

April 4, 2025

GROUP 8

DATA MINING AND WAREHOUSING

Comp 361

UNIVERSITY OF ENERGY AND NATURAL RESOURCES

FINAL DOCUMENTATION REPORT

Developing a Comprehensive Data Mining and Warehousing Solution for Hurricane and Cyclone Tracking

**Group Members:**

ELIJAH ATO BAIDEN UEB3517522

SAMUEL ADZAHO UEB3510522

LOUISA OFOSUHEMAA TUTU UEB3507922

OPOKU ADDAI GYASIWAAA VENDA UEB3513222

PRINCE NATHANIEL DANQUAH UEB3528424

FATI MUSAH UEB3517424

**TABLE OF CONTENTS**

1. Abstract - 2
2. Introduction - 3
3. Weekly Breakdown - 4
4. Methodology, Results and Analysis - 29
5. Conclusion - 30
6. References - 31

# ABSTRACT

This report details the development of a data mining and warehousing solution for hurricane and cyclone tracking over the Atlantic Region. Our project aimed to analyse historical data to improve forecasting accuracy, predicting trajectories, classify storm categories, and identify patterns using some data mining techniques.

Important methodologies included supervised learning, unsupervised learning, and association rule mining. Our data was collected from NOAA, and Kaggle, then pre-processed using cleaning, normalization and feature selections. With our regression model, we achieved a low RMSE of 0.68 for wind speed prediction, while classification model accurately categorized hurricanes overcoming initial concerns about overfitting. Using clustering techniques, we identified three (3) distinct geographical and meteorological patterns, association rules revealed time-dependent relationships between storm characteristics.

A Star Schema was then designed and implemented in MySQL to facilitate efficient OLAP analysis. The project successfully demonstrated the application of data mining and warehousing to enhance hurricane understanding and prediction.

# INTRODUCTION

Hurricanes and cyclones represent some of the most destructive natural disaster globally, causing extensive damage to life, infrastructure, and property. A hurricane is a powerful tropical storm characterized by high wind speeds (over 74 mph), significant rainfall, and low atmospheric pressure, typically forming over warm ocean waters especially Atlantic and Pacific Ocean waters. Accurate prediction of their path, intensity, and patterns is crucial for effective disaster preparedness, enabling timely warnings and responses from governments, emergency services, and the public.

While meteorologists employ various technologies, Machine Learning (ML) and Data Mining offer promising ways to enhance forecasting accuracy. This project leverages data mining techniques to develop a data-driven approach for hurricane tracking and analysis. The primary goal is to analyse historical hurricane data to improve forecasting capabilities. These objectives include:

* Predicting hurricane trajectories (wind speed - mph) to enhance early warning systems
* Classifying hurricanes based on wind speed into categories (Tropical Depression TD – Category 5 hurricanes)
* Identifying patterns and trends in historical data using clustering techniques to aid risk assessment.

By applying these data mining methodologies, our project aims to contribute significantly to disaster preparedness efforts.

Weekly Breakdown

# Week 1: Machine Learning Concepts and Strategies

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to learn from data and make predictions or decisions without being explicitly programmed. ML models recognise pattens, analyse trends, and improve performance over time.

ML plays a crucial role in hurricane tracking by improving prediction accuracy and decision-making which can help the people, government and emergency responders.

**PROBLEM STATEMENT**

Our project aimed to leverage data mining techniques to develop a data-driven approach for hurricane tracking. The project sought to analyse historical hurricane data to improve forecasting accuracy. Specifically, it would:

1. Predict the trajectory of hurricanes to improve early warning systems
2. Classify hurricanes into Category TD - cat5 based on wind speed.
3. Identify patterns and trends in historical hurricane data to aid in risk assessment using clustering techniques.

By using data mining techniques, we aimed to contribute to disaster preparedness efforts and improve early warning systems.

**PROPOSED DATA MINING TECHNIQUES**

To achieve our objectives, we were to be exploring the following approaches:

1. Supervised Learning

* Regression: Predicting wind speed, pressure and overall intensity. eg Random Forest, Linear Regression etc.
* Classification: Categorizing storms into different categories using Decision Tree

1. Unsupervised Learning

* Clustering: Identifying similarities in hurricane paths and strength. Eg K-Means.

# Week 2: Data Mining Basics and Differentiation from ML

In week 2, Our goal was to explain the data mining process for our identified problem, highlighting challenges and potential applications.

**DATA MINING PROCESS FOR HURRICANE & CYCLONE PREDICTION**

The project followed a structured data mining process to extract insights from hurricane data.

Data Collection: Step 1

Historical hurricane data was to be gathered from reputable, publicly available sources:

* **NOAA HURDAT2 Dataset**: Providing core historical data on hurricane paths, wind speeds, and pressure measurements.
* **NASA Climate Data**: Offering supplementary data on atmospheric conditions influencing storm formation.
* **Kaggle Hurricane Datasets**: Serving as an additional source, particularly useful for ML applications, often derived from NOAA data.

Data Preprocessing: Step 2

In order to ensure our dataset was of high quality, we were to:

* Handle missing values using mean/mode imputation.
* Normalize numerical data (wind speed, and pressure) for better model performance.
* Convert categorical variables (hurricane categories) into numerical labels

Data Mining Techniques: Step 3

To achieve our aim, we planned to use the following mining techniques:

* Regression – Predicting wind speed.
* Classification – Categorizing hurricanes into Category TD - 5 based on intensity.
* Clustering – Identifying storm formation patterns in different regions.

Model Evaluation and Storage: Step 4

We evaluate model performance using accuracy, R square score, confusion matrix, and store the cleaned data into a data warehouse thus MySQL for future analysis.

**CHALLENGES WHICH WERE TO BE FACED IN THE PROJECT**

While implementing data mining techniques for hurricane tracking, we anticipated the following challenges:

* Data inconsistencies: Missing weather reading or inaccurate records in our dataset.
* Feature selection: Identifying which weather parameters impacted hurricane intensity.
* Computational Complexity: Handling our dataset efficiently.
* Predictive Accuracy: Ensuring that our model provided reliable forecasts.

To mitigate these challenges, we planned to use advanced data preprocessing techniques, to optimize our machine learning models using hyperparameter tuning.

# Week 3: Data Types, Visualization, and Similarity Measures

In week 3, we focused on exploring our datasets to understand key weather patterns affecting storms and ensure the data is suitable for learning models. Our aim was to provide a statistical summary and visualization of our datasets.

**DATASET COLLECTION**

We collected storm data from the Kaggle Hurricane dataset which was also taken from the National Oceanic and Atmosphere Administration (NOAA) HURDAT2 Database.

Link - <https://www.kaggle.com/datasets/thedevastator/atlantic-and-eastern-pacific-hurricane-data>

Identified Key Features

* Date - The date of the hurricane event.
* Time – The time of the recorded observation (HH:MM)
* Latitude - Geographic latitude of the hurricane.
* Longitude - Geographic longitude of the hurricane.
* Wind Speed (mph) - Maximum sustained wind speed.
* Pressure (mb) - Atmospheric pressure at the storm centre
* Category - Hurricane intensity classification.

**DATA EXPLORATION AND VISUALISATION**

We performed a statistical analysis using python (Pandas, matplotlib, seaborn frameworks) to obtain useful data from the dataset.

Note: the following was the output from our terminal mode from python

Identification of missing values: Output

**Columns in dataset**: Index (['Key', 'Name', 'Date', 'Time', 'Latitude', 'Longitude', 'Wind Speed',

'Pressure', 'Temperature', 'NE34', 'SE34', 'SW34', 'NW34', 'NE50',

'SE50', 'SW50', 'NW50', 'NE64', 'SE64', 'SW64', 'NW64', 'Category'],

dtype='object')

**Counting of Columns**

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 49689 entries, 0 to 49688**

**Data columns (total 22 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 Key 49689 non-null object**

**1 Name 49689 non-null object**

**2 Date 49689 non-null object**

**3 Time 49689 non-null object**

**4 Latitude 49689 non-null float64**

**5 Longitude 49689 non-null float64**

**6 Wind Speed 49689 non-null int64**

**7 Pressure 49562 non-null float64**

**8 Temperature 49562 non-null float64**

**9 NE34 6507 non-null float64**

**10 SE34 6507 non-null float64**

**11 SW34 6507 non-null float64**

**12 NW34 6507 non-null float64**

**13 NE50 6507 non-null float64**

**14 SE50 6507 non-null float64**

**15 SW50 6507 non-null float64**

**16 NW50 6507 non-null float64**

**17 NE64 6507 non-null float64**

**18 SE64 6507 non-null float64**

**19 SW64 6507 non-null float64**

**20 NW64 6507 non-null float64**

**21 Category 49689 non-null object**

**dtypes: float64(16), int64(1), object(5)**

**memory usage: 8.3+ MB**

**Statistical Summary Info**

**Latitude Longitude ... SW64 NW64**

**count** 49689.000000 49689.000000 ... 6507.000000 6507.000000

**mean** 27.038580 -65.622762 ... 5.190564 6.291686

**std** 10.070211 19.601482 ... 14.138406 17.173767

**min** 7.200000 -109.500000 ... 0.000000 0.000000

**25%** 19.100000 -81.000000 ... 0.000000 0.000000

**50%** 26.400000 -67.900000 ... 0.000000 0.000000

**75%** 33.100000 -52.400000 ... 0.000000 0.000000

**max** 81.000000 63.000000 ... 150.000000 300.000000

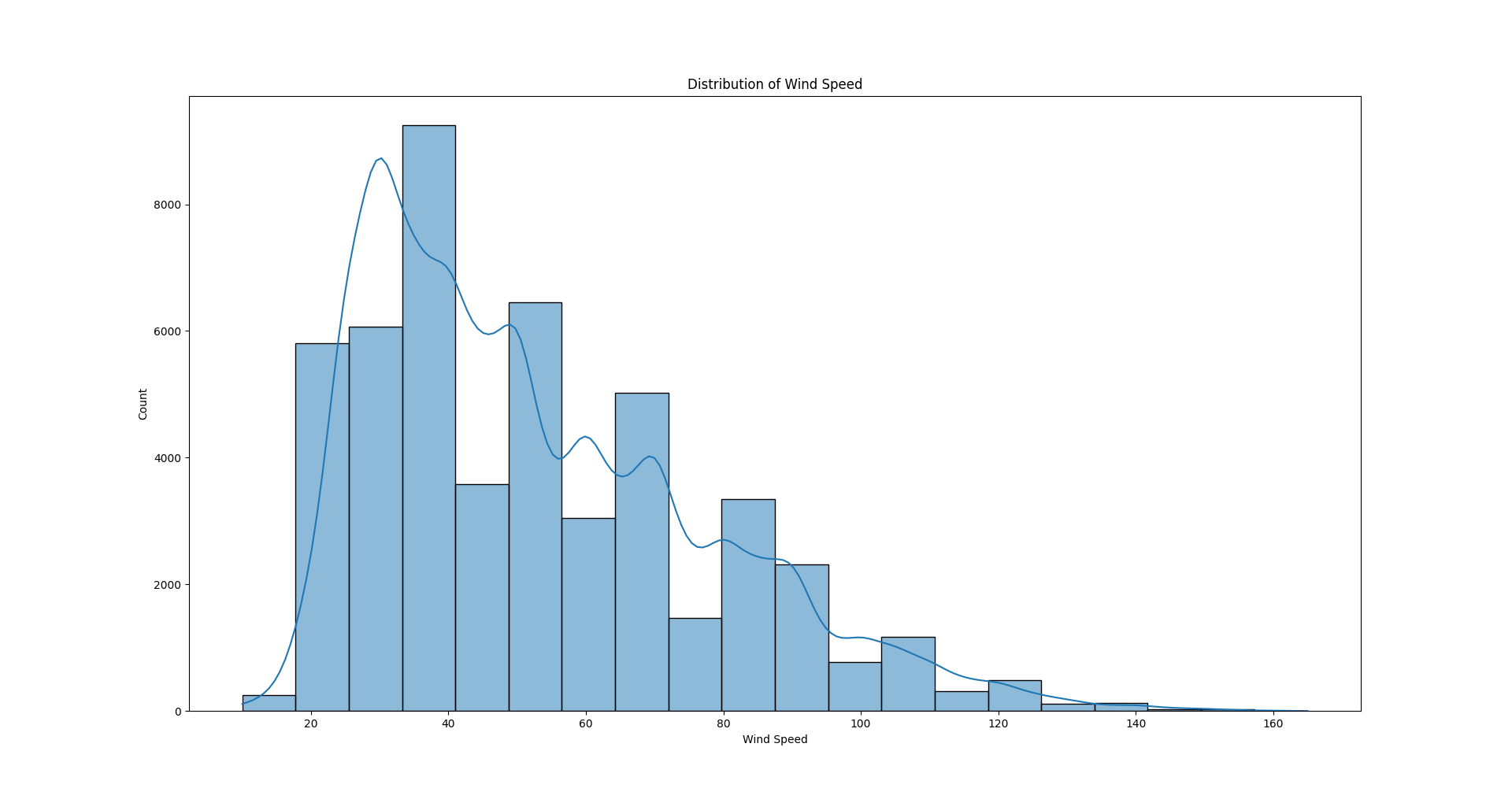
[8 rows x 17 columns]

Insight obtained was, there were some missing values in temperature and pressure features.

**DATA VISUALISATIONS**

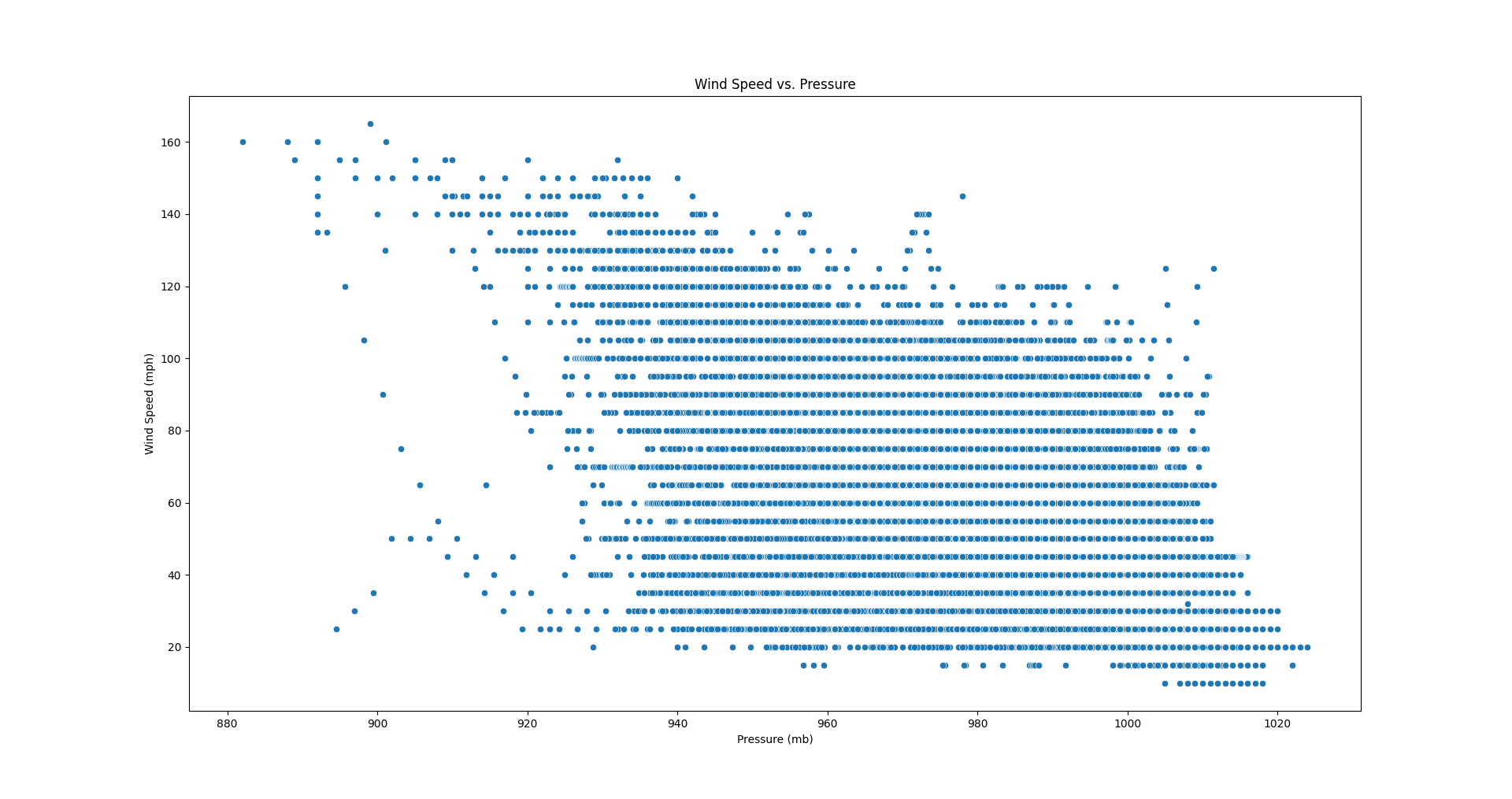
We generated key visualisations for analysis of our cyclone data.

Wind Speed Distribution



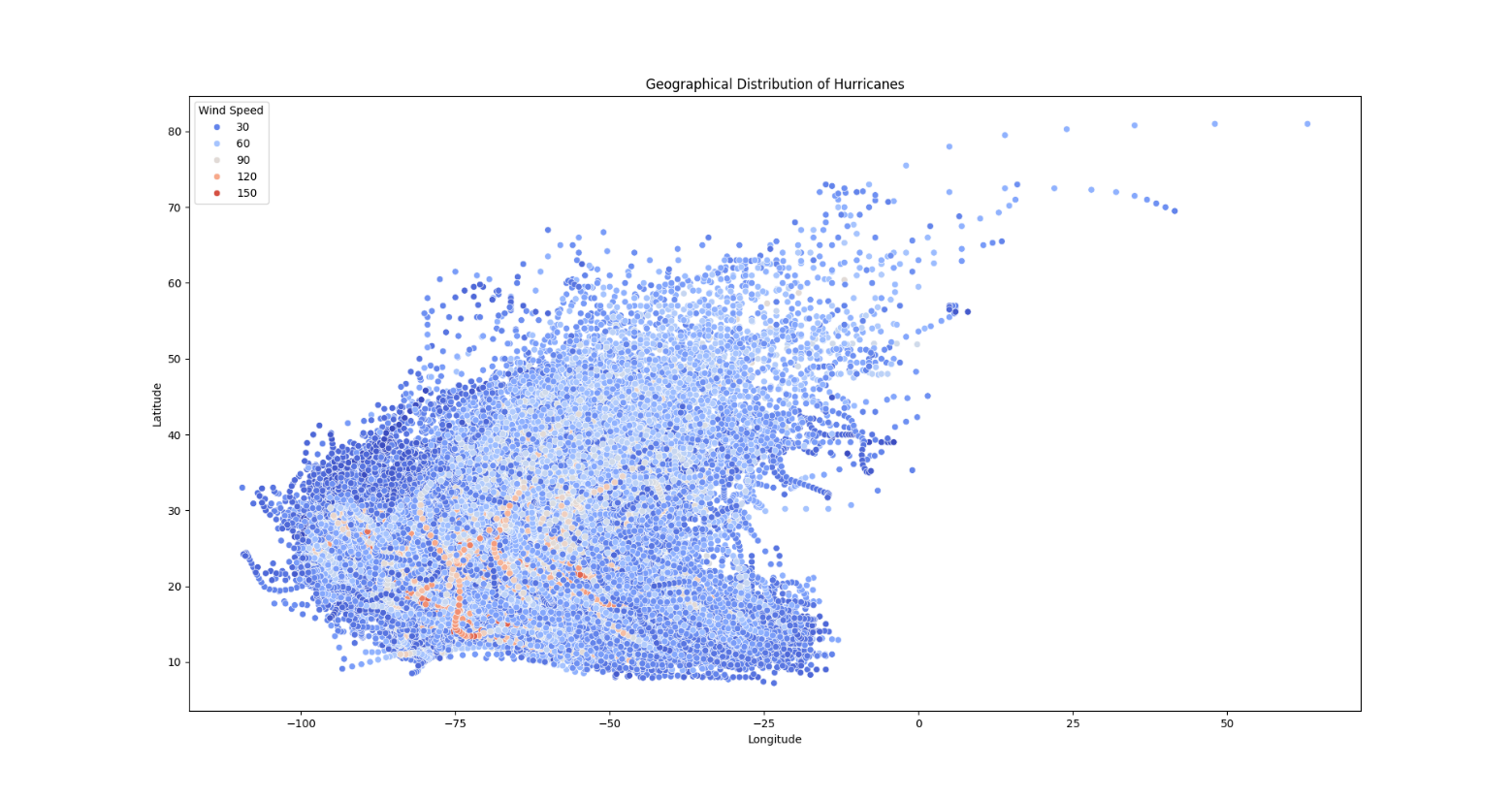
Here, the analysis was that most hurricanes have a wind speed between 30mph – 120mph with extreme ones above 140mph.

Wind Speed vs Pressure (Scatter Plot Diagram)



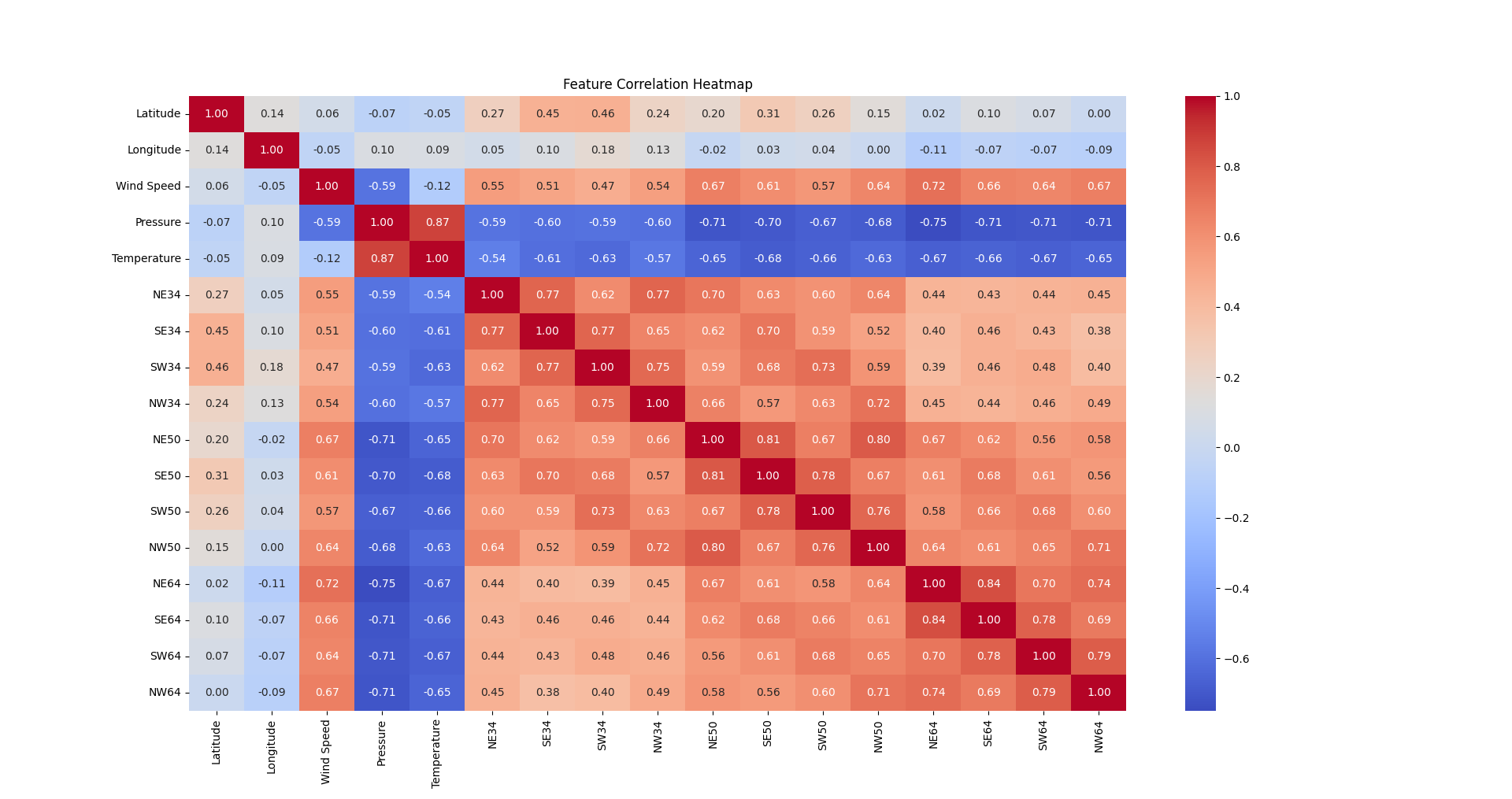
**Insight: Lower pressure correlates with higher wind speeds**, confirming pressure as a key predictor.

Geographical Location Plot (Latitude vs Longitude)



**Insight**: Hurricanes **occur mostly in specific geographic regions**, which helps in storm pattern detection.

Feature Heatmap Correlation



**Insight:**

**Wind Speed & Pressure correlation = -0.60** → Strong negative correlation.

**Latitude & Longitude have weak correlation with wind speed**, indicating location alone doesn’t determine storm strength.

**KEY TAKEAWAYS**

* **Wind speed and pressure are highly correlated**, making pressure a strong predictor.
* **Geographic location showed regional storm patterns**, which would help cluster storms.

# Week 4: Data Cleaning, Integration, Transformation and Reduction

Preprocessing was essential to prepare the data for modelling. This involved several steps performed using Weka and Python:

**DATA CLEANING**

Missing values identified during exploration in week 2, primarily the ‘Pressure’ Feature, were handled by replacing them with the median value using Weka’s ReplaceMissingValues filter.

Other key features like latitude, longitude and wind speed had complete data.

**DATA TRANSFORMATION**

Normalization was applied to numeric features with varying scales (Latitude, Longitude, Pressure) to bring them into a common range 0 – 1, primarily using Weka’s Normalize filter.

Wind speed, being a target variable for regression, was intentionally left unnormalized as tree-based models used later do not strictly require target normalization.

**Normalization Method**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Before Normalization** | **After Normalization** | **Tool Used** |
| **Latitude** | -128.34 to 50.12 | 0 to 1 | Weka: Normalize filter |
| **Longitude** | -180.00 to 180.00 | 0 to 1 | Weka: Normalize filter |
| **Pressure (mb)** | 920 to 1020 | 0 to 1 | Weka: Normalize filter |
| **Wind Speed (mph)** | **Not normalized (10 to 165)** | Left unnormalized for regression | - |

**DATA REDUCTION**

Feature selection techniques were employed to improve model efficiency and performance by removing irrelevant or redundant attributes. Both Weka (using CfsSubsetEval with BestFirst search) and Python (using SelectKBest with f\_regression) were utilized to identify and retain the most relevant features for prediction.

**FINAL PREPROCESSED DATASET**

The final pre-processed dataset, free of missing values and with normalized and selected features, was saved in ARFF (for Weka) and CSV (for Python) formats for subsequent modelling stages.

# Week 5: Classification Techniques and Decision Trees

In week 5, we focused on implementing and evaluation supervised learning models (Regression and Classification) using Python.

Several data mining techniques were applied to address the project objectives.

**REGRESSION: PREDICTING WIND SPEED**

A Random Forest Regressor model was trained using Python to predict hurricane wind speed (mph) based on Latitude, Longitude, Temperature, and Pressure. Hyperparameter tuning was performed using Grid Search CV for optimization.

Evaluation

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Best Parameters | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 200} |
| Mean Squared Error (MSE) | 0.47 |
| Root Mean Squared Error (RMSE) | 0.68 |

The model demonstrated strong performance, achieving a Root Mean Squared Error (RMSE) of 0.68 mph. This low error rate indicated that the predictions were very close to the actual wind speed values, suggesting good generalisation ability.

**CLASSIFICATION: PREDICTING HURRICANE CATEGORY**

A Random Forest Classifier was developed to predict the hurricane category (ranging from Tropical Depression – Category5) using Latitude, Longitude, Temperature, Pressure, and Wind Speed as features. Grid Search CV was again used for hyperparameter tuning.

Evaluation

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Best Hyperparameters | { 'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 200} |
| Accuracy | 98% |
| Precision, Recall, F1-Score | 0.98 |

Initial high accuracy of 98% prompted overfitting checks. Tests involving reduced training data (accuracy remained high) and randomised labels (accuracy dropped significantly) confirmed the model was learning real patterns and not overfitting. Feature importance analysis revealed that wind speed heavily dominated predictions (85% importance). To address this and improve robustness, adjustments were made, including further hyperparameter tuning (adjusting max\_depth, min\_samples\_leaf) and dataset balancing using SMOTE.

| **Feature** | **Importance Score** |
| --- | --- |
| Wind Speed | **85%** |
| Pressure | **10%** |
| Temperature | **3%** |
| Latitude, Longitude | **Negligible** |

**TESTING**

The refined model was tested with real-world sample parameters and successfully predicted the correct category

**Sample Prediction**

Enter real-world hurricane parameters:

Latitude: 15

Longitude: -45

Pressure: 965 mb

Temperature: 26°C

**Hurricane Prediction Results**

* Predicted Wind Speed: 78.45 mph
* Predicted Hurricane Category: Category 2

# Week 6: Clustering Concepts and Algorithms

In this week, we were to apply unsupervised learning (clustering) to identify patterns and relationships to better understand cyclone behaviour, geographical distributions and intensity classifications.

K-Means clustering was employed using Python (Scikit-learn) to identify natural groupings within the hurricane data based on Latitude, Longitude, Pressure, Temperature, and Wind Speed.

Method

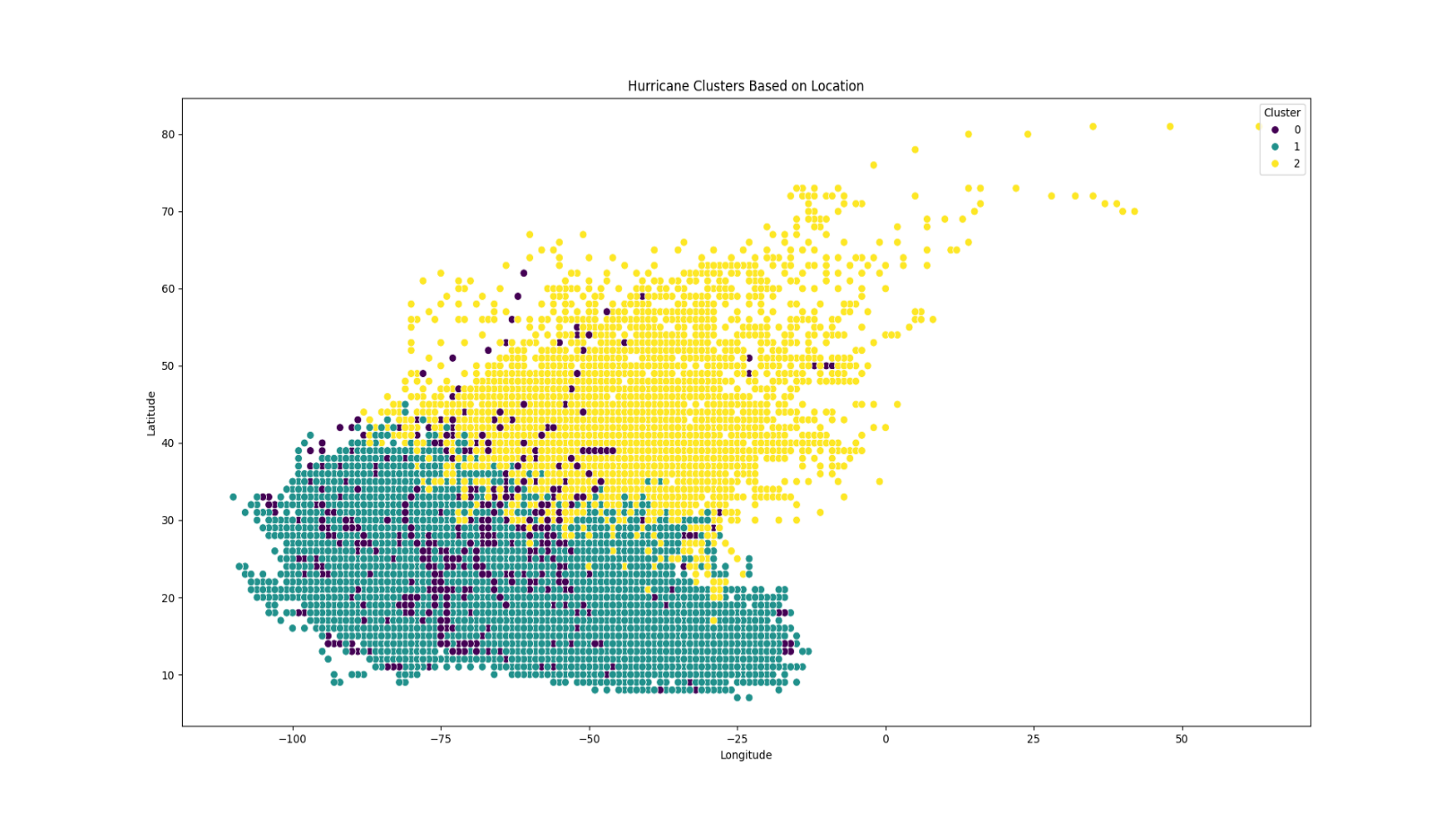
The Elbow Method was used to determine the optimal number of clusters by analysing the Within-Cluster Sum of Squares (WCSS), resulting in the selection of 3 clusters.

**INSIGHTS AND ANALYSIS**

The analysis successfully grouped hurricanes into three distinct clusters, primarily based on geographic location and wind speed. This suggests geographically distinct regions where hurricanes form and behave differently. Visualizations (2D and 3D scatter plots) highlighted these geographic patterns and relationships between meteorological factors (e.g., lower pressure correlating with higher wind speeds, higher temperatures associated with stronger hurricanes).

**Geographic Distribution of Hurricanes - 2D**

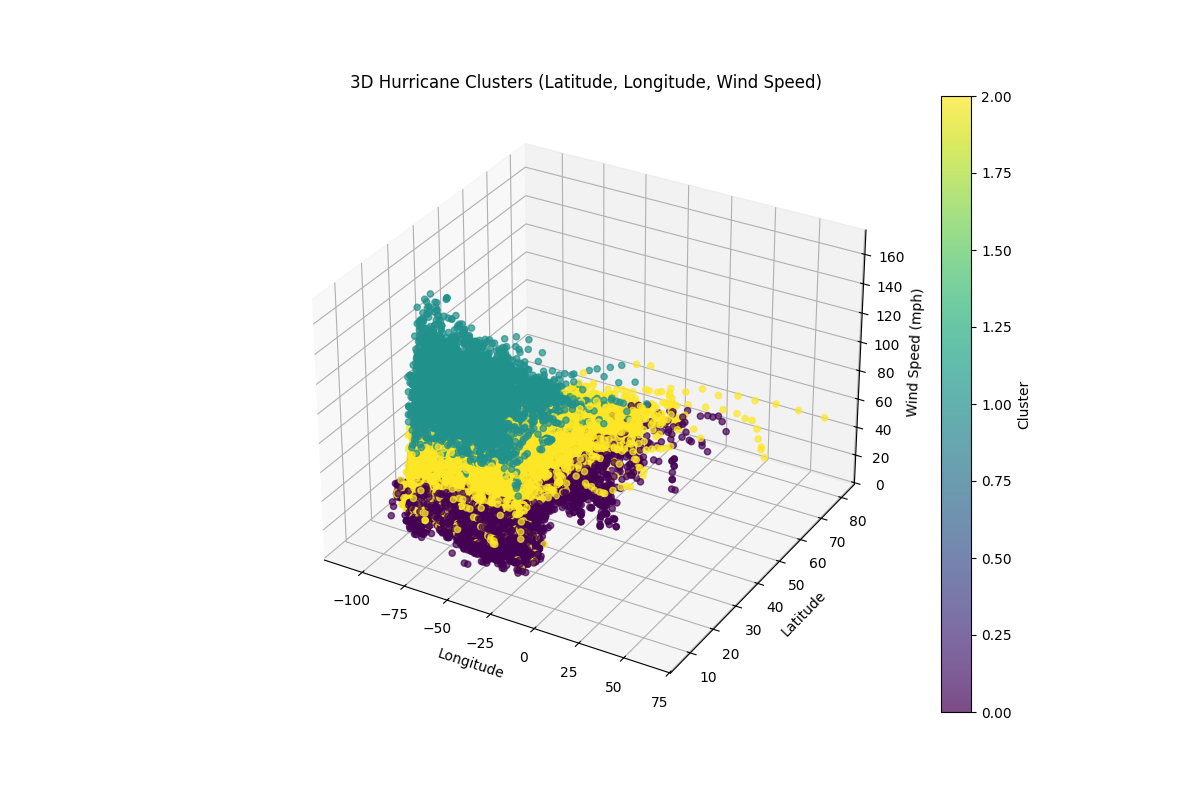
* The clusters indicated that hurricanes **followed specific paths**, most likely driven by **oceanic** and **atmospheric** conditions.
* **Western Hemisphere Dominance:** Most hurricanes occur **in the Atlantic and Pacific regions**, as seen from the longitude range.
* **Latitude Differences**: Some hurricanes occur at **lower latitudes (closer to the equator)**, while others move toward **higher latitudes**.

****

**Insight**: This visualization helps identify patterns in hurricane locations, showing where hurricanes tend to form and how they are spatially distributed.

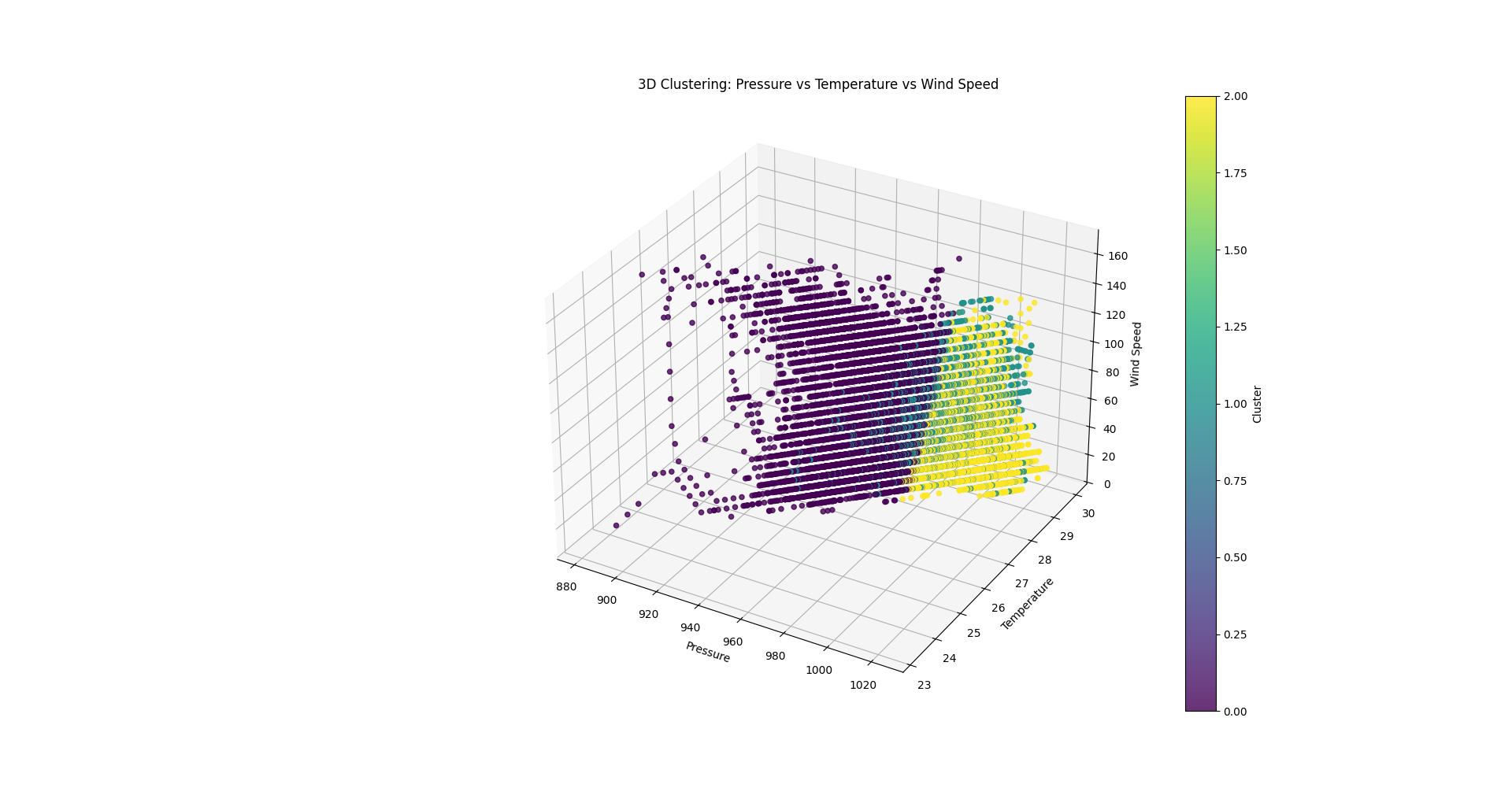
**Geographic Clusters (Latitude, Longitude, Wind Speed) – 3D**

* The **3D scatter plot** of **Latitude, Longitude, and Wind Speed** shows distinct clusters of hurricanes forming in different regions.
* **Key Observations:** 
  + **Cluster 1 (Low Wind Speed)** – Represents tropical storms and weak hurricanes. These are likely to be forming or dissipating. This represents hurricanes occurring primarily in **low-latitude tropical regions**.
  + **Cluster 2 (Moderate Wind Speed)** – Includes storms that are intensifying or sustaining strength as they move across regions. Some hurricanes are more **scattered across different longitudes**, showing variation in formation zones.
  + **Cluster 3 (High Wind Speed)** – Represents powerful hurricanes with **extreme wind speeds**, typically associated with major hurricanes.

****

**Meteorological Clusters (Pressure, Temperature, Wind Speed)**

* Another **3D plot of Pressure, Temperature, and Wind Speed** reveals:
  + **Low-pressure systems** are correlated with **higher wind speeds**, supporting the meteorological principle that hurricanes intensify as pressure decreases.
  + **Higher temperatures** are also seen in stronger hurricanes, reinforcing the role of warm ocean waters in storm development.
  + The clustering pattern suggests that storms with similar temperature-pressure conditions form in specific regions.



Applications

These clustering insights can aid in improving predive models, enhancing regional disaster preparedness, and analysing long-term climate change impacts on cyclone behaviour.

# Week 7: Association Rule Mining and Apriori Algorithm

At this stage, our focus on discovering relationships between hurricane characteristics using association rule mining with Python.

Numerical features (Wind Speed, Pressure, Temperature) were discretised into categories (Low, Medium and High), and Time was categorised (Morning, Afternoon, Evening, Night). The data was transformed into a transactional format. Apriori was applier with a minimum support of 0.1 and minimum confidence of 0.6 to find frequent and strong rules. Lift metric was also used to measure the rule significance.

**RESULTS AND INTERPRETATION**

**Top Association Rules found within our dataset**

The table below summarizes the most relevant association rules discovered:

| **Antecedents (If Condition)** | **Consequents (Then Condition)** | **Support** | **Confidence** | **Lift** | **Interpretation** |
| --- | --- | --- | --- | --- | --- |
| **Morning** | **Medium Wind Speed, Low Pressure** | **17%** | **67%** | **1.007** | In the **morning**, there is a **67% chance** that the wind speed is medium and the pressure is low. |
| **Morning, Warm Temperature** | **Medium Wind Speed, Low Pressure** | **17%** | **67%** | **1.007** | If it is **morning and warm**, there is a **67% chance** of medium wind speed and low pressure. |
| **Morning** | **Low Wind Speed** | **21%** | **83%** | **1.006** | If it is **morning**, there is an **83% chance** the wind speed is low. |
| **Morning, Warm Temperature** | **Low Wind Speed** | **21%** | **83%** | **1.006** | If it is **morning and warm**, there is an **83% chance** of low wind speed. |
| **Night** | **Warm Temperature, Medium Wind Speed** | **40%** | **83%** | **1.001** | At **night**, there is an **83% chance** of warm temperature and medium wind speed. |
| **Afternoon** | **Low Wind Speed** | **20%** | **83%** | **1.000** | In the **afternoon**, there is an **83% chance** the wind speed is low. |

Findings

Several interesting rules were discovered, revealing time-dependent patterns. For instance, event ‘Morning’ were associated with 'Medium Wind Speed' and 'Low Pressure' (67% confidence). 'Night' events showed a link to 'Warm Temperature' and 'Medium Wind Speed' (83% confidence). These rules suggest hurricane characteristics can vary significantly depending on the time of day.

# Week 8: Data Warehouse and OLAP Systems

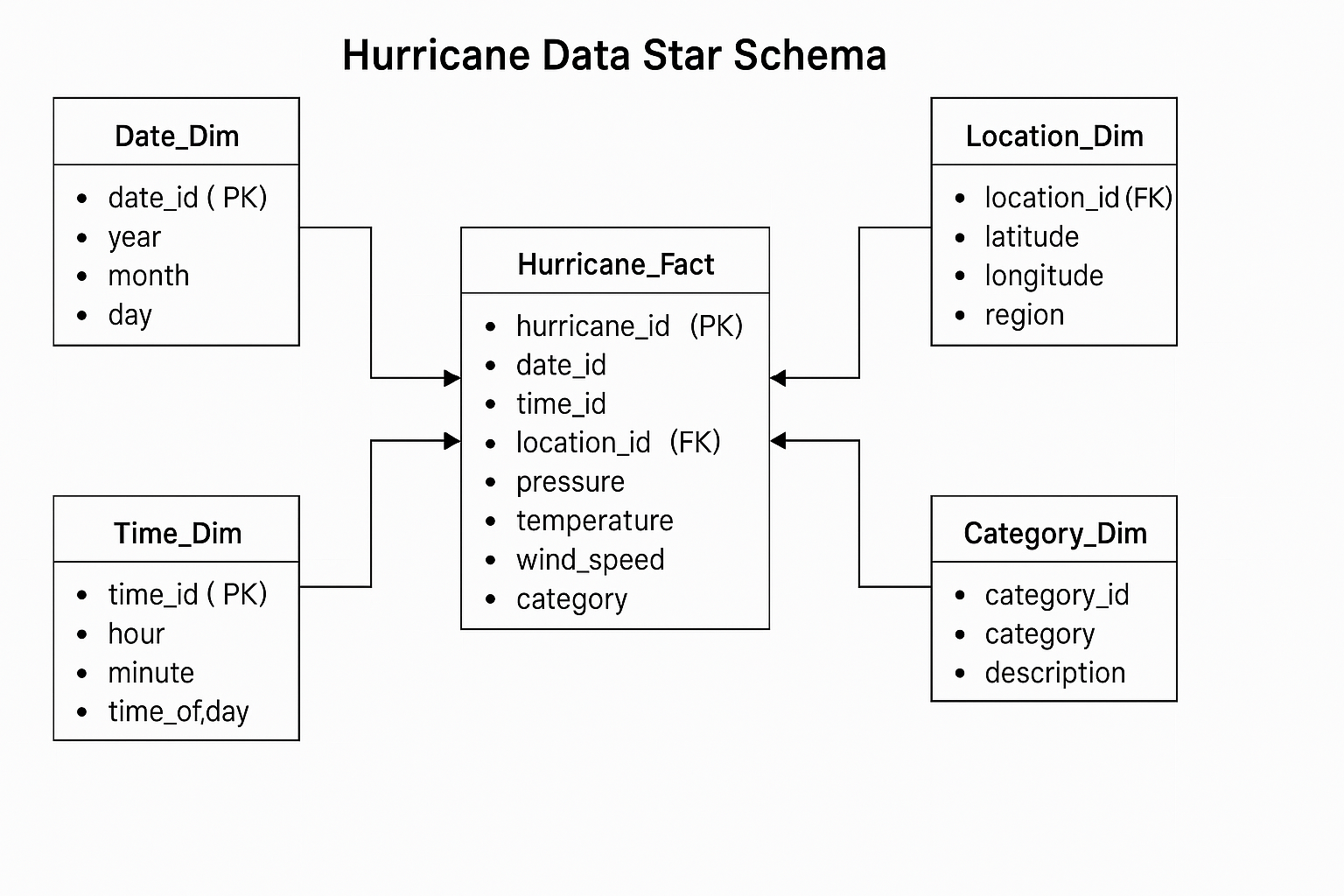
Designing the architecture for our data warehouse to store and analyse cyclone data to facilitate efficient storage, querying, and analysis of the historical hurricane data was our focus here.

**SCHEMA DESIGN**

A Star Schema architecture was chosen for the data warehouse due to its simplicity, query performance, and suitability for OLAP (Online Analytical Processing) operations.

Structure

The schema consists of a central Hurricane\_Fact table containing quantitative measures (pressure, temperature, wind\_speed) and foreign keys linking to several dimension tables.



**EXPLANATION OF SCHEMA COMPONENTS**

**Fact Table: Hurricane\_Fact**

The **central table** in the schema stores key hurricane parameters. Each record represents a unique hurricane event.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| hurricane\_id | INT (PK) | Unique ID for each hurricane event |
| date\_id | INT (FK) | Links to Date\_Dim |
| time\_id | VARCHAR(5) (FK) | Links to Time\_Dim |
| location\_id | INT (FK) | Links to Location\_Dim |
| pressure | INT | Atmospheric pressure in mb |
| temperature | FLOAT | Temperature in °C |
| wind\_speed | FLOAT | Wind speed in mph |
| category\_id | INT (FK) | Links to Category\_Dim |

**Dimension Table: Date\_Dim**

Stores the **date information** of hurricanes.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| date\_id | INT (PK) | Unique date identifier |
| year | INT | Year of the event |
| month | INT | Month of the event |
| day | INT | Day of the event |

**Dimension Table: Time\_Dim**

Captures **time-specific details** for hurricanes.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| time\_id | VARCHAR(5) (PK) | Unique time identifier (HH:MM) |
| hour | INT | Hour of the event |
| minute | INT | Minute of the event |
| time\_of\_day | VARCHAR(20) | Morning, Afternoon, Night |

**Dimension Table: Location\_Dim**

Holds **geographical information** about hurricane locations.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| location\_id | INT (PK) | Unique location identifier |
| latitude | FLOAT | Latitude coordinate |
| longitude | FLOAT | Longitude coordinate |
| region | VARCHAR(50) | Geographic region |

**Dimension Table: Category\_Dim**

Defines hurricane **categories based on intensity**.

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| category\_id | INT (PK) | Unique category identifier |
| category | VARCHAR(20) | Hurricane category name |
| description | VARCHAR(100) | Description of severity |

**JUSTIFICATION**

This design allows for fast joins, easy scalability (adding more dimensions), and is optimised for analytical tasks like slicing, dicing, and roll-up analysis.

Thus, its’;

* **Fast Query Performance**
* **Easy to Scale**
* **Optimized for OLAP**

It also supports aggregation and enhances reporting.

# Week 9: Implementing Schemas in Real-World Applications

This was the final step where we implemented the design star schema in MySQL. This implementation enabled for efficient querying and analysis of historical hurricane trends, wind speed variations, and atmospheric conditions.

This involved:

* Creating tables for the **fact and dimension tables**.
* Establishing relationships between tables using **primary keys (PK) and foreign keys (FK)**.
* Populating the tables with sample data.
* Running SQL queries to test the data flow.

**PROCESS**

SQL scripts were written to create the fact and dimension tables, defining primary and foreign key relationships to enforce data integrity. The tables were then populated with sample data.

SQL scripts

-- Fact Table: Hurricane\_Fact  
CREATE TABLE Hurricane\_Fact (  
 hurricane\_id INT PRIMARY KEY AUTO\_INCREMENT,  
 date\_id INT,  
 time\_id VARCHAR(5),  
 location\_id INT,  
 pressure INT,  
 temperature FLOAT,  
 wind\_speed FLOAT,  
 category\_id INT,  
 FOREIGN KEY (date\_id) REFERENCES Date\_Dim(date\_id),  
 FOREIGN KEY (time\_id) REFERENCES Time\_Dim(time\_id),  
 FOREIGN KEY (location\_id) REFERENCES Location\_Dim(location\_id),  
 FOREIGN KEY (category\_id) REFERENCES Category\_Dim(category\_id)  
);

-- Date Dimension Table  
CREATE TABLE Date\_Dim (  
 date\_id INT PRIMARY KEY AUTO\_INCREMENT,  
 year INT,  
 month INT,  
 day INT  
);

-- Time Dimension Table  
CREATE TABLE Time\_Dim (  
 time\_id VARCHAR(5) PRIMARY KEY, -- Format HH:MM  
 hour INT,  
 minute INT,  
 time\_of\_day VARCHAR(20) -- Morning, Afternoon, Night  
);

-- Location Dimension Table  
CREATE TABLE Location\_Dim (  
 location\_id INT PRIMARY KEY AUTO\_INCREMENT,  
 latitude FLOAT,  
 longitude FLOAT,  
 region VARCHAR(50)  
);

-- Category Dimension Table  
CREATE TABLE Category\_Dim (  
 category\_id INT PRIMARY KEY AUTO\_INCREMENT,  
 category VARCHAR(20),  
 description VARCHAR(100)  
);

Sample data populating

-- Insert sample dates  
INSERT INTO Date\_Dim (year, month, day) VALUES  
(2025, 3, 30),  
(2025, 4, 1);

-- Insert sample times  
INSERT INTO Time\_Dim (time\_id, hour, minute, time\_of\_day) VALUES  
('08:00', 8, 0, 'Morning'),  
('15:30', 15, 30, 'Afternoon');

-- Insert sample locations  
INSERT INTO Location\_Dim (latitude, longitude, region) VALUES  
(25.6, -80.2, 'Florida'),  
(13.4, -60.5, 'Caribbean');

-- Insert sample hurricane categories  
INSERT INTO Category\_Dim (category, description) VALUES  
('Tropical Storm', 'Slightly Weak Hurricane'),  
('Category 1', 'Weak Hurricane');

-- Insert sample hurricane events  
INSERT INTO Hurricane\_Fact (date\_id, time\_id, location\_id, pressure, temperature, wind\_speed, category\_id) VALUES  
(20250330, '08:00', 1, 980, 28.5, 65, 1),  
(20250401, '15:30', 2, 995, 27.0, 50, 2);

**VERIFICATION**

SQL queries were executed to test the data flow and ensure the relationships were correctly established. Example queries joined the fact table with dimension tables to retrieve comprehensive hurricane event details (e.g., retrieving events by region, time, or category). The implementation confirmed the schema's ability to support efficient querying and OLAP operations like roll-up and drill-down analysis.

**Query: Retrieved Hurricane Events with Details**

SELECT H.hurricane\_id, D.year, D.month, D.day, T.time\_of\_day, L.region, H.pressure, H.wind\_speed, C.category  
FROM Hurricane\_Fact H  
JOIN Date\_Dim D ON H.date\_id = D.date\_id  
JOIN Time\_Dim T ON H.time\_id = T.time\_id  
JOIN Location\_Dim L ON H.location\_id = L.location\_id  
JOIN Category\_Dim C ON H.category\_id = C.category\_id;

**Expected Output:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **hurricane\_id** | **year** | **month** | **day** | **time\_of\_day** | **region** | **pressure** | **wind\_speed** | **category** |
| 1 | 2025 | 3 | 30 | Morning | Florida | 980 | 65 | Tropical Storm |
| 2 | 2025 | 4 | 1 | Afternoon | Caribbean | 995 | 50 | Category 1 |

**OBSERVATION**

The schema allowed for quick querying to find hurricanes by region, time, or category. It also supports OLAP operations like role-up (grouping hurricanes by year) or drill-down (analysing individual hurricanes).

# METHODOLOGY, RESULTS AND ANALYSIS

This project employed data mining and warehousing to analyze and predict hurricane characteristics;

* **Predictive Modeling:** Using supervised learning, a Random Forest Regressor accurately predicted hurricane wind speeds (RMSE: 0.68 mph) based on location, temperature, and pressure. A Random Forest Classifier identified hurricane intensity levels, determining wind speed as the most influential predictor (85% importance); model robustness was improved via Grid Search CV and SMOTE.
* **Pattern Identification:** Unsupervised K-Means clustering grouped hurricanes into three distinct meteorological types, revealing that lower pressure and higher temperatures often correlate with stronger regional hurricanes. Association rule mining (Apriori) uncovered time-dependent patterns, associating morning events with medium wind/low pressure (67% confidence) and night events with warm temperatures/medium wind (83% confidence).
* **Data Management:** A Star Schema data warehouse implemented in MySQL provided structured storage, facilitating efficient OLAP queries for hurricane data analysis.

Collectively, these methods delivered valuable insights into hurricane behavior, enhancing tracking and forecasting capabilities.

# CONCLUSION

Our project successfully applied a range of data mining and warehousing techniques to analyse historical hurricane data. Beginning with data collection and thorough preprocessing, we developed predictive models for wind speed (regression) and hurricane category (classification), achieving promising results after refinement and testing. Clustering analysis revealed distinct geographical and meteorological patterns in hurricane behaviour, while association rule mining uncovered time-dependent relationships between weather conditions.

Finally, the design and implementation of a Star Schema data warehouse in MySQL provide a robust platform for ongoing storage, querying, and analysis of hurricane data. The combined insights and tools developed contribute to a better understanding of hurricanes and can support improved forecasting and disaster preparedness efforts.

**Future Work could include:**

* Integrating real-time data streams for dynamic prediction updates.
* Exploring more advanced algorithms (e.g., Deep Learning for trajectory prediction).
* Expanding the feature set with additional atmospheric or oceanic variables.
* Developing interactive visualization dashboards connected to the data warehouse.
* Refining the data warehouse schema further based on evolving analytical needs.

# REFERENCES

* **Data Sources**:
  + National Oceanic and Atmospheric Administration (NOAA) HURDAT2 Dataset.
  + NASA Climate Data.
  + Kaggle Hurricane Datasets (often derived from NOAA data).
* **Tools & Libraries**:
  + Python (Programming Language)
  + Pandas (Data manipulation and analysis)
  + Scikit-learn (Machine learning: Regression, Classification, Clustering)
  + Matplotlib & Seaborn (Data visualization)
  + mlxtend (Association rule mining)
  + Weka (Data preprocessing and analysis software)
  + MySQL (Database management system for data warehouse implementation)
* **Hurricane / Cyclone Categories:**

The first two are generally considered as precursors to hurricanes.

* + **Tropical Depression:**
    - Sustained Winds: Up to 38 mph (Up to 62 km/h; Up to 33 knots)
  + **Tropical Storm:**
    - Sustained Winds: 39 - 73 mph (63 - 118 km/h; 34 - 63 knots)

***(Note: Storms are typically named when they reach this intensity)***

* + **Category 1 Hurricane:**
    - Sustained Winds: 74 - 95 mph (119 - 153 km/h; 64 - 82 knots)
  + **Category 2 Hurricane:**
    - Sustained Winds: 96 - 110 mph (154 - 177 km/h; 83 - 95 knots)
  + **Category 3 Hurricane (Major Hurricane):**
    - * Sustained Winds: 111 - 129 mph (178 - 208 km/h; 96 - 112 knots)
  + **Category 4 Hurricane (Major Hurricane):**
    - Sustained Winds: 130 - 156 mph (209 - 251 km/h; 113 - 136 knots)
  + **Category 5 Hurricane (Major Hurricane):**
    - Sustained Winds: 157 mph or higher (252 km/h or higher; 137 knots or higher)