A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment: Predicting Cricket Team Winning Probability**

**Problem Statement:**

A sports analytics company wants to predict the probability of a cricket team winning a match based on various factors such as player performance, match conditions, and historical data. Accurate predictions can help in strategic decision-making during matches. Your task is to develop a machine learning model that predicts the winning probability of a cricket team given these attributes.

**Objective:**

Build a predictive model to estimate the probability of a cricket team winning a match. Evaluate the model using appropriate metrics and provide insights into the factors that most influence match outcomes.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **MatchID**: Unique identifier for each match.
2. **Team1**: The name of the first team.
3. **Team2**: The name of the second team.
4. **Venue**: The location where the match is being played (e.g., Stadium name, City).
5. **TossWinner**: The team that won the toss (Team1/Team2).
6. **BatFirst**: The team that chose to bat first (Team1/Team2).
7. **Team1Score**: The total runs scored by Team1.
8. **Team2Score**: The total runs scored by Team2.
9. **Team1WicketsLost**: The number of wickets lost by Team1.
10. **Team2WicketsLost**: The number of wickets lost by Team2.
11. **Team1RunRate**: The run rate of Team1.
12. **Team2RunRate**: The run rate of Team2.
13. **WeatherCondition**: The weather condition during the match (e.g., Clear, Rainy, Overcast).
14. **PitchCondition**: The condition of the pitch (e.g., Dry, Wet, Green).
15. **WinningTeam**: The target variable indicating which team won the match (Team1/Team2).

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding or Label Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as net run rate or average score per over.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different classification models (e.g., Logistic Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to match outcomes.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence match outcomes.
   * Suggest actionable strategies for cricket teams to optimize their performance based on the findings.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for improving match-winning strategies.

This assignment problem focuses on predicting the winning probability of a cricket team, a classification problem in the sports analytics domain. The problem involves data preprocessing, feature engineering, model training, and evaluation, with a focus on understanding the factors that influence match outcomes and helping teams make data-driven decisions.