Date	1 oct 2025
Team ID	LTVIP2025TMIDS40838
Project Name	Weather-Based Prediction of Wind Turbine Energy Output: A Next-Generation Approach to Renewable Energy
Maximum Marks	3 Marks

Phase 2: Data Collection and Preprocessing (Enhanced Version)

---

1. Data Collection

Sources:

- Weather APIs (e.g., OpenWeatherMap, Weather Underground) for real-time data
- Historical wind farm datasets from government or private renewable energy portals
- Public datasets (e.g., UCI Machine Learning Repository) for weather and energy generation

Parameters Collected:

- Wind speed (m/s)
- Wind direction (degrees)
- Air temperature (°C)
- Air pressure (hPa)

- Humidity (%)
- Time and date
- Turbine energy output (kWh or MW)

Data Format & Storage:

- CSV and JSON for structured storage
- Relational database (MySQL / SQLite) for large datasets
- Version-controlled datasets for reproducibility

2. Data Preprocessing

2.1 Handling Missing Values:

- Interpolation for continuous data (e.g., wind speed)
- Deletion of rows with excessive missing fields
- Imputation using mean/median for numeric columns

2.2 Removing Outliers:

- Use Interquartile Range (IQR) or Z-score methods
- Identify anomalous wind speeds or energy outputs that are physically impossible

2.3 Normalization and Standardization:

- Scaling features to a common range (0–1) for better model convergence
- Standardization to ensure zero mean and unit variance

2.4 Feature Engineering:

- Wind power coefficient: Calculated using wind speed and turbine parameters
- Rolling averages for wind speed to smooth fluctuations
- Extract time-based features: hour, day, month for seasonal patterns
- Combining correlated features to reduce dimensionality

2.5 Dataset Splitting:

- Training set (70%)
- Validation set (15%)
- Test set (15%)
- Ensures unbiased model evaluation

3. Data Quality Analysis

- **Statistical Summary:** Mean, median, standard deviation for each feature
- **Correlation Analysis:** Identify relationships between weather parameters and turbine output
- **Missing Data Report:** Percentage of missing values per feature
- **Visualization:** Histograms, boxplots, and scatterplots for exploratory analysis

4. Tools and Technologies Used

- Python libraries: **Pandas**, **NumPy**, **Matplotlib**, **Seaborn**, **Scikit-learn**
- Jupyter Notebook for preprocessing and analysis
- APIs for data fetching and real-time data collection

5. Expected Outcome

- Clean, structured, and feature-engineered dataset
- Ready for model training and evaluation
- Reduced noise and outliers for better prediction accuracy
- Insights from correlation and visualization to guide feature selection