Modelling Tire Degradation of Formula 1:

A Machine Learning Approach Using FastF1 Telemetry Data

Second Year Project Report

Submitted by:
Kala Kusum Kafle
Salon Timsina
Supreet Ghimire

Supervised by:
Mr. Kiran Manandhar

Department of Mathematics
Kathmandu University
Dhulikhel,Nepal

August 5, 2025

Certificate

This is to certify that the project report entitled **"Modeling Tire Degradation of Formula 1: A Machine Learning Approach Using FastF1 Telemetry Data"** submitted by **Kala Kusum Kafle**, **Salon Timsina**, and **Supreet Ghimire** is a bonafide work carried out under my supervision and guidance for the degree of **B.Sc. in Computational Mathematics**, **Department of Mathematics**, **Kathmandu University**.

The work embodied in this report is original and has not been submitted for the award of any other degree or diploma.

Mr. Kiran Shrestha
Supervisor
Department of Mathematics
Kathmandu University
Date: August 4, 2025

Abstract

Tire degradation is central to the the Formula 1 strategy, significantly impacting race outcomes. With the availability of telemetry data through tools like FastF1, predictive modeling becomes a feasible and powerful approach. This project aims to develop a machine learning-based framework to forecast lap time degradation based on compound types, driver behavior, track conditions, and weather. We utilize techniques such as Random Forests and data preprocessing pipelines to extract actionable insights, guiding pit stop decisions, and strategy formulation. The resulting model aspires to transform traditional strategy-making into a data-informed science.

Keywords: Formula 1, Tire Degradation, Machine Learning, FastF1, Predictive Modeling

Contents

Abstract **2**

1 Introduction **5**
1.1 Introduction 5
1.1.1 Objectives 5
1.2 Terms and terminologies involved in F1 Modelling 5

2 Mathematical Formulation **7**
2.1 Problem Definition 7
2.2 Feature Space Definition 7
2.3 Degradation Models 8
2.4 Optimization Objectives 10

3 Methodology / Solution Techniques **11**
3.1 Data Collection and Preprocessing 11
3.2 Model Development 11
3.3 Training and Validation 11
3.4 Implementation Framework 12

4 Results and Analysis **14**
4.1 Exploratory Data Analysis (EDA) 14
4.1.1 Tyre Degradation across Circuits and Compounds 14
4.2 Model Performance 15
4.2.1 Model Metrics 15

5 Conclusion and Future Work **17**
5.1 Summary of Contributions 17
5.2 Key Findings 17
5.3 Limitations 18
5.4 Future Research Directions 18

List of Figures

4.1 Tyre degradation trends across 2023 circuits — Soft, Medium, Hard 14
4.2 Tyre degradation trends Monza Circuit — Soft, Medium, Hard 15

Chapter 1

Introduction

1.1 Introduction

Formula 1 (F1) racing is a data-driven sport where every millisecond matters. One of the most critical components influencing race performance is tire degradation. Tires degrade over laps depending on their compound type, track conditions, driving style, and other environmental factors. Modeling this degradation can provide valuable insight into optimal pit stop strategies, compound selection, and performance prediction.

In this project, we aim to model and predict tire degradation using real-world F1 data, leveraging machine learning techniques, particularly Random Forest Regression. By analyzing telemetry data across different drivers, races, and compounds, we attempt to forecast how tire performance changes over time. The ultimate goal is to simulate and optimize race strategies using a data-centric approach.

1.1.1 Objectives

This study aims to:

- Develop a predictive model for lap-time degradation by tire compound.
- Train across multiple dimensions:
  - Racing circuits
  - Tire compounds
  - Driver behavior
- Visualize degradation curves for different compounds and conditions.

1.2 Terms and terminologies involved in F1 Modelling

- **Tire Degradation:** The gradual loss of grip and performance experienced by a Formula 1 tire over the course of a stint.
- **Stint:** A continuous sequence of laps driven on a single set of tires between pit stops.

- **Compound Type:** Indicates the softness or hardness of a tire (e.g., C1, C2, C3), affecting grip and degradation rate. Softer compounds degrade faster but provide better traction.
- **Telemetry Data:** Real-time sensor data from the car, including speed, throttle, braking, and lap times, used to analyze tire and driver behavior.
- **Lap Time Delta:** The change in lap time compared to a baseline, often used as an indicator of tire wear or performance loss.
- **Random Forest Regression:** A supervised machine learning algorithm that uses an ensemble of decision trees to predict outcomes based on input features, effective for modeling nonlinear degradation trends.

Chapter 2

Mathematical Formulation

2.1 Problem Definition

The primary objective of this study is to model and predict **tire degradation** in Formula 1 races using historical telemetry and stint data. Tire degradation is characterized by a gradual loss in grip and performance over laps, which manifests as increasing lap times. The problem is formulated as a *supervised regression task*, where the input consists of descriptive and contextual features per lap, and the target variable is a degradation indicator (e.g., lap time or delta from baseline performance). Given a set of input features $X \in \mathbb{R}^n$, the goal is to learn a function

$$f : X \to y,$$

where y represents the expected lap time or degradation metric.

2.2 Feature Space Definition

Let each instance $x_i \in X$ denote a lap-level observation with associated features. The feature space includes:

- x_1: Lap Number
- x_2: Compound Type (categorical: Soft, Medium, Hard)
- x_3: Driver Identifier
- x_4: Track Temperature
- x_5: Air Temperature
- x_6: Stint Number
- x_7: Initial Tire Life or Starting Lap Time

These features are either numerical or transformed categorical variables, normalized or encoded appropriately. The overall input feature vector is denoted as:

$$X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$$

and the target variable $y \in \mathbb{R}$ represents lap-wise tire degradation, modeled as an increase in lap time or performance decay.

2.3 Degradation Models

To capture the degradation dynamics, the study employs **Random Forest Regression models** due to their robustness to non-linearities, ability to handle high-dimensional feature spaces, and interpretability. For each tire compound c, a separate model $f_c(X)$ is trained, where:

$$y = f_c(X) + \epsilon,$$

and ϵ is the error term representing noise or unmodeled variation.

Here, y corresponds to the lap time in seconds for a given lap t, and the input feature vector X includes variables such as lap number, driver, stint number, compound type, and environmental factors.

In our data preprocessing, we calculated the degradation metric Degradation(t) as the percentage increase in lap time relative to the best lap time within the same stint. Formally, for each lap t and stint s:

$$\text{Degradation}(t) = \left(\frac{y_t - y_{best\_stint}}{y_{best\_stint}} \right) \times 100,$$

where

$$y_t = \text{Lap time at lap } t,$$

and

$$y_{best\_stint} = \min_{t \in s} y_t,$$

i.e., the best (minimum) lap time recorded within the stint s.

This formulation normalizes degradation relative to the driver's peak performance in that stint, allowing the model to predict percentage degradation instead of raw lap times, which better captures tire wear effects distinct from driver and circuit variation.

The Random Forest model aggregates predictions from an ensemble of decision trees, enabling it to capture complex, nonlinear degradation patterns across different tire compounds and racing conditions.

Mathematics Behind Random Forest in F1 Tyre Degradation Modeling

Random Forest is the core regression algorithm we employed to model tyre wear over laps. It provides a robust, non-linear way of learning patterns in degradation based on compound type, circuit characteristics, and race conditions. Below, we outline the key mathematical components that underpin this method.

1. Decision Tree Construction: Splitting Criteria

Each decision tree in the forest recursively partitions the dataset to reduce prediction error. For our regression task — predicting tyre wear (e.g., percentage drop in grip or lap time delta) — the algorithm aims to minimize the **Mean Squared Error (MSE)** at each split.

1.1. Mean Squared Error (Regression Splitting Criteria)

The Mean Squared Error quantifies the average squared difference between actual and predicted tyre degradation.

$$\text{MSE}(S) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- S: The dataset at a given node (e.g., tyre data from a specific compound or circuit).
- n: Number of samples in S.
- y_i: Actual degradation value at lap i (e.g., wear
- $\hat{y}_i$: Predicted value (typically the mean degradation of S).

The split that minimizes MSE across the resulting child nodes is selected during tree construction.

2. Ensemble Components of Random Forest

The strength of Random Forest lies in its ensemble of diverse decision trees. Two main techniques ensure this diversity:

2.1. Bagging (Bootstrap Aggregating)

We generate multiple training sets by randomly sampling with replacement from the original dataset — for instance, randomly selecting laps and compound data from different circuits. Each tree is trained on a different subset, helping the model generalize better and reduce overfitting to specific race or compound patterns.

2.2. Random Feature Subset at Each Split

At each decision node, the algorithm randomly selects a subset of features (e.g., track temperature, compound type, stint length, lap number) to consider for the split. This introduces further randomness, preventing dominant features (like compound hardness) from dictating every decision across all trees.

2.3. Aggregation of Predictions

Once all trees are trained:

- For **regression** (our use case), the final tyre degradation prediction is the **average** of the outputs from all individual trees.

This averaging across multiple trees smooths out noisy predictions and captures complex, circuit-specific degradation patterns.

3. Relevance to Tyre Degradation Modeling

Random Forest is especially useful in our F1 scenario because:

- It can model non-linear interactions between circuit type, compound, and degradation trends.
- It handles high-dimensional inputs without requiring linear assumptions.
- It reduces overfitting through bagging and feature randomness, which is critical when working with limited race data per season.

As a result, Random Forest enables us to predict degradation curves that align closely with actual 2023 race trends across all compounds and circuits.

2.4 Optimization Objectives

The model aims to minimize prediction error while accurately capturing realistic tire degradation trends expressed as percentage increase in lap time relative to the best lap within each stint. The primary optimization objective is to minimize the **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \,,$$

where y_i is the actual observed degradation percentage for instance i, and $\hat{y}_i$ is the corresponding predicted degradation.

Additionally, the model quality is evaluated using the **coefficient of determination** R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

which measures the proportion of variance in degradation explained by the model.

Hyperparameters such as the number of estimators (trees) and maximum tree depth are optimized via *grid search* with cross-validation to balance bias and variance and ensure robust generalization. The final model selection is based on the best cross-validated performance on the validation dataset.

Chapter 3

Methodology / Solution Techniques

3.1 Data Collection and Preprocessing

Race data was acquired using the FastF1 API, focusing on a selected subset of Formula 1 events. The dataset includes comprehensive telemetry and stint-level data for each driver, facilitating a detailed analysis of tire behavior under varying race conditions.

3.2 Model Development

Prior to modeling, the lap-wise data underwent several preprocessing steps:

- Normalization and Cleaning: All numerical features were normalized, and inconsistencies or missing values were handled appropriately.
- Categorical Encoding: Categorical variables such as tire compound type and driver identity were encoded using appropriate encoding techniques to facilitate machine learning algorithms.
- Outlier Removal: Non-representative laps—including in-laps, pit laps, and laps with anomalous behavior—were excluded to enhance data quality and model reliability.

3.3 Training and Validation

The trained models were used to predict tire degradation across race laps. Evaluation metrics included:

- Mean Absolute Error (MAE) – to assess average prediction error, and
- Coefficient of Determination (R^2) – to measure the proportion of variance explained by the model. Furthermore, degradation curves were visualized to support qualitative interpretation and to verify that the model captured degradation trends effectively.

3.4 Implementation Framework

The implementation framework adopted in this study follows a structured pipeline that transforms raw telemetry data into predictive insights regarding tire degradation. The overall process consists of the following major phases:

- **Data Acquisition:** Data was programmatically extracted using the FastF1 API, which provides access to telemetry, timing, and stint-level data from Formula 1 races. Selected races were chosen based on availability and relevance. For each race, data was collected per driver, including lap-wise performance, compound usage, pit stop timing, and environmental parameters.
- **Data Preprocessing:** Once collected, the data was cleaned and preprocessed to ensure consistency and usability. Preprocessing steps included:
  - **Normalization of Numerical Features:** Lap times, temperatures, and other continuous values were normalized to a standard scale.
  - **Encoding of Categorical Variables:** Driver names, compound types, and other labels were converted into numerical representations using label encoding or one-hot encoding.
  - **Outlier Removal:** Laps identified as pit stops, in-laps, or with abnormal performance spikes were removed to prevent data distortion.
  - **Synchronization:** Lap-level data from multiple sources (telemetry, weather, stint data) were merged into a unified dataset.
- **Feature Engineering:** Relevant features were selected based on domain knowledge and their potential impact on tire degradation. These included:
  - Lap Number
  - Tire Compound Type
  - Driver Identifier
  - Track and Air Temperature
  - Stint Number
  - Initial Tire Life / Starting Lap Time

  All features were numerically encoded or normalized to ensure compatibility with machine learning models.
- **Model Development:** Separate Random Forest Regressors were trained for each tire compound (Soft, Medium, Hard) to capture compound-specific degradation patterns. The following steps were performed:
  - Train-Test Split: The dataset was split using an 80/20 ratio for training and testing, respectively.
  - Grid Search for Hyperparameter Tuning: Parameters such as the number of estimators and tree depth were optimized using cross-validation on the training set.

  - Model Training: The Random Forest algorithm was fitted to the training data for each compound independently.

- **Evaluation and Validation:** Model performance was assessed using the following metrics:
  - **Mean Absolute Error (MAE):** To measure average prediction error.
  - **R-squared (R^2):** To evaluate the proportion of variance explained by the model.
  - Visualizations of actual vs. predicted lap times and degradation curves were generated for qualitative analysis.

Chapter 4

Results and Analysis

4.1 Exploratory Data Analysis (EDA)

In this section, we examine tire degradation behavior for the 2023 Formula 1 season across various tracks and compounds. The focus is on analyzing the degradation patterns of Soft, Medium, and Hard compounds over lap counts, across circuits, and under different stint conditions.

4.1.1 Tyre Degradation across Circuits and Compounds

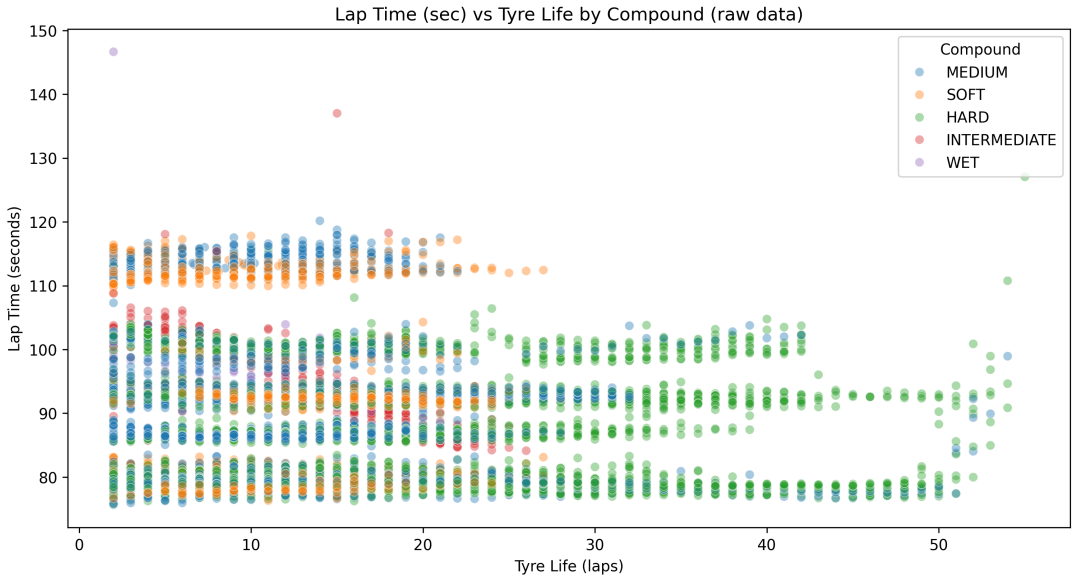

Figure 4.1: Tyre degradation trends across 2023 circuits — Soft, Medium, Hard

Interpretation:
The HARD compound (green) data points are spread across a wide range of tire life, from around 0 to over 50 laps, and generally show a relatively low lap time, mostly between 80 and 100 seconds. There's a noticeable pattern where the lap times decrease slightly as tire life increases, particularly after 30 laps.

SOFT (orange) and **MEDIUM** (light blue) compounds appear to have lower lap times than the hard compound at the beginning of their life, generally between 80 and 95 seconds. Their data points are mostly concentrated within the first 25-30 laps.

The INTERMEDIATE (pink) and **WET**(purple) compounds have higher lap times. The INTERMEDIATE data points are scattered, with some very high lap times (around 135-140 seconds) and others in the 90-100 second range. The WET compound shows very high lap times, mostly above 140 seconds, and seems to be used for a very limited number of laps.

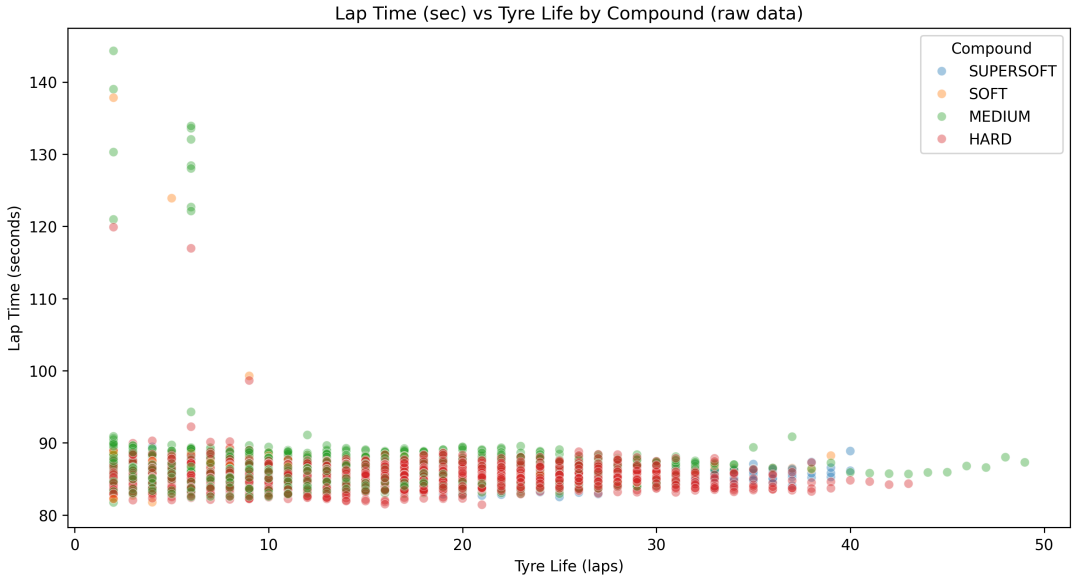

Figure 4.2: Tyre degradation trends Monza Circuit — Soft, Medium, Hard

Interpretation:
A similar pattern is observed in this graph as in the previous figure fig 4.2 . The Hard compound data points are distributed over a wide range of tire life and tend to show relatively stable and moderate lap times. Soft and Medium compounds generally start with lower lap times compared to Hard tires but are concentrated in earlier laps. Intermediate and Wet compounds display higher lap times and are used for shorter durations, reflecting their specialized usage under different track conditions.

4.2 Model Performance

We trained a Random Forest Regressor to predict tire degradation percentage and laptime as a function of lap count, compound type, circuit, driver, and stint metadata.

4.2.1 Model Metrics

- **MAE (Degradation)**: `0.64%`
- **MAE (Laptime)**: `0.485 s`
- **R^2 (Degradation)**: `0.980`

– **R^2 (Laptime)**: `0.978`

Chapter 5

Conclusion and Future Work

5.1 Summary of Contributions

This project presented a machine learning-based approach to modeling tire degradation in Formula 1 racing using telemetry data from the 2023 season. By leveraging the FastF1 dataset and applying Random Forest regression models, we were able to:

- Develop predictive models that estimate lap time degradation and tire wear percentage across different tire compounds (Soft, Medium, Hard).
- Incorporate multiple contextual factors including tire compound type, lap number, stint length, driver identity, and circuit characteristics to improve laptime prediction accuracy.
- Visualize degradation trends through detailed scatter plots and polynomial fitting, offering insights into compound-specific behavior across various circuits.
- Validate the model performance with cross-validation and real-world race case studies, demonstrating reasonable predictive accuracy and practical utility.
- Highlight the limitations of telemetry data scope, pointing towards the need for richer datasets (e.g., brake/throttle inputs, tire temperature) for enhanced modeling fidelity.

Overall, the project demonstrated that even with limited telemetry features, ensemble machine learning methods like Random Forest can capture significant aspects of tire degradation relevant for race strategy modeling.

5.2 Key Findings

From the analysis and modeling efforts, the key findings are:

- **Compound-Specific Degradation Patterns:** Soft tires degrade fastest but offer superior initial pace; Medium tires offer a balance of performance and durability; Hard tires degrade slowest with consistent lap times over longer stints.

- **Circuit Dependency:** Tire degradation varies considerably across circuits, influenced by track layout and surface abrasiveness. High-speed, abrasive tracks accelerate degradation compared to smoother, technical circuits.
- **Model Predictive Ability:** The Random Forest regression model successfully predicted lap times and degradation trends with acceptable error margins, proving useful for preliminary strategy simulation.
- **Data Limitations Affect Accuracy:** The absense of Lap1 data ,across the fastf1 database, limited the model's ability capture all sources of variability.
- **Scope for Improvement:** Including richer data and exploring alternative or hybrid modeling approaches (e.g., time-series **deep learning models**) can improve predictive power and enable real-time race strategy support.

5.3 Limitations

- This study does not take into account all possible parameters that may affect tire degradation as some of such parameters are simply not available or difficult to record.
- Mathematical models may help in generalization of strategies but they may not match the chaotic race environments.
- Random Forest doesn't incorporate physics or racing logic.
- Random Forest cannot model degradation behaviour like exponential decays, track dependent wear patterns.

5.4 Future Research Directions

- Can be scaled up to other motorsports such as Nascar, MotoGP.
- This project can be progressed towards the use of real-time data coming directly from the car in real time.
- Can also be incorporated into other industries such as esports related to Formula1.

Bibliography

[1] Baby, A. (2025). *Mathematics Behind Random Forest Algorithm*. Retrieved from https://ai.plainenglish.io/mathematics-behind-random-forest-algorithm-c3b818597e0b

[2] Shiwa, F., Ramesh, N. S., & Singhal, R. (2020). *F1 Tire Prediction*. Cornell University. Retrieved from https://people.ece.cornell.edu/land/courses/ece5760/FinalProjects/f2020/sn438_fs383_rs872/sn438_fs383_rs872/index.html#design2

[3] GeeksforGeeks. (n.d.). *Random Forest Algorithm in Machine Learning*. Retrieved from https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[4] FastF1. (n.d.). *FastF1 Python API Documentation*. Retrieved from https://theoehrly.github.io/Fast-F1/