

Classification Case Study

1. Data Cleaning and Preprocessing

- **Missing Data Handling:** Identify and handle missing values in critical columns such as **RainTomorrow**, **Rainfall**, **Humidity9am**, and **WindGustSpeed** using appropriate techniques like imputation or removal.
- **Data Type Conversion:** Convert date and categorical variables to appropriate formats. The date might be segmented into year, month, and day components for more detailed analysis.
- **Outlier Detection:** Examine the data for any outliers and handle them as necessary to avoid skewed model results.

2. Feature Engineering

- **Weather Change Indicators:** Create features like temperature change from morning to afternoon (**Temp3pm - Temp9am**) and similar metrics for humidity and pressure.

3. Model Development

- **Building the Model:** Use the scikit-learn library to implement a Decision Tree classifier. Start with default settings to establish a baseline performance.
- **Parameter Tuning:** Employ grid search with cross-validation to find the optimal parameters for the decision tree, such as **max_depth**, **min_samples_split**, and **min_samples_leaf**, aiming to prevent overfitting and improve prediction accuracy.

4. Insights and Recommendations

- **Feature Importance:** Analyze which features are most influential in predicting rain to provide insights into which weather conditions are most indicative of rainy days.