges our understanding of the natural world [^3]. With the advent of technology and data analysis techniques, a deeper exploration of these complex processes is now possible. [^4]. The project, titled "Weather Data Analysis: A Regression and Classification Approach on the ERA5 Dataset", aims to contribute to this ongoing exploration by examining how different variables interact to create complex weather phenomena.

Geographic Focus: Bancroft, Canada
While this project leverages the ERA5 dataset, Bancroft, Ontario in Canada is chosen to specifically focus on, for an in-depth analysis. The geographic location of Bancroft can be seen under "../references/images/complete_map.png". This decision is rooted in the unique climatic and meteorological characteristics of the region, which offer a distinctive case study for understanding complex weather phenomena. The geographical and climatic characteristics of Bancroft (strongly influenced by the so called 'lake-effect') create an excellent case study for analyzing the relationships among different atmospheric elements, which also leads to interest in that area by IBM and IBM clients [^5]. The work presented is limited to the evaluation of the years 2015 to 2022 in order to include only the current and latest climate developments and to record the current status without falsifying the results due to strongly deviating values from observations further in the past.

Focusing on Bancroft enables a targeted application of regression and classification techniques, facilitating a detailed analysis of localized weather patterns and their impacts [^6].

1.2 Motivation for Research Question

The Impetus for Investigation
The motivation of this project is rooted in historical interest and present-day necessity. As discussed in a systematic review titled "The contribution of weather forecast information to agriculture, water, and energy sectors in East and West Africa," accurate weather prediction is crucial for effective agriculture, disaster management, and urban planning, particularly in the context of climate change risks[^7]. An example of the complexity in weather systems is the Lake Effect, explained in "A Hybrid Dataset of Historical Cool-Season Lake Effects From the Eastern Great Lakes of North America". The Lake Effect is a phenomenon where cold winds moving over warmer lake waters, leads to significant changes in weather patterns nearby[^8]. The European Centre for Medium-Range Weather Forecasts (ECMWF) and its ERA5 dataset provide an great opportunity to analyze detailed weather parameters over several decades. This data, examined for its comprehensiveness in the "Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China," is a example for researchers seeking to understand weather patterns and atmospheric dynamics[^9].

1.3 Research Questions

The research is guided by four pivotal questions, addressed through regression and classification analyses.

Regression Analysis:
Is it possible to accurately predict the temperature and its relationship with wind characteristics using historical data? This question is inspired by the current understanding of atmospheric variables and their interactions and specified as follows:

- *Temperature Prediction:* Is it possible to build an accurate regression model to predict temperature based on historical data? This involves exploring the relationships between temperature and windspeed over time (daytime, day, season, etc.). The result of this regression analysis is the identification of a trend in the weather data as well the prediction of the temperature for the next x days, weeks, etc. based on the historical weather.
- *Temperature and Wind Modeling:* Is it possible to find a correlation or causation between the temperature and windspeed, windgust or winddirection using regression techniques? This involves investigating the impact of independent variables (wind features) on the temperature. The result of this regression analysis is the identification of the correlation or causation between wind features and the temperature based on the historical weather. As an outlook to this analysis it could also be analyzed whether certain winddirections result in a certain type of weather event using logarithmic regression or classification.
- *Multivariate Temperature Prediction / Linear Regression of Temperature with Multiple Predictors:* How does the incorporation of multiple atmospheric predictors, such as windspeed, windgust, winddirection, air pressure, snow and ice parameters enhance the accuracy of temperature prediction compared to a model solely based on windspeed? This question delves into the complex interplay of various atmospheric variables and their combined effect on temperature changes over different time scales (daytime, day, season, etc.). The outcome of this analysis will contribute to a nuanced model that considers a spectrum of atmospheric condi-

tions for more precise temperature forecasting. This tasks also includes the regression model to uncover interactions and synergies among different atmospheric predictors in influencing temperature variations. This question seeks to identify potential non-linear relationships or combined effects of multiple predictors on temperature. Understanding these interactions is crucial for refining the predictive model and gaining insights into the intricate dynamics of the atmosphere. Furthermore temporal dynamics are to be analyzed. How does the temporal aspect, including diurnal patterns and seasonal variations, impact the relationship between temperature and multiple predictors? This question aims to capture the temporal dynamics of the atmosphere and assess how the predictive power of the model varies across different timeframes. Examining temporal patterns will contribute to a more nuanced understanding of how atmospheric predictors influence temperature fluctuations under varying conditions.

- *Extreme Weather Event Prediction by Temperature in Logistic Regression:* Can logistic regression effectively classify and predict the occurrence of extreme or normal weather events based on temperature (or alternatively windspeed) ranges? This question involves categorizing temperature into distinct classes and determining whether certain temperature ranges are indicative of extreme weather events. The outcome of this analysis will contribute to weather an observation is an extreme weather event, providing valuable insights into the correlation between temperature and specific meteorological phenomena. To achieve this, this tasks also includes to find out, what the critical temperature thresholds associated with the onset or intensification of different weather events are. This question aims to identify temperature breakpoints that signify transitions between different weather conditions. Understanding these thresholds is essential for developing early warning systems and enhancing preparedness for weather-related events.

Classification Analysis:
Can effective classification and prediction of weather events, including extreme occurrences, be achieved based on multivariate weather data? This aspect of the research aims to develop methods for accurate prediction of weather events, acknowledging the complexity and variability of weather patterns and is specified as follows:

- *Extreme Weather Events in Binary Classification:* Is it possible to classify and predict extreme weather events such as storms? This involves training a binary classification model to identify patterns indicative of extreme events. The result of this classification analysis is the prediction of extreme weather events based on the current weather data and a model that was trained on historical weather data.
- *Weather Event and Pattern Classification in Multiclass Classification:* Is it possible to categorize and predict different extreme weather events based on multivariate weather data? This involves using multiclass classification algorithms. The results of this classification analysis is the prediction of certain weather events based on the current weather data and a model that was trained on historical weather data.

1.4 Hypotheses Regarding Question

The hypotheses are twofold, corresponding to the projects dual analytical approach:

In Regression: There exists a significant correlation between temperature and wind characteristics, which can be modeled to predict future temperature trends and variations. This hypothesis is based on the premise that atmospheric variables are interconnected and can be analyzed to forecast

weather conditions.

In Classification: Specific patterns in the weather data can accurately predict various weather events, including extreme conditions. This hypothesis is informed by the need for effective prediction models in the face of increasingly frequent and severe weather events.

1.5 Data structure and description of sample

- **Source:** ERA5 dataset by ECMWF (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5)
- **Temporal Coverage:** Multiple decades (2015-2022) (https://confluence.ecmwf.int/display/CKB/ERA5%3
- **Spatial Resolution:** Approximately 31 km (https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+
- **Variables:** Includes but not limited to temperature, precipitation, wind speed, atmospheric pressure, etc. The variables are grouped in the average, minimum and maximum of the measured variable for the observed hour. (https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Parameterlistings)
- **Labels:** The data has been labelled in weather event profiles by meteorologists and data scientists from IBM and The Weather Company.
- **Observations:** The ERA5 weather dataset provides comprehensive atmospheric information on a global scale, with each observation (row) in the dataset representing a set of meteorological parameters at a specific location and time.

The ERA5 dataset is provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). It is a product of atmospheric reanalysis, a process that involves assimilating observational data from various sources into a numerical weather prediction model.

- **Temporal Aspect:** The dataset provides temporal information at hourly intervals, covering multiple years, allowing for both short-term and long-term analyses.
- **Spatial Aspect:** With a spatial resolution of around 31 km, the dataset provides a comprehensive global coverage, facilitating region-specific studies. In this project a geographical resolution for the region Bancroft in Ontario, Canada is chosen (as this area is of particular interest for IBM).
- **Multivariate Nature:** Multiple meteorological parameters are available, enabling a comprehensive analysis of weather patterns.
- **Quality:** ERA5 is known for its high quality and precision, making it suitable for detailed analyses and modeling.

1.6 Variables

The ERA5 dataset is comprehensive, providing a wide range of meteorological variables that cover various aspects of the Earth's atmosphere. The observations and general characteristics measured in the ERA5 dataset are crucial for understanding and analyzing weather and climate patterns. Some of the key variables included in the dataset, which are most important for the analysis of this assignment, are[^18]: 1. **Temperature:** ERA5 provides information on air temperature at different levels of the atmosphere. This includes 2-meter temperature, which is often used as a representative of surface temperature. 2. **Wind:** Wind speed, gust and direction are essential meteorological parameters. ERA5 includes information on both zonal (east-west) and meridional (north-south) components of the wind at different altitudes. 3. **Precipitation:** Precipitation data includes rainfall and snowfall. This information is crucial for understanding water cycles and is vital

for various applications, including hydrology and agriculture. 4. **Pressure:** Atmospheric pressure at different levels is provided in the dataset. Changes in atmospheric pressure are associated with weather patterns and can influence local weather conditions. 5. **Snow and ice:** Snow density as well as cumulative snow and ice are included.

These variables collectively provide a detailed snapshot of the Earth's atmosphere and are fundamental for conducting both regression and classification analyses. Researchers and meteorologists often leverage these variables to study climate trends, weather patterns, and environmental changes. The high spatial and temporal resolution of the ERA5 dataset enhances its utility for diverse applications, including climate research, environmental monitoring, and weather forecasting [^19].

1.7 Overview of data and structure

At first the data is loaded from the .csv file and the head of the data is printed to have a first look at the columns (variables) and rows (observations) of the data. The dataset consists of 65345 observations, 184 columns (including the useful and unique predictor variables and reasonable response variable mentioned in chapter "1.6 Variables", identifier variables "Unnamed: 0" (this might previously be used as an index and will be dropped for the further analysis becuase of redundancy), run_datetime and valid_datetime (which contain the datetime for the observation, which is not only an identifier variable but can also be used in the analysis e.g., time series analysis)). In a next step the distribution and statistics of the variables will be taken a closer look at by analyzing their quantiles, minima, maxima and maxima.

[2]:

```
   Unnamed: 0 substation         run_datetime       valid_datetime  horizon  \
0           0   Bancroft  2015-07-15 00:00:00  2015-07-15 00:00:00      0.0
1           1   Bancroft  2015-07-15 01:00:00  2015-07-15 01:00:00      0.0
2           2   Bancroft  2015-07-15 02:00:00  2015-07-15 02:00:00      0.0
3           3   Bancroft  2015-07-15 03:00:00  2015-07-15 03:00:00      0.0
4           4   Bancroft  2015-07-15 04:00:00  2015-07-15 04:00:00      0.0

     avg_temp  avg_windspd  avg_windgust  avg_pressure_change  avg_snow  …  \
0  287.389224     3.386380     11.136197            52.892217       0.0  …
1  287.378997     3.326687     11.002795            50.256685       0.0  …
2  287.388845     3.243494     10.700595            47.944054       0.0  …
3  287.427324     3.145505     10.323983            45.855264       0.0  …
4  287.489158     3.047607      9.921157            44.823453       0.0  …

   MA_avg_winddir_month  MA_avg_temp_change_month  label0  label1  label2  \
0             80.302464                       NaN       0       1     NaN
1             78.584418                 -0.010227       0       1     NaN
2             77.809235                 -0.000189       3       1     NaN
3             77.931830                  0.012700       2       1     NaN
4             79.272034                  0.024983       2       1     NaN

   label3           wep  storm_id2  year  month
0       1  Blue sky day        NaN  2015      7
1       1  Blue sky day        NaN  2015      7
2       1  Blue sky day        NaN  2015      7
```

```
3       1  Blue sky day        NaN  2015      7
4       1  Blue sky day        NaN  2015      7

[5 rows x 184 columns]

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65345 entries, 0 to 65344
Columns: 184 entries, Unnamed: 0 to month
dtypes: float64(166), int64(8), object(10)
memory usage: 91.7+ MB
None
Unnamed: 0          int64
substation         object
run_datetime       object
valid_datetime     object
horizon           float64
                   …
label3              int64
wep                object
storm_id2          object
year                int64
month               int64
Length: 184, dtype: object
```

[4]:

```
             run_datetime            wep    avg_temp  avg_temp_celsius  \
0     2015-07-15 00:00:00   Blue sky day  287.389224         14.239224
1     2015-07-15 01:00:00   Blue sky day  287.378997         14.228997
2     2015-07-15 02:00:00   Blue sky day  287.388845         14.238845
3     2015-07-15 03:00:00   Blue sky day  287.427324         14.277324
4     2015-07-15 04:00:00   Blue sky day  287.489158         14.339158
…                    …            …          …                 …
65340 2022-12-27 04:00:00  Moderate rain  264.241641         -8.908359
65341 2022-12-27 05:00:00   Blue sky day  264.115391         -9.034609
65342 2022-12-27 06:00:00   Blue sky day  264.024853         -9.125147
65343 2022-12-27 07:00:00   Blue sky day  264.048368         -9.101632
65344 2022-12-27 08:00:00   Blue sky day  263.918722         -9.231278

       min_wet_bulb_temp  avg_dewpoint  avg_temp_change  avg_windspd  \
0             280.809506    280.735246              NaN     3.386380
1             280.809506    280.414058        -0.010227     3.326687
2             280.809506    280.187074         0.009848     3.243494
3             280.809506    280.049330         0.038479     3.145505
4             280.809506    279.980697         0.061834     3.047607
…                    …            …                …            …
65340         260.284794    262.061976        -0.124561     1.962197
65341         260.284794    262.114357        -0.126250     1.978823
65342         260.284794    262.206179        -0.090537     2.005855
65343         260.284794    262.350025         0.023514     2.040978
```

```
65344         260.284794    262.512490        -0.129646     2.078741

       max_windgust  avg_winddir  …  avg_winddir_cos  wind_direction_label  \
0         14.899891    80.302464  …         0.190676                  East
1         14.899891    76.866373  …         0.102466                  East
2         14.899891    76.258867  …         0.651950                  East
3         14.899891    78.299616  …        -0.971290                  East
4         14.702229    84.632852  …        -0.981976                  East
…             …          …  …              …                     …
65340      8.444256   232.606824  …         0.991695             Southwest
65341      7.475906   229.938704  …        -0.823955             Southwest
65342      7.305549   227.024163  …         0.675251             Southwest
65343      7.305549   223.900355  …        -0.662027             Southwest
65344      6.818578   220.894487  …         0.554528             Southwest

      max_cumulative_precip  max_snow_density_6  max_cumulative_snow  \
0                     2.009                 0.0                0.000
1                     1.209                 0.0                0.000
2                     0.400                 0.0                0.000
3                     0.000                 0.0                0.000
4                     0.000                 0.0                0.000
…                     …                 …                    …
65340                 2.126                 0.0               25.643
65341                 2.226                 0.0               21.161
65342                 2.426                 0.0               16.430
65343                 2.826                 0.0               10.859
65344                 3.426                 0.0                5.640

       max_cumulative_ice  avg_pressure_change  label0  label1  label2
0                     0.0            52.892217       0       1     NaN
1                     0.0            50.256685       0       1     NaN
2                     0.0            47.944054       3       1     NaN
3                     0.0            45.855264       2       1     NaN
4                     0.0            44.823453       2       1     NaN
…                   …                  …     …     …     …
65340                 0.0                  NaN       5       0     3.0
65341                 0.0                  NaN       5       1     NaN
65342                 0.0                  NaN       5       1     NaN
65343                 0.0                  NaN       5       1     NaN
65344                 0.0                  NaN       0       1     NaN

[65345 rows x 21 columns]
```

1.8 Variable lists and overview

In the pursuit of extracting meaningful insights and constructing predictive models from a given dataset, it is imperative to establish a comprehensive understanding of the variables involved. This section delineates the variables within the dataset, providing a structured overview of each

variable's key characteristics through a data dictionary. The data dictionary encapsulates essential information for each variable, facilitating a streamlined comprehension of their names, descriptions, roles, types, and formats.

While not explicitly used as predictors or responses in modeling, the inclusion of ID columns can offer valuable insights for data exploration and understanding. These columns serve as unique identifiers and are crucial for tracking and referencing specific records within the dataset.

The meticulous documentation provided in this variable list serves as a foundation for subsequent analyses, ensuring transparency, reproducibility, and informed decision-making throughout the data exploration and modeling processes. A detailed data dictionary can be seen in the apendix.

1.9 Response Variables

Temperature: The variable temperature is a numerical continous variable, which is measured in units of Kelvin. Temperature is a measure of the thermal energy in a substance or system. It is a fundamental physical quantity. Temperature is a measure of the average kinetic energy of the particles in a substance or system. As the temperature of a substance increases, the particles in the substance gain kinetic energy and move faster. In the dataframe there are different kinds of temperature given.

Weather Event: The weather event is a categorical nominal variable with the following values: **Blue Sky Day:** A day with clear and cloudless skies. There are no significant weather phenomena such as rain, snow, or storms. **Mild Snowfall:** A light snowfall, where small amounts of snow fall, but not enough to be considered moderate or heavy snowfall. **Moderate Snowfall:** A moderate snowfall, where larger amounts of snow fall compared to mild snowfall but not as much as heavy snowfall. **Moderate Rain:** A moderate rainfall, where a moderate amount of precipitation occurs in the form of rain. **Heavy Snowfall with Accumulated Snow:** Intense snowfall accompanied by a significant accumulation of snow on the ground. **Continuous Freezing Rain:** Continuous freezing rain, where rain falls on frozen surfaces and immediately turns into ice. **Storm with Freezing Rain / Heavy Snow- and Icestorm:** A storm with freezing rain or heavy snowfall, accompanied by the formation of ice. This can lead to hazardous conditions. **Snowstorm with High Precipitation:** A snowstorm with intense snowfall and a high amount of precipitation in the form of snow.

2 Methodology

REMOVE THE FOLLOWING TEXT

This section includes a brief description of your modeling process.

Explain the reasoning for the type of model you're fitting, predictor variables considered for the model.

Additionally, show how you arrived at the final model by describing the model selection process, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

2.1 Methodological Insights and Impact

This project employs a range of statistical and machine learning techniques to analyze the ERA5 dataset. The outcomes of this research are expected to not only enhance meteorological understanding but also provide practical tools for weather prediction, with implications for environmental policy and public safety. By harnessing the power of the ERA5 dataset [^10] and advanced analytical techniques [^11], this project aims to contribute valuable insights to the field of meteorology and support informed decision-making in a world increasingly affected by weather-related challenges.

2.2 Methodology Overview

This project uses a multifaceted approach, primarily utilizing Design Science Research (DSR) [^12] in line with Hesse's framework[^13]. This methodological framework focuses on the creation and critical evaluation of artifacts to address specific problems. In this DSR approach, complex weather phenomena are identified as problems to be addressed. Additionally, iterative prototyping [^14] is employed, enabling systematic refinement of models and methods based on continuous evaluation and integration of data-driven insights[^15]. This project combines the DSR cycle by Gregor / Hevner, with iterative prototyping by Wilde / Hess and Goldman / Narayanaswamy. This integration fosters a dynamic environment where each prototype's development and evaluation progressively inform subsequent cycles of design and analysis. This leads into a cycle of artifact creation (in this case, models and algorithms) specifically tailored to analyze weather patterns in Bancroft using the ERA5 dataset. These models will be refined continuously through iterative prototyping, where each iteration's outcomes inform the next cycle, ensuring they are increasingly effective and accurate. The artifacts are then rigorously evaluated against the research questions. The results are evaluated in every iteration using cross-validation by Shao 1993 and Browne 2000[^16]. This process is enriched by a comprehensive literature review by Webster / Watson [^17], conducted before and during the implementation, ensuring the methods and analyses remain aligned with current meteorological and data science advancements. The amalgamation of DSR, iterative prototyping, cross-validation and literature research forms the foundation of this approach, ensuring a thorough and robust analysis that is well-suited to address the complexities in atmospheric data analysis.

For the analysis the data was split into training and testing datasets according to [^22] and [^23]. [^22]: Scikit-learn (2023a): https://scikit-learn.org/stable/documentation.html [^23]: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.

```
<Figure size 640x480 with 0 Axes>
```

<Figure size 640x480 with 0 Axes>

<Figure size 640x480 with 0 Axes>

Bancroft Distribution of Weather Features by Weather Event Profiles in Distograms
Avg temperature in Kelvin
Min wet bulb temperature in Kelvin
Avg dewpoint in Kelvin
Avg temperature change in Kelvin
Avg wind speed in m/s
Max windgust in m/s
Avg winddirection in degrees
Max cumulative precipitation in mm
Max snow density (when snow exceeds 6.35mm in kg/m^3)
Max cumulative snow accretion in mm
Max cumulative ice accretion in mm
Avg pressure change in Pa
Density
Blue sky day: 52633
Frondurchlauf / Continuous freezing rain: 934
Heavy snowfall with accumulated snow: 1040
Mild snowfall: 4497
Moderate rain: 2630
Moderate snowfall: 2920
Snowstorm with high precipitation: 261
Storm with freezing rain / Heavy snow- and icestorm: 430

```
<Figure size 640x480 with 0 Axes>
```

```
<Figure size 640x480 with 0 Axes>
```

3 Classification Results

3.1 Binary Classification of Extreme Weather Events

The following visualization displays four confusion matrices, each representing the performance of a different binary classification model: ExtraTrees, XGBoost, LightGBM (LGBM), and RandomForest. While all models demonstrate high accuracy, with a significant majority of instances correctly classified, which is indicative of their ability to discriminate between the two classes effectively. The LGBM classifier shows the least number of Type II errors, signifying its strength in identifying true extreme weather events with minimal misses. Conversely, the XGBoost classifier presents with the lowest Type I errors, suggesting it is more conservative in predicting extreme weather, thus minimizing false alarms. In practical applications, Type I errors can be particularly critical as they represent missed predictions of extreme weather, which are crucial for timely warnings and safety measures. Therefore, the XGBoost classifier might be preferred in scenarios where the cost of missing an actual extreme weather event is high. Each of these models offers a trade-off between sensitivity to detecting true events and specificity in avoiding false alarms, which needs to be carefully balanced according to the application's requirements and the consequences of prediction errors.

```
ExtraTrees Accuracy: 0.9930369576861274
XGBoost Accuracy: 0.9946438136047134
LGBM Accuracy: 0.9935725763256561
```

```
RandomForest Accuracy: 0.9932665085316398
```

```
<Figure size 640x480 with 0 Axes>
```

The classification reports provide an evaluation of the performance of different models.

ExtraTrees: The ExtraTrees model demonstrates high precision and recall for both classes, achieving an accuracy of 99.30%. The precision, recall, and F1-score for both extreme weather events (0) and blue sky events (1) are consistently high, indicating robust performance across both classes.

XGBoost: The XGBoost model exhibits excellent precision, recall, and F1-score for both classes, resulting in an overall accuracy of 99.40%. Similar to ExtraTrees, it shows strong performance in correctly classifying both extreme weather and blue sky events.

LightGBM: The LightGBM model achieves a high accuracy of 99.37%, with impressive precision, recall, and F1-score for both classes. Notably, it maintains a high recall for extreme weather events (0), ensuring that a significant proportion of these events are correctly identified.

RandomForest: The RandomForest model performs well, achieving an accuracy of 99.32%. It shows strong precision, recall, and F1-score for both extreme weather events (0) and blue sky events (1), indicating reliable performance across different weather scenarios.

In summary, all four models—ExtraTrees, XGBoost, LightGBM, and RandomForest—demonstrate robust performance in classifying weather events, with high accuracy and consistent precision and recall metrics across the evaluated classes.

3.2 Multiclass Classification of Various Extreme Weather Events

After successfully predicting extreme weather and blue sky day weather events, a key result of this research should be the prediction of specific extreme weather events. Once it is determined, that an observation is an extreme weather event, it's important to analyse what specific kind of extreme weather event it is. These results can then be used by scientists and governmental institutions to take countermeasures to prevent damage and minimize the risk for a weather event to be hazardous. The analysis for the classification of specific weather events and patterns is conducted using multiclass classification techniques. The research question to be answered ist: Is it possible to categorize and predict different extreme weather events based on multivariate weather data? This involves using multiclass classification algorithms. The results of this classification analysis is the prediction of certain weather events based on the current weather data and a model that was trained on historical weather data.

K-Nearest Neighbors Classification Report
Support Vector Machine Classification Report
Decision Tree Classifier Classification Report
Gradient Boosting Classifier Classification Report
Classes
Metrics
precision
recall
f1-score
3.0
5.0
0.0
1.0
6.0
4.0
2.0
accuracy
macro avg
weighted avg

The multiclass classification with K-Nearest Neighbors generally yields good results, with the actual outcomes closely aligning with the predictions. The classification report provides a comprehensive overview of the model's performance across multiple classes. Precision, representing the accuracy of positive predictions, is generally high, ranging from 78% to 100%. Recall, which measures the ability of the model to capture all relevant instances, shows consistent performance, with values ranging from 78% to 100%. The F1-score, which considers both precision and recall, is strong across most classes, indicating a balanced trade-off between precision and recall. The macro average F1-score is 0.88, suggesting a good overall performance. The weighted average F1-score, considering class imbalance, is also 0.92, demonstrating the model's effectiveness in making accurate predictions across the entire dataset. The high accuracy of 92% further supports the model's reliability in classifying instances from diverse classes. Overall, the model exhibits robust performance across various metrics, showcasing its ability to handle multiclass classification tasks effectively.

The Support Vector Machine (SVM) exhibits a higher rate of misclassifying events compared to the K-Nearest Neighbors (KNN) algorithm. Specifically, Class 0 experiences a frequent misprediction, often being classified as Class 1 by the SVM. This discrepancy suggests a higher tendency for the SVM to incorrectly assign instances from Class 0 to Class 1, potentially indicating challenges in distinguishing between these two classes. The performance disparity between SVM and KNN highlights the importance of considering the specific characteristics of the dataset and the nature of the classes when selecting an appropriate classification algorithm.
The classification report reveals some variations in the model's performance across different classes compared to the previous report. For class 0.0, precision has slightly decreased to 0.81, indicating that the positive predictions for this class are now at a slightly lower accuracy. Recall for class 1.0 has improved to 0.69, signifying an enhancement in capturing more relevant instances, and the F1-score for this class has also increased to 0.52. Class 2.0 shows an improvement in precision (0.78), but a decrease in recall (0.55), resulting in a slightly lower F1-score of 0.64. Class 4.0 exhibits a notable increase in precision (0.70) and a slight decrease in recall (0.94), leading to a higher F1-score of 0.80. Overall, the macro-average precision and recall remain relatively consistent at 0.75 and 0.77, respectively, contributing to a macro-average F1-score of 0.75. The weighted average F1-score stands at 0.82, indicating an overall improvement in the model's ability to balance precision and recall across the dataset, with an accuracy of 82%.

The Decision Tree Classifier predicts all weather events well. It exhibits commendable performance in classifying weather events across multiple metrics. It achieves high precision, recall, and F1-score for most weather event classes, indicating its effectiveness in making accurate predictions. The classifier demonstrates particularly strong performance for classes 3.0, 4.0, and 5.0, achieving perfect precision and recall values. The overall accuracy of 95% underscores the classifier's ability to correctly classify the majority of instances. The Decision Tree's interpretability and simplicity make it a valuable model, especially since it's results can be visualised as a decision tree plot for better transparency and comprehensibility, but its performance may be surpassed by more sophisticated models in certain scenarios and decision trees are known to be prone to overfitting.

The Confusion Matrix for the Gradient Boosting Classifier shows very well performance as most labels are predicted correctly. The Classification Report demonstrates remarkable performance across multiple metrics. It achieves high precision, recall, and F1-score for most weather event classes. Notably, it maintains precision rates above 94% for all classes, indicating a low false positive rate. The recall values, reflecting the ability to correctly identify instances of each class, are consistently high, ranging from 92% to 100%. The overall accuracy of 98% further emphasizes the

classifier's proficiency in correctly classifying instances. Compared to the preceding models, the Gradient Boosting Classifier stands out with superior accuracy and balanced performance across various weather event classes. Since it combines mutliple decision trees, similar to a random forrest, it also profits from very good interpretability and simplicity, better transparency and comprehensibility, without the proneness for overfitting. Its ability to handle complex relationships within the data and make accurate predictions makes it a robust choice for this classification task.

The analysis of the classification reports provides valuable insights into the performance of different classifiers across multiple weather event labels. The Extra Trees, XGBoost, and Random Forest classifiers consistently demonstrate high precision, recall, and F1-score across various weather event categories, showcasing their effectiveness in accurately predicting events. The Support Vector Machine (SVM) tends to misclassify events more frequently, especially for class 0, often predicting it as class 1. On the other hand, the K-Nearest Neighbors (KNN) algorithm generally performs well, with actual results closely aligning with predictions. The Gradient Boosting Classifier outshines other models, exhibiting exceptional precision, recall, and F1-score for most classes and achieving an impressive overall accuracy of 98%. The Decision Tree Classifier also performs well, with high precision and recall values for most classes, resulting in an accuracy of 95%. Comparatively, the SVM faces challenges in achieving robust performance across all classes, while the KNN algorithm proves to be reliable. The Gradient Boosting and Decision Tree classifiers emerge as top performers, providing accurate predictions across a diverse range of weather event labels.

Generally, the results for the multiclass classification analysis are excellent, proofing that extreme weather events can be predicted with a very high accuracy using mutliclass classification techniques.

4 Discussion and Conclusion

In this exploratory data analysis project, titled "Weather Data Analysis: A Regression and Classification Approach on the ERA5 Dataset," an endeavor was made to understand and predict weather patterns in the Bancroft region of Ontario, Canada, utilizing the ERA5 dataset from 2015 to 2022. The project was motivated by the significant role that accurate weather prediction plays in various sectors and the unique climate characteristics of the region, influenced by the lake effect. The complex task of weather prediction was approached through regression and classification analyses, with the aim of modeling weather parameters and predicting weather events, including extreme conditions.

4.1 Regression Analysis Findings

The regression analyses in chapter 5.3.4. aimed to forecast the temperature with historical data. The research question "Is it possible to build an accurate regression model to predict temperature based on historical data?" can be answered with yes. The results indicated that while an exact daily weather forecast remains challenging, satisfactory levels of accuracy in predicting the temperature forecast for the general temperature trends over the year using linear regression models were achieved. Support Vector Regressor emerged as the most effective model in this context.

However, predicting the temperature using the wind speed variable in a linear regression analysis (5.3.1.) or a paramerer mix of different variables in a multiple predictor linear regression analysis (5.3.2.) did not proof to be successfull. The research question "Is it possible to find a correlation or causation between the temperature and windspeed, windgust or winddirection using regression techniques?" can therefore be answered with no. Using the given data, linear regression models

cannot be successfully, with a satisfactory accuracy, be used to predict the temperature with the wind variables or a mix of variables. The analyses proofed, that the cause of this is for one the mostly non-linear relationship of the parameters, which proofs to be a lot more complex. The correlation of the parameters also was not sufficiently high for a linear regression analysis. For these reasons it was decided that the further approach is to continue the analysis with logistic regression and classification techniques.

Also, the SARIMAX model used for temperature and wind modeling showed a systematic bias, consistently overestimating temperatures. This highlighted the limitations of using SARIMAX for predicting temperature variations based on wind, prompting the need for further refinement and exploration of alternative modeling approaches.

The last regression analysis was using logistic regression techniques to classify extreme weather and bluse sky day events in a binary classification. The research question was "Can logistic regression effectively classify and predict the occurrence of extreme or normal weather events based on temperature (or alternatively windspeed) ranges?". The approach to predict extreme weather events based solely on temperature using logistic regression's techniques (5.3.5) was insufficient, often misclassifying extreme events as normal. This underscored the necessity for more complex or multivariate approaches to accurately anticipate hazardous weather conditions. Instead of further optimizing the logistic regression approach (e.g., hyperparameter tuning or multiple predictor regression), it was decided that the more promising approach would be to identify further binary classifiers as presented in the classification analysis chapters.

4.2 Classification Analysis Findings

In binary classification the goal was to predict whether an observation was an extreme weather or blue sky day event (5.4.1.). The asked research question was "Is it possible to classify and predict extreme weather events such as storms?". It was identified that extreme weather events can indeed very accurately be separated from blue sky day events and both classes can be predicted with a very high accuracy, precision and recall. "ExtraTreesClassifier," "XGBClassifier," "RandomForestClassifier," and "LGBMClassifier" are the top-performing classifiers based on LazyClassifier's assessment. Each demonstrated high accuracy, with XGBoost slightly leading the pack. These models proved effective in categorizing and predicting weather events from the given data, providing valuable tools for future weather prediction endeavors. The results of this analysis could then be used in multiclass classification, to determine the specific type of extreme weather event.

The multiclass classification further nuanced the understanding of various weather events (5.4.2.). The goal was to determine and classify the specific type of extreme weather event, answering the research question "Is it possible to categorize and predict different extreme weather events based on multivariate weather data?". The research question can be answered with yes, the prediction and categorization of various extreme weather events is possible with a very high accuracy, precision and recall. Gradient boosting emerged as a particularly potent method, achieving high precision, recall, and F1-scores across all classes. This success illustrates the potential of sophisticated classification algorithms in deciphering complex weather patterns and predicting diverse weather events. This knowlegde can then also be used by scientists for further research for governmental institutions, e.g., when it comes to taking countermeasures to prevent damage from certain extreme weather events and minimize the risks and dangers.

4.3 Critical reflection and outlook

This project's journey through regression and classification analyses of weather data has provided a deeper insight into the atmospheric dynamics of Bancroft, Ontario. While profound analyses and research in temperature trend prediction and weather event classification was made, the challenges and inaccuracies encountered remind of the complexities inherent in meteorological studies. Predicting the weather accurately remains a daunting task, demanding continual refinement of models and methods.

Critically reflecting on the used methodology and procedure, it can be said, that after conducting the EDA, which showed, that wind and temperature variables do not correlate or show association, further regression analyses of these parameters could have been discontinued there. But since scientific literature discussed, that such an analysis proofs to be valuable and other researchers already made successfull aproaches with those techniques on this data, the approach was continued, to further evaluate patterns and characteristics of temperature and wind parameters and their correlations. Especially the mutliple predictor analysis promised to deliver interesting results and after all, the research question and hypotheses could succesfully be refuted, which in the end is a very valuable contribution to science.

Furthermore, the original ERA5 dataset was reduced using PCA and feature forward selection / feature backward deletion to conduct analysis on feature relevance. Only a reduced dataset of parameters was used in the final analyses. In further research other variables and their association / correlation could be analyzed. For example, literature suggests that cloud cover has a strong influence on air temperature. Using the same methodology and implemenation as in this assignment, simply switching the parameters, additional research with valuable insights into meteorological data can be conducted, resulting in better results.

Moreover, the generalizability of the data has to be scrutinized, as the analyses are based on regional data and meteorological data is always very biased towards regional effects. Conducting the same analysis on data for regions in e.g., African countries might result in completely different results.

The unpredictability and irregularity of weather and temperature patterns is a crucial aspect that adds complexity to the analysis. Acknowledging the irregular nature of meteorological phenomena emphasizes the inherent challenges in making precise predictions, even with advanced analytical techniques.

While the analyses presented valuable insights, the potential for further optimization, including hyperparameter tuning, remains. Exploring these avenues could enhance the performance of the models and provide more accurate predictions, especially in the context of fine-tuning parameters for machine learning algorithms.

The exploration of weather patterns could be expanded to include a dedicated analysis of trends related to climate change. Understanding the long-term impacts on temperature and weather events could provide valuable insights into the broader environmental context.

Acknowledging that certain factors, such as climate change and omitted variable bias, might not have been explicitly addressed in the analysis, highlights the potential sources of variance and errors. Future research could delve deeper into these factors to refine models and improve predictive accuracy. It's essential to recognize that certain aspects, possibly including external factors or variables not considered in the analysis, might influence the weather and temperature trends. Acknowledging what falls out of the scope of the current study adds a layer of humility to the

findings and encourages future researchers to explore additional dimensions.

In summary, while the project has contributed significantly to understanding and predicting weather patterns, there exist additional dimensions and challenges that warrant exploration. The complexities of meteorological studies, coupled with the acknowledgment of unpredictable weather dynamics, call for continual refinement, optimization, and consideration of broader environmental factors for a comprehensive understanding of weather phenomena. Our findings contribute to the broader discourse on weather prediction, emphasizing the need for multidimensional approaches and the potential of machine learning techniques. As climate variability continues to present profound challenges, the insights garnered here offer a stepping stone towards more accurate, reliable, and comprehensive weather forecasting methods. Moving forward, integrating more diverse datasets, refining models, and exploring new methodologies will be crucial in enhancing the predictive capabilities and understanding of weather patterns. This endeavor not only aids in better forecasting but also in strategic planning and preparedness for the diverse impacts of weather and climate change across sectors.

5 Apendix

5.1 Data Dictionary

[15]:

	Name	Description	Role	Type
0	run_datetime	Date and time when the weather observations we…	ID / predictor	numerical continuous / ID
1	wep	Weather Event Type (WEP) is a categorization o…	response	categorical nominal
2	avg_temp	The average temperature measured at two meters…	response / predictor	numerical continuous
3	min_wet_bulb_temp	Minimum wet bulb temperature recorded during t…	predictor	numerical continuous
4	avg_dewpoint	Average dewpoint temperature observed during t…	predictor	numerical continuous
5	avg_temp_change	Average change in temperature during the obser…	predictor	numerical continuous
6	avg_windspd	Average wind speed measured during the recordi…	predictor	numerical continuous
7	max_windgust	Maximum wind gust observed during the recordin…	predictor	numerical continuous
8	avg_winddir	Average wind direction (in degree) observed du…	predictor	numerical continuous
9	wind_direction_label	Wind direction (in cardinal direction) observe…	predictor	categorical ordinal
10	max_cumulative_precip	Maximum cumulative precipitation recorded, con…		
11	max_snow_density_6	Maximum snow density at a depth of 6 inches, c…		
12	max_cumulative_snow	Maximum cumulative snow recorded, considering …		
13	max_cumulative_ice	Maximum cumulative ice recorded, considering a…		
14	avg_pressure_change	Average change in atmospheric pressure during …		

```
10            predictor       numerical continuous
11            predictor       numerical continuous
12            predictor       numerical continuous
13            predictor       numerical continuous
14            predictor       numerical continuous

                                                 Format
0   <class 'pandas._libs.tslibs.timestamps.Timesta…
1                                        <class 'str'>
2                              <class 'numpy.float64'>
3                              <class 'numpy.float64'>
4                              <class 'numpy.float64'>
5                              <class 'numpy.float64'>
6                              <class 'numpy.float64'>
7                              <class 'numpy.float64'>
8                              <class 'numpy.float64'>
9                                        <class 'str'>
10                             <class 'numpy.float64'>
11                             <class 'numpy.float64'>
12                             <class 'numpy.float64'>
13                             <class 'numpy.float64'>
14                             <class 'numpy.float64'>
```

5.2 Sources