**✅ ML Project Review Notes (10-10-2025)**

**1. Dataset Description**

**Dataset Name**: IPL Players' Performance Statistics

**Description**:

* Contains season-wise performance statistics of IPL players.
* Includes **batting**, **bowling**, and **fielding** data such as:
  + Runs\_Scored, Batting\_Average, Wickets\_Taken, Economy\_Rate, Catches\_Taken, etc.
* Player names are **repeated for different seasons**, enabling per-season analysis.

**Use Case**:

* Predict a player’s performance in terms of:
  + **Batting**: Runs scored
  + **Bowling**: Wickets taken

**2. Preprocessing Techniques Used and Why**

✅ You chose **Data Cleaning** as the main preprocessing technique. Here's how you explain it:

**✨ Techniques Used:**

| **Technique** | **Why It's Needed** |
| --- | --- |
| **Missing Value Handling** | Some values were either missing or labeled as "No stats". You replaced them with NaN and dropped them where necessary to avoid skewing results. |
| **Data Type Conversion** | Columns like Runs\_Scored, Wickets\_Taken, etc. may have been read as strings. You converted them to numeric types to enable ML modeling. |
| **Feature Selection** | Selected only relevant features (batting/bowling stats) to train separate models for predicting runs and wickets. |
| **Dropping Incomplete Rows** | Used .dropna() to exclude rows with missing data from training. This helps improve model accuracy and avoids errors. |

**🛠 Optional Improvement (for future):**

* You could **impute** missing values instead of dropping them entirely using mean/median (especially if your dataset gets bigger).

**3. Machine Learning Technique Used and Why**

✅ **Model**: RandomForestRegressor

**🎯 Why Random Forest?**

| **Reason** | **Explanation** |
| --- | --- |
| **Works well with tabular data** | Your dataset is structured, with numeric input features. |
| **Handles non-linearity** | Player performance isn't always linear — RF captures complex relationships. |
| **Robust to overfitting** | Due to ensemble nature (uses multiple decision trees). |
| **Minimal preprocessing** | Doesn't require strict feature scaling (like normalization). |

**📌 You built two models:**

* **Batting Model** → Predicts Runs\_Scored
* **Bowling Model** → Predicts Wickets\_Taken

**4. Estimated Performance Measures**

✅ You can calculate performance on test data using:

**📏 Metrics Used:**

| **Metric** | **Purpose** |
| --- | --- |
| **Mean Squared Error (MSE)** | Measures average squared difference between actual and predicted values. Lower = better. |
| **R² Score (Coefficient of Determination)** | Indicates how well predictions match actual values (1 = perfect fit). |

**➕ Add this in your code before review to show results:**

# Batting model evaluation

y\_bat\_pred = bat\_model.predict(Xb\_test)

print("🏏 Batting Model:")

print("MSE:", mean\_squared\_error(yb\_test, y\_bat\_pred))

print("R² Score:", r2\_score(yb\_test, y\_bat\_pred))

# Bowling model evaluation

y\_bowl\_pred = bowl\_model.predict(Xw\_test)

print("\n🎯 Bowling Model:")

print("MSE:", mean\_squared\_error(yw\_test, y\_bowl\_pred))

print("R² Score:", r2\_score(yw\_test, y\_bowl\_pred))

**5. Execution (Demonstration)**

✅ Your code allows **interactive prediction**:

* User inputs part/full player name.
* The system:
  + Finds all matching players.
  + Predicts:
    - **Runs Scored** (using batting model)
    - **Wickets Taken** (using bowling model)
* Handles missing data gracefully and gives proper warnings if not enough data.

📌 **Demo Tip**: Try these player names during review:

* MS Dhoni
* Ruturaj
* Devon  
  (They exist in your dataset and have some clean records.)

**📋 Final Summary for Review**

| **Section** | **Key Points** |
| --- | --- |
| **Dataset** | IPL Player Statistics — season-wise data with batting & bowling metrics |
| **Preprocessing** | Data cleaning: missing values, type conversion, feature selection |
| **ML Algorithm** | RandomForestRegressor — good for non-linear tabular data |
| **Performance Metrics** | MSE, R² Score |
| **Execution** | CLI-based interactive predictions (batting & bowling) |