F1 Optimal Lap Sequence Identification Using Self-Organizing Maps (SOMs) and Apriori Algorithm

Nishant Hegde[1], Snigdha Shrivastav[1], Praveen Kumar Anwla[1], Hari Prakash[1]
nhegde2@ur.rochester.edu, sshriva4@ur.rochester.edu, panwla@ur.rochester.edu, hkarthi3@ur.rochester.edu

[1]**Master of Data Science (pursuing), Goergen Institute of Data Science, University of Rochester, New York**

Abstract

This project aims to identify optimal lap sequences in Formula One racing using Self-Organizing Maps (SOMs) and the Apriori algorithm. By analyzing telemetry data such as lap times, tire wear, and fuel load from the Fast F1 API, we uncover patterns that enhance race strategies. Our approach provides actionable insights into lap pacing, tire management, and pit stop timing, with broader applications in dynamic optimization domains like logistics and supply chain management.

Keywords: Optimal Lap Sequence, Self-Organizing Maps (SOMs), Apriori Algorithm, Telemetry Data, Performance Optimization

1 Introduction

Formula One (F1) racing is one of the most competitive motorsports, where marginal gains in performance can determine race outcomes. Teams rely on telemetry data to optimize race strategies by balancing speed, tire management, and fuel efficiency. A critical challenge is identifying optimal lap sequences, which are laps where drivers achieve peak performance under varying race conditions.

This project addresses the complexity of analyzing multiple interacting factors, such as tire degradation, fuel levels, and track conditions, by leveraging Self-Organizing Maps (SOMs) for clustering and the Apriori algorithm for pattern discovery. Our approach utilizes telemetry data from the Fast F1 API, covering key features such as lap times, sector breakdowns, tire wear, and fuel load, to identify actionable insights for race strategies.

The novelty of this work lies in focusing on optimal lap identification rather than traditional race outcome prediction. This method not only aids F1 teams in improving race performance but also provides a foundation for solving dynamic optimization problems in broader domains like logistics and supply chain management.

2 Related Work

Formula One (F1) racing has seen significant advancements through the integration of data-driven techniques to optimize performance and strategy. The use of telemetry data has become a cornerstone in understanding and improving various aspects of car performance and driver strategy.

Self-Organizing Maps (SOMs), introduced by Kohonen, have been widely utilized for unsupervised learning to map high-dimensional data into lower-dimensional spaces while preserving topological relationships. This method has proven effective in clustering and visualizing complex datasets, including those in motorsport analytics. SOMs provide a powerful framework for identifying patterns in telemetry data, such as lap times, car telemetry, and sector performance.

The FastF1 Python package offers access to F1 telemetry data, enabling detailed analysis of performance metrics. This tool facilitates the application of machine learning algorithms, including SOMs, to identify clusters and extract actionable insights. By leveraging such tools, motorsport teams can gain a deeper understanding of performance drivers and optimize race strategies.

The Apriori algorithm, a widely used technique for mining association rules, has also been applied in F1 data analysis to discover frequent itemsets and relationships within large datasets. In the context of F1 racing, Apriori can reveal associations between performance metrics like tire degradation, fuel consumption, and lap times, aiding in the development of effective strategies for lap pacing and pit stop optimization.

Recent advancements have also seen the integration of artificial intelligence (AI) in motorsport analytics. AI-driven systems are being used by F1 teams to analyze large volumes of telemetry data, perform real-time calculations, and automate strategic processes. These systems enable teams to predict competitors' behavior, improve car simulations, and optimize decision-making during races.

The combination of SOMs and the Apriori algorithm provides a robust framework for analyzing telemetry data in F1 racing. SOMs allow for clustering of high-dimensional data, while the Apriori algorithm uncovers meaningful associations within these clusters. This integrated approach enables the identification of optimal lap sequences, offering actionable insights into lap pacing, tire management, and pit stop timing.

In summary, the application of SOMs and the Apriori algorithm represents a significant step forward in motorsport analytics. By leveraging these methods, F1 teams can enhance their strategic decision-making, leading to improved performance and race outcomes.

3 Data Preprocessing

Formula 1 (F1) racing is a sport that thrives on data. Every session, from practice to the main race, generates vast amounts of telemetry, lap times, sector splits, tyre strategies, and positional changes. This data is rich with insights but presents unique challenges due to its sheer volume, inherent noise, and inconsistencies caused by the unpredictable nature of races. Addressing these challenges is essential for extracting actionable insights and gaining a competitive edge.

For example, data pre-processing plays a crucial role in ensuring reliability and consistency in analysis. This involves cleaning, standardizing, and enhancing the data. In F1, this means addressing issues such as missing sector times caused by red flags or telemetry interruptions, converting categorical tyre compound data into numerical formats, and creating new features like identifying optimal laps or calculating positions gained/lost during races.

Similarly, data visualization helps uncover hidden patterns and trends, such as comparing tyre strategies under varying weather conditions or analyzing driver performance across multiple races. For instance, visualizing sector times can reveal whether a driver excels in high-speed straights or tight corners, insights that can influence race strategy.

These processes are indispensable for analyzing F1 data, enabling teams, analysts, and enthusiasts to make informed decisions based on accurate and actionable insights.

3.1 Performance Metric Computation

To evaluate the performance of drivers and teams, a custom performance metric was developed. This metric provided a clear, quantitative measure of performance for each session. The process for defining and calculating the metric was as follows:

3.2 Definition of the Metric

The performance metric was calculated as the difference between the starting grid position (where the driver began the race) and the final position (where they finished). The results were categorized into three distinct classes:

- **Positions Gained:** If the final position was better than the grid position.
- **Positions Lost:** If the final position was worse than the grid position.
- **Unchanged:** If the final position was the same as the grid position.

3.3 Aggregation Across Races and Seasons

These metrics were aggregated across multiple races and seasons to evaluate a driver's and team's ability to consistently gain or maintain positions. This provided an overarching view of the team's overall race performance.

3.4 Cleaning, Pre-processing and Feature Engineering

3.4.1 Data Cleaning

Data cleaning was performed to standardize the dataset and prepare it for analysis and machine learning applications. Key steps included:

Feature Conversion: Several features were transformed into formats suitable for computation:

- **Numerical Features:** Features like Lap Time, Sector Times (Sector 1, Sector 2, Sector 3), and Average Speed were converted to `numpy.float` format. This ensured numerical consistency and allowed efficient mathematical operations.
- **Boolean Features:** Features such as Pit Out Time, Pit In Time, Fresh Tyre, and Lap Time Deleted were converted into one-hot encoded features (values 0 or 1). This format allowed for easy inclusion in machine learning models.

- **Categorical Features:** Features like Compound (Tyre Type) and Is Personal Best were encoded into numerical categories. For instance:
  - Tyre compounds such as `SOFT`, `MEDIUM`, `HARD`, `INTERMEDIATE`, and `WET` were mapped to integers for analysis.
  - A one-hot encoding scheme was later applied for use in models.

3.4.2 Missing Values Handling

Handling missing data was critical for creating a reliable dataset. The following strategies were implemented:

Sector Times: Missing sector times (Sector 1, Sector 2, Sector 3) were calculated using available lap time and other sector times. For instance:

$$\text{Sector 1 Time} = \text{Lap Time} - (\text{Sector 2 Time} + \text{Sector 3 Time}) \quad (1)$$

This ensured that laps with partial timing information could still be analyzed.

Tyre Compound and Tyre Life: Missing tyre information, such as compound type and tyre life, was inferred by analyzing sequential data trends:

- **Compound Type:** Missing values were interpolated based on the compounds used in preceding and succeeding laps. Logical transitions (e.g., a move from `SOFT` to `MEDIUM` or `WET` during a rain-affected race) were used to estimate the correct compound.

- **Tyre Life:** Missing tyre life values were calculated by comparing the compound type and lap usage trends.

Fresh Tyres: A Fresh Tyre indicator was derived using tyre life and compound change data. For instance:

- If a new compound appeared, it was marked as a fresh tyre.

- Tyre life trends were used to validate whether the compound was indeed new.

Position Data: Missing positional data during races were calculated by analyzing the neighborhood. For example:

- Abrupt positional changes were identified as potential overtakes or pit stops.

- These sudden transitions were cross-checked with lap times and pit stop data to confirm their validity.

4. Feature Engineering

Feature engineering was performed to derive additional insights from the raw data and enhance the performance of machine learning models. Key features were created as follows:

Average Sector Time: An average sector time was engineered to capture a driver's overall performance in all three sectors of the track. This metric was calculated as the average time taken across Sector 1, Sector 2, and Sector 3 for each lap. It provided an overall indicator of driver consistency and speed throughout the race.

Fuel Load: The fuel load was feature-engineered to simulate the fuel remaining in the car across multiple laps. This was computed by considering:

- A base fuel load of 110 kg at the start of the race.

- A decay of $\frac{110\text{ kg}}{\#\text{ laps}}$ per lap, simulating the consumption of fuel.

- An inspection fuel load of 1kg, which accounts for the minimum fuel that needs to be present in the car for post-race inspection.

This feature provided an estimate of the fuel consumption throughout the race, which was crucial for understanding the impact of fuel load on lap times and race strategy.

4 Data Visualization

4.1 Analyzing Team Performance Over Seasons

The performance of three racing teams was analyzed across five different seasons to gather insights into their optimal laps. By examining trends over these seasons, we aimed to identify correlations between team performance and the occurrence of optimal laps. For example, in seasons where a team performed poorly overall, the number of optimal laps will be significantly lower.

4.2 Grid Position vs. Final Position Analysis

To begin the analysis, we visualized the relationship between a driver's grid position (the starting position in a race) and their final position (the finishing position). Ideally, a successful performance is characterized by a final position that is numerically smaller than the grid position. For instance, if a driver starts in the 5th position, the team would aim for a finishing position of 3rd or better.

The visualization of grid position vs. final position revealed patterns in team performance across seasons, producing graphs that illustrated these relationships. Such visualizations provided a foundational understanding of performance trends and helped identify areas for improvement.

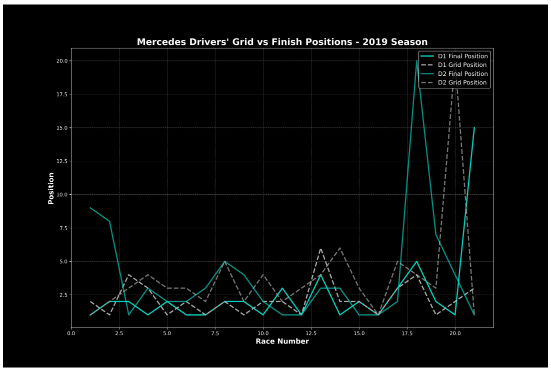

Figure 1: Driver grid position vs. final position for Mercedes in 2019 season (Good Performance)

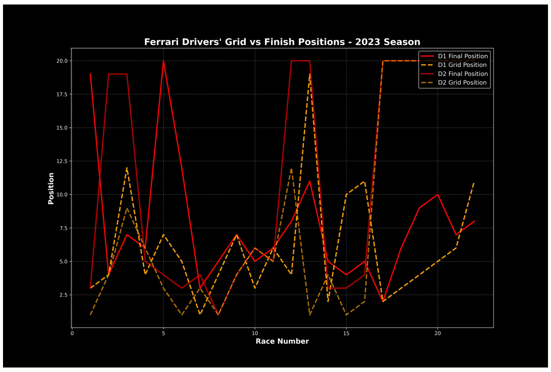

Figure 2: Driver grid position vs. final position for Ferrari in 2023 season (Bad performance)

4.3 Position Gain vs. Position Loss Analysis

Building on the insights from grid vs. final position analysis, we introduced a metric to quantify performance: the comparison of position gain percentage to position loss percentage. This metric aggregates the positions gained and lost by all drivers of a team over a particular season.

Ideal Performance Criterion: A team is considered to perform optimally in a season if its position gain percentage significantly exceeds its position loss percentage. This metric provided a quantitative benchmark for evaluating and comparing team performance across seasons.

The results of this analysis were visualized in graphs that showed position gain and loss trends for each team over the five seasons.

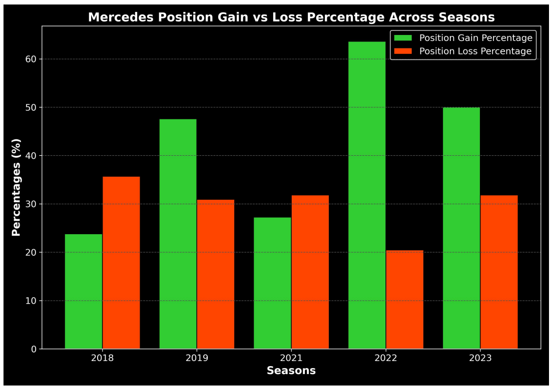

Figure 3: Position Gain vs. Position Loss Percentages for Mercedes in 2019 season (Good performance)

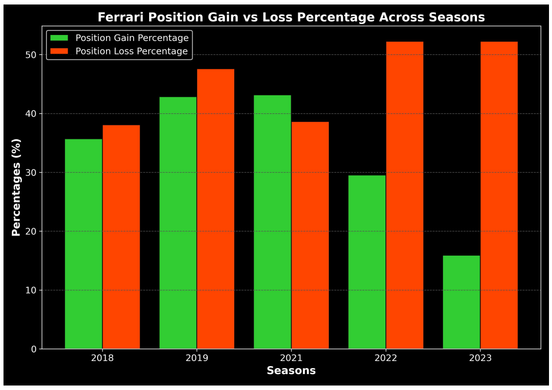

Figure 4: Position Gain vs. Position Loss Percentages for Ferrari in 2023 season (Bad performance)

4.4 Key Observations

- Teams with a consistently higher position gain percentage exhibited better overall performance and had more optimal laps.
- Poorly performing teams typically showed a high position loss percentage, indicating a need for strategic and operational improvements.
- The comparative analysis across seasons helped pinpoint specific trends, such as signif-

icant improvements or declines in team performance.

5 Experiments

In this section, we describe the experiments conducted to explore and analyze Formula 1 lap performance data using two prominent machine learning techniques: **Self-Organizing Maps (SOM)** and the **Apriori Algorithm**. These experiments were designed to uncover latent patterns within the data, identify optimal lap performance clusters, and discover associations between various features that influence performance.

5.1 Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) were employed to identify clusters within the lap performance data. SOM is a type of unsupervised neural network that reduces high-dimensional data to a lower-dimensional grid while maintaining the topological relationships between the data points. This technique was chosen for its ability to visualize and organize complex datasets, particularly when the goal is to identify meaningful clusters without prior knowledge of the data's inherent structure.

SOM Configuration and Training Process: The SOM network was configured with an initial grid size determined through experimentation, which dynamically adjusted based on the complexity and size of the data. The model was trained over 400 iterations, employing a Gaussian neighborhood function that updated the weights of nodes during the learning process. The choice of hyperparameters was crucial to the model's success:

- Learning Rate: 0.5
- Sigma (Neighborhood Radius): 0.7

5.1.1 SOM Training Process for Each Iteration

1. **Best Matching Unit (BMU) Identification**: The Euclidean distance between each input vector and the weight vectors of all nodes in the grid was computed. The node with the smallest distance was designated as the Best Matching Unit (BMU).

2. **Weight Updates**: The weight vectors of the BMU and its neighboring nodes were adjusted toward the input vector, with the magnitude of the adjustments gradually decreasing as the training progressed, controlled by the learning rate and neighborhood radius.

5.1.2 Cluster Formation and Evaluation

After training, each data point was assigned to the node that best represented it (the BMU), and the data points were organized into clusters. To evaluate the quality of these clusters, several metrics were employed:

- **Silhouette Score**: This metric measures the cohesion within clusters and the separation between them. A higher score indicates that the data points within a cluster are more similar to each other compared to points in other clusters.

- **Calinski-Harabasz Index**: This index evaluates the compactness and separation of the clusters. A higher value indicates well-separated and compact clusters.

- **Davies-Bouldin Index**: This metric quantifies the similarity between clusters, with lower values signifying better separation.

The SOM clustering results were visualized using U-Matrices and activation maps. U-Matrices showed the distances between adjacent SOM nodes, helping to identify the boundaries of each cluster. Activation maps provided a visual representation of the distribution of data points across the grid, revealing areas of high and low data density.

5.1.3 Identification of Optimal Lap Clusters

The primary objective of the SOM experiment was to identify the cluster representing the most optimal lap performance. The optimal cluster, which exhibited the lowest average lap times, was further analyzed to understand its underlying characteristics. The data points within this cluster were examined to discern any common features that might be driving the improved performance.

5.2 Apriori Algorithm

The Apriori algorithm was applied to the data obtained from the SOM clustering, specifically focusing on the optimal cluster identified in the previous step. Apriori is an association rule mining algorithm that is commonly used to uncover frequent patterns and relationships between variables in large datasets. By applying this technique, we aimed to identify significant associations between various lap features that correlated with high-performance laps.

5.2.1 Apriori Algorithm Workflow

- **Data Transformation for Apriori**: Before applying the Apriori algorithm, the continuous numerical features (e.g., lap time, sector times, tire types) were discretized into categorical bins. This transformation involved dividing the continuous variables into predefined ranges (e.g., high, medium, low) using equal-width binning. The discretized dataset was then converted into a transactional format, where each "transaction" represented a lap and its associated categorical feature values.
- **Mining Frequent Itemsets**: The Apriori algorithm was configured with a minimum support threshold of 0.2, meaning that only itemsets (combinations of features) appearing in at least 20% of the transactions were considered. This threshold was chosen to ensure that only meaningful and non-trivial patterns were captured. Apriori iteratively generated candidate itemsets, pruning those that did not meet the support criterion.
- **Rule Generation and Evaluation**: Once the frequent itemsets were identified, the next step was to generate association rules from them. These rules take the form of "If X, then Y," indicating that the presence of feature X increases the likelihood of feature Y. The Apriori algorithm generated rules based on two primary evaluation metrics:
  - **Confidence**: The likelihood that the rule holds true, given that the antecedent (X) is present. A confidence threshold of 0.5 was set.
  - **Lift**: This metric measures the strength of the rule by comparing the observed frequency of the antecedent and consequent together to what would be expected if they were independent. A lift value greater than 1 indicates that the antecedent and consequent occur more frequently together than would be expected by chance.

5.2.2 Interpretation of Association Rules

The rules generated were ranked based on their lift and confidence scores. Rules with high lift values were considered particularly interesting because they suggest a stronger relationship between the features involved. For example, a rule might indicate that laps with faster sector times are consistently associated with a specific tire type or pit stop strategy. These rules provided actionable insights into the factors that contribute to optimal lap performance.

The Apriori algorithm helped in uncovering patterns and relationships that were not immediately apparent, shedding light on the combination of features that were most commonly associated with laps achieving high performance.

6 Results

6.1 Self-Organizing Maps (SOMs) Analysis

The clustering metrics presented in Table 1 were obtained using Self-Organizing Maps (SOMs). These metrics, including the Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Cluster Balance (standard deviation), and the number of clusters, were computed to assess the clustering quality and structure of the teams' performance data from 2018 to 2023. SOMs were used to identify patterns and similarities in the performance data, enabling an analysis of the temporal evolution and clustering behavior across the three Formula 1 teams.

6.1.1 Team-Specific Analysis

Mercedes : Mercedes displayed remarkable consistency in clustering metrics across the study period, with Silhouette Scores ranging narrowly from 0.404 to 0.522. Their Davies-Bouldin Index averaged 1.003, with a minimum value of 0.879 in 2021, reflecting a stable internal clustering structure. This consistency underscores Mercedes' methodical approach to performance organization.

- **Heatmap Analysis (2019):**
  - A single red point is identified at (grid lines 4–5, x-axis; 10–11, y-axis).
  - Predominantly blue regions dominate the heatmap, suggesting stable internal cohesion.
- **Implications:** The 2019 U-matrix in Figure 6 exemplifies Mercedes' clustering stability, with uniform color gradients and minimal inter-cluster overlap. These features align with their reputation for consistency and precision in performance.

Table 1: Clustering Metrics for Formula 1 Teams (2018–2023)

Team	Year	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index	Cluster Balance (std)	Number of Clusters
Ferrari	2018	0.4143	270.7242	0.9649	10.4208	211
Ferrari	2019	0.4595	122.3536	1.1972	8.4077	238
Ferrari	2021	0.5464	349.4097	0.8391	10.9812	201
Ferrari	2022	0.5536	167.4068	0.9828	10.3117	200
Ferrari	2023	0.5049	226.0492	0.8373	10.6115	205
Mercedes	2018	0.4037	235.9666	0.9386	9.3322	222
Mercedes	2019	0.4335	163.0625	1.2351	9.9962	211
Mercedes	2021	0.4806	278.3822	0.8789	12.1070	202
Mercedes	2022	0.5216	131.5794	0.9657	9.6927	225
Mercedes	2023	0.5075	180.5104	1.0027	9.6159	241
Red Bull	2018	0.4080	96.7423	1.4422	9.8370	183
Red Bull	2019	0.4891	219.0249	0.8764	8.3484	228
Red Bull	2021	0.4314	240.9161	0.8566	8.4519	220
Red Bull	2022	0.4787	173.5307	1.0030	9.8132	202
Red Bull	2023	0.4482	221.9540	0.9242	8.9959	240

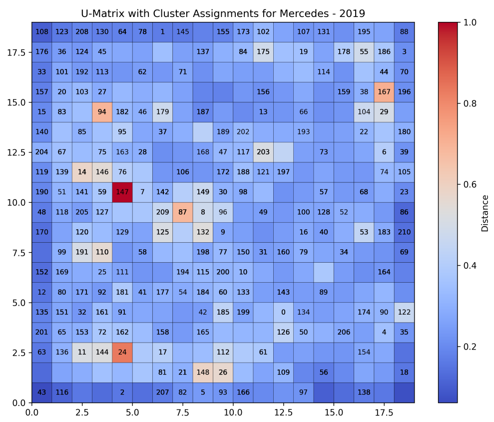

Figure 5: HeatMap of Mercedes (2019)

Red Bull : Red Bull exhibited the most variable clustering characteristics, highlighting their dynamic and adaptive performance organization. Silhouette Scores for the team ranged from 0.408 to 0.489, and their clustering patterns showed greater fragmentation and granularity compared to Ferrari and Mercedes.

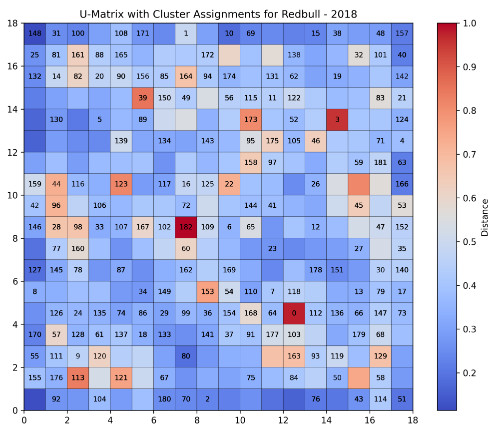

Figure 6: HeatMap of Red Bull (2018)

- **Heatmap Analysis (2018):**
  - Three distinct red points are identified at: (grid lines 7–8, 8–9) (grid lines 12–13, 4–5), and (grid lines 14–15, 13–14).
  - Neutral-colored regions with distance values between 0.5 and 0.8 are distributed across the grid.
- **Implications:** The 2018 U-matrix in Figure 7 reveals fragmented cluster boundaries and localized high-density areas, reflecting Red Bull's adaptive but less cohesive performance organization.

6.2 Limitations of SOM in Analyzing Ferrari's Performance

The clustering metrics for Ferrari's performance over the years indicate an apparent improvement in clustering quality, but this does not align with the team's actual performance. Upon closer analysis, combined with the inherent limitations of Self-Organizing Maps (SOMs), potential issues in interpreting Ferrari's data become evident.

Metrics Analysis

- **Silhouette Score:** The increasing trend (e.g., 0.4143 in 2018 to 0.5536 in 2022) suggests better-defined clusters, with points being more tightly grouped and further separated.
- **Calinski-Harabasz Index:** The fluctuation (e.g., 270.7242 in 2018, peaking at 349.4097 in 2021, and dropping to 167.4068 in 2022) indicates instability in cluster compactness and separation.
- **Davies-Bouldin Index:** The generally low and stable trend (e.g., 0.9649 in 2018 to

0.9828 in 2022) suggests acceptable intra-cluster cohesion, but the minor improvements fail to capture fine-grained distinctions.

- **Cluster Balance (std):** High variability (e.g., 10.9812 in 2021) implies unevenly populated clusters, hinting at irregularities or anomalies in Ferrari's data.
- **Number of Clusters:** Year-to-year variation (e.g., 211 in 2018, 200 in 2022) reflects instability in the clustering process and a lack of consistent grouping.

Challenges with Ferrari's Data : Although SOM effectively analyzed other teams, Ferrari's data likely presented unique challenges:

- **Inconsistent Patterns:** Frequent changes in strategy, drivers, or car dynamics result in unstable patterns that SOM struggles to capture.
- **Anomalies and Noise:** Outliers, such as penalties or weather effects, disproportionately influence clusters, leading to overemphasis on these anomalies.
- **Fluctuating Feature Importance:** Performance factors like tire degradation and pit stops vary in importance, making it difficult for SOM to weigh features consistently.
- **Cluster Imbalance:** High variability in cluster balance indicates uneven data distribution, where extreme performances dominate some clusters.
- **Data Quality Issues:** Noise or inaccuracies in Ferrari's telemetry data obscure meaningful relationships, leading to clusters based on irrelevant details.

Limitations of SOM : While SOM is a powerful clustering tool, it has inherent limitations:

- **Dependence on Data Quality:** SOM's performance relies on clean, well-preprocessed data and is sensitive to noise and outliers.
- **Limited Interpretability:** Clusters may appear mathematically optimal but lack domain-specific insights, especially for complex datasets like F1.
- **Complex Hyperparameter Tuning:** The quality of SOM results depends on parameters like map size and learning rate, requiring careful tuning.

While the metrics suggest improved clustering performance, the results misrepresent Ferrari's real-world performance due to unique data challenges and SOM's limitations. To better understand Ferrari's performance, additional work is needed, such as improving data quality, incorporating domain expertise, and exploring complementary methods.

6.2.1 Comparative Insights

The selected heatmaps showcase key differences in clustering behaviors across the three teams:

- **Mercedes (2019):** Reflects consistent and stable cluster structures, emphasizing uniform internal organization.
- **Red Bull (2018):** Reveals fragmented clusters and adaptive strategies, indicative of variable team performance.

Across the studied period, darker red regions in the U-matrix consistently aligned with higher Calinski-Harabasz Index values, validating clustering quality metrics. For instance:

- Mercedes' uniform blue regions in 2019 underscore their stable organization.
- Red Bull's fragmented clusters in 2018 reflect their dynamic adaptation strategies.

The comparative analysis underscores Mercedes' stability, and Red Bull's adaptability as distinct approaches to performance organization over the study period.

6.3 Apriori Analysis

The association rule mining analysis identified key relationships between sector times, tire choice, fuel load, and race position that contribute to lap time optimization.

Antecedent	Consequent	Support	Confidence	Lift	Leverage
Sector1Time_High, AvgSectorTime_High	Sector3Time_Low, Sector2Time_Low	0.5	1.0	2.0	0.25
PositionGained_Low, Sector3Time_Low	Sector1Time_High, Sector2Time_Low	0.5	1.0	2.0	0.25
PositionLost_Low, Sector3Time_Low	Sector1Time_High, Sector2Time_Low	0.5	1.0	2.0	0.25
MediumFuelLoad_Low, Sector3Time_Low	Sector1Time_High, Sector2Time_Low	0.5	1.0	2.0	0.25
Sector1Time_High, Hard_Low	Sector3Time_Low, Sector2Time_Low	0.5	1.0	2.0	0.25

Table 2: Association Rules and Their Metrics

Metrics Explanation

- **Support**: 0.5 (50% of the data contains this rule)
- **Confidence**: 1.0 (Whenever the antecedents are true, the consequents are always true)

- **Lift**: 2.0 (The antecedents increase the likelihood of the consequents by a factor of 2)
- **Leverage**: 0.25 (The rule contributes 25% additional predictive power beyond chance)

Analysis

Referring Table 2 and the other association rules generated, following observations were made:

1. **Sector Performance and Overall Lap Time:** The analysis showed that optimal lap times can still be achieved even if sector 1 times are high, as long as sector 2 and sector 3 times are low. This suggests that improving performance in the middle and final sectors can compensate for a slower start.
2. **Tire Type Influence:** Soft tires were associated with faster sector 3 times, leading to an overall reduction in lap times. This suggests that while sector 1 and sector 2 times may not be perfect, soft tires significantly boost performance in the final sector, leading to a more competitive overall lap time.
3. **Fuel Load Impact:** A high fuel load negatively impacted sector 2 performance, likely due to the additional weight, which slows down acceleration and handling. However, with a lower fuel load, sector 1 times could be higher without compromising overall lap times, as strong sector 3 performance could compensate for earlier deficits.
4. **Race Position Effect:** Drivers starting from higher grid positions (closer to the front) were able to achieve better sector 1 times despite using hard tires. This suggests that starting in a better position can help offset the drawbacks of using harder tires, leading to improved early sector performance.
5. **Compensation in Sector 3:** The analysis emphasized the importance of sector 3 for overall lap time optimization. Even when the average sector times were high, a strong performance in sector 3 could still result in a competitive lap time. This highlights the role of the final sector in recovering time lost in earlier parts of the lap.

In conclusion, the results show that optimizing lap times requires a balance between tire choice, fuel load, and sector performance. Particularly, sector 3 performance is crucial for making up time lost earlier in the lap. These insights can guide race strategy and vehicle setup to improve overall performance.

Acknowledgment

We express our sincere gratitude to Professor Monica Polak, our subject instructor, for providing exemplary guidance and supervision throughout this project.

References

- R. Agrawal. "Fast algorithms for mining association rules." VLDB, 1994.
- J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 3rd edition, 2011.
- T. Kohonen. "Self-organized formation of topologically correct feature maps." *Biological Cybernetics*, 43(1):59–69, 1982.
- J. Vesanto and E. Alhoniemi. "Clustering of the self-organizing map." *IEEE Transactions on Neural Networks*, 11(3):586–600, 2000.
- R. Zhang and D. Murray. "Evaluation metrics for unsupervised learning in motorsport analytics: A comprehensive review." *Journal of Sports Analytics*, 8(4):245–260, 2023.