

Name: Sriman Vidhyarth. S

Roll No:713322AD102

Dept: AI&DS

Task 4: Data Science Lifecycle Example

1. Problem Definition:

- **Objective:** The goal is to predict the performance of football players in upcoming matches, which can help in making strategic decisions for team selection, training, and transfers.
- **Stakeholders:** Coaches, team managers, scouts, and analysts.

2. Data Collection:

- **Data Sources:** Gather data from various sources such as match statistics, player biometrics, training sessions, injury records, and social media.
- **Data Types:** Player attributes (e.g., age, height, weight), match performance metrics (e.g., goals, assists, passes), physical metrics (e.g., speed, endurance), and external factors (e.g., weather, opponent strength).

3. Data Preparation:

- **Cleaning:** Handle missing values, remove duplicates, and ensure data consistency.
- **Transformation:** Convert categorical variables (e.g., player positions) into numerical formats, normalize data, and create new features (e.g., average goals per match, injury frequency).

4. Exploratory Data Analysis (EDA):

- **Visualization:** Use charts and graphs to explore data patterns (e.g., performance trends over time, impact of injuries on performance).
- **Statistical Analysis:** Calculate summary statistics, correlation matrices, and perform hypothesis testing to understand relationships between variables (e.g., the correlation between training intensity and match performance).

5. Feature Engineering:

- **Feature Selection:** Identify the most relevant features that influence player performance (e.g., training intensity, match experience, recent injuries).

- **Feature Creation:** Create new features from existing data (e.g., a fitness score based on various physical metrics, a form index based on recent match performances).

6. Model Building:

- **Algorithm Selection:** Choose appropriate machine learning algorithms (e.g., regression analysis, decision trees, neural networks).
- **Training:** Split the data into training and testing sets. Train the model on the training set using selected features and target variable (performance score).

7. Model Evaluation:

- **Metrics:** Evaluate model performance using metrics such as accuracy, mean absolute error (MAE), root mean squared error (RMSE), and R-squared.
- **Validation:** Perform cross-validation to ensure the model's robustness and generalization.

8. Model Deployment:

- **Integration:** Deploy the model into the team's analytics platform, integrating it with other data sources and applications used by coaches and analysts.
- **Real-Time Predictions:** Use the model to generate real-time predictions on player performance, allowing the team to make informed decisions for upcoming matches.

9. Monitoring and Maintenance:

- **Performance Tracking:** Continuously monitor the model's performance and update it with new data to maintain accuracy.
- **Feedback Loop:** Gather feedback from coaches and analysts, refine the model based on new data, and adapt to changing conditions (e.g., new training methods, changes in player roles).