Project Documentation

Data Scientist Machine Test

Weather Classification using Meteorological Parameters-A Machine Learning Approach on Weather Classification Dataset

Name: Rosemary Raphael

Table of Contents

15. Challenges faced and how it was addressed

16. Conclusion

1. Introduction

Weather classification is a crucial task that influences numerous sectors, including agriculture, transportation, energy, and urban planning. Accurate weather classification not only aids in efficient resource allocation but also helps mitigate risks associated with extreme weather events. This project utilizes historical meteorological data to classify weather patterns using clustering techniques and supervised machine learning models. The combination of unsupervised and supervised approaches enhances understanding of weather conditions, paving the way for better decision-making and predictive capabilities.

2. Aim

The project aims to achieve two objectives: first, to classify weather patterns using clustering techniques to segment data into meaningful groups, and second, to predict weather conditions using supervised machine learning models. By incorporating hyperparameter tuning and cross-validation, the project ensures accurate and practical results that can be applied in real-world scenarios.

3. Business Problem / Problem Statement

Weather variability poses challenges for industries reliant on environmental conditions. For example, agriculture must contend with droughts or heavy rains, while logistics face disruptions caused by extreme weather. Despite the availability of extensive meteorological data, its complexity often limits actionable insights. This project addresses these challenges by using clustering and predictive modeling to classify and forecast weather patterns, enabling businesses to minimize uncertainties and improve planning.

4. Project Workflow

The project follows a structured workflow: understanding the data, cleaning and preprocessing it to ensure quality, engineering derived features, conducting exploratory data analysis (EDA), applying clustering techniques, and using supervised machine learning models for prediction. Clustering results are evaluated using metrics like silhouette scores, while predictive models are optimized through hyperparameter tuning and validated using cross-validation. Finally, insights and recommendations are presented to stakeholders.

5. Data Understanding

The dataset consists of **13,200 rows and 11 columns** that capture various meteorological parameters for weather classification. The dataset includes continuous variables such as **temperature, humidity, wind speed, and atmospheric pressure**, along with categorical variables like **weather type**. Initial exploration revealed that there are no missing values in the dataset and outliers are not dropped as temperatures and other factors vary with each region and also it is necessary for increasing the model's accuracy. These variables were crucial in identifying patterns and creating meaningful segments for analysis and prediction.

6. Data Cleaning

Data cleaning involves addressing missing values, detecting and handling outliers using the interquartile range (IQR) method or Z-Score. Outliers can be are not dropped as temperatures and other factors vary with each region. In the dataset no inconsistent values were found. Since all the columns are important no columns were removed for clustering and classification.

7. Obtaining Derived Metrics

Derived metrics such as **visibility** were categorized into "**Moderate", "Clear", "Poor"** to

enrich the analysis by creating a new column named as the **Visibility Category.** This method is called **Feature Engineering**.

8. Filtering Data for Analysis, Scaling & Encoding

The dataset was filtered to retain only high-quality, relevant records. Duplicate entries were removed, and irrelevant time periods or regions were excluded. This ensured that the clustering and predictive results were reliable and meaningful. The **categorical columns** like Cloud Cover","Season","Location","Visibility Category" & "Weather Type" was encoded using **One-hot encoding and Lebel encoding** along with **scaling numerical columns** like "Temperature", "Humidity","Wind Speed","Precipitation (%)","Atmospheric Pressure","UV Index" using **MinMax Scaler and Standard Scaler** for Supervised and Unsupervised Machine Learning.

9. Exploratory Data Analysis (EDA)

Univariate analysis was conducted to understand the distribution of variables, with **histograms and count plots** used for continuous features like temperature and humidity. Segmented univariate analysis explored temperature distributions and wind speed variations for specific weather types, with **box plots and bar plots**. Bivariate analysis used **scatter plots, box plots, line plot and count plot** to uncover relationships between variables. Multivariate analysis applied **correlation heatmap** to identify weather patterns based on temperature, humidity, and pressure. **Pair plots** were used to visualize the relationships between multiple variables in the dataset.

10. Supervised Machine Learning

Various supervised algorithms, **including Logistic Regression – 0.85% accuracy, Random Forest – 0.91% accuracy, SVM – 0.88% accuracy, KNN – 0.88% accuracy, and XGBoost – 0.91% accuracy**, were implemented to classify weather patterns. The dataset is

split into train and Test with a random state of 42. Hyperparameter tuning using **random search** for improved model performance, while **k-fold cross-validation** ensured robust validation. Evaluation metrics such as **accuracy, precision, recall, F1-score** were used to compare model performance, with XGBoost achieving the **highest accuracy – 0.91%.**

11. Unsupervised Machine Learning

Unsupervised machine learning, uses techniques like the **K-Mean clustering** along with the **Elbow Method**. The Elbow Method helps to identify the optimal number of clusters by plotting the inertia for different cluster values and locating the "elbow" point. The optimal number of clusters is determined i.e., **4 clusters**, K-Means is applied to group the data. To evaluate the quality of the clustering, the **Silhouette Score**, which measures the cohesion and separation of the clusters is used. The score is 0.21% which indicates the model performed moderately well. The results are visualized through a clustering graph, providing insights into the data's structure and the effectiveness of the clustering solution. Which classified weather type into high temperature, moderate temperature, low temperature and clear skies.

12. Overall Insights Obtained from Analysis

The dataset reveals several interesting weather patterns and relationships. Temperature is right-skewed, peaking around 20-30°C, with warmer conditions linked to lower humidity and higher UV index. Humidity is normally distributed, with the highest levels in winter, while precipitation is bimodal, with common low and high levels. Weather conditions like clear skies show higher temperatures, whereas overcast skies are cooler. Wind speed negatively affects visibility, and UV index increases with temperature. Seasonal and weather type variations significantly impact humidity, precipitation, and temperature, with correlations suggesting that higher humidity often signals rain and stronger winds, while UV exposure is higher in clear skies.

13. Recommendations

The recommendations focus on improving weather forecasting, public safety, and preparedness for various conditions. Monitoring temperature patterns can help predict drier conditions and higher UV exposure, while tracking humidity levels in winter aids in forecasting rain or snow. UV exposure awareness is crucial, especially in sunny, inland, and mountain regions. Wind and precipitation should be monitored to anticipate storms and reduced visibility, and seasonal adjustments for heating and hydration are recommended. Location-specific forecasts for coastal and inland areas can guide public health measures, while cloud cover and temperature trends can inform outdoor activities. Additionally, attention to precipitation, wind speed, and atmospheric pressure helps with storm preparedness and emergency planning.

14. Key Findings

The key findings highlight distinct weather patterns and their interactions. Warmer temperatures are associated with drier conditions and higher UV exposure, while winter sees higher humidity and increased precipitation, especially in areas at risk for rain or snow. Wind speed and precipitation often correlate with reduced visibility and stronger storms. Clear skies bring warmer temperatures, whereas overcast skies are cooler. Seasonal trends affect both energy use and health risks, with winter requiring more heating and the warmer seasons necessitating sun protection. Location-specific weather patterns, such as sunny conditions in inland areas and snow in mountainous regions, further emphasize the need for tailored forecasting and public safety measures.

15. Challenges Faced and How They Were Addressed

The challenges faced in this weather analysis included handling outliers, as well as dealing with the complexity of multiple weather variables with varying distributions and correlations.

The complexity of variable interactions was tackled by performing thorough exploratory data analysis (EDA) to identify correlations and patterns, which informed the selection of appropriate models and features for prediction. Additionally, some weather variables showed outliers, especially in extreme conditions, which were managed by using robust methods that minimized their impact on the analysis. Finally, the dataset's skewed distributions, such as temperature, were addressed through normalization and transformation techniques like Minmax Scaler, Standard Scaler, One-hot encoding and Label encoding to improve model performance and ensure more accurate predictions. Regularization techniques and cross-validation were employed to reduce model overfitting and ensure robust performance.

16. Conclusion

In conclusion, this project effectively used both supervised and unsupervised machine learning techniques to classify and predict weather patterns. By analyzing variables such as temperature, humidity, and UV index, and applying clustering methods like K-Means and models like XGBoost, the project provided valuable insights into weather trends and correlations. Despite challenges with outliers and complex variable interactions, techniques like feature engineering, hyperparameter tuning, and cross-validation ensured reliable results. The findings offer actionable recommendations for improving weather forecasting and preparedness, ultimately aiding decision-making in sectors impacted by weather variability.