Module 1:

ADS-599 Capstone Project Proposal

Fill out this form and submit it by the end of Module 1 in Canvas.

Student Information:

Team Leader/Representative:
Maria Mora Mora / Mauricio / Gabriel Mancillas
Full Names of Team Members:

1. Maria Mora Mora
2. Mauricio Espinoza
3. Gabriel Mancillas

Capstone Project Information:

Title of your Capstone Project: Soccer Performance Intelligence System

Name of Your Selected Dataset: Multi-source Soccer Analytics Dataset (API-Sport, Social Media, Wikipedia)

Description of Your Selected Dataset: The dataset will be aggregated from **four comprehensive sources** to provide deep insights into European football. It incorporates over five years of detailed match data, player statistics, team performance metrics, and tournament results from the Champions League and major European league competitions via **API-Football records**. We also integrate advanced statistical data from **FBref.com**, including Expected Goals (xG), Expected Assists (xA), defensive actions, possession metrics, and crucial formation-specific performance analytics across the Premier League, La Liga, Serie A, Bundesliga, and Ligue 1. To capture the pulse of public opinion, social media sentiment data from **Twitter discussions** about teams, players, and matches is included, complemented by extended historical records from the **SportMonks API**. This combined dataset will encompass approximately **750+ variables** and an estimated **15,000+ records**. These variables comprehensively cover player performance (e.g., goals, assists, xG/xA, defensive stats), team statistics (e.g., tactical metrics, venue analytics, financial indicators), match-level data (e.g., results, contextual factors), sentiment scores, and historical tournament outcomes. This rich data foundation is designed to support comprehensive Shapley value analysis and the development of sophisticated formation-specific tactical intelligence.

Provide your team GitHub link here: https://github.com/mmoramora/ADS599_Capstone

How many times have your team member(s) met during Module 1: Once/Twice

What was the agreed upon method of communication? Are you using any teamwork project management software: Yes slack and zoom meeting, google doc for shared deliverables, github for code sharing

Attach a typed proposal (two-page minimum) in your own words for the work you intend to complete towards the following components to satisfy the requirements for your capstone project.

Project Proposal

Purpose / The Problem

Traditional soccer prediction models rely primarily on statistical analysis of historical match data, missing the rich contextual information available in news sentiment, player narratives, and complex relationship networks between players, teams, and competitions. Current prediction systems fail to incorporate the nuanced factors that influence team performance, such as media pressure, player relationships, injury networks, and evolving team dynamics that are captured in natural language sources. Moreover, existing models cannot identify the best contributors within specific positions or tactical systems, making it difficult to determine which players provide the highest value in different formations or playing styles.

Background

Recent advances in cooperative game theory, particularly Shapley values from their original mathematical foundation, provide a theoretically grounded method for fairly distributing contribution credit among players within tactical systems rather than just teams. Shapley values, when applied to tactical systems, reveal the marginal contribution of each player to the system's overall effectiveness, accounting for how player value changes based on the specific tactical setup, formation, and playing style rather than simple team membership.

Your Motivation

Soccer analytics represents a rapidly growing field where traditional statistical models are being enhanced by game theory approaches that can identify system-specific contributions rather than team-based metrics. The Champions League provides an ideal testing ground due to its high-stakes environment, extensive media coverage, and rich historical data. By applying Shapley values to tactical systems and combining with RAG capabilities, we can identify the best contributors within specific formations and playing styles, understanding which players

provide maximum marginal value when the tactical system changes, not just when team composition changes.

Your Working Hypothesis

Recent studies by Pappalardo et al. (2019) achieved 68% accuracy in match outcome prediction using traditional statistical models, while Huang & Chen (2023) reported 72% accuracy with deep learning approaches on Premier League data. Our hypothesis posits that incorporating system-specific Shapley values and multi-modal data integration will achieve >78% accuracy in Champions League performance prediction, representing a meaningful improvement over current state-of-the-art methods. Furthermore, we hypothesize that tactical system-based Shapley values will provide superior explanatory power for player contribution analysis compared to team-based attribution methods, enabling more accurate identification of optimal formation selection and tactical decision-making based on available squad members' marginal contributions to specific tactical systems.

Data Science Objectives

1. Implement multi-source data integration from API-Sport, social media, and Wikipedia APIs with tactical formation classification from match data
2. Develop game theory framework calculating Shapley values for player contributions to tactical system effectiveness rather than team outcomes
3. Create sentiment analysis pipeline using fine-tuned models specifically trained on tactical and formation-related soccer content
4. Build RAG system enabling queries about system-specific contributors like "Which center-backs contribute most to high-pressing systems?"
5. Implement system-based Shapley value calculations with statistical significance testing for formation-specific contributions
6. Develop tactical system taxonomy using unsupervised clustering on formation and playing style data
7. Create formation-specific player ranking systems based on marginal contribution to tactical effectiveness
8. Benchmark system-specific predictions against actual tactical success metrics and formation performance

Planned Methodology

Implement automated formation detection from positional data using computer vision and clustering algorithms, with manual validation for tactical style classification. **Game Theory Implementation:** Deploy Shapley value calculations for each tactical system independently, measuring player marginal contributions to system-specific outcomes like pressing effectiveness,

possession dominance, or defensive solidity. **System-Specific Analysis:** Apply cooperative game theory framework where coalitions represent tactical formations, calculating fair value distribution among players based on their contribution to tactical system success rather than team results. **RAG Integration:** Develop natural language interface enabling tactical queries like "Which players contribute most to counter-attacking systems?" with Shapley-based explanations. **Validation Framework:** Implement rigorous testing comparing system-specific Shapley values against actual tactical performance metrics and formation success rates.

Real-World Impact and Planned Deliverables

Deliverable: A theoretically grounded RAG system that identifies and ranks the best contributors within specific tactical systems using game theory Shapley values, enabling queries like "Which midfielders add most value to possession-based systems?" with mathematical explanations of marginal tactical contributions. **Secondary Deliverables:** Peer-reviewable methodology paper on applying cooperative game theory to tactical systems, open-source implementation of system-based Shapley value calculations, tactical adaptability metrics for player evaluation, and comprehensive analysis of formation-specific contributor rankings. **Real-World Applications:** Formation-specific player recruitment identifying specialists for tactical systems, tactical optimization based on available squad members and their system-specific Shapley values, playing style adaptation guided by player contributions to different tactical approaches, and strategic planning using game theory insights about optimal tactical system selection given current squad composition.

Additional Comments / Roadblocks

Technical Challenges: Entity resolution across different data sources may require significant preprocessing, and API rate limits could constrain data collection timeline. Graph construction complexity may require iterative refinement of schema design. **Data Quality:** Social media data may contain noise requiring robust filtering mechanisms, and ensuring balanced representation across different teams and leagues. **Computational Requirements:** Knowledge graph storage and querying may require significant computational resources, potentially necessitating cloud infrastructure. **Timeline Considerations:** The project scope is ambitious and may require prioritization of core components if time constraints emerge.

Main Data Sources:

- **FBref.com** - Free tier with registration, provides comprehensive match data, team statistics, and league information
- **Twitter API v2** - Free tier available with developer account for social media sentiment

- https://www.sportmonks.com/football-api/ API - Free data base with rating and comprehensive
- API_Football "https://dashboard.api-football.com" - PAID Comprehensive soccer data API providing real-time and historical match data

References:

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6, 236. https://doi.org/10.1038/s41597-019-0247-7

I assert that this proposal is the result of my own work, and all applicable sources have been properly cited and acknowledged. I understand that my capstone project will be reviewed by the Program Director as part of the approval process and successful completion of this project requires approval from the Program Director and/or Course Instructor. In addition, I understand that I am strongly encouraged to submit my finished project for publication.

Student Signature	**Date**
Maria Mora Mora	**06/27/2025**
Student Signature	**Date**
Gabriel Mancillas Gallardo	**06/30/2025**
Student Signature (if applicable)	**Date**
_____	_____