Rainfall Prediction Model – Interview Preparation Guide

Section 1: Interview Questions Based on the Rainfall Prediction Model (With Answers)

1. What problem does your Rainfall Prediction model solve?

This model predicts whether it will rain today based on historical weather data. It helps in early warning systems, agricultural planning, and operational decision-making.

2. What dataset did you use?

I used historical weather data for the Melbourne region containing temperature, humidity, pressure, wind, cloud cover, and rainfall indicators.

3. What was your target variable?

The target variable was RainToday, a binary classification label with values Yes or No.

4. How did you handle categorical and numerical features?

Numerical features were standardized using StandardScaler, and categorical features were one-hot encoded using OneHotEncoder within a pipeline.

5. Why did you use a Pipeline?

Pipelines ensure consistent preprocessing during training and testing, prevent data leakage, and simplify model experimentation.

6. How did you handle class imbalance?

I used stratified train-test splitting and evaluated models using F1-score and recall rather than accuracy alone.

7. Why did accuracy differ from F1-score?

Accuracy was high due to class imbalance, but F1-score better reflected the model's ability to detect rainy days.

8. Why did Logistic Regression perform better for rainfall detection?

Logistic Regression had a higher recall for the rain class, reducing false negatives.

9. How did you evaluate the model?

I used GridSearchCV with StratifiedKFold cross-validation, classification reports, confusion matrices, and feature importance analysis.

10. What features were most important?

Humidity, cloud cover, pressure, and recent rainfall indicators were among the most influential features.

Section 2: Interviewer-Style Questions You May Be Asked

1. Why is recall more important than accuracy in rainfall prediction?
2. What would be the business impact of false negatives in this model?
3. How would you improve the model's performance further?
4. Why did you try Random Forest and Logistic Regression specifically?
5. How would this model behave in a different geographical region?
6. How would you deploy this model into production?
7. How would you monitor model performance over time?
8. What steps would you take if the data distribution changes?
9. How do you handle missing data in your pipeline?
10. How would you explain this model's predictions to non-technical stakeholders?
11. Why did you choose F1-score as a metric during hyperparameter tuning?
12. What trade-offs exist between model interpretability and performance?
13. How would you make this model cost-sensitive?
14. What assumptions does Logistic Regression make about the data?
15. If rainfall becomes rarer, how would that affect your model?