Project Module 2

Taha Hasnain Raza
BSCE20027

System Architecture:

1. **Data Collection Module**:
   - Responsible for gathering cricket match data from various sources such as APIs, databases, or CSV files.
   - Converts raw data into a structured format suitable for further processing.
2. **Preprocessing Module**:
   - Cleans the collected data by handling missing values, outliers, and inconsistencies.
   - Performs feature engineering to extract relevant features and transform data into suitable formats.
3. **Model Training Module**:
   - Utilizes machine learning algorithms to train predictive models based on historical cricket match data.
   - Includes algorithms such as Linear Regression, Decision Trees, Random Forest, Gradient Boosting, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), CatBoost, etc.
   - Incorporates hyperparameter tuning and cross-validation techniques to optimize model performance.
4. **Evaluation Module**:
   - Assesses the trained models' performance using evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and others.
   - Compares the performance of different algorithms to select the best-performing model.
5. **Prediction Module**:
   - Takes input features such as teams, venue, runs, wickets, overs, etc., for an ongoing or upcoming match.
   - Utilizes the trained model to predict the total score or outcome of the match.
   - Provides real-time predictions for live matches or upcoming fixtures.
6. **User Interface (UI)**:
   - Offers an interactive interface for users to input match details and view predictions.

- Displays match predictions along with confidence intervals and other relevant information.
- May include features like historical match data visualization, team performance analysis, etc.

7. **Database Integration**:
    - Stores collected data, trained models, and prediction results in a database for future reference and analysis.
    - Supports functionalities like data retrieval, storage, and update operations.
8. **Deployment and Integration**:
    - Integrates all modules into a cohesive system architecture.
    - Deploys the system on a suitable platform such as a local environment.
    - Ensures scalability, reliability, and performance optimization.

Planned Schedule for Integration:

1. **Week 1-2**:
    - Data collection and preprocessing module implementation.
    - Dataset gathering and cleaning.
2. **Week 3-4**:
    - Model training module development.
    - Initial model selection and evaluation.
3. **Week 5-6**:
    - UI design and development.
    - Database integration planning.
4. **Week 7-8**:
    - Final model training and evaluation.
    - Database implementation and integration.
5. **Week 9-10**:
    - System integration and testing.
6. **Week 11-12**:
    - User interface refinement and testing

Finalized Algorithms:

1. Linear Regression
2. Decision Trees
3. Random Forest
4. Gradient Boosting
5. Support Vector Regression (SVR)
6. K-Nearest Neighbors (KNN)
7. CatBoost

Dataset Collection Details:

- Sources: Cricket APIs (e.g., ESPN Cricinfo API, Cricket Australia API), cricket databases, public datasets, web scraping from cricket websites.
- Data Fields: Match details (teams, venue, date), innings summary (runs, wickets, overs), player statistics (batsman, bowler performance), match outcome.
- Data Preprocessing: Handle missing values, remove duplicates, standardize formats, perform feature engineering.