Project Plan

Project Title: Predicting IPL Match Outcomes Using LightGBM Classifier and Feature Engineering

Research Question: Can machine learning, specifically the LightGBM classifier, along with feature engineering, be leveraged to accurately predict the winner of IPL matches.

Project Objectives:

1. Exploratory data analysis (EDA) will be performed to understand the dataset and identify relevant features.

2. New features will be engineered to enhance model performance.

3. A LightGBM model will be trained and evaluated for match outcome prediction.

4. The model's accuracy will be assessed, and feature importance will be interpreted.

Summary of Project and Background:

- The Indian Premier League (IPL) is a major T20 cricket league with significant fan interest and betting potential. Accurate prediction of match outcomes is valuable for both fans and stakeholders.

- A comprehensive IPL dataset (Patrick B, 2020) containing match details, player statistics, and outcomes from 2008-2020 will be utilized.

- The approach involves a LightGBM model due to its efficiency and proven effectiveness in classification tasks (Ke et al., 2017).

Reference:

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).

- Patrick B (2020). IPL Complete Dataset (2008-2020). Kaggle. Available at: https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020/data [Accessed 10 June 2024].

- S. Agrawal, S. P. Singh and J. K. Sharma, "Predicting Results of Indian Premier League T-20 Matches using Machine Learning," *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, 2018, pp. 67-71, doi: 10.1109/CSNT.2018.8820235.

- Gour, Prabodh & Khan, Mohd. Faheem. (2024). Utilizing Machine Learning for Comprehensive Analysis and Predictive Modelling of IPL-T20 Cricket Matches. Indian Journal Of Science And Technology. 17. 592-597. 10.17485/IJST/v17i7.2944.

Task List and Project Timeline

Timeline

This project will be conducted over a six-week period for now, with tasks overlapping to ensure efficient progress.

- **Week 1 (June 1-15):** Data acquisition and cleaning. The IPL dataset will be downloaded and preprocessed to handle missing values, inconsistencies, and irrelevant information.

- **Week 2 (June 16-20):** Exploratory data analysis (EDA). The cleaned dataset will be explored to understand the distributions of variables, relationships between features, and potential patterns relevant to match outcomes.

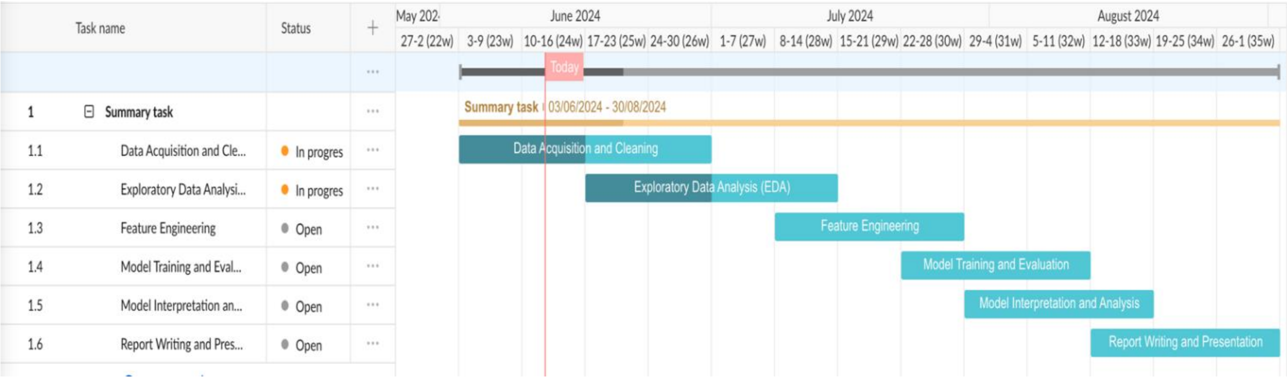
- **Week 3 (July 21-28):** Feature engineering. New features will be derived from the existing data to potentially improve the model's predictive power. This may involve aggregating statistics, creating interaction terms, or encoding categorical variables.

- **Week 4 (July 29-Aug 10):** Model training and evaluation. The LightGBM classifier will be trained on the engineered dataset, and its performance will be evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score.

- **Week 5 (Aug 11-18):** Model interpretation and analysis. The trained model will be analyzed to understand the importance of different features in predicting match outcomes. This will involve examining feature importance scores and potentially visualizing decision boundaries.

- **Week 6 (Aug 19-26):** Report writing and presentation. The findings of the analysis will be compiled into a comprehensive report, including visualizations and interpretations. The project will conclude with a presentation summarizing the key results and insights.

Gantt Chart:



	Task name	Status
1	⊟ Summary task	
1.1	Data Acquisition and Cle...	● In progres
1.2	Exploratory Data Analysi...	● In progres
1.3	Feature Engineering	● Open
1.4	Model Training and Eval...	● Open
1.5	Model Interpretation an...	● Open
1.6	Report Writing and Pres...	● Open

Data Management Plan

Overview of the Dataset

- The dataset is a comprehensive collection of Indian Premier League (IPL) match data from 2008 to 2020.
- It includes information on matches, deliveries (ball-by-ball data), teams, and players.
- The original purpose of collecting this data was likely to record and analyze the performance of teams and players in the IPL.

Data Collection

- The dataset will be collected from Kaggle, a platform for data science competitions and datasets. https://cricsheet.org/matches/
- The cricsheet is the original source from where the data has been collected.
- The specific link to the dataset is: https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020/data

Summary of Data

- **Format:** The dataset is provided in CSV (Comma-Separated Values) format.
- **Size:** The total size of the dataset is approximately 1 MB.
- **Records:** The dataset contains thousands of records, including details of each ball bowled and various match statistics.
- **Expected Size of Data and Code Files:** The size of the data and code files is expected to remain relatively small (a few MBs) as the analysis primarily involves data manipulation and model training.

Document Control

- **GitHub:** The project's code and associated files will be managed using a private GitHub repository.
- **Version Control:** Git will be used for version control, allowing for tracking changes and reverting to previous versions if needed.
- **File Naming:** A clear and consistent file naming convention will be followed, using descriptive names and version numbers (e.g., "ipl_data_cleaning_v1.ipynb," "ipl_feature_engineering_v2.ipynb").

Metadata

- A detailed README file will be included in the GitHub repository.
- The README file will provide information on the dataset, the project's objectives, instructions for running the code, and explanations of the analysis steps.

Security and Storage

- The dataset and code files will be stored locally on a secure machine.
- Regular backups will be created and stored on an external hard drive and cloud storage (e.g., Google Drive, One drive) to prevent data loss.
-

Ethical Requirements

1. **GDPR Requirements:** The dataset does not contain any personally identifiable information (PII) and therefore complies with GDPR(General Data Protection Regulation) requirements.
2. **UH Ethical Policies:** The project will adhere to all relevant UH ethical policies regarding data usage and research conduct.
3. **Permission to Use Data:** The dataset is publicly available on Kaggle, and its license permits its use for research and analysis purposes.
4. **Ethical Data Collection:** While there is no explicit information on the original data collection process, it can be reasonably assumed that the data was collected ethically as it is sourced from a reputable platform (Kaggle) and is widely used for research purposes.