**Data Scientist Assessment – Technical Skills**

**Question:** You have been provided with 4 datasets from IPL 2021 and WCT20 2021. These are:

1. **IPL 2021 Fantasy Points:** This dataset contains fantasy points data for various players during the IPL 2021 matches. Key columns include Date, Player Name, Position, Credits, Selection, and Points.
2. **IPL 2021 Performance:** This dataset provides performance data for players during IPL 2021 matches. It includes columns like Date, Player Name, Match, Runs Scored, Balls Played, Batting Strike Rate, and various bowling metrics.
3. **WCT20 2021 Fantasy Points:** This dataset has similar data as the IPL 2021 Fantasy Points dataset but for the WCT20 2021 matches.
4. **WCT20 2021 Performance**: This dataset is analogous to the IPL 2021 Performance dataset but for WCT20 2021 matches.

Your task is to:

1. Combine the datasets to form a comprehensive dataset. Load and preprocess the dataset using Python.
2. Perform exploratory data analysis (EDA) to identify trends and insights.
3. Create visualizations that effectively communicate player performance patterns over the course of the tournaments.
4. Based on the available features, train a model to predict a player's performance in batting and bowling, and their fantasy points for the next match.
5. Evaluate the performance of your model and explain which features are most influential in predicting fantasy points.
6. Based on the available features, develop a classification model that helps categorize players as ‘High’, ‘Moderate’ and ‘Low’ performers based on their fantasy league points
7. Evaluate the model's accuracy, precision, recall, and F1-score.

**Assessment Instructions**

**1. Submission Format:**

* Ensure all your code, outputs, and visualizations are clearly documented in a Jupyter notebook or a similar interactive environment.
* If you use any external scripts or tools, include them in your submission.
* The notebook/script should have:
  1. Clear section headings for each task (e.g., Data Loading, EDA, Model Training).
  2. Comments and markdown cells explaining the approach.
  3. All code cells, outputs, and visualizations.
  4. Conclusions or insights derived from each task.

**2. Dataset:**

* All necessary datasets have been provided. Do not seek external data to supplement this assessment.
* If you've created intermediate or processed versions of the original datasets, include them. Files should be in standard formats like CSV or Excel.

**3. Coding:**

* Ensure your code is well-commented, explaining your logic and approach where necessary.

**4. Data Preprocessing:**

* Document any data cleaning, preprocessing, or transformation decisions.
* Justify any rows or columns you choose to exclude.

**5. Model Building:**

* Clearly state the algorithms you choose and why.
* Document any hyperparameter tuning or cross-validation processes.
* If you create new features, explain the rationale behind them.
* Save models using libraries like pickle, and provide instructions in the main notebook/script on how to load and use these models.

**6. Evaluation Metrics:**

Ensure you provide relevant evaluation metrics when building predictive models (e.g., RMSE, MAE for regression tasks; accuracy, F1-score, precision, recall for classification tasks).

**7. Visualizations:**

* All plots should have appropriate titles, axis labels, and legends.
* Briefly describe any insights or patterns each visualization reveals.

**Submission Instructions Using GitHub**

1. **Repository Setup:**

* Create a new GitHub repository specifically for this assessment.
* Name the repository as: **DataScientist-TechnicalAssessment**.

1. **Main Jupyter Notebook/Python Script:**

* Commit the primary notebook or script where you've performed all tasks to the root of the repository.
* Ensure that all code, visualizations, and markdown comments are clearly visible in the GitHub-rendered notebook/script.

1. **Data Files:**

* If you've created intermediate or processed datasets, add them to a folder named **data** in the repository.

1. **Model Files:**

* Store trained machine learning models in a folder named **models**.
* Include instructions in your main notebook/script on how to load and use these models.

1. **Visualizations:**

* If there are separate visualization files, store them in a folder named **visualizations**.

1. **Documentation:**

* If you've created a separate summary document, add it to a folder named **docs**.

1. **README File:**

* Your repository should have a README.md file at the root level.
* This README should:
* Briefly describe the purpose of the repository.
* Provide a directory structure explaining where to find key files.
* Offer instructions on how to run the main notebook/script.
* List any special libraries or tools you've used.

1. **Committing and Pushing:**

* Make meaningful commits with clear commit messages.
* Push all changes to your GitHub repository.

1. **Submission:**

* Once you've added all necessary files and are satisfied with your work, make sure your repository is public.
* Share the link to your GitHub repository as your assessment submission.