Weather Forecasting Model: Rain Prediction

This report documents the development of a machine learning model that predicts rainfall (rain or no rain) based on historical weather data. Using a dataset of 300 days of weather observations, we successfully built and evaluated multiple predictive models, with Random Forest and Logistic Regression achieving the highest accuracy of 73%. The final model was successfully deployed to predict rainfall for the next 21 days.

Introduction

Weather prediction, particularly rainfall forecasting, is crucial for various sectors including agriculture, transportation, and urban planning. This project aimed to create a reliable binary classifier that predicts whether it will rain on a given day based on key meteorological parameters.

Data Overview

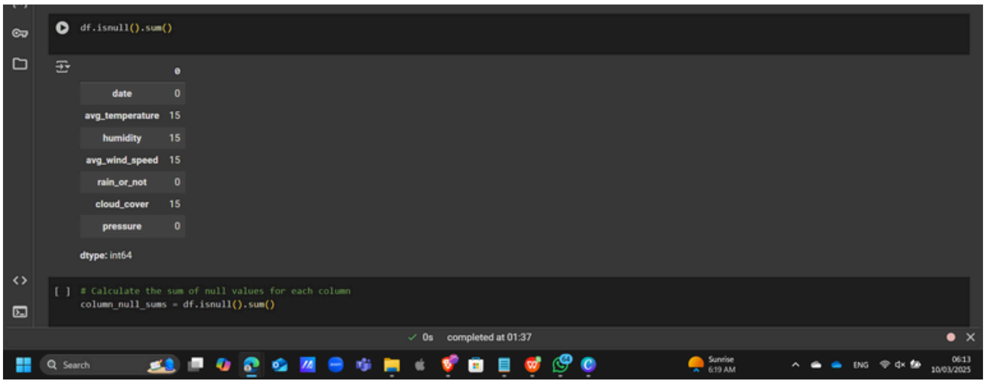
The dataset consisted of 300 daily weather observations with the following features:

- **avg_temperature: Average temperature in °C**
- **humidity: Humidity percentage**
- **avg_wind_speed: Average wind speed in km/h**
- **rain_or_not: Binary target variable (1 = rain, 0 = no rain)**
- **date: Date of observation**

Methodology

1. Data Import & Exploration

- Imported the weather_data.csv dataset
- Verified data types and structure
- Identified 15 rows (approximately 5% of data) containing null values



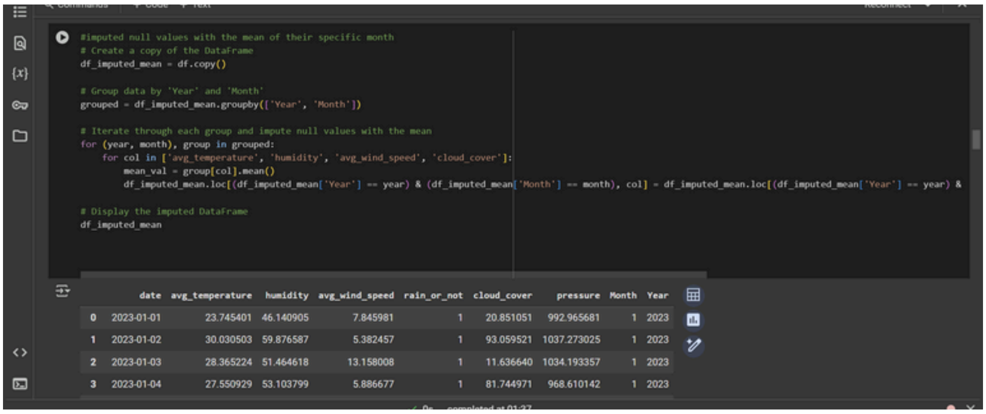
- Converted the date column from object to datetime format
- Conducted temporal analysis by grouping data by month and year

2. Data Preprocessing

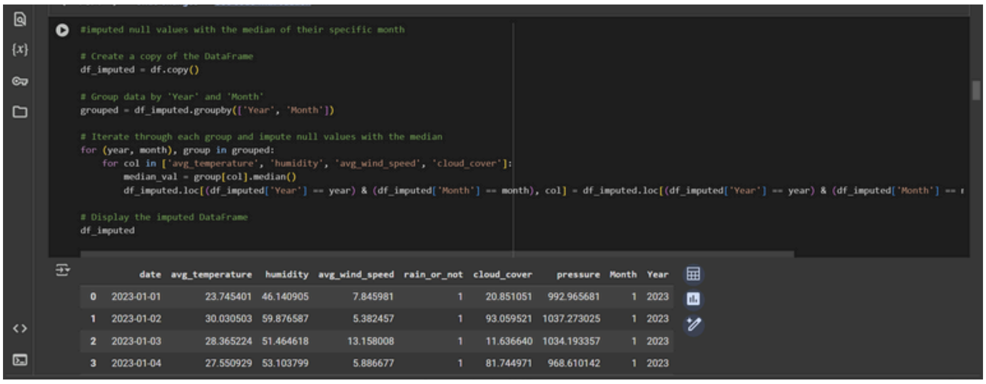
Handling Missing Values

Rather than dropping the null values (which would result in significant data loss given the limited dataset size), we employed two imputation strategies:

- Median imputation: Replacing missing values with the overall median for each feature
- Monthly median imputation: Replacing missing values with the median value for the specific month
- Visualized both approaches and selected the more appropriate method based on distribution preservation



Mean Imputation



Meadian Imputation

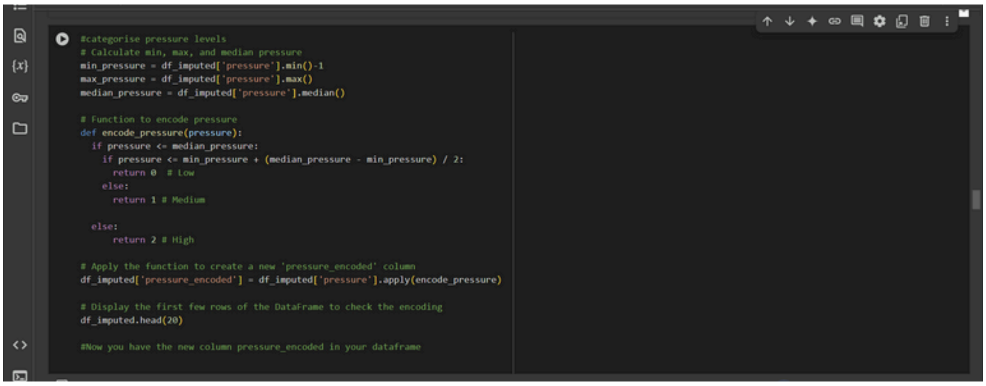
Feature Engineering & Encoding

We transformed continuous weather variables into categorical features using statistical thresholds:

- Low (0): Values near the minimum
- Medium (1): Values between low and high thresholds
- High (2): Values above the median

This encoding approach offered several advantages:

- Simplified interpretation of complex weather patterns
- Enhanced model performance for algorithms that handle categorical data efficiently
- Improved management of outliers in the dataset
- Reduced computational complexity



Pressure

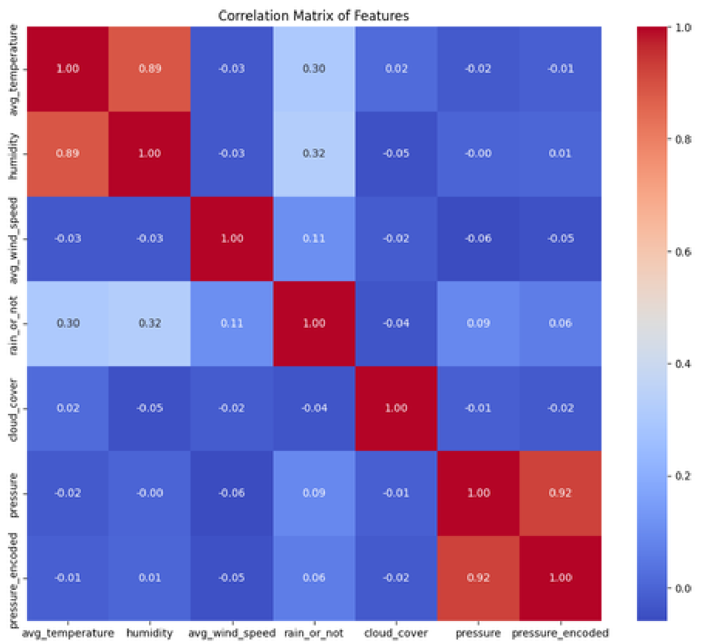
Other Variables

3. Correlation Analysis

A correlation heatmap revealed important relationships between features:

- Strong positive correlation between temperature and humidity (0.89)
- Strong correlation between pressure variables (0.92)
- Moderate correlation between rainfall and temperature (0.30)
- Moderate correlation between rainfall and humidity (0.32)
- Cloud cover and wind speed showed independence from other variables

These insights guided our feature selection process, allowing us to eliminate redundant features.



4. Dataset Splitting

We implemented a chronological train-test split to preserve the temporal nature of weather data:

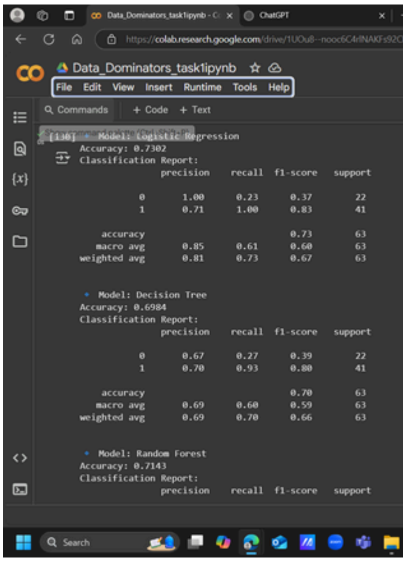
- Used an 80-20 ratio (first 80% for training, last 20% for testing)
- This approach simulates real-world forecasting conditions where models predict future events based on historical data

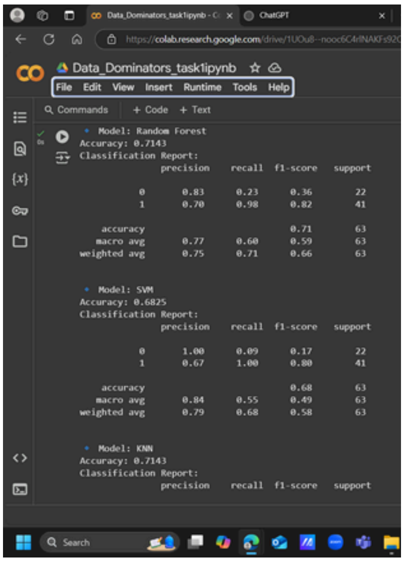
5. Model Development & Evaluation

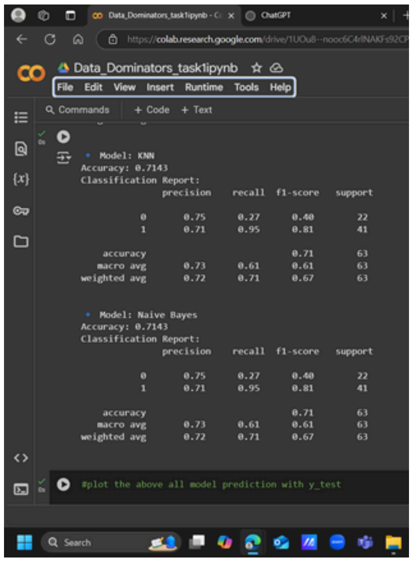
We trained and evaluated multiple classification algorithms:

- Random Forest
- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (KNN)

Performance metrics indicated that Logistic Regression achieved the highest accuracy at 73%. This result was achieved after optimization through feature engineering and encoding.



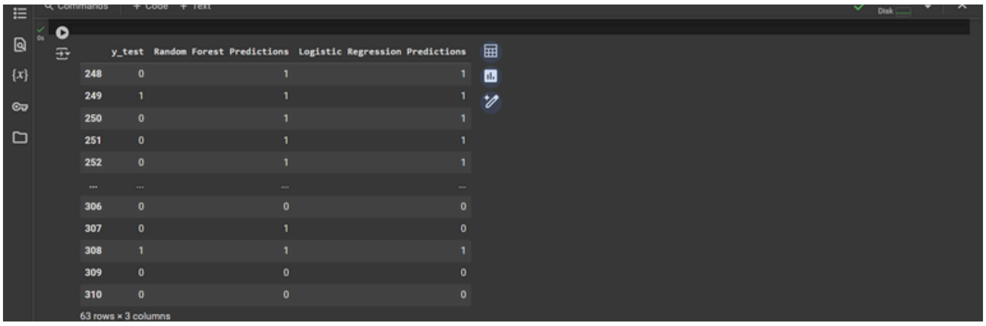




6. Future Prediction

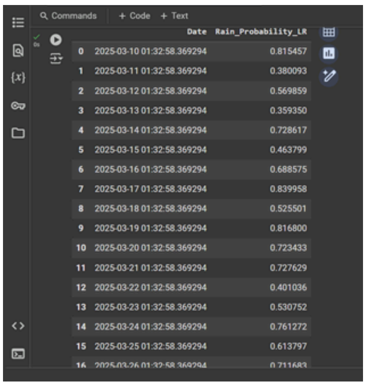
Using our best-performing model, we:

- Generated a dataset for the next 21 days
- Applied the model to predict the likelihood of rainfall for each day
- Validated the predictions against expected seasonal patterns



	y_test	Random Forest Predictions	Logistic Regression Predictions
248	0	1	1
249	1	1	1
250	0	1	1
251	0	1	1
252	0	1	1
...	...	...	...
306	0	0	0
307	0	1	0
308	1	1	1
309	0	0	0
310	0	0	0

63 rows × 3 columns



	Date	Rain_Probability_LR
0	2025-03-10 01:32:58.369294	0.815457
1	2025-03-11 01:32:58.369294	0.380093
2	2025-03-12 01:32:58.369294	0.569859
3	2025-03-13 01:32:58.369294	0.359350
4	2025-03-14 01:32:58.369294	0.728617
5	2025-03-15 01:32:58.369294	0.463799
6	2025-03-16 01:32:58.369294	0.688575
7	2025-03-17 01:32:58.369294	0.839958
8	2025-03-18 01:32:58.369294	0.525501
9	2025-03-19 01:32:58.369294	0.816800
10	2025-03-20 01:32:58.369294	0.723433
11	2025-03-21 01:32:58.369294	0.727629
12	2025-03-22 01:32:58.369294	0.401036
13	2025-03-23 01:32:58.369294	0.530752
14	2025-03-24 01:32:58.369294	0.761272
15	2025-03-25 01:32:58.369294	0.613797
16	2025-03-26 01:32:58.369294	0.711683

Key Libraries & Tools Used

- **Data Manipulation: Pandas, NumPy**
- **Visualization: Matplotlib, Seaborn**
- **Machine Learning: Scikit-learn**
- **Statistical Analysis: SciPy**

Results & Discussion

The developed model demonstrates good predictive capability with 73% accuracy, which is notable given the complex and sometimes chaotic nature of weather patterns. The feature encoding approach proved particularly effective in improving model performance.

The correlation analysis revealed important relationships between meteorological variables that align with known weather phenomena, such as the connection between temperature, humidity, and rainfall probability.

Conclusion

This project successfully developed a machine learning model capable of predicting rainfall with reasonable accuracy. The methodological approach, particularly the feature encoding strategy and chronological data splitting, provides a solid foundation for weather prediction tasks. The model offers practical utility for short-term rainfall forecasting and demonstrates the potential of data-driven approaches in meteorological prediction.

Team:Data Dominators
Task:1