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 | Date: 29.12.2023 | Name: Julian Erath, Furkan Saygin, Sofie Pischl | Group: 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]. Technological advancements now allow for a more profound exploration of these processes [^2]. Accurate weather prediction is crucial for agriculture, disaster management, and urban planning, particularly in the context of climate change risks [^3]. 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. "The study leverages the ERA5, a high-quality global atmospheric reanalysis dataset covering multiple decades [^4]. 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 [^6]. This provides an excellent case study for analyzing relationships among different atmospheric elements [^7]. The evaluation is limited to the years 2015 to 2022 to capture the latest climate developments.

1.2 Data

Data description of sample 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. Various meteorological parameters such as temperature, precipitation, wind speed, and atmospheric pressure, grouped by average, minimum, and maximum values for the observed hour, 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, with each observation representing a set of meteorological parameters at a specific location and time. Recognized for its high quality and precision, the ERA5 dataset's enhanced spatial and temporal resolution makes it well-suited for detailed analyses and modeling across diverse applications, including climate research, environmental monitoring, and weather forecasting [^10].

Variables The dataset, crucial for this analysis, encompasses key variables such as air temperature, wind speed and direction, precipitation (rainfall and snowfall), atmospheric pressure, snow

density, cumulative snow, cumulative ice, and weather events. Temperature is measured in Kelvin, while wind information includes zonal and meridional components. Precipitation data is essential for hydrology and agriculture, and atmospheric pressure variations are associated with weather patterns. 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 [^11].

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 xy. The dataset comprises 65,345 observations and 184 columns, including unique predictor variables and a response variable. Identifier variables like "Unnamed: 0" are identified and dropped due to redundancy, while 'run_datetime' and 'valid_datetime' are transformed into datetime format. A new column, 'avg_temp_celsius,' is created by converting temperatures from Kelvin to Celsius. Wind directions in 'avg_winddir' are then categorized into cardinal directions using a function, resulting in the 'wind_direction_label' column. Subsequently, a new dataframe is formed by selecting specific columns for 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 [^12][^13].

1.3 Research Questions

Regression Analysis: 1. Is it possible to accurately predict temperature based on historical data? 2. Can a correlation between temperature and wind features be identified using regression techniques? 3. How does incorporating multiple atmospheric predictors enhance the accuracy of temperature prediction?

Classification Analysis: 1. Can extreme weather events be classified and predicted based on multivariate weather data? 2. Is it possible to categorize and predict different extreme weather events using multiclass classification algorithms?

1.4 Exploratory Data Analysis (EDA)

As mentioned before, the dataset comprises 65,345 entries with 184 columns, including features like substation, timestamps, weather-related parameters, and various labels. Am Most variables are of the "float64" data type (166), 8 variables are of type "int," and 10 are of type "object."

In terms of temperature, the 'avg_temp' column displays a mean of 279.57°C with a standard deviation of 11.38°C. The temperature ranges from a minimum of 243.85°C to a maximum of 300.93°C. The 'avg_windspd' column indicates an average wind speed of 2.44 m/s, ranging from 0.64 m/s to 6.47 m/s, while 'avg_windgust' exhibits an average gust speed of 7.64 m/s, with values ranging from 1.98 m/s to 19.91 m/s.

Exploring labels, 'label1' is binary with a mean of 0.81, suggesting an imbalanced distribution towards class 1. 'label2' is present in 12,712 entries, ranging from 0 to 6, and 'label3' has a mean of 1.18, indicating a slight skew towards lower values.

First, the variable "avg_temp" is examined. This includes depicting the temperature trend over time, as well as displaying the box plot and histogram. Next, the occurrences of weather events and their corresponding box plots are shown.

Next, the occurrences of weather events and their corresponding box plots are shown.

2 Methodology

2.1 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.

2.2 Analysis Steps

Die erste Phase der Methologie konzentrierte sich auf die umfassende Vorbereitung und Aufbereitung des ERA5-Datensatzes, um eine solide Basis für die anschließende Analyse sicherzustellen. Diese Phase hat das Ziel die Datenqualität zu gewährleisten und die Genauigkeit der Modelle zu maximieren.

Zunächst wurde der ERA5-Datensatz importiert und anschließend ausgelesen. Dafür wurden die ersten Zeilen sowie die Metadaten des Datensatzes betrachtet, um den Datensatz für die weitere Analyse vorzubereiten. Dieser Prozess beinhaltete die Auswahl relevanter meteorologischer Variablen sowie das schaffen neuer Variablen, die für unsere Analysezwecke von Bedeutung waren. Zuerst wird die Datums- und Zeitangaben im Datensatz in ein standardisiertes Datumsformat umgewandelt. Anschließend wird die durchschnittliche Temperatur von Kelvin in Celsius umgerechnet. Schließlich werden die Windrichtungsdaten, die in Grad angegeben waren, in kardinale Richtungen (wie Nordost, Ost usw.) umgewandelt, um die Daten anschaulicher und die Interpretation verständlicher zu gestalten. Als Ergebnis dieser Phase wurde das Dataframe mit folgenden Variablen zur weiteren Analyse in betracht gezogen: - Laufzeit (run_datetime): Das Datum und die Uhrzeit der Wetteraufzeichnungen. - Wetterereignistyp (wep): Eine Klassifizierung der Wetterbedingungen anhand spezifischer Parameter zu einem bestimmten Zeitpunkt und Ort. - Durchschnittstemperatur (avg_temp): Die durchschnittliche Temperatur, gemessen in Kelvin, basierend auf allen Sensoren für die Dauer einer Stunde. - Durchschnittstemperatur in Celsius (avg_temp_celsius): Umrechnung der Durchschnittstemperatur von Kelvin in Celsius. - Minimale Nassbulb-Temperatur (min_wet_bulb_temp): Die niedrigste Nassbulb-Temperatur während der Beobachtungsperiode. - Durchschnittlicher Taupunkt (avg_dewpoint): Der durchschnittliche

Taupunkt während der Beobachtungsperiode. - Temperaturänderung (avg_temp_change): Die durchschnittliche Temperaturveränderung während der Beobachtungsperiode. - Durchschnittliche Windgeschwindigkeit (avg_windspd): Die durchschnittliche Windgeschwindigkeit während der Beobachtungsperiode. - Maximale Windböe (max_windgust): Die höchste beobachtete Windböe während der Beobachtungsperiode. - Durchschnittliche Windrichtung (avg_winddir): Die durchschnittliche Windrichtung in Grad während der Beobachtungsperiode. - Sinus der Windrichtung (avg_winddir_sin): Sinus-Transformation der durchschnittlichen Windrichtung. - Cosinus der Windrichtung (avg_winddir_cos): Cosinus-Transformation der durchschnittlichen Windrichtung. - Kardinale Windrichtung (wind_direction_label): Umrechnung der durchschnittlichen Windrichtung in kardinale Richtungen. - Maximaler kumulativer Niederschlag (max_cumulative_precip): Die höchste kumulierte Niederschlagsmenge während der Beobachtungsperiode. - Maximale Schneedichte (max_snow_density_6): Die höchste Schneedichte in einer Tiefe von 6 Zoll während der Beobachtungsperiode. - Maximaler kumulativer Schnee (max_cumulative_snow): Die höchste kumulierte Schneemenge während der Beobachtungsperiode. - Maximaler kumulativer Eis (max_cumulative_ice): Die höchste kumulierte Eismenge während der Beobachtungsperiode. - Durchschnittliche Druckänderung (avg_pressure_change): Die durchschnittliche Veränderung des atmosphärischen Drucks während der Beobachtungsperiode. - Zusätzliche WEP-Labels (label0, label1, label2): Zusätzliche Kategorien zur Differenzierung verschiedener Wetterereignisse, einschließlich blauer Himmel und extremer Wetterbedingungen

2.3 Analysis and Visualisation

Die zweite Phase der Methologie umfasst eine tiefgehende Analyse und Visualisierung der Daten, um Erkenntnisse zu sammeln die Entscheidend für die Zielsetzung sein könnten.

Eine nähere Betrachtung der Wetterereignistypen hat gezeigt, dass ´Blue sky day´ mit einer Häufigkeit von 42106 am meisten vertreten ist, anschließend folgt mild snowfall mit 3598, moderate snowfall mit 2336 und moderate rainfall mit 2104. Extremwetterereignisse sind relativ wenig vertreten. Storm with freezing rain / heavy snow- and icestorm ist lediglich 69 mal, continuous freezing rain 37 mal, storm with freezing rain / heavy snow- and icestorm 17 mal und snowstorm with high precipitation 10 mal vorgekommen.

Anschließend wird die zeitliche Komponente untersucht in dem sinngemäß einzigartige Variablen über die Zeit geplottet werden.

- Saisonale Temperaturmuster: Deutlich sichtbar ist eine jährliche Zyklizität in den Temperaturparametern, mit höheren Werten im Sommer und niedrigeren im Winter. Die gleitenden Durchschnitte, sowohl auf wöchentlicher als auch auf monatlicher Basis, glätten die täglichen Schwankungen und verstärken die beobachteten saisonalen Trends.
- Windcharakteristika: Windgeschwindigkeit und -böen zeigen eine hohe tägliche Variabilität ohne erkennbare saisonale Muster. Dies lässt darauf schließen, dass die Windmuster in Bancroft komplexen und vielschichtigen Wetterdynamiken unterliegen und nicht nur einfachen saisonalen Einflüssen.
- Variabilität der Windrichtung: Ähnlich wie bei der Windgeschwindigkeit zeigt die Windrichtung eine hohe Variabilität ohne erkennbaren saisonalen Trend, was darauf hinweist, dass lokale geografische und meteorologische Komplexitäten die Windmuster erheblich beeinflussen.
- Schneefallmuster und -dichte: Die Schneedaten zeigen ein ausgeprägtes saisonales Muster, das invers mit den Temperaturdaten korreliert ist. Spitzen während der Wintermonate

korrespondieren mit den Temperaturtiefs, und die Schneedichte scheint mit nachfolgenden Schneefällen zuzunehmen, was auf eine Verdichtung im Laufe der Zeit hindeutet.

- Druckänderungen: Die durchschnittliche Druckänderung spiegelt das Diagramm der Temperaturänderung eng wider und deutet auf eine starke Beziehung zwischen atmosphärischem Druck und Temperatur hin, wobei saisonale Faktoren einen bedeutenden Einfluss auf diese Variationen zu haben scheinen. Die Ergebnisse können in der appendix xy betrachtet werden.

Eine Häufigkeitsverteilung der Wetterparameter in Form eines Histogramms erlaubt es zusätzliche Erkenntnisse durch statistische Methoden zu erlangen: - Temperaturen: Die bimodale Verteilung der Temperaturen mit Spitzen, die den Sommer- und Wintermonaten entsprechen, deutet auf ein klar definiertes saisonales Klima hin. Die Linksschiefe der Verteilung könnte auf eine längere Dauer oder eine größere Häufigkeit von kühleren Perioden im Jahresverlauf hindeuten, während die Symmetrie der Temperaturänderungen auf eine relative Stabilität des Klimas ohne drastische tägliche Schwankungen hinweist. - Wind: Die Rechtsschiefe bei Windgeschwindigkeit und Windböen lässt auf eine klimatische Norm mit überwiegend mäßigen Windbedingungen schließen, wobei gelegentliche stärkere Böen die Ausnahmen bilden. Diese Schiefe könnte darauf hinweisen, dass extreme Windereignisse zwar auftreten, aber nicht dominant sind. - Niederschlag: Ein stark rechtsschiefes Muster bei der Niederschlagsmenge zeigt, dass geringfügige Niederschlagsereignisse die Norm sind, während starke Niederschläge eher selten auftreten. Diese Verteilung könnte bedeuten, dass Bancroft vorwiegend trocken ist, mit sporadischen, intensiven Regenfällen. - Schnee und Eis: Die extreme Rechtsschiefe in der Verteilung von Schnee und Eis spiegelt wider, dass signifikante Akkumulationen ungewöhnlich sind, was auf ein Klima hindeutet, in dem solche Ereignisse zwar selten, aber potenziell intensiv sind. - Atmosphärischer Druck: Die geringe Variation und Schiefe bei den Druckänderungen deutet auf ein konsistentes Klima mit wenig Fluktuation hin, was für die Vorhersagbarkeit des Wetters in der Region vorteilhaft sein könnte.

Ein Boxplot bietet die Möglichkeit Parameter auf wichtige statistische Kennzahlen wie Median, Quartile, Interquartilabstand (IQR), Ausreißer und Verteilung zu untersuchen. Für alle genannten Wetterparameter wurde ein Boxplot erstellt, welche als Ergebnis die vorherigen Untersuchungen der Time Series und des Histogramms bestätigt. Klare saisonale Schwankungen in den Temperaturen und geringe tägliche Variabilität. Windgeschwindigkeiten sind meist niedrig, mit gelegentlichen Spitzen. Niederschlagsmuster sind überwiegend gering, mit seltenen schweren Ausreißern. Schnee- und Eisakkumulationen treten selten auf, während der atmosphärische Druck überwiegend stabil bleibt. Diese Ergebnisse deuten auf ein Klima, das regelmäßige saisonale Veränderungen ausgesetzt ist, mit gelegentlichen Extremwetterereignissen.

Das Umkehren der Sicht und Gruppieren der Daten in Wetterereignisse ermöglicht Einblicke in den Einfluss der einzelnen Wetterparameter auf die ausgewählten Wetterereignisse. Die Distogramme der Wetterdaten aus Bancroft enthüllen mehrere Schlüsselmuster: - Höhere Temperaturen führen eher zu ´Blue sky days´ während niedrige Temperaturen für einen milden Schneefall bis Schneesturm führen. Jedes Ergeignis ist um den Median zentriert, was bedeutet, dass alle Ereignisse bei der Median Temperatur vorkommen könnten. - eine Konzentration der Temperaturänderungen nahe Null und die beinahme Symmetrische Form, deutet auf eine schlechte Wahl als predictor Variable hin, da jedes Ereignisse unabhängig der Temperaturänderung vorkommen kann; - die Winddaten zeigen, dass die Wetterereignisse schlecht durch die Windgeschwindigkeit oder Windböen seperiert werden können. Die durchschnittliche Windrichtung weist Muster auf, bestimmte Wetterereignisse scheinen häufiger bei bestimmten Werten vorzukommen. Als Prediktor Variable alleine eher nicht geeignet, da zwar Muster erkannbar, trotzdem schlecht seperierbar sind - Die Verteilung der Niederschlags-, Schnee- und Eismengen ist rechtsschief, was häufige leichte

Ereignisse und seltene intensive Vorkommnisse anzeigt, während die Druckverteilung relativ normal erscheint, was auf ein stabiles atmosphärisches Umfeld schließen lässt. Die Variablen lassen bestimmte Wetterereignisse besser seperieren, da sie sinngemäß verbunden sind wie z. B. ´max cumulative snow accreation in mm´ und Schneefal bzw. Schneesturm.

Das gleiche wurde nochmal als Boxplot visualisiert und hat folgende Erkenntnisse gebracht: - **Clear Skies (Blue):** High average temperatures with low precipitation, typical of clear conditions. - **Continuous Freezing Rain (Orange):** Tight temperature ranges and variable pressure changes indicative of freezing rain conditions. - **Light to Heavy Snowfall (Red to Green):** Low temperatures with moderate to high snow and ice accumulations, typical of varying snowfall intensities. - **Moderate Rain and Snow (Purple and Brown):** Demonstrates variability in temperature changes and precipitation amounts. - **Severe Storms (Pink and Grey):** Marked by extreme precipitation and snow amounts, significant temperature and pressure fluctuations, and variable wind conditions.

Selbstverständlich sind bestimmte Wetterereignisse seltener als andere, weshalb es notwendig ist sich die Verteilung anzuschauen. Ein Kuchendiagramm welches erst alle Wetterereignisse abbildet und anschließend die Extremwetterereignisse, hat folgende Ergebnisse gebracht: - Gesamtverteilung der Wetterereignisse: Der größte Anteil der Beobachtungen fällt unter “Blue Sky Day”, was klareres Wetter ohne extreme Bedingungen bedeutet. Dieser Zustand stellt über 77.9% der gesamten Beobachtungen dar. Andere Wetterereignisse wie mäßiger Regen, leichter und mäßiger Schneefall sind ebenfalls hervorgehoben, aber weniger häufig. - Verteilung extremer Wetterereignisse: Moderate snowfall is the most common, making up 35.4% of non-clear sky observations. Moderate snowfall (23.0 %) and moderate rain (20.7 %) are the next most common events. Those three accumulated make up to 79.1 percent, the remaining 20.9 percent are the more extreme events. The more extreme events include:

Heavy snowfall with accumulated snow

Frontdurchlauf / Continuous freezing rain

Storm with freezing rain / Heavy snow- and icestorm

Snowstorm with high precipitation,

sorted by their frequency.

Anschließend wurden Assoziationen und Korrelationen zwischen verschiedenen meteorologischen Parametern untersucht. Dazu wurden Scatterplots für alle Paare relevanter Parameter erstellt und die Korrelationskoeffizienten berechnet. Die wichtigsten Erkenntnisse:

- Starke Korrelationen zwischen ähnlich skalierten Variablen: Parameter mit ähnlichen Skalen und Maßeinheiten zeigten tendenziell stärkere Korrelationen. Beispiele sind die durchschnittliche Temperatur, Taupunkt, Temperaturänderung und die minimale Feuchtkugeltemperatur sowie durchschnittliche Windgeschwindigkeit und maximale Windböen.
- Dynamische Natur der Korrelationen: Die Korrelationen zwischen den Parametern änderten sich in Reaktion auf extreme Wetterereignisse. Diese Ereignisse scheinen die Beziehungen zwischen den Parametern maßgeblich zu beeinflussen.

Einbeziehung von Temperatur- und Windparametern in Regressionsanalysen: Obwohl keine direkte Korrelation zwischen Temperatur- und Windparametern festgestellt wurde, wurden sie dennoch in die Regressionsanalysen einbezogen. Dies beruht auf der Annahme, dass ihre Beziehung möglicher-

weise nichtlinear ist oder von weiteren Faktoren beeinflusst wird, die durch lineare Korrelation nicht erfasst werden.

Die Ergebnisse zeigen, dass eine umfassende Betrachtung von Korrelationen und Assoziationen zwischen meteorologischen Parametern notwendig ist, um komplexe Interaktionen und den Einfluss extremer Wetterereignisse auf diese Beziehungen zu verstehen. Die Studie legt nahe, dass für ein vollständiges Verständnis der atmosphärischen Dynamik in Bancroft eine Kombination aus linearen und nichtlinearen Analysemethoden erforderlich ist.

In einer weiteren Analyse wurde der Zusammenhang zwischen Windrichtung, Windgeschwindigkeit und Durchschnittstemperatur untersucht. Die Hauptergebnisse sind:

- Windgeschwindigkeit: Die durchschnittliche Windgeschwindigkeit variiert je nach Windrichtung. Die höchsten Windgeschwindigkeiten wurden bei Nord- und Westwinden beobachtet, während die niedrigsten Geschwindigkeiten bei Südost- und Ostwinden auftraten.
- Durchschnittstemperatur: Die Farbe der Balken im Diagramm repräsentiert die Durchschnittstemperatur. Wärmere Temperaturen sind durch dunklere Rottöne gekennzeichnet, kühlere Temperaturen durch dunklere Blautöne. Winde aus südwestlicher Richtung bringen die höchsten Temperaturen, während nordgerichtete Winde mit den niedrigsten Temperaturen verbunden sind.

Interpretation: Die Studie zeigt, dass Windgeschwindigkeit und -richtung variieren und mit unterschiedlichen Temperaturen assoziiert sind. Südwestwinde korrelieren tendenziell mit wärmeren Temperaturen, während Nordwinde kältere Luftmassen mitbringen. Höhere Windgeschwindigkeiten bei Nord- und Westwinden deuten auf stärkere Windereignisse oder eine generelle Tendenz zu höheren Windgeschwindigkeiten aus diesen Richtungen hin.

Die Ergebnisse verdeutlichen, dass die Windrichtung einen signifikanten Einfluss auf Windgeschwindigkeit und Temperatur hat. Diese Erkenntnisse sind wichtig für die Wettervorhersage und Bereiche wie die Energieproduktion, wo Windenergie und Temperaturmanagement entscheidende Faktoren sind. Die Analyse legt nahe, dass ein Zusammenhang oder sogar eine kausale Verbindung zwischen Wind und Temperatur besteht, da südliche Winde milder und wärmer sind, während nördliche Winde stärker und kälter sind.

A transition is now made to a more abstract yet insightfuThe multi-dimensional weather data will be condensed into three principal components, providing a visual exploration of the intrinsic structure and variability of the data. By plotting this 3D PCA scatter plot, hidden patterns, clusters, or anomalies across the weather events are anticipated to be uncovered.

Die PCA-Analyse (Principal Component Analysis) des Wetterdatensatzes hat folgende Schlüsselerkenntnisse geliefert:

- Verteilung der Datenpunkte: Die Datenpunkte verteilen sich ungleichmäßig und bilden eine ausgeprägte, kegelförmige Struktur, die sich hauptsächlich entlang der PC1- und PC2-Achsen erstreckt. Dies deutet darauf hin, dass diese Komponenten den Großteil der Variabilität in den Wetterbedingungen erfassen.
- Ausprägung einzelner Wetterereignisse: Besondere Wetterereignisse wie “Sturm mit gefrierendem Regen / Schwerer Schnee- und Eissturm” zeigen eine deutliche Ausdehnung im PCA-Raum, insbesondere auf den Achsen PC1 und PC2, was sie als Ausreißer kennzeichnet. Im Gegensatz dazu erstreckt sich “Blue Sky Day” vorwiegend unterhalb von 0 auf PC1 und -5 bis 10 auf PC2, was die breite Variation in meteorologischen Merkmalen hervorhebt.

- Dichte Clusterbildung: Ein dichter Cluster um den Ursprung zeigt, dass die häufigsten Wetterbedingungen in Schlüsselmerkmalen wie Durchschnittstemperatur und Niederschlag eine hohe Überschneidung aufweisen. Dieser dichte Kern steht im Kontrast zu dem ausgedehnten Schwanz und den Ausreißern in den PCA-Plots und deutet auf ein Spektrum von häufigen bis hin zu seltenen und extremen Wetterereignissen hin.
- Herausforderungen bei der Interpretation: Die Trennung der Cluster, besonders innerhalb von PC3, ist aufgrund der Feinheit, die diese Komponente einfängt, schwierig. Die erhebliche Überschneidung von Bedingungen innerhalb von PC1 zeigt die Komplexität bei der Unterscheidung verschiedener Wettermuster.

2.4 Model

2.4.1 Regression Analysis Temperature and Wind

Als erstes ist es notwendig ein geeignetes Modell zu wählen. Dafür wurde eine lineare Regression, ein GradientBoosting, ein SGD Regressor und ein Support Vector Regressor trainiert. Als Prediktoren wurden sämtliche Variablen verwendet, die auch für die EDA in Betracht gezogen wurden. Die Response Variable ist in diesem Fall die durchschnittlche Windgeschwindigkeit. Das Verhältnis der Trainingsdaten zu den Testdaten wurde auf 80 zu 20 festgelegt. Evaluiert wurden die Modelle nach dem Mean Squared Error(MSE) und dem Mean Absolute Error (MAE), wobei niedrigere Werte ein präziseres Modell signalisieren. Alle Modelle haben ähnliche Resultate gezeigt, weshalb eine Visualisierung der Ergebnisse vorgenommen wurde. Key Insights:

- Weak Relationship Between Wind Speed and Temperature: The data points are broadly scattered in the plots, indicating a weak relationship between wind speed and temperature. This is further supported by the relatively high MSE and MAE values across all models, suggesting limited accuracy in predicting temperature based solely on wind speed.
- Residual Analysis: The residual plots, which show differences between actual and predicted values, are centered around zero but have a wide spread. This indicates substantial prediction errors. Notably, the Support Vector Regression model shows a negative mean residual, hinting at a tendency to underpredict temperatures.
- Model Performance and Bias: All models have mean residuals close to zero, except for the Support Vector Regression. This suggests no significant bias in over or underestimation for most models.
- Complexity of Temperature Patterns: The results highlight the complexity of temperature patterns, which are influenced by various climatic factors. This complexity cannot be adequately captured through regression with wind speed as the only predictor, indicating the necessity for multiple regression analysis incorporating additional relevant variables.
- Linear Relationship in Specific Cases: The reference to a subsequent linear regression analysis between wind speed and wind gusts suggests that in cases where two variables are closely linked and exhibit a clear linear relationship, simple regression can be effective, as indicated by lower MSE and MAE values. However, such analyses might be limited in scientific value if the variables are inherently correlated and do not offer unique insights into the system under study. Overall, the analysis underscores the importance of considering multiple factors and the limitations of using a single predictor in complex systems like weather patterns. It also highlights the need for careful interpretation of model results, especially in the context of inherent correlations and the complex nature of environmental data.

2.4.2 Regression Analysis Multiple Linear Regression

2.4.3 Regression Analysis Temperature Forecast

2.4.4 Logistic Regression

2.4.5 Binary Classification

2.4.6 Multiclass Classification

For the analysis the data was split into training and testing datasets according to [^22] and [^23].

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 the wind parameter average wind speed and the average temperature is investigated. In this context, models such as the Linear Regression Model, Gradient Boosting Model, Stochastic Gradient Descent Model, and Support Vector Regression Model are utilized to depict the correlation. These models predicting average temperature from average wind speed using various regression techniques 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.

Feature Selection Before initializing a Linear Regression Model with Multiple predictors, the best variables are examined. Therefore, the variables are checked for multicollinearity in the first step. Multicollinearity exists if two independent variables exhibit a high degree of correlation. High correlations indicate a robust relationship, implying that only one of the two variables is necessary for regression analysis. The variables that do not have high collinearity are written to a new df (variables: avg_temp, avg_windspd, avg_winddir, avg_winddir_sin, avg_winddir_cos, max_cumulative_precip, max_snow_density_6, max_cumulative_snow, max_cumulative_ice, label1). Next, a forward and backward selection is being made on these variables, but it did not yield significant advantages in this study. Prior to the commencement of the Bachelor's thesis, a thorough evaluation of variables was conducted based on existing literature and relevant metrics. This meticulous pre-selection process ensured that only essential variables were included in the dataset, aligning with established theoretical foundations and methodological considerations.

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. 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. SARIMAX (Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors) models are among the most widely used statistical models for forecasting, with excellent forecasting performance [ˆ35]. 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), Akaike Information Criterion (AIC), 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.

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. 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 xy 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.

The classification reports found in appendix xy 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.

3.2.2 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 x to y.

The multiclass classification with KNN 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 SVM exhibits a higher rate of misclassifying events compared to the 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 DTC 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 GBC 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 SVM tends to misclassify events more frequently, especially for class 0, often predicting it as class 1. On the other hand, the KNN algorithm generally performs well, with actual results closely aligning with predictions. The GBC outshines other models, exhibiting exceptional precision, recall, and F1-score for most classes and achieving an impressive overall accuracy of 98%. The DTC 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 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

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 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

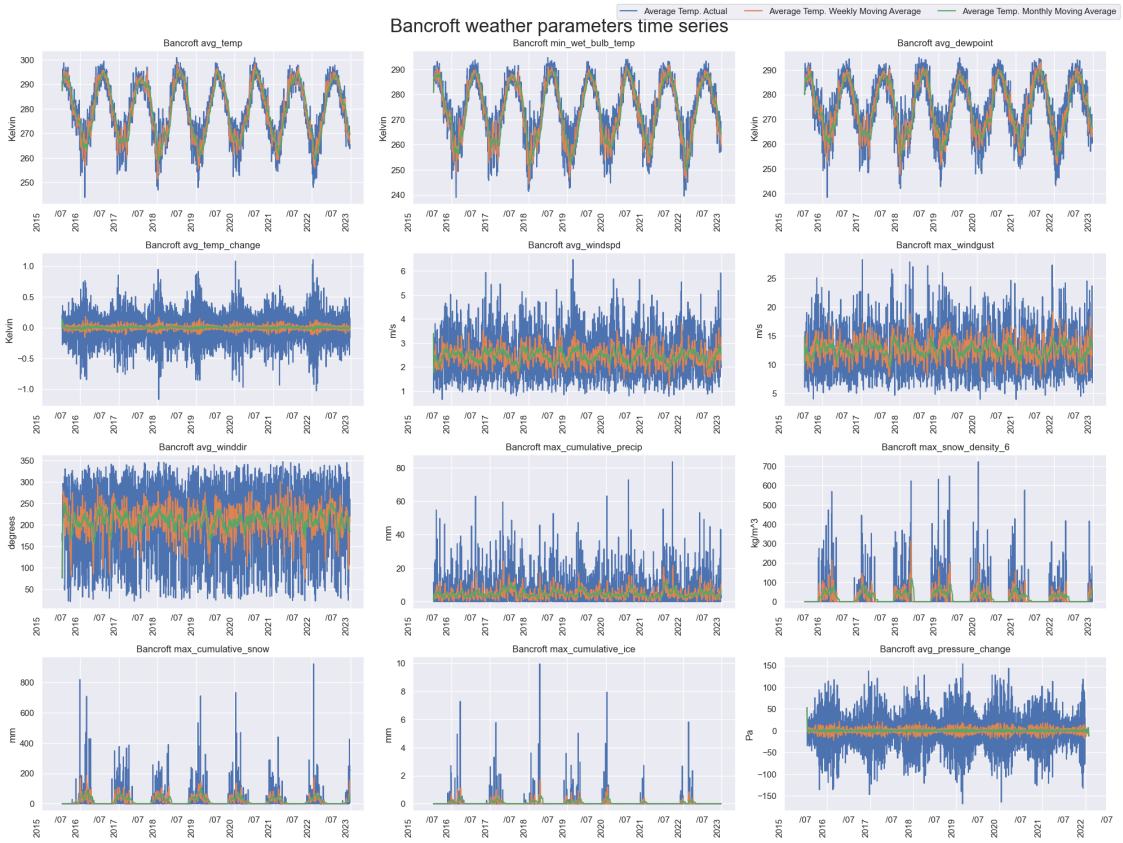
[5]:

	Name	Description	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…	predictor	numerical continuous
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 …		

11	predictor	numerical	continuous
12	predictor	numerical	continuous
13	predictor	numerical	continuous
14	predictor	numerical	continuous

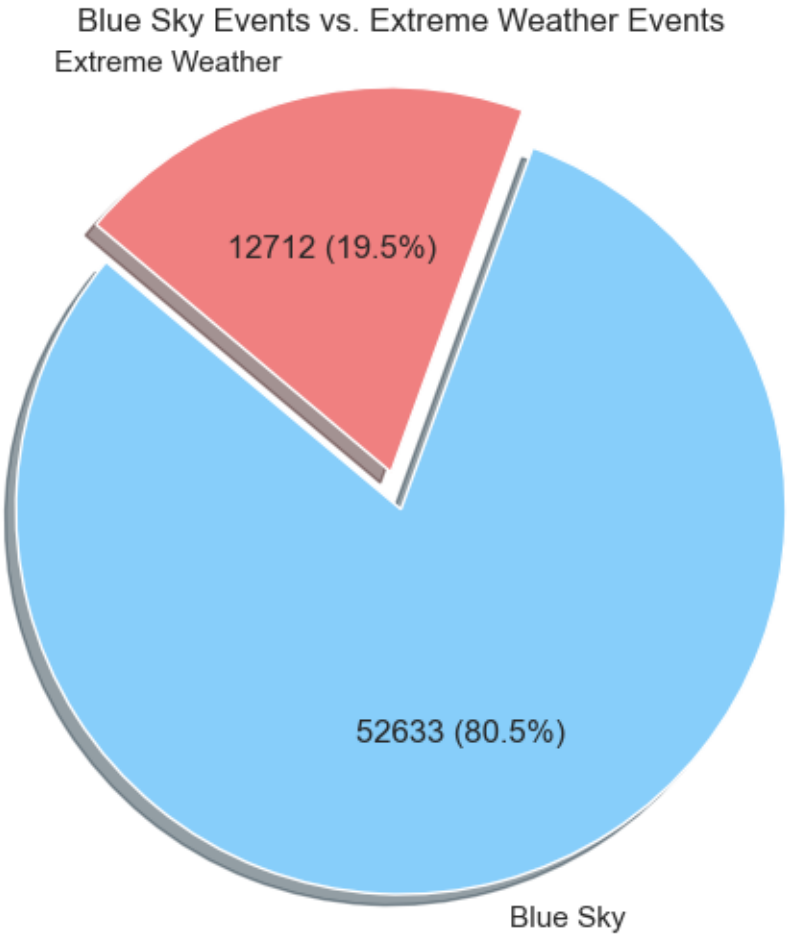
	Format
0	<class 'pandas._libs.tslibs.timestamps.Timesta…
1	<class 'str'>
2	<class 'numpy.float64'>
3	<class 'numpy.float64'>
4	<class 'numpy.float64'>
5	<class 'numpy.float64'>
6	<class 'numpy.float64'>
7	<class 'numpy.float64'>
8	<class 'numpy.float64'>
9	<class 'str'>
10	<class 'numpy.float64'>
11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

5.4 Time series



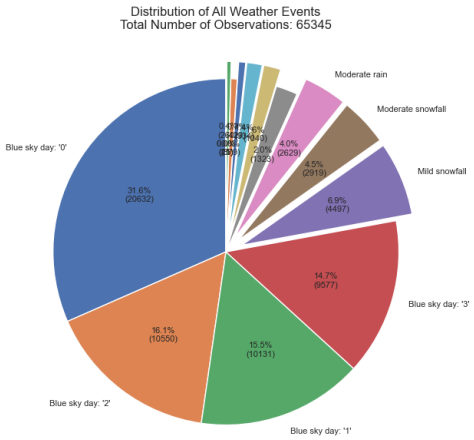
<Figure size 2000x1500 with 0 Axes>

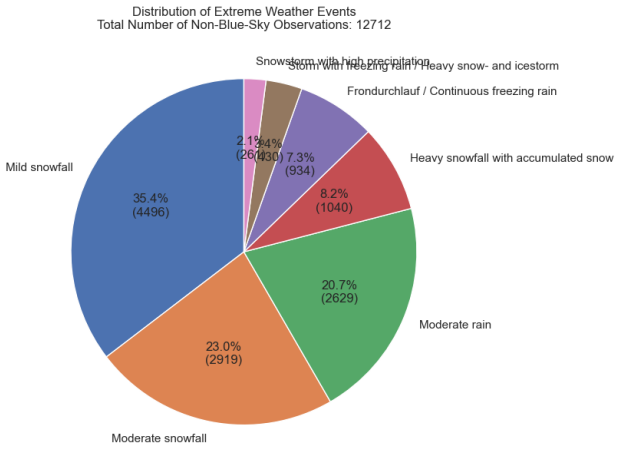
5.5 Class Distribution of Blue Sky and Extreme Weather Events



<Figure size 2000x1500 with 0 Axes>

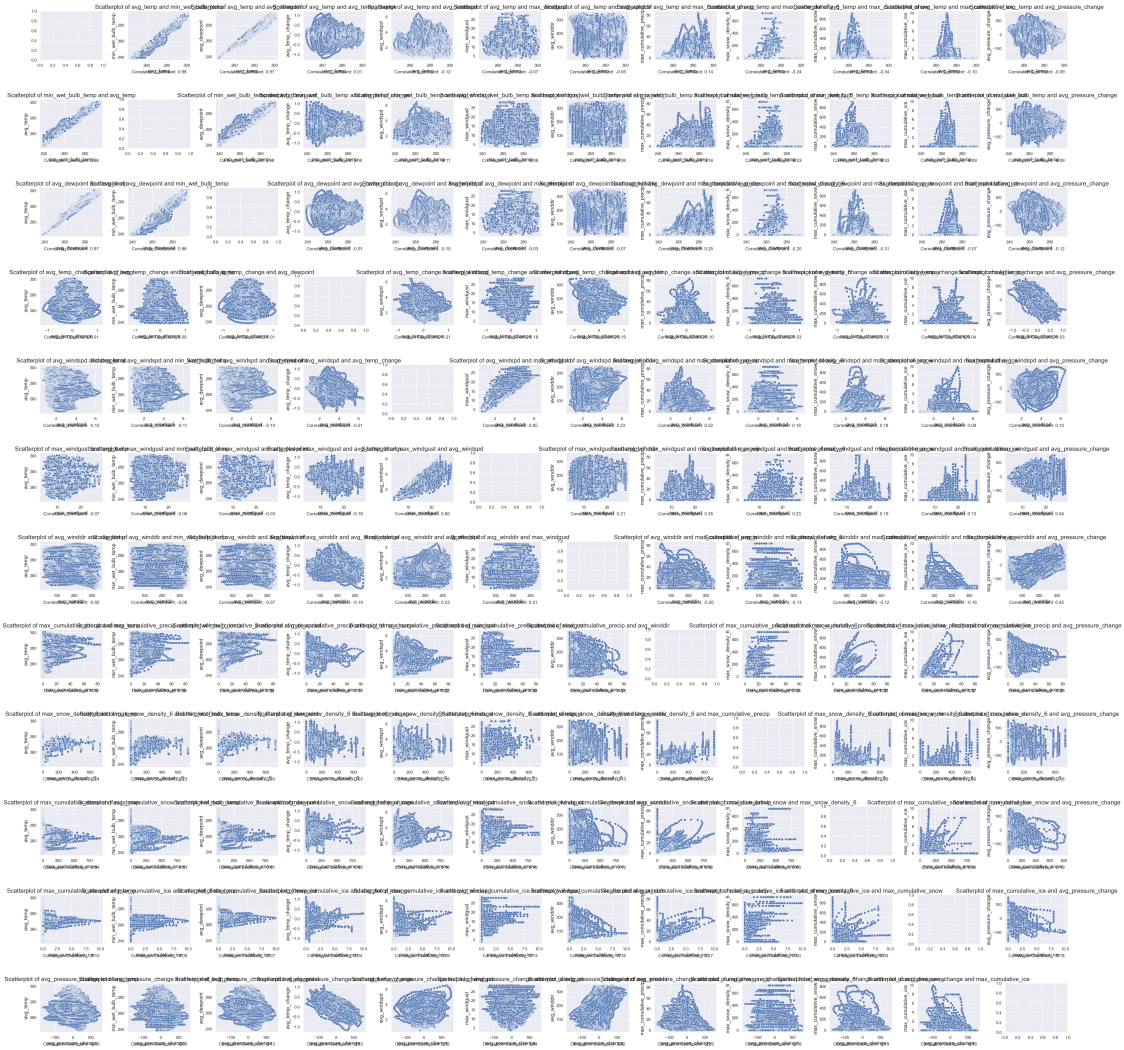
5.6 Distribution of All Weather Events





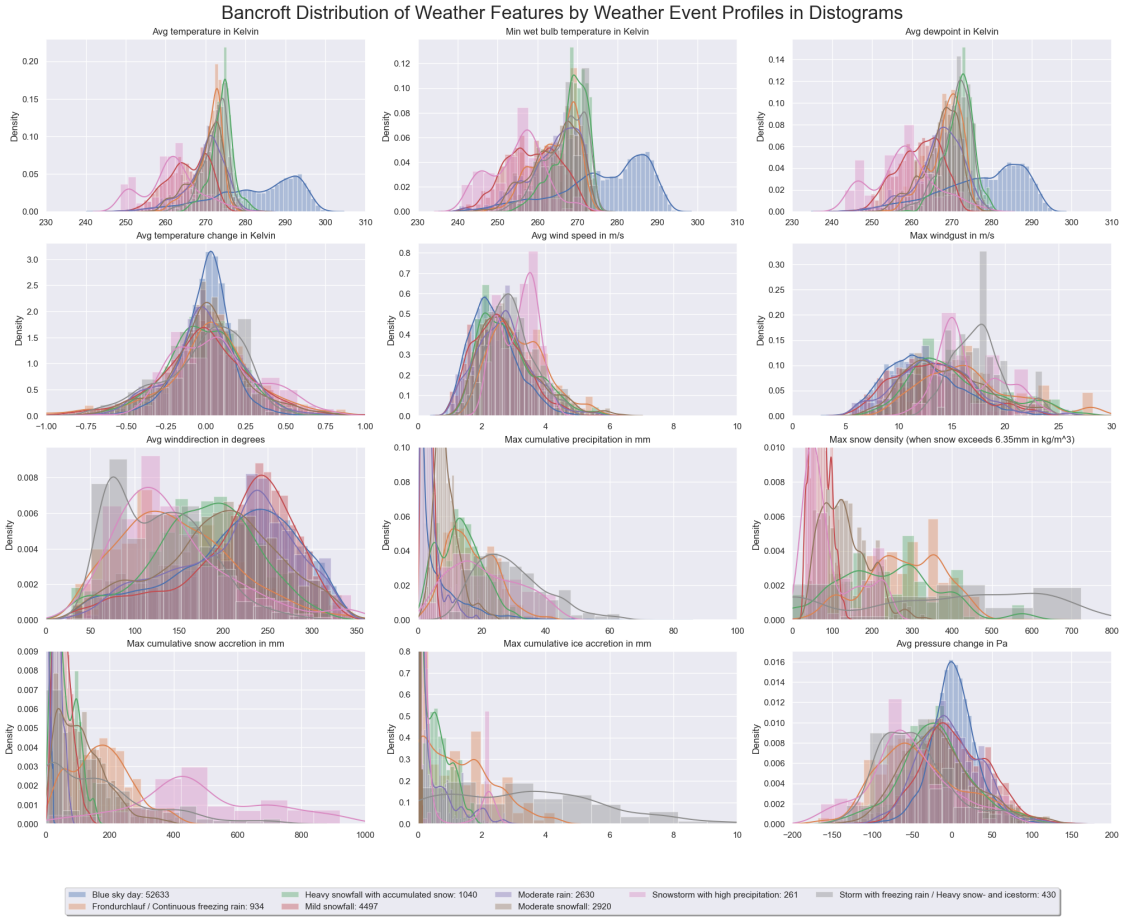
<Figure size 2000x1500 with 0 Axes>

5.7 Association Plots and Correlation Analysis

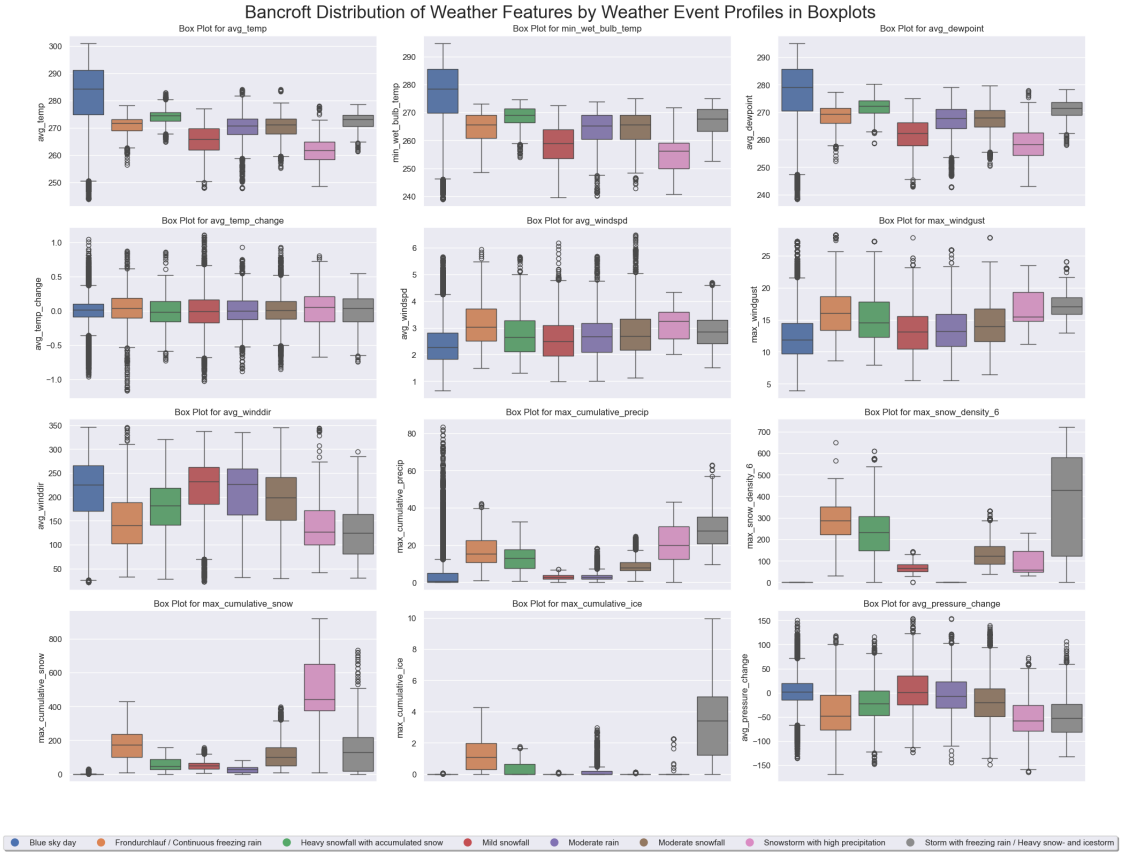


<Figure size 2000x1500 with 0 Axes>

5.8 Distribution of Weather Features by Weather Event Profiles in Distograms

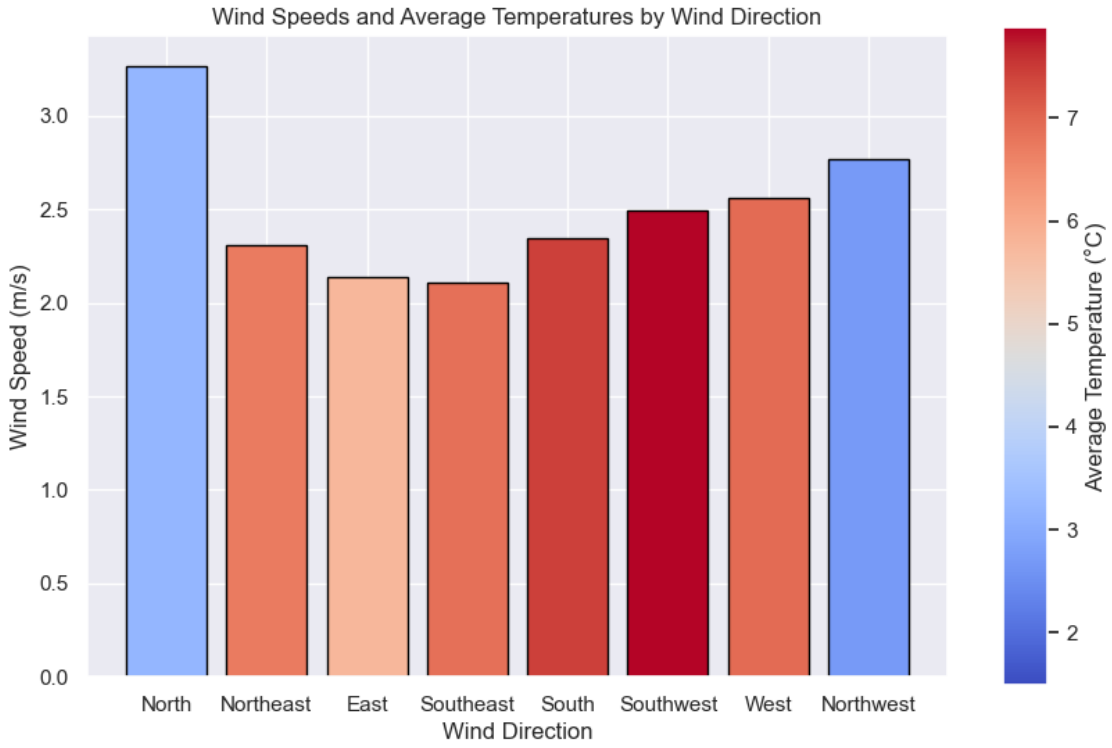


5.9 Distribution of Weather Features by Weather Event Profiles in Boxplots



<Figure size 2000x1500 with 0 Axes>

5.10 Analysis of Wind Speeds and Average Temperatures by Wind Direction

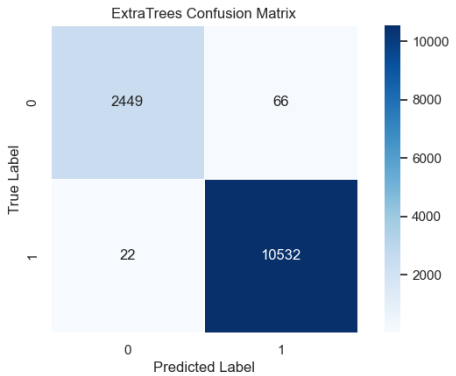


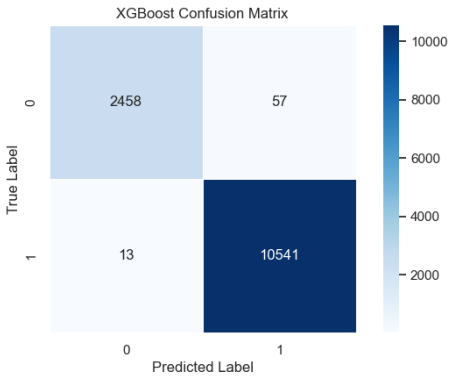

```
<Figure size 2000x1500 with 0 Axes>
```

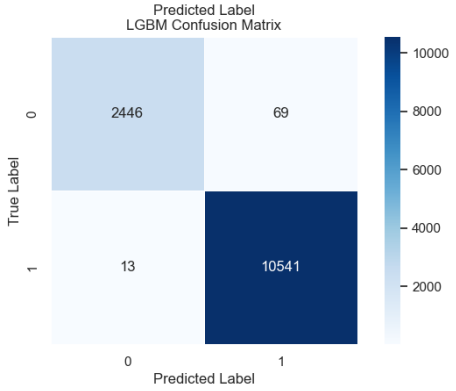
5.11 3D plot of all weather observations using PCA

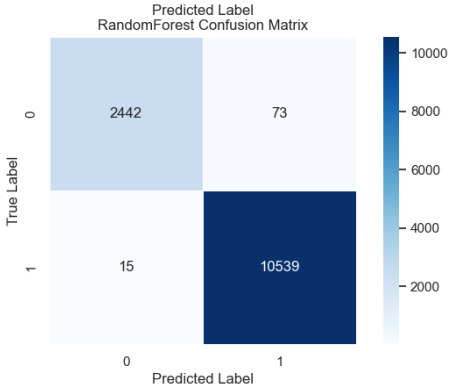
5.12 Methodology and Results Binary Classification

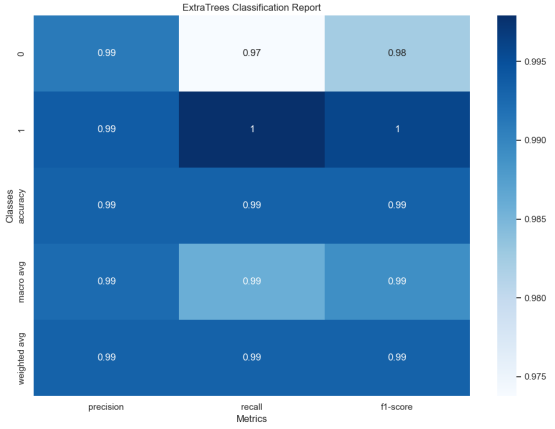
```
ExtraTrees Accuracy: 0.9930369576861274
XGBoost Accuracy: 0.9946438136047134
LGBM Accuracy: 0.9935725763256561
RandomForest Accuracy: 0.9931134746346316
```

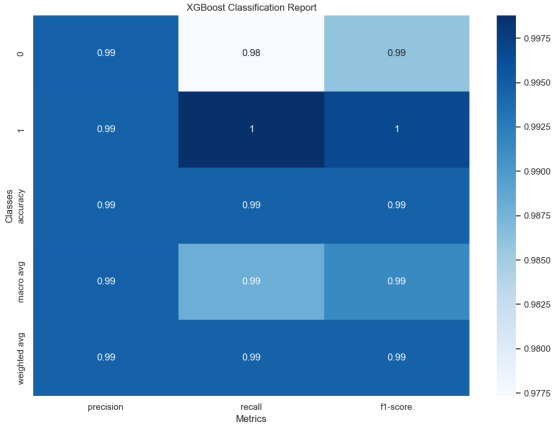

ExtraTrees Confusion Matrix
2449
66
22
10532
True Label
Predicted Label

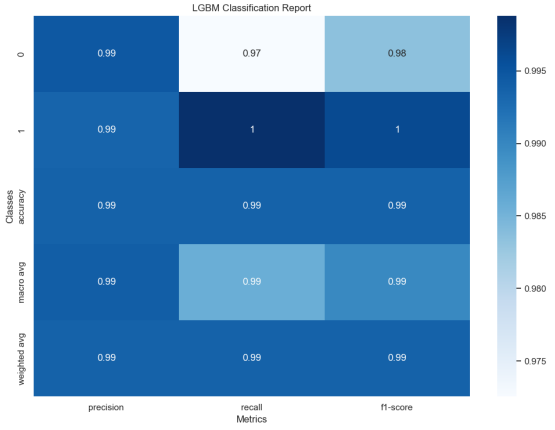

XGBoost Confusion Matrix
2458
57
13
10541
True Label
Predicted Label

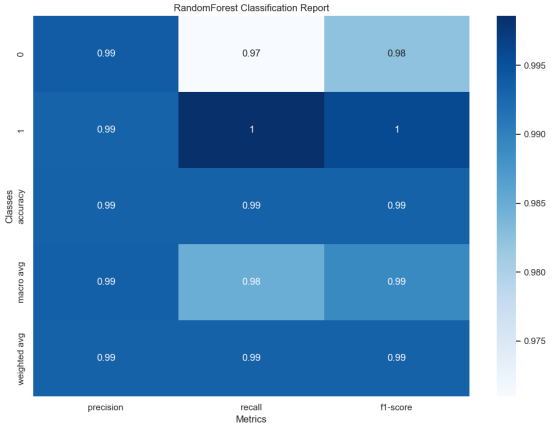

LGBM Confusion Matrix
2446
69
13
10541
True Label
Predicted Label


RandomForest Confusion Matrix
2442
73
15
10539
True Label
Predicted Label


ExtraTrees Classification Report
Classes
Metrics
precision
recall
f1-score

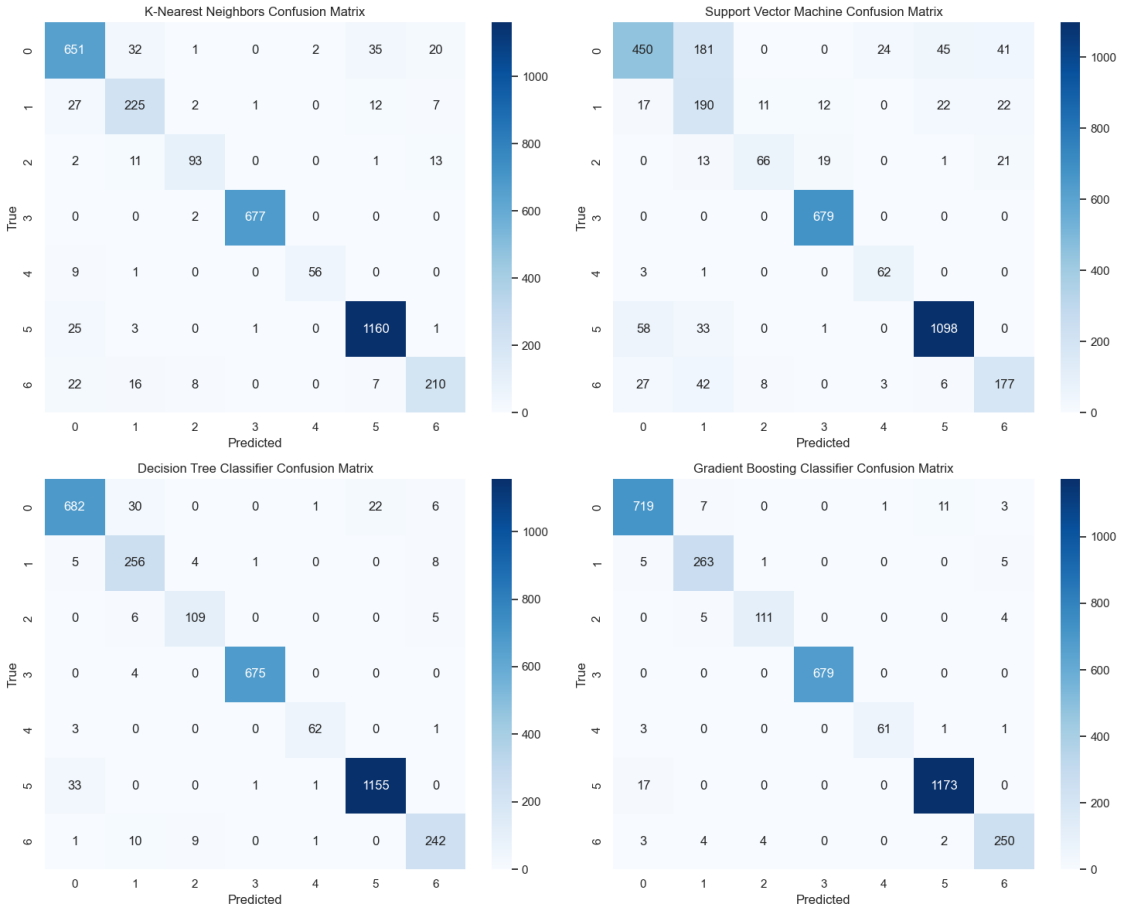

XGBoost Classification Report
Classes
Metrics
precision
recall
f1-score

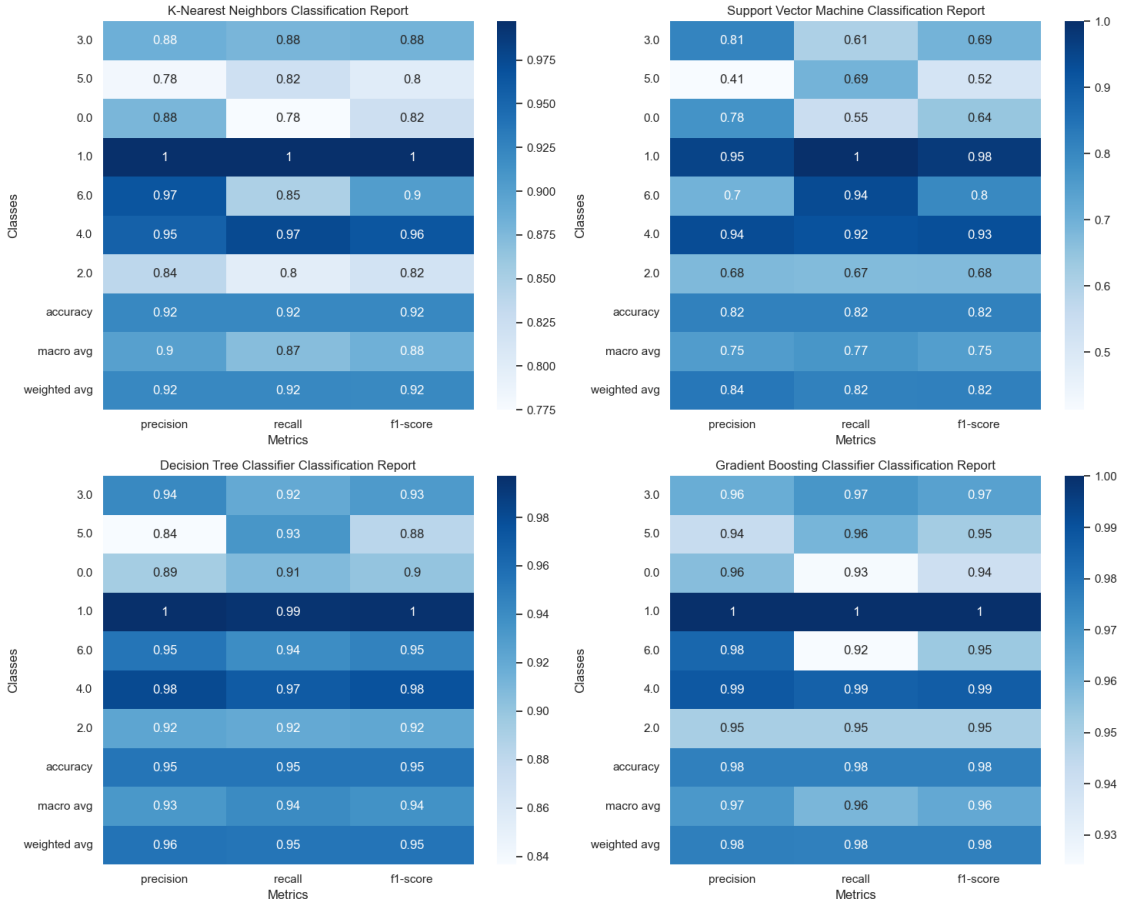

LGBM Classification Report
Classes
Metrics
precision
recall
f1-score

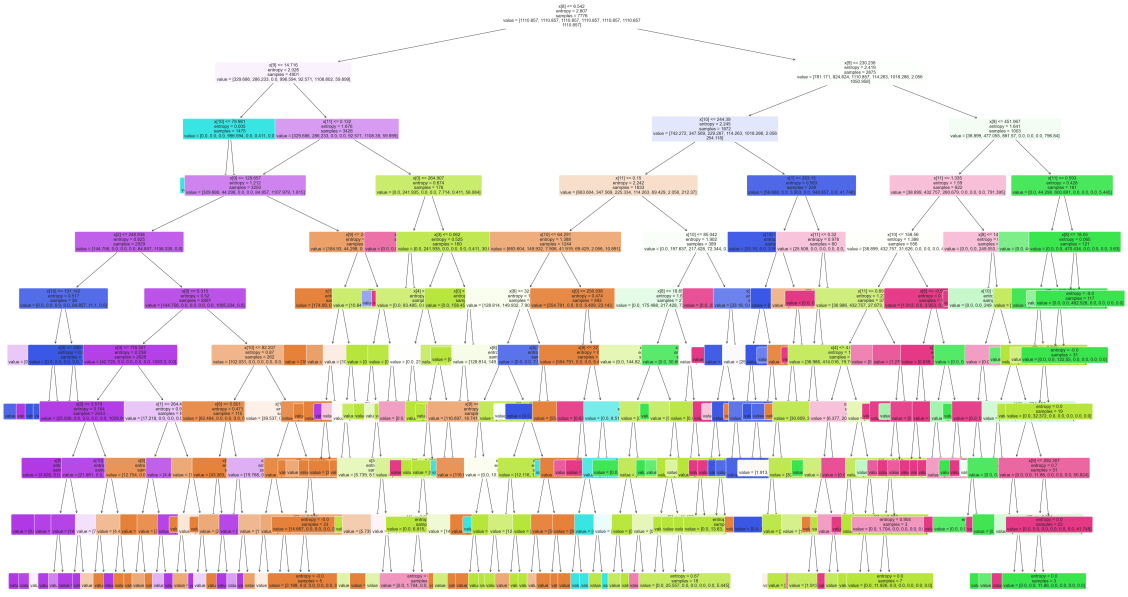

RandomForest Classification Report
Classes
Metrics
precision
recall
f1-score

<Figure size 2000x1500 with 0 Axes>

5.13 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



5.14 Sources