# Final Report

# Introduction

The purpose of this comprehensive report is to present the findings from multiple sprints of analyzing betting data from industrial bet logs. The objective is to extract insights and patterns regarding user behavior, preferences in sports, teams, types of bets, and the amount of bets placed. Each sprint focused on different aspects of the data to build a holistic understanding of the betting landscape.

# Sprint 1: Initial Analysis and Key Insights

#### **Introduction**

The purpose of this report is to analyze betting data from industrial bet logs to extract insights and patterns regarding user behavior, preferences in sports, teams, types of bets, and the amount of bets placed.

#### **Key Findings**

In the analysis of industrial bet logs data, several key insights have emerged, shedding light on user behavior, preferences, and betting trends. The dataset comprises a diverse array of users engaged in betting activities, with a total of 550 unique users. This varied user base forms the foundation of our analysis, providing a rich tapestry of betting behavior to explore.

Users with multiple bets were singled out for focused analysis, as this subgroup offers valuable insights into recurring betting patterns and preferences. This analysis is significant as it potentially serves as a cornerstone for targeted marketing efforts and personalized user engagement strategies.

An examination of betting preferences showed intriguing insights into the sports, teams, tournaments, and market types favored by users. The analysis of the most popular sports for betting included metrics such as total stake, number of bets, and median stake, providing invaluable guidance for platform optimization and content curation.

Furthermore, a deep dive into betting trends surrounding popular teams revealed betting behavior associated with both home and away teams. By scrutinizing metrics such as total stake, number of bets, and median stake for each team, we identified which teams held the greatest appeal among bettors. Similarly, an exploration of popular tournaments unveiled those commanding the highest betting activity. By dissecting metrics across different tournaments, we gained a clearer picture of user preferences in tournament betting.

The analysis also extended to market type preferences, illuminating the distribution of bets across various market types. This granular understanding of user preferences within different market segments enables platform operators to tailor their offerings effectively to align with user expectations.

#### **Data Preprocessing**

The initial step involved loading the JSON data into a Pandas DataFrame using the pd.json\_normalize() function, which flattens the nested structure of the JSON data into a tabular format. After loading the data, relevant fields were extracted from the DataFrame to facilitate analysis, including match details such as match ID, date of match, home team, away team, sport, category, and tournament. Additionally, information about the market type and odds associated with each bet was extracted.

Timestamp columns, namely 'time' and 'match\_dateofmatch', were converted from string format to datetime objects using the pd.to\_datetime() function. This conversion enabled efficient time-based analysis, such as tracking betting trends over time and comparing betting activity across different periods. An investigation was conducted to identify missing values in the 'market\_specialoddsvalue' column, which could indicate various scenarios, such as bets placed on standard odds or certain types of markets where special odds were not applicable.

#### **Analysis of the Data**

In the analysis of the betting dataset, comprehensive descriptive statistics and summaries were generated to provide foundational insights into user engagement and betting behaviors. Unique user counts were computed, revealing the diversity and scale of the user base involved in betting activities. Value counts for key categorical variables, including match\_sport, match\_category, and match\_tournament, were calculated to identify the most popular sports, categories, and tournaments among bettors.

Detailed aggregations of stake amounts across various dimensions such as users, sports, teams, and tournaments were performed, including calculating the mean, median, minimum, maximum, and sum of stakes. This statistical groundwork provided a nuanced view of betting habits, guiding strategic decisions to optimize user engagement and betting experiences.

#### **Visualization Component**

Several graphical representations were employed to distill and communicate complex patterns in a digestible format. Bar charts were used to highlight the most popular tournaments, focusing on those with more than 100 entries. Pie charts demonstrated the top five sports by popularity and by total stakes invested, offering insights into user preferences and high-stakes areas. A time series plot of daily total bets over the past year was created to observe trends and fluctuations in betting activity. These visualizations form a cornerstone of our analytical reporting, making the data accessible and actionable for strategic decision-making.

#### **Exploratory Analysis**

In the exploratory analysis phase, we delved into the dataset to uncover patterns and preferences not immediately apparent from straightforward descriptive statistics. A key focus was on the 'market\_specialoddsvalue' attribute, which provides insights into specific betting conditions and strategies employed by users. By counting the occurrences of various entries in this field, we identified prevalent betting preferences, such as a common inclination towards betting on more than 0.5 goals. This analysis sheds light on the strategic choices bettors make based on odds, offering a deeper understanding of risk preferences and betting tactics within the user base.

#### **Conclusion**

Our first sprint has successfully revealed critical insights into betting behaviors and user preferences within the industrial bet logs. The diligent analysis of 550 unique users uncovered significant patterns, particularly in terms of sports and teams that captivate our audience’s interest. This depth of understanding aims to refine the platform to better cater to user needs. The discovery of missing values in the 'market\_specialoddsvalue' provided a unique opportunity to delve into specific betting conditions, enhancing our understanding of user strategies and preferences.

Furthermore, the concentrated analysis on geographical preferences showed a high concentration of betting activities in Portugal and among International Clubs, showcasing regional preferences that could guide targeted marketing strategies. The obvious popularity of soccer, highlighted by both the number of bets and total stakes, underscores the sport's global appeal and potential for promotional efforts.

In the following sprint, these insights will be crucial in optimizing our platform and recommending strategies to boost user engagement and overall satisfaction. The first sprint has provided a strong foundation for understanding our users and set the stage for proactive future developments.

# Summary of Sprint 2 Progress Report - Group 1 (Marios, Pavlos, David)

This report analyzes a dataset of user betting behavior to identify patterns, segment users, and generate personalized betting recommendations. Hierarchical clustering was used to create user profiles based on features like average stake, total stake, bet count, favorite sport, and favorite tournament. Three main clusters were identified, with Cluster 1 being the largest. Clusters were evaluated using metrics like the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, showing reasonably well-defined clusters.

Personalized recommendations were generated based on user clusters, focusing on categories such as match sport, match category, tournament, and odd field type. Single user recommendations showed high accuracy in the odd field type category with a hit rate of 1.0, while recommendations for all users varied, with the highest average hit rate of 0.55 in the odd field type category.

The results indicate that clustering and recommendation systems can provide valuable insights and personalized suggestions, with potential for further refinement to improve accuracy. Future work could explore additional features to enhance the system's performance.

# Summary of Sprint 2 Progress Report - Group 2 (Alketa, Nasos, Dimitris, Nick)

In the initial phase of Sprint 2, the team focused on developing a recommendation algorithm for a betting service. The primary objective was to implement a clustering algorithm to enhance personalized recommendations. The team chose K-means as the primary method due to its simplicity and effectiveness in handling large datasets. K-means partitions data into distinct clusters based on similarity, which is useful for segmenting user data and tailoring recommendations.

An extensive literature review on clustering techniques was conducted, comparing different methods, including collaborative filtering and sequential pattern mining. This research highlighted the strengths and limitations of various algorithms and ultimately guided the decision to use K-means for its computational efficiency and straightforward implementation. However, K-means has limitations, such as sensitivity to the initial placement of centroids and the need to specify the number of clusters beforehand.

In the context of the betting recommendation system, K-means was chosen for its ability to segment users based on their betting behaviors and preferences. This clustering approach aimed to identify distinct user segments, such as high-frequency bettors or those with particular sports preferences, to provide targeted recommendations and improve user engagement and satisfaction.

The implementation involved feature engineering, extracting relevant features from the dataset, and normalizing them using StandardScaler. The elbow method was used to determine the optimal number of clusters, resulting in four clusters. These clusters were then used to segment users and develop personalized recommendations based on cluster membership.

However, the results from the K-means clustering were inconclusive, leading to a decision to pivot to collaborative filtering techniques. The silhouette analysis and other evaluation metrics indicated overlapping and poorly defined clusters, suggesting that K-means might not be the optimal method for the dataset. Collaborative filtering, particularly the Alternating Least Squares (ALS) model, was identified as a better approach due to its effectiveness in recommendation systems and its ability to predict user preferences based on user behavior patterns.

The focus shifted to collaborative filtering, specifically the ALS model, known for its efficacy in dealing with sparse datasets and scalability. ALS leverages user-item interaction data to predict individual user preferences, dynamically adapting to users' betting behaviors and preferences. This approach aims to develop a robust recommendation system tailored to the betting industry.

The implementation of the ALS model involved several steps. Data preprocessing was crucial, requiring the transformation of raw data into a structured format suitable for ALS. The nested JSON data was flattened, and essential columns like user ID, match ID, and stake were extracted. The interaction data was transformed into a binary format, indicating whether a user had engaged with an item.

The ALS model was initialized with appropriate parameters, and the data was split into training and test sets. The model was trained on the training set, and predictions were generated on the test set. The accuracy of the recommendations was evaluated using metrics such as RMSE, MAE, precision, recall, and F Beta Score. These metrics provided insights into the model's performance and highlighted areas for further refinement.

Future work will focus on advanced preprocessing techniques, feature engineering, fine-tuning the ALS model, and hyperparameter tuning to optimize performance. Cross-validation will be employed to ensure reliability and robustness. Additionally, integrating content-based filtering will create a hybrid recommendation system, enhancing the diversity and relevance of the recommendations.

# Summary of Sprint 3 Progress Report - Group 1 (Marios, Pavlos, David)

This sprint laid the foundation for developing a personalized recommendation system for the betting company, with a shift from clustering to collaborative filtering, aiming to improve user engagement and satisfaction.

After creating a good model our team developed a robust recommendation system API for sports betting using the FastAPI framework. This API processes user betting data, applies the KMeans clustering algorithm for data segmentation, and generates personalized betting recommendations.

**API Endpoints:**

Users can upload new datasets in JSON format. The endpoint validates the file type, merges data with existing datasets, preprocesses it, performs KMeans clustering, and generates user profiles for future use.

Provides betting recommendations for all users based on preprocessed data and user profiles. Recommendations are generated from clustering results and returned in a structured format.

Offers user-specific betting recommendations. It retrieves the user profile and preprocessed data, generates personalized recommendations based on their cluster, and returns them.

**Key Components:**

We also focused on developing a robust recommendation system API for sports betting using FastAPI. Key components include data preprocessing to clean and transform data for consistency, KMeans clustering to segment users based on betting behaviors for accurate grouping, and recommendation generation that leverages clustering results to offer personalized betting suggestions. This approach ensures our API delivers tailored recommendations aligned with user preferences and historical betting patterns, enhancing overall user experience and engagement in sports betting.

Moreover, our implementation revolves around leveraging FastAPI for its high-performance capabilities, asynchronous handling, automatic documentation generation using Swagger UI and ReDoc, and robust data validation through Pedantic models. We utilize FastAPI to develop a sophisticated recommendation system API for sports betting, ensuring reliable functionality and seamless user interaction. Postman serves as our tool for comprehensive API testing, validating endpoints and ensuring consistent and accurate responses. By integrating advanced data processing techniques and KMeans clustering, we enhance the precision of our personalized betting recommendations, ultimately boosting user engagement and satisfaction in the sports betting domain.

The JSON data that are received after processing the uploaded document contains a preview of the processed data, along with a message indicating successful processing. The output consists of a list of dictionaries, where each dictionary represents a user's betting activity. Each user entry contains the following key-value pairs:

**time**: The timestamp indicating when the bet was placed.

**bet**: A list containing details of the user's bets. Each bet includes information such as the stake amount, the type of event (live or pre-match), details of the match (including teams, sport, category, and tournament), market information (e.g., total goals, 1x2), and odds offered.

**userid**: The unique identifier of the user who placed the bet.

The entity of the output as shown in the Figure 1.1 is the following: **Figure 1.1 POST Request**

The user with ID 30131 placed a bet with a stake of $50. The bet was on a live soccer match between Club Libertad Asuncion and Deportivo Tachira in the Copa Libertadores tournament. The user selected an option related to the total goals scored in the first half, with odds of 1.7.

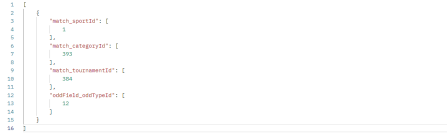
The JSON snippet that is in the Figure 1.2 represents a list of dictionaries, where each dictionary contains keys corresponding to different attributes of the user's betting activity. The key represents for the user ID 30131 are the following:

**match\_sportId**: This key indicates the sport IDs of the matches the user placed bets on. In this case, the user placed bets on matches belonging to Sport ID 1, which typically represents soccer.

**match\_categoryId**: It signifies the category IDs of the matches the user placed bets on. Here, the user placed bets on matches categorized under Category ID 393, which could represent "International Clubs" in the context of soccer tournaments.

**match\_tournamentId**: This key indicates the tournament IDs of the matches the user placed bets on. The user placed bets on matches that are part of Tournament ID 384, which could represent a specific soccer tournament like the "Copa Libertadores."

**oddField\_oddTypeId**: It denotes the odd types associated with the user's bets. In this case, the user placed bets with Odd Type ID 12, which could represent an over/under type of betting option.

**Figure 1.2 GET Request**

The Figure 1.3 contains recommendations for multiple users, both new and old, existing in the dataset. Each recommendation entry includes the following information:

**userid**: Represents the unique identifier of the user for whom the recommendations are generated.

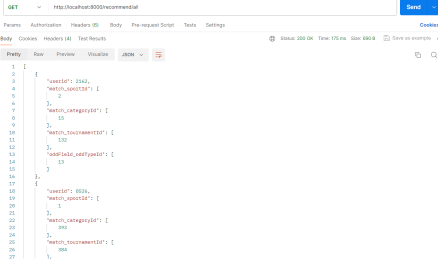
**match\_sportId**: Indicates the sport IDs associated with the recommended matches for the user. This suggests the type of sports events recommended for the user to consider betting on.

**match\_categoryId**: Reflects the category IDs of the recommended matches, providing further context about the types of events recommended.

**match\_tournamentId**: Specifies the tournament IDs of the recommended matches, giving insights into specific tournaments or competitions the user might be interested in betting on.

**oddField\_oddTypeId**: Represents the odd types associated with the recommended bets. This provides information about the types of betting options suggested for the user.

Overall, this JSON snippet offers personalized recommendations tailored to each user's preferences and betting history, encompassing various aspects such as sports, categories, tournaments, and types of bets.

**Figure 1.3 GET Request for all the users**

# Summary of Sprint 3 Progress Report, 1st approach.

In the latest sprint, significant progress was made in several key areas. The primary focus was on hyperparameter tuning to optimize the model's performance. Extensive experimentation with various hyperparameter settings was conducted using techniques like grid search and random search. This process also involved implementing cross-validation to ensure the robustness of the model and to prevent overfitting. As a result, there were notable improvements in both the accuracy and reliability of the model.

Dataset management was another critical area of development. The objective was to enable continuous updates and maintain the dataset's relevance. To achieve this, a mechanism was developed to upload new parts of the dataset seamlessly. This integration ensured that new data could be added without causing disruptions to the existing dataset. Additionally, data validation checks were implemented to maintain high data quality. This continuous update mechanism ensures that the model is always trained on the most recent and relevant data.

The implementation of content-based filtering aimed to personalize recommendations based on user preferences and historical data. A content-based filtering algorithm was designed and integrated with user profiles and historical data. This algorithm was fine-tuned to improve the relevance of the recommendations. As a result, the system now provides personalized betting recommendations that better align with user preferences, enhancing user satisfaction and engagement.

Technical Implementation:

The technical implementation involved setting up an environment for processing betting data using a combination of Python libraries and PySpark. Initially, necessary libraries such as pandas, numpy, pyspark, and scikit-learn were installed and imported to facilitate data manipulation and machine learning tasks. A function to flatten nested JSON structures was defined, making it easier to analyze and process detailed match properties.

The process started by mounting Google Drive to access the data, loading existing betting data from a JSON file into a Pandas DataFrame. Newly uploaded JSON data was then combined with the existing data into a single DataFrame. This updated DataFrame was saved back to Google Drive, ensuring that the dataset remained comprehensive and up-to-date.

A Spark session was initiated to handle large-scale data processing. The updated JSON data was read into a Spark DataFrame and flattened, transforming nested betting data into a more manageable format. The data was then prepared for the Alternating Least Squares (ALS) model, with interactions defined as stakes greater than zero. The ALS model was trained on the data, and a portion was reserved for testing. After training, the model generated predictions, which were converted to a Pandas DataFrame.

To provide actionable insights, top-K recommendations for each user were generated based on the ALS model's predictions. These recommendations were saved to a CSV file. Additionally, in response to client feedback, content-based filtering was incorporated to provide more context beyond just match IDs. A mapping for match properties like sports, categories, and tournaments was created. The ALS predictions were enriched with these properties, extracting the top-K sports, categories, and tournaments for each user. This hybrid recommendation approach ensures users receive personalized recommendations for specific matches and broader categories they are interested in. The final hybrid recommendations were saved to a CSV file, providing a comprehensive recommendation system that meets client requirements.

Next Iteration Goals:

The next iteration aims to refine the model's output, ensuring higher quality and accuracy in recommendations. The current recommendation outputs will be analyzed to identify areas for improvement, and enhancements will be implemented to refine the accuracy of predictions. User testing will be conducted to gather feedback and further refine the outputs, resulting in more accurate and relevant recommendations that users can rely on for making informed betting decisions.

Ensuring consistency in the model's output is another priority. The goal is to address any fluctuations or variability in the recommendation results. Stability measures and regular monitoring will be implemented to maintain consistent performance, leading to stable and reliable recommendations that build user trust.

Enhancing user experience is also a key objective. Comprehensive documentation on using the data extraction tool and scripts will be created. A tool for backend developers to extract data from the model will be developed, including examples and use cases to demonstrate integration into backend workflows. This will result in a more intuitive and engaging user experience, leading to higher user satisfaction and better adoption of the model.

By focusing on these goals, the team aims to enhance the performance, consistency, and usability of the betting recommendation model, ensuring it delivers high-quality recommendations and provides a robust tool for backend developers to extract and use the data effectively.

### 

# Summary of Sprint 4 Progress Report, 2nd approach.

In the earlier sprints, significant strides were made in optimizing the model's performance and enhancing dataset management processes. Through extensive hyperparameter tuning, using grid search and random search techniques, the team achieved notable improvements in model accuracy and reliability. Additionally, a robust mechanism for continuous dataset updates was developed, ensuring the model was consistently trained on the most recent and relevant data. Data validation checks were also implemented to maintain high data quality. Furthermore, a content-based filtering algorithm was designed and integrated to personalize recommendations based on user preferences and historical data, significantly enhancing user satisfaction and engagement.

**Technical Implementation:**

The technical implementation involved setting up an environment for processing betting data using Python libraries and PySpark. The process began with mounting Google Drive to access data, followed by loading existing and new betting data, combining them into a single DataFrame, and saving the updated dataset back to Google Drive. A Spark session was then initiated to handle large-scale data processing, transforming nested betting data into a manageable format and preparing it for the Alternating Least Squares (ALS) model. The ALS model was trained on this data, and predictions were generated and enriched with contextual information, resulting in a comprehensive recommendation system.

**Sprint 4: Integration and Project Conclusion**

Based on feedback from the professor, the focus for Sprint 4 initially shifted towards integrating the recommendation system with the API to enhance the system’s functionality and streamline its deployment. However, due to unforeseen technical challenges and limited time, full integration with the API could not be achieved. This technical hurdle necessitated a strategic pivot. As a result, the decision was made to merge efforts with the other team to successfully connect and conclude the project.

Given the shift in focus, the primary objective for this sprint was to combine the ALS-based recommendation system with the components developed by the other team. This collaborative effort aimed to create a unified and robust recommendation system, leveraging the strengths and progress of both teams.

**Objectives:**

The primary objective was to document the progress made on the API integration and outline the challenges faced and potential solutions for future work. The focus then shifted to merging codebases, and methodologies developed by both teams to create a comprehensive recommendation system. This included addressing any remaining technical debts and refining the existing functionalities based on collaborative feedback.

**Planned Actions:**

The initial step involved documenting the current state of API integration, including the challenges encountered and potential future solutions. This documentation aimed to provide a clear path for completing the integration when technical constraints are resolved. Next, the unified recommendation system was developed by integrating the datasets and models from both teams, enhancing accuracy and robustness. Thorough testing was conducted to ensure the system met project requirements and performed reliably. Any final refinements were implemented based on testing results and stakeholder feedback. Finally, a comprehensive final report documenting the entire project, including methodologies, challenges, solutions, and future work recommendations, was prepared. A final presentation was developed to showcase the completed project, highlighting collaborative efforts and outcomes.

**Expected Outcomes:**

The expected outcomes of Sprint 4 included an enhanced recommendation system that leveraged the strengths of both teams' approaches and datasets. Detailed documentation of the API integration efforts was provided, offering a clear path for future work. A complete final report encapsulating the project's scope, methodologies, and outcomes was also expected. Ultimately, a unified and functional recommendation system, ready for deployment or further development, was the primary goal.

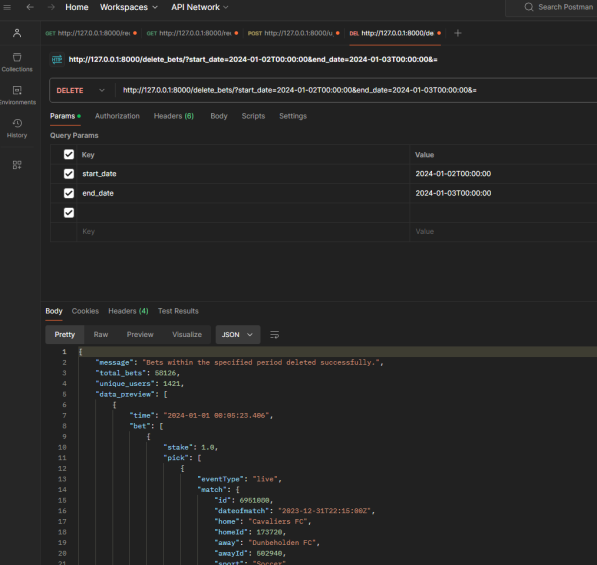
# Summary of Sprint 4 Progress Report - Group 1 (Marios, Pavlos, David)

Οur system provides detailed descriptive statistics and functionalities for managing betting data effectively. It captures and analyzes betting activities over specific periods, revealing user engagement levels and total betting volume. We track additions of new bets and categorize users based on their betting behaviors, offering insights into user segments and their preferences.

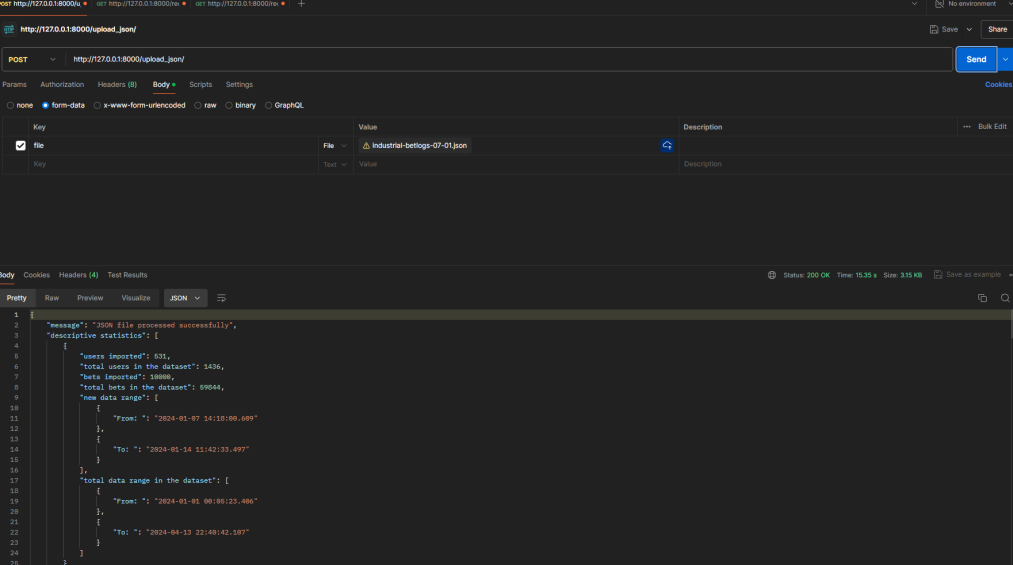
Additionally, the system features robust functionality for removing results:

* Users can delete individual bets, allowing for personalized management of betting histories.
* Time-based data removal capabilities enable users to clean up data within specified date ranges. Recently, we successfully implemented the removal of bets placed between January 1st and February 1st, showcasing the system's efficiency in handling such tasks.

These capabilities as shown in empower users to maintain accurate and organized betting records while enhancing overall data management and user satisfaction.

****

In conclusion, our project successfully developed a robust and efficient betting recommendation system utilizing FastAPI and KMeans clustering. The system's ability to preprocess data, identify user clusters, and generate personalized recommendations has significantly improved the relevance and accuracy of the suggestions provided to users. Additionally, the inclusion of detailed descriptive statistics offers users deeper insights into their betting patterns, while the functionality to remove specific bets or data within defined time frames enhances the system's flexibility and user control. By thoroughly testing the API endpoints with Postman, we ensured the system's reliability and performance. This comprehensive approach has resulted in a powerful and user-centric betting recommendation system that effectively meets the diverse needs of its users and sets a high standard for future projects in this domain.



### **General Conclusions From Both Teams**

In this comprehensive project, we successfully developed a robust betting recommendation system that leverages advanced data processing techniques, machine learning algorithms, and API integrations. Through multiple sprints, our team analyzed industrial bet logs to extract meaningful insights into user behavior, betting preferences, and patterns. This analysis provided a strong foundation for creating a personalized recommendation system aimed at enhancing user engagement and satisfaction.

**Key Achievements:**

1. **User Behavior Analysis:** By examining the betting logs of 550 unique users, we gained valuable insights into user behavior, preferences for sports, teams, tournaments, and market types. This helped in identifying popular sports like soccer and regions with high betting activity, such as Portugal and International Clubs.
2. **Clustering and Personalization:** Using hierarchical clustering and K-means algorithms, we segmented users based on betting behaviors. This allowed for the creation of personalized recommendations tailored to different user profiles. Despite initial challenges with clustering, the pivot to collaborative filtering techniques like ALS proved effective in generating accurate and relevant betting suggestions.
3. **API Development:** We developed a sophisticated recommendation system API using FastAPI, capable of processing betting data, segmenting users, and providing personalized recommendations. The API's endpoints were rigorously tested using Postman to ensure reliability and performance, enhancing user interaction and experience.
4. **Data Management and Update Mechanism:** A robust mechanism for continuous dataset updates was implemented, ensuring the model remained trained on the most current data. This was complemented by data validation checks to maintain high data quality, critical for accurate recommendations.
5. **Hybrid Recommendation System:** Integrating content-based filtering with ALS-based collaborative filtering enriched the recommendation system, providing users with diverse and contextually relevant suggestions. This hybrid approach improved the overall quality.
6. **Sprint Collaboration and Integration:** The final sprint focused on integrating efforts from both teams, culminating in a unified and comprehensive recommendation system. Despite technical challenges, the collaborative effort ensured a robust final product ready for deployment or further development.

**Challenges and Solutions:**

* **Technical Integration:** While full integration with the API faced unforeseen technical challenges, detailed documentation of the progress and potential solutions was provided. This sets a clear path for future work to complete the integration seamlessly.
* **Model Refinement:** Continuous hyperparameter tuning and feature engineering improved the model's performance, ensuring high accuracy and reliability in recommendations.

**Future Work:**

The project lays a solid foundation for future enhancements. Key areas for future work include:

* **Complete API Integration:** Resolve technical issues to fully integrate the recommendation system with the API, enabling seamless deployment.
* **Advanced Feature Engineering:** Explore additional features and advanced preprocessing techniques to further refine the recommendation accuracy.
* **User Testing and Feedback:** Conduct user testing to gather insights and refine the recommendations based on real-world feedback, ensuring they meet user expectations and needs.

**Summary:**

This project demonstrates the successful development of a sophisticated betting recommendation system that effectively meets the diverse needs of its users. By combining detailed user behavior analysis, advanced clustering techniques, collaborative filtering, and robust API development, we created a powerful tool that enhances user engagement and satisfaction in the betting domain. The comprehensive approach and collaborative effort set a high standard for future projects, providing valuable insights and methodologies for continued innovation and improvement in personalized recommendation systems.