final_report

January 10, 2024

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

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

1 Introduction and data

1.1 Motivation

Weather, an age-old Earth phenomenon, captivates human interest due to its intricate blend of temperature, wind, and precipitation, molding our surroundings and challenging our understanding of the natural world [^1]. Accurate weather prediction is crucial for agriculture, disaster management, and urban planning, particularly in the context of climate change risks [^2]. The project, titled "Weather Data Analysis: A Regression and Classification Approach on the ERA5 Dataset" aims to contribute to this exploration by examining how different variables interact to create complex weather phenomena.

1.2 Data

Data description of sample
The study leverages the ERA5 dataset, sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF), is comprised of atmospheric reanalysis data spanning multiple decades (2015-2022) at hourly intervals and characterized by a spatial resolution of approximately 31 km [^3]. The data is collected through reanalysis, assimilating observational data from satellites, weather stations, and other sources into a numerical weather prediction model. Focusing on the region of Bancroft in Ontario, Canada, the project explores the unique climatic and meteorological characteristics of the area, influenced by the 'lake-effect' phenomenon [^4]. Various meteorological parameters as described below are included in the dataset. The data, labeled by meteorologists and data scientists from IBM and The Weather Company, offers comprehensive global-scale atmospheric information, making it well-suited for detailed analyses and modeling, including climate research, environmental monitoring, and weather forecasting [^5], [^6].

Variables
The dataset, encompasses key variables such as air temperature, wind speed and direction, precipitation, atmospheric pressure, snow density, cumulative snow, cumulative ice, and weather events. The dataset also includes categorical weather events such as 'Blue Sky Day', 'Mild Snowfall', and 'Storm with Freezing Rain'. These variables form the foundation for the assignment's comprehensive analysis [^7].

Overview of data
Initially, the .csv file is loaded, and the data's head is printed for an initial overview of columns (variables) and rows (observations), as can be seen in appendix 5.2 "Display of the Used Dataframe". The dataset comprises 65,345 observations and 186 columns, including unique predictor variables and a response variable. A new dataframe is formed by selecting specific columns and transforming columns (based on analysis for feature relevance and literature) to achieve optimized resource usage. This dataframe is later split into training, testing, and validation sets, underlining the foundational role of proper data splitting for reliable machine learning model development and generalization to new data [^8], [^9].

1.3 Research Questions

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

Regression Hypothesis: 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. The hypothesis will be examined through the following questions: Is it possible to build an accurate regression model to predict temperature based on historical data? Is it possible to find a correlation or causation between the temperature and the wind features using regression techniques? How does the incorporation of multiple atmospheric predictors enhance the accuracy of temperature prediction compared to a model solely based on windspeed? Can logistic regression effectively classify and predict the occurrence of extreme or normal weather events based on temperature ranges?

Classification Hypthesis: 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. The following questions will help to evaluate this hypothesis: Is it possible to binary classify and predict extreme weather events such as storms? Is it possible to categorize and predict different extreme weather events based on multivariate weather data?

1.4 Exploratory Data Analysis (EDA)

The dataset includes features such as the substation (Bancroft), timestamps, weather-related parameters, and various labels for the corresponding weather events. As revealed in the appendix 5.3 "Data Dictionary" most variables are of the "float64" data type (167), 8 variables are of type "int64", 9 are of type "object", and 2 are "datetime64". First, the variable "avg_temp" is examined. This includes depicting the temperature trend over time (seen in 5.4 "Time series"), as well as displaying the box plot and histogram as shown in appendix 5.8 "Distribution of Weather Features by Weather Event Profiles in Distograms" and 5.9 "Distribution of Weather Features by Weather Event Profiles in Boxplots".

2 Methodology

The first phase of methodology is focused on the comprehensive preparation and processing of the ERA5 dataset to ensure a solid basis for the subsequent analysis. This phase aims to ensure data quality and maximize the accuracy of the models.

Data acquisition
After import and inspection, the date and time information in the dataset is converted into a standardized date format. Some data correction is performed. As a result of this phase, the dataframe with the variables as seen in 5.2 "Display of the Used Dataframe" is created and used in the further analysis.

2.1 Analysis and Visualisation

Weather event analysis shows 'blue sky day' as the most common, followed by mild and moderate snowfall, then moderate rainfall. Extreme events like storms with freezing rain and heavy snow, as well as high precipitation snowstorms, are significantly less frequent.

The weather data time series analysis for Bancroft shows (appendix 5.4 "Time series"): - Temperature: Distinct annual cycle with high summer and low winter temperatures, and moving averages smoothing daily fluctuations. - Wind: High daily variability in speed and direction, with no clear seasonal patterns, indicating complex local weather dynamics. - Snowfall: Clear seasonal trend, inversely related to temperature, with increasing density over time. - Pressure: Mirrors temperature changes, suggesting a strong seasonal relationship.

Histogram analysis reveals (appendix 5.8 "Distribution of Weather Features by Weather Event Profiles in Distograms"): - Temperature: Bimodal distribution with more frequent cooler periods. - Wind: Predominantly moderate conditions with occasional strong gusts. - Precipitation: Mostly minor events, with rare heavy rainfall. - Snow and Ice: Significant accumulations are rare. - Pressure: Consistent, with little fluctuation.

Boxplot analysis of weather parameters confirms previous findings (appendix 5.9 "Distribution of Weather Features by Weather Event Profiles in Boxplots"): clear seasonal temperature fluctuations, mostly low wind speeds with occasional peaks, generally low precipitation with rare high outliers, infrequent snow and ice accumulations, and stable atmospheric pressure.

Distogram and box plot analyses of weather events in Bancroft yield insights into weather parameter influences: - Temperature: 'Blue sky days' correlate with higher temperatures, while colder temperatures are associated with snowfall events. All events can occur at median temperatures. - Temperature Changes: Symmetric distribution around zero suggests temperature changes are not reliable predictors for weather events. - Wind: Wind speed and gusts poorly differentiate events, but certain patterns are noted in average wind direction. Wind alone is not a strong predictor. - Precipitation, Snow, Ice: Skewed distribution towards light events with rare intense occurrences. Related variables like snow accumulation help distinguish events like snowfall. - Atmospheric Pressure: Relatively normal distribution, indicating a stable environment.

Box plot findings include: - Clear Skies: High average temperatures, low precipitation. - Continuous Freezing Rain: Narrow temperature ranges, variable pressure changes. - Snowfall Variations: Low temperatures with varying snow and ice accumulations. - Moderate Rain and Snow: Variable temperatures and precipitation amounts. -Severe Storms: Extreme precipitation, significant temperature and pressure fluctuations, variable wind conditions.

Pie-chart findings (appendix 5.6 "Distribution of All Weather Events"): - Overall Weather Event Distribution: 'Blue Sky Day' dominates with 77.9% of observations, indicating prevalent clear weather. Moderate rain, light and moderate snowfall are less common but notable. - Extreme Weather Event Distribution: Among non-clear sky observations, moderate snowfall is most frequent (35.4%). Moderate snowfall (23.0%) and moderate rain (20.7%) follow, together accounting for

79.1% of these events. The remaining 20.9% comprises rarer, more extreme events like heavy snowfall, continuous freezing rain, and storms with freezing rain or heavy snow, listed in order of frequency.

Scatterplot analysis and correlation coefficients revealed (appendix 5.7 "Association Plots and Correlation Analysis"): - Similar-Scale Correlations: Stronger correlations were found between variables with similar scales and units, like average temperature, dew point, temperature change, - minimum wet bulb temperature, and between average wind speed and maximum wind gusts. - Correlation Variability: Correlations between parameters varied significantly during extreme weather events, indicating their substantial impact on these relationships.

In Bancroft's climate study (appendix 5.10 "Analysis of Wind Speeds and Average Temperatures by Wind Direction"): - Temperature and Wind in Regression: Temperature and wind parameters were included in regression analyses despite no direct linear correlation, acknowledging possible non-linear influences. - Key Findings: Wind speed varies by direction, with higher speeds in north/west winds and lower in southeast/east. Southwesterly winds bring warmer temperatures, while northerly winds are cooler.

The Principal Component Analysis (PCA) of Bancroft's weather data revealed (5.11 "3D plot of all weather observations using PCA"): - Data Point Distribution: A cone-shaped structure along the PC1 and PC2 axes suggests these components capture most weather variability. - Event Characterization: Specific events like "Storm with freezing rain / Heavy snow and ice storm" are outliers, extending prominently on PC1 and PC2, while "Blue Sky Day" shows wide variability, clustering below 0 on PC1 and between -5 and 10 on PC2. - Clustering: A dense cluster near the origin indicates that frequent weather conditions share similar features like average temperature and precipitation, contrasting with the extended tail and outliers that represent rare and extreme events. - Interpretation Challenges: Difficulty in separating clusters, especially within PC3, due to subtle differences, and a significant overlap in PC1, highlighting the complexity of differentiating weather patterns.

2.2 Model

2.2.1 Regression Analysis Temperature and Wind

A linear regression, gradient boosting, SGD regressor, and support vector regressor were trained to predict average wind speed using variables from the EDA. The training-to-test data ratio was 80:20, with model evaluation based on MSE and MAE.

Key Findings: - Weak Relationship Between Wind Speed and Temperature: Scattered data points and high MSE/MAE values across models suggest a limited predictive accuracy of temperature from wind speed alone. - Residual Analysis: Wide spread of residuals, especially in the Support Vector Regression model, indicates substantial prediction errors and a tendency to underpredict temperatures. - Model Performance and Bias: Most models show no significant bias, but the complexity of temperature patterns requires multiple regression with additional variables. - Linear Relationship in Specific Cases: Simple regression between closely linked variables like wind speed and gusts shows lower MSE and MAE, but may offer limited scientific value due to inherent correlations. The analysis emphasizes the need for multiple factors in modeling complex systems like weather patterns and highlights the importance of interpreting model results carefully, considering the environmental data's complexity.

2.2.2 Regression Analysis Multiple Linear Regression

The objective is to predict temperature using multiple predictors, with a focus on feature selection to ensure insights and minimal correlation between variables. Key findings from the correlation analysis include:

High correlation among temperature-related variables like 'avg_temp', 'avg_temp_celsius', 'min_wet_bulb_temp', and 'avg_dewpoint'. Moderate correlation between 'avg_windspd' and 'max_windgust', but low correlation with temperature variables. Low correlation for wind direction variables like 'avg_winddir'. Moderate to high correlation among precipitation variables such as 'max_cumulative_precip', 'max_snow_density_6', and 'max_cumulative_snow'. Negative correlation between 'avg_pressure_change', 'avg_temp', and 'label1'. Redundant variables were removed, and a feature forward selection identified nine key variables, including 'max_snow_density_6', 'avg_temp', and 'avg_windspd', enhancing model accuracy. Backward feature elimination retained all selected features, achieving an accuracy of 99.969%. Training with the optimized dataset included a linear regression model using 'avg_windspd' and 'avg_winddir' to predict average temperature. Another model utilized all variables for performance comparison. The target variable was then changed to predict average wind direction using all available variables, with results discussed in the Results chapter, using metrics like MSE, MAE, and R-squared.

2.2.3 Regression Analysis Temperature Forecast

A SARIMAX model visualizes avg_temp, avg_winddir, avg_windspd, and avg_windgust, including time series, trend, seasonal, and residual components. The Augmented Dickey-Fuller test ensures these time series are stationary, critical for SARIMAX model accuracy. Model selection relies on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to balance fit and complexity, with lower values indicating better models. AutoARIMA optimizes ARIMA model parameters (p, d, q) for this dataset, focusing on daily seasonality (m=24) and model simplicity. It aims to model data patterns effectively without overfitting. The diagnostic phase assesses the SARIMAX model's fit, confirming its ability to capture data patterns and validating assumptions like the normal distribution of residuals and no autocorrelation for robustness and reliability. Lastly, the model is refitted with revised parameters and evaluated on the test dataset using AIC, BIC, MSE, and SSE values. The optimal model for this project is identified using Lazypredict, which compares multiple models' performance. The XGBRegressor is chosen for predicting average temperature from wind speed and direction. TimeSeriesSplit is used for cross-validation to assess the model's effectiveness on unseen data, with hyperparameters fine-tuned for a balance between complexity and accuracy. The model's performance is evaluated using MSE, MAE, and visualizations of actual versus predicted values. Next, a regression model is developed to predict temperature based on historical data, including daytime, day, and season. Multiple models - linear regression, gradient boosting, SGD regressor, and SVR regressor - are considered, with the train-test split based on the year of data. These models are evaluated by their residuals and visualized for actual versus predicted values, with detailed discussions in the 'results' chapter.

2.2.4 Logistic Regression

To predict extreme weather events with the temperature variable, the project uses logistic regression. First it is required to normalize the data and choose a constant, which which is done by the statsmodels in this case. After the fit of the data into the model we can evaluate it using the aic and the confusion matrix. A final visualization is helpful to understand the results und to find

techniques to improve the performance.

2.2.5 Binary Classification

The goal is to be able to predict extreme weather events by any variables. For that we first customize our datframe by drop columns, which are not numeric and not needed. After that ´label1´ is choosen as the dependend variable and lazypredict library is used to find the best model for this case. 'ExtraTrees', 'XGBoost', 'LGBM', 'RandomForest' were choosen as the best models for this case, which is why all of them are implemented. The individual confusion matrix are used to evaluate the model performance in combination with the precision, recall and f1-score metrics.

2.2.6 Multiclass Classification

Now it is important to classify which extreme weather event it is in particular. Other classifier need to be trained, for this usecase knn, svm, dtc and gbc were chosen. And again the individual confusion matrix is used to evaluate the model performance in combination with the, precision, recall and f1-score metrics.

3 Results

3.1 Regression Results

Is there a significant correlation between temperature and wind characteristics, which can be modeled to predict future temperature trends and variations? This question was addressed within the scope of this project. Various regression techniques were employed, and different sub-questions were examined.

Temperature and Wind Modeling
In the first step, the relationship between wind speed and temperature is investigated. In this context, models such as the Linear Regression Model (LRM), Gradient Boosting Model, Stochastic Gradient Descent Model, and Support Vector Regression Model are utilized to depict the correlation (appendix 5.12 "Linear Regression Analysis Temperature and Wind Modeling Results"). These models are predicting temperature from wind speed using various regression techniques and are compared with each other. Results show a weak correlation, high MSE, and MAE across all models, indicating poor prediction. Outliers and dispersed residuals suggest significant deviations. Support Vector Regression tends to underpredict. Findings suggest the need for multiple regression with additional variables. A subsequent linear regression analysis on wind gusts reinforces the idea that correlated variables may yield successful models but lack scientific value. Multiple regressor analysis is proposed to enhance temperature prediction due to the limited effectiveness of wind speed alone.

Linear Regression Analysis with Multiple Predictors
In the initial phase of the Temperature and Wind Modeling over Time analysis, a Multiple Linear Regression is introduced, as introduced in the lecture. Based on that, the temperature variable is now predicted with improved accuracy using linear regression with multiple predictor variables, addressing the research question of how the incorporation of various atmospheric predictors enhances temperature prediction over different time scales, uncovering interactions and synergies among predictors, and analyzing temporal dynamics to refine the predictive model. The temperature is predicted on windspeed and wind direction in the first step. In the next step, the temperature is

predicted using the before-utilized variables (appendix 5.13 “Linear Regression Analysis Multiple Predictors Correlation Matrix of Variables”). For this analysis seasonality and trend for the temperature are also analysed (appendix 5.14 “Linear Regression Analysis Multiple Predictors Seasonality and Trend”). After implementing the Multiple Linear Regression (MLR) model, there can be a lack of accuracy in predicting average temperature from wind speed and direction, as well as from the remaining variables. The overall conclusion underscores the need for further refinement, potentially involving additional features or non-linear models, to enhance predictive accuracy, especially in accurately predicting extreme temperatures.

SARIMAX MODEL

After successfully predicting the temperature parameter through multiple predictor linear regression, the focus shifts to forecasting the temperature parameter with a statistical SARIMAX approach (appendix 5.15 “Linear Regression Analysis Multiple Predictors SARIMAX Forecast Results”). SARIMAX models are among the most widely used statistical models for forecasting, with excellent forecasting performance [^16]. To keep the model’s complexity low and avoid lengthy computation times later on, only wind variables are used for an initial approach here. The analysis of Trend and Seasonality revealed a slight variability with some periods showing a gentle rise or fall and a consistent and expected cyclical pattern corresponding to the seasons. The augmented Dickey-Fuller Test (ADF) [^17], Akaike Information Criterion (AIC) [^18], and Bayesian Information Criterion (BIC) are performed on the data. The ADF Test indicated stationarity, the AIC and BIC showed that windspeed and winddirection are the most suitable predictors. After that, the actual SARIMAX Model is created. The evaluation reveals the model’s limitations in capturing short-term fluctuations, particularly missing sharp peaks, and consistently overestimating temperatures, indicating a systematic bias and the need for further refinement or alternative modeling approaches to enhance accuracy.

XGBoost

After implementing the SARIMAX as a popular approach for time series analysis, the Lazy Regressor library from sklearn was utilized to find the best-performing regressor. The Lazy Regressor showed that all Regression Models have a rather low R-Squared Value. The XGBoost Regressor is determined as the best-performing Model with an R-Squared Value of 0.13. Based on that, the XGBoost Model is used. The evaluation of the model shows a moderate level of predictive accuracy, with the model following the general temperature trend but exhibiting discrepancies in magnitude and timing, supported by reported Mean Squared Error (MSE) and Mean Absolute Error (MAE) values, suggesting potential for improvement through model tuning and additional feature exploration.

Temporal Prediction

In the next step, the relationship between temperature and time is explored. A Linear Regression Model, Gradient Boosting Regressor, an SGD Regressor, and a Support Vector Regressor are used here. The Evaluation of the plots presents that the Gradient Boosting Regressor demonstrates a promising ability to closely track temperature changes with fewer deviations and a tighter distribution of residuals, supporting the conclusion that linear regression models, while not perfect, can provide valuable forecasts for temperature trends in Bancroft, Canada. The results can be seen in appendix 5.16 “Linear Regression Analysis Prediction Forecast Results”.

Temporal Logistic Regression

Logistic regression, placed between linear regression and classification chapters, serves as a bridge to better understand the data story, where blue dots represent actual labels, red dots indicate predicted probabilities, and the orange curve reflects the probability of extreme weather events

based on temperature alone (appendix 5.17 "Logistic Regression Analysis Predicting WEP by Temperature Results"). The graph reveals significant overlap in temperature ranges for different event types, leading to high false positives and low recall. Consequently, logistic regression with temperature as the sole predictor is deemed insufficient for this classification task, suggesting the potential need for additional predictors, hyperparameter tuning, or alternative modeling approaches for improved performance.

Conclusion

In conclusion, the investigation into the correlation between temperature and wind characteristics, with the aim of modeling future temperature trends and variations, has yielded valuable insights within the scope of this project. Employing various regression techniques, the exploration delved into different sub-questions surrounding this overarching hypothesis. The results indicate that while initial models, particularly those based solely on wind parameters, exhibited limitations in predictive accuracy, the incorporation of multiple predictors through advanced regression analyses showcased a promising avenue for refinement. The comprehensive evaluation underscores the complexity of the relationship between temperature and wind characteristics, emphasizing the need for nuanced modeling approaches and consideration of additional factors to enhance the precision of temperature predictions over diverse temporal scales. Overall, this study provides a foundation for future research endeavors seeking to unravel the intricate dynamics between meteorological variables and advance our understanding of climate forecasting.

3.2 Classification Results

3.2.1 Binary Classification of Extreme Weather Events

The visualization of the results of the binary classification can be found in appendix 5.18 "Methodology and Results Binary Classification" and displays four confusion matrices, each representing the performance of a different binary classification model: ExtraTrees, XGBoost, LightGBM, 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.

The classification reports found in appendix 5.18 provide an evaluation of the performance of different models. 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. 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. 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. 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.

3.2.2 Multiclass Classification of Various Extreme Weather Events

After successfully predicting extreme weather and blue sky day weather events, a key result of this research is 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.

The multiclass classification is conducted using the models K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree Classifier (DTC) and Gradient Boosting Classifier (GBC). These models have fundamentally different functionality so that the different model types can be compared with each other and strengths and weaknesses in the application to weather data can be assessed for each model type. The detailed results and visualisations for each model can be found in the appendix 5.19 "Methodology and Results Multiclass Classification".

KNN's multiclass classification performs well, aligning actual outcomes closely with predictions. The classification report highlights high precision and consistent recall, both with values between 78%-100% through all labels. The F1-score is strong for most classes, with a macro average of 0.88 and a weighted average of 0.92, demonstrating effectiveness despite class imbalance. The model's 92% accuracy underscores its reliability across diverse classes, showcasing robust performance in multiclass classification tasks.

The SVM displays a higher misclassification rate than KNN, particularly misclassifying Class 0 as Class 1. This discrepancy suggests challenges in distinguishing between these classes. The performance gap underscores the need to consider dataset characteristics when selecting a classification algorithm. The classification report indicates some performance variations. Precision for Class 0.0 decreases to 0.81, while recall for Class 1.0 improves to 0.69, leading to an increased F1-score of 0.52. Class 2.0 shows improved precision (0.78) but decreased recall (0.55), resulting in a slightly lower F1-score of 0.64. Class 4.0 sees increased precision (0.70) and a slight recall decrease (0.94), yielding a higher F1-score of 0.80. Macro-average precision and recall remain consistent at 0.75 and 0.77, contributing to a macro-average F1-score of 0.75. The weighted average F1-score is 0.82, indicating an overall improvement in balancing precision and recall with 82% accuracy.

The DTC excels in predicting various weather events, showing impressive performance across multiple metrics with high precision, recall, and F1-score. Particularly noteworthy is its perfect precision and recall for classes 3.0, 4.0, and 5.0. The overall accuracy of 95% highlights its effectiveness in

classifying most instances. The Decision Tree's interpretability and simplicity, visualized through a decision tree plot, enhance transparency. However, in some scenarios, more advanced models may outperform it, and decision trees can be susceptible to overfitting.

The GBC Confusion Matrix highlights excellent performance with accurate predictions for most labels. The Classification Report demonstrates impressive precision, recall, and F1-score across diverse weather event classes, maintaining precision rates above 94%. Recall values consistently range from 92% to 100%, showcasing the classifier's ability to identify instances accurately. The 98% overall accuracy underscores its proficiency in classification. Compared to prior models, the Gradient Boosting Classifier excels in accuracy and balanced performance. Its use of multiple decision trees, akin to a random forest, enhances interpretability and simplicity while avoiding overfitting. Its capacity to handle complex relationships within the data 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 SVM tends to misclassify events more frequently. The GBC and DTC 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

4.1 Regression Analysis Findings

The regression analyses aimed to predict temperature using historical data, achieving satisfactory accuracy in general temperature trend forecasts for the year with linear regression models. The Support Vector Regressor emerged as the most effective model. However, attempts to predict temperature with wind speed in linear regression or a mix of variables in multiple predictor linear regression were unsuccessful. The non-linear relationship and insufficient correlation between temperature and wind variables led to the decision to explore logistic regression and classification techniques. The SARIMAX model used for temperature and wind modeling exhibited a consistent bias, overestimating temperatures, highlighting limitations and prompting the need for alternative modeling approaches. The final regression analysis employed logistic regression to classify extreme weather and clear sky events. However, the approach based solely on temperature was insufficient, emphasizing the need for more complex or multivariate methods to accurately predict hazardous weather conditions. Instead of optimizing logistic regression further, the focus shifted to identifying additional binary classifiers in subsequent classification analyses.

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. The 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. 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 delved into regression and classification analyses of weather data in Bancroft, Ontario, offering insights into atmospheric dynamics. Despite challenges and complexities in meteorological studies, the pursuit of accurate weather prediction demands ongoing model refinement. The absence of linear correlation between wind and temperature variables, as revealed in the EDA, could have led to discontinuation, but the value found in literature influenced the decision to persist. The approach, including PCA and feature selection, provided interesting results, adding value to the scientific discourse. However, the regional bias in the data and the irregular nature of meteorological phenomena emphasize the challenges in making precise predictions. While the analyses presented valuable insights, further optimization, including hyperparameter tuning, remains a potential avenue. Exploring weather patterns and their relationship to climate change could expand understanding, acknowledging potential sources of variance and errors. Recognizing the limitations and external factors influencing weather trends adds humility to the findings, urging future researchers to explore additional dimensions. Despite the contributions to weather prediction, the complexities of meteorological studies and unpredictable weather dynamics necessitate continual refinement and consideration of broader environmental factors. In summary, this project contributes to weather prediction discourse, highlighting the need for multidimensional approaches and the potential of machine learning techniques. As climate variability poses challenges, these insights pave the way for more accurate and comprehensive forecasting methods. Integrating diverse datasets, refining models, and exploring new methodologies are crucial for better forecasting, strategic planning, and preparedness across sectors in the face of weather and climate change impacts.

5 Apendix

5.1 Simple Exploratory Data Analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65345 entries, 0 to 65344
Columns: 186 entries, Unnamed: 0 to wind_direction_label
dtypes: datetime64[ns](2), float64(167), int64(8), object(9)
memory usage: 92.7+ MB
```

[3]:

```
                    count                          mean                  min  \
Unnamed: 0        65345.0                   32685.658321                  0.0
run_datetime        65345  2019-04-06 14:09:11.362766848  2015-07-15 00:00:00
valid_datetime      65345  2019-04-06 14:09:11.362766848  2015-07-15 00:00:00
horizon           65345.0                           0.0                  0.0
avg_temp          65345.0                     279.574328           243.849393
…                   …                             …                    …
label2            12712.0                       3.06191                  0.0
label3            65345.0                        1.1811                  0.0
year              65345.0                    2018.745535               2015.0
month             65345.0                       6.711852                  1.0
avg_temp_celsius  65345.0                       6.424328           -29.300607

                                  25%                  50%  \
Unnamed: 0                    16343.0              32689.0
run_datetime      2017-05-25 23:00:00  2019-04-07 01:00:00
valid_datetime    2017-05-25 23:00:00  2019-04-07 01:00:00
horizon                           0.0                  0.0
avg_temp                   271.114219           279.882735
…                                …                    …
label2                            1.0                  3.0
label3                            1.0                  1.0
year                           2017.0               2019.0
month                             4.0                  7.0
avg_temp_celsius             -2.035781             6.732735

                                  75%                  max           std
Unnamed: 0                    49025.0              65361.0  18867.701277
run_datetime      2021-02-14 16:00:00  2022-12-27 08:00:00           NaN
valid_datetime    2021-02-14 16:00:00  2022-12-27 08:00:00           NaN
horizon                           0.0                  0.0           0.0
avg_temp                   289.903226           300.934144     11.383325
…                                …                    …            …
label2                            5.0                  6.0      2.126446
label3                            2.0                  3.0      0.740687
year                           2021.0               2022.0      2.162032
month                            10.0                 12.0      3.446477
avg_temp_celsius             16.753226            27.784144     11.383325

[177 rows x 8 columns]
```

5.2 Display of the Used Dataframe

[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]
```

5.3 Data Dictionary

[5]:

```
                     Name                                        Description  \
0            run_datetime  Date and time when the weather observations we…
1                     wep  Weather Event Type (WEP) is a categorization o…
2                avg_temp  The average temperature measured at two meters…
3       min_wet_bulb_temp  Minimum wet bulb temperature recorded during t…
4            avg_dewpoint  Average dewpoint temperature observed during t…
5         avg_temp_change  Average change in temperature during the obser…
6             avg_windspd  Average wind speed measured during the recordi…
7            max_windgust  Maximum wind gust observed during the recordin…
8             avg_winddir  Average wind direction (in degree) observed du…
9    wind_direction_label  Wind direction (in cardinal direction) observe…
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 …

                    Role                       Type  \
0         ID / predictor  numerical continuous / ID
1               response        categorical nominal
2   response / predictor       numerical continuous
3              predictor       numerical continuous
4              predictor       numerical continuous
5              predictor       numerical continuous
6              predictor       numerical continuous
7              predictor       numerical continuous
8              predictor       numerical continuous
9              predictor        categorical ordinal
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'>
```

5.4 Time series

```
<Figure size 2000x1500 with 0 Axes>
```

5.5 Class Distribution of Blue Sky and Extreme Weather Events

```
<Figure size 2000x1500 with 0 Axes>
```

5.6 Distribution of All Weather Events

```
<Figure size 2000x1500 with 0 Axes>
```

5.7 Association Plots and Correlation Analysis

```
<Figure size 2000x1500 with 0 Axes>
```

5.8 Distribution of Weather Features by Weather Event Profiles in Distograms

5.9 Distribution of Weather Features by Weather Event Profiles in Boxplots

```
<Figure size 2000x1500 with 0 Axes>
```

5.10 Analysis of Wind Speeds and Average Temperatures by Wind Direction

```
<Figure size 2000x1500 with 0 Axes>
```

5.11 3D plot of all weather observations using PCA

5.12 Linear Regression Analysis Temperature and Wind Modeling Results

```
X_train shape: (52276, 1)
X_test shape: (13069, 1)
y_train shape: (52276,)
y_test shape: (13069,)
Linear Regression Model:
Mean Squared Error: 127.78
Mean Absolute Error: 9.63

Gradient Boosting Model:
Mean Squared Error: 127.72
Mean Absolute Error: 9.63

Stochastic Gradient Descent Model:
Mean Squared Error: 127.78
Mean Absolute Error: 9.63
```

```
Support Vector Regression Model:
Mean Squared Error: 129.85
Mean Absolute Error: 9.57
```

```
<Figure size 2000x1500 with 0 Axes>
```

5.13 Linear Regression Analysis Multiple Predictors Correlation Matrix of Variables

```
<Figure size 2000x1500 with 0 Axes>
```

5.14 Linear Regression Analysis Multiple Predictors Seasonality and Trend

```
<Figure size 2000x1500 with 0 Axes>
```

5.15 Linear Regression Analysis Multiple Predictors SARIMAX Forecast Results

```
RUNNING THE L-BFGS-B CODE

           * * *

Machine precision = 2.220D-16
 N =            9     M =           10

At X0         0 variables are exactly at the bounds

At iterate    0    f= -2.33934D+00    |proj g|=  2.27869D+01

 This problem is unconstrained.
```

```
At iterate    5    f= -2.35851D+00    |proj g|=  5.07954D+00

At iterate   10    f= -2.42141D+00    |proj g|=  1.55207D-01

At iterate   15    f= -2.42143D+00    |proj g|=  2.57911D-01

At iterate   20    f= -2.42185D+00    |proj g|=  8.96676D-01

At iterate   25    f= -2.42207D+00    |proj g|=  1.29386D-02

At iterate   30    f= -2.42237D+00    |proj g|=  1.15344D+00

At iterate   35    f= -2.42282D+00    |proj g|=  1.45321D-01

At iterate   40    f= -2.42283D+00    |proj g|=  2.66534D-01

At iterate   45    f= -2.42324D+00    |proj g|=  7.45443D-01

At iterate   50    f= -2.42344D+00    |proj g|=  9.01914D-02

           * * *

Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
    9     50     59      1     0     0   9.019D-02  -2.423D+00
  F =  -2.4234367498076992

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT
```

```
<Figure size 2000x1500 with 0 Axes>
```

5.16 Linear Regression Analysis Prediction Forecast Results

```
<Figure size 2000x1500 with 0 Axes>
```

5.17 Logistic Regression Analysis Predicting WEP by Temperature Results

```
Optimization terminated successfully.
         Current function value: 0.375341
         Iterations 7
AIC: 39246.63967824996
Label 0: Extreme Weather Event
 Label 1: Blue Sky Day
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.39      0.20      0.26      2515
           1       0.83      0.93      0.88     10554

    accuracy                           0.79     13069
   macro avg       0.61      0.56      0.57     13069
weighted avg       0.75      0.79      0.76     13069

<Figure size 800x550 with 0 Axes>
```

```
<Figure size 800x550 with 0 Axes>
```

5.18 Methodology and Results Binary Classification

```
ExtraTrees Accuracy: 0.993496059377152
XGBoost Accuracy: 0.9946438136047134
LGBM Accuracy: 0.9935725763256561
RandomForest Accuracy: 0.9933430254801439
```

ExtraTrees Confusion Matrix
2448
67
25
10529
True Label
Predicted Label
0
1
10000
8000
6000
4000
2000

XGBoost Confusion Matrix
2458
57
13
10541
True Label
Predicted Label
0
1
10000
8000
6000
4000
2000

LGBM Confusion Matrix
2446
69
13
10541
True Label
Predicted Label
0
1
10000
8000
6000
4000
2000

RandomForest Confusion Matrix
2441
74
15
10539
True Label
Predicted Label
0
1
10000
8000
6000
4000
2000

```
<Figure size 800x550 with 0 Axes>
```

5.19 Methodology and Results Multiclass Classification

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

5.20 Sources

[1]: Liljequist, G.H. / Cehak, K. (1984): Allgemeine Meteorologie. 3. Auflage, Springer-Verlag. Engineering 29.2 (2022, Springer): 1247–1275
[2]: The contribution of weather forecast information to agriculture, water, and energy sectors in East and West Africa
[3]: ECMWF (2023a): ERA5: data documentation. URL: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation
[4]: A Hybrid Dataset of Historical Cool-Season Lake Effects From the Eastern Great Lakes of North America
[5]: Hjelmfelt, M.R. (1990): Numerical study of the influence of environmental conditions on lake-effect snowstorms over Lake Michigan, in: Monthly Weather Review, 118(1), pp.138-150.
[6]: de Lima, Glauston, R.T. / Stephan, S. (2013): A new classification approach for detecting severe weather patterns, in: Computers & geosciences 57 (2013): 158-165.
[7]: ECMWF (2023b): ERA5: data documentation parameterlistings. URL: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Parameterlistings
[8]: Scikit-learn (2023): https://scikit-learn.org/stable/documentation.html
[9]: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.
[10]: Gregor, S. / Hevner, A.R. (2013): Positioning and Presenting Design Science Research for Maximum Impact, in: MIS Quarterly, Jg. 37, Nr. 2, S. 337-355; Hevner, A. / Chatterjee, S. (2010): Design Research in Information Systems, Theory and Practice. Hrsg. von R. Sharda/S. Voß. Bd. 22. Integrated Series in Information Systems. New York, NY, USA: Springer New York, NY.; Hevner, A. / March, S.T. / Park, J. / Ram, S. (2004): Design Science in Information Systems Research, in: MIS Quaterly 28.1, S. 75–105.
[11]: Wilde, T. and Hess, T., 2007. Forschungsmethoden der wirtschaftsinformatik. Wirtschaftsinformatik, 4(49), pp.280-287.; Goldman, N. and Narayanaswamy, K., 1992, June. Software evolution through iterative prototyping. In Proceedings of the 14th international conference on Software engineering (pp. 158-172).
[12]: Reflective physical prototyping through integrated design, test, and analysis
[13]: Design Science in Information Systems Research.
[14]: Shao, J., 1993. Linear model selection by cross-validation. Journal of the American statistical Association, pp.486-494.; Browne, M.W., 2000. Cross-validation methods. Journal of mathematical psychology, 44(1), pp.108-132.
[15]: Webster, J. / Watson, R.T. (2002): Analyzing the past to prepare for the future: Writing a literature review, in: MIS quarterly. Jun 1: xiii-xiii.
[16]: Ortiz, Joaquin Amat Rodrigo and Javier Escobar (n.d.): Forecasting SARIMAX and ARIMA models - Skforecast Docs, [online] https://joaquinamatrodrigo.github.io/skforecast/0.7.0/user_guides/forecasting-sarimax-arima.html#.
[17]: Prabhakaran, Selva (2022): Augmented Dickey Fuller Test (ADF Test) – must read guide, Machine Learning Plus, [online] https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/.
[18]: Zach (2021): How to calculate AIC of regression models in Python, Statology, [online] https://www.statology.org/aic-in-python/.

Fathi, M. / Haghi Kashani, M. / Jameii, S. M. / Mahdipour, E. (2022): Big Data Analytics in Weather Forecasting: A Systematic Review, in: Archives of Computational Methods in Engineering 29.2 (2022, Springer): 1247–1275

Ghirardelli, J.E. (2005): An Overview of the Redeveloped Localized Aviation Mos Program (Lamp) For Short-Range Forecasting.