

Formula 1 World Driver Champion 2025 Prediction Using XGBoost

Auliya Vishwakarma Hestia¹, Aditiyawan^{1,2}

¹Universitas Multimedia Nusantara

²Badan Riset dan Inovasi Nasional (BRIN)

*Email aditiyawan@lecturer.umn.ac.id

Abstract—Formula 1 (F1) is one of the most prestigious motorsport events in the world, combining speed, strategy, and high technology. Predicting the World Driver Champion (WDC) is an intriguing challenge in the era of modern sports analytics. This research aims to build a model to predict the WDC for the 2025 season using the Extreme Gradient Boosting (XGBoost) algorithm, known for its superiority in handling complex and tabular data. The dataset is obtained from the Jolpi API, covering historical F1 data from the 1950 to 2025 seasons, including race results, qualifying sessions, driver standings, and team data. The data is then processed through merging, cleaning, feature engineering, and transformation stages, and then split into training and testing data. The model is developed and optimally adjusted through hyperparameter tuning using the Optuna library to find the best parameter combinations that increase prediction accuracy. The model evaluation was conducted using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score metrics. The evaluation results for the 2025 season data indicate that the model has an MAE value of 1.0610, an RMSE of 1.3634, and an R^2 score of 0.9441. These results demonstrate that the XGBoost model is capable of predicting drivers' final positions with high accuracy. This research is expected to contribute to the development of data-driven prediction systems for motorsports, particularly Formula 1.

Index Terms—Formula 1, Prediction, World Driver Champion, XGBoost

I. INTRODUCTION

Formula 1 (F1) is the highest class of international motorsport, combining speed, strategy, and cutting-edge technology [16], [18]. Since its inception in 1950, F1 has evolved into one of the most data-driven and technologically complex sports. Modern teams rely on real-time telemetry from thousands of car sensors to monitor tire wear, fuel consumption, aerodynamic performance, and component degradation, enabling optimized decision-making throughout the race [10], [19], [20].

Throughout an F1 season, drivers accumulate points based on their finishing positions in each race [17]. The driver with the highest total at the end of the season becomes the World Driver Champion (WDC). However, predicting the WDC is challenging due to the dynamic nature of the sport, involving variables such as vehicle performance, team strategy, and evolving race regulations [9]. Technical innovations like the Drag Reduction System (DRS) have increased overtaking

opportunities, making race outcomes more unpredictable [8], [17].

The recent dominance of Max Verstappen, who secured four consecutive WDC titles from 2021 to 2024, reflects the significance of consistency and adaptability in modern F1 [14], [15]. His continued performance makes him the top contender for the 2025 season, generating widespread interest in predictive analytics for motorsport outcomes.

The role of data extends beyond the racetrack. In professional sports, data-driven strategies are reshaping management decisions, as seen in major financial transactions like the \$10 billion acquisition of the Los Angeles Lakers [?], [13]. These developments reflect a growing demand for accurate predictive models powered by machine learning (ML) techniques.

Extreme Gradient Boosting (XGBoost) has shown strong performance in predicting sports outcomes, including cricket [7], tennis [6], and horse racing [5]. Additionally, Patil et al. [4] demonstrated that statistical models can uncover technical factors in F1 cars that strongly correlate with a driver's season points using regression and principal component analysis.

This research aims to leverage the XGBoost algorithm to predict the 2025 Formula 1 World Driver Champion, based on data such as race and qualification results, driver standings, and car/team attributes. The model utilizes data up to Round 10 of the 2025 season, retrieved via the official Jolpi API. The outcome of this research could support fans, analysts, sponsors, and broadcasters in understanding mid-season performance trends in the competitive landscape of F1.

II. THEORETICAL BASIS

A. Formula 1

Formula 1 (F1) is the pinnacle of single-seater motorsport, featuring high-performance cars, advanced technology, and elite drivers from around the world [3]. The championship awards points based on finishing positions in two race formats: the *feature race* (main race) and the *sprint race*, each with distinct scoring systems [26].

Table I summarizes the current points allocation for both race types. The feature race awards a maximum of 25 points, reflecting its status as the primary event of the weekend. The sprint race, introduced in 2021 to increase weekend

excitement, awards fewer points and has a shorter distance without mandatory pit stops [25], [34].

Position	Feature Race	Sprint Race
1st	25	8
2nd	18	7
3rd	15	6
4th	12	5
5th	10	4
6th	8	3
7th	6	2
8th	4	1
9th	2	0
10th	1	0

TABLE I

POINTS DISTRIBUTION FOR FEATURE AND SPRINT RACES.

The F1 scoring system has evolved over time to ensure competitive fairness and maintain fan engagement. Major changes include expanding point allocations (e.g., from top 6 to top 10), adding fastest lap points (2019–2024), and trialing double points for the final race in 2014 [28], [35]. However, fastest lap points were removed again in 2025 [29].

F1 also employs a multi-phase qualifying format (Q1, Q2, Q3) to determine starting positions for the feature race. Each session progressively eliminates the slowest drivers, culminating in a shootout among the top 10 for pole position [30], [31]. The 107% rule applies in Q1 to ensure competitive entry into the race [36].

On sprint weekends, an alternate Sprint Qualifying format is used, comprising shorter timed sessions (SQ1–SQ3), which set the grid for the sprint race. Grid penalties from engine component changes or infractions can also affect final race positions, making qualification strategy as crucial as race pace.

B. Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that enables systems to automatically learn from data, identify patterns, and make predictions or decisions with minimal human intervention. Unlike traditional programming, which relies on explicit rules, ML develops models based on historical data and improves performance through iterative training [2].

According to Badillo *et al.* [1], ML employs statistical and computational methods to extract knowledge from data for tasks such as classification and prediction. It is particularly effective in processing large and complex datasets and adapting to dynamic input patterns. A further subfield of ML is Deep Learning (DL), which utilizes multi-layered artificial neural networks to process unstructured data such as images and audio.

Fig. 1 illustrates the hierarchical relationship between AI, ML, and DL. AI represents the broadest category, encompassing all intelligent computational systems. ML is a subset of AI focused on learning from data, while DL is a more specific subset of ML that uses deep neural architectures to solve more complex problems [32].

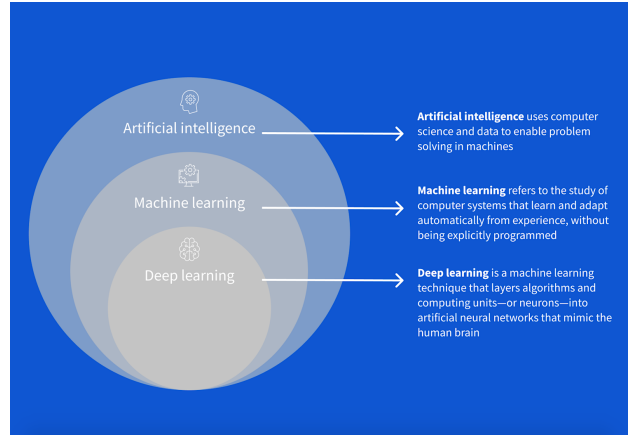


Fig. 1. The relationship between AI, Machine Learning, and Deep Learning.

C. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a high-performance machine learning algorithm based on the gradient boosting decision tree (GBDT) framework [11], [12]. It has gained significant popularity for tabular data prediction tasks and competitive data science applications due to its efficiency, scalability, and predictive accuracy.

Unlike traditional decision trees, which are prone to overfitting and limited generalization [22], XGBoost builds an ensemble of shallow trees sequentially, where each new tree corrects the residual errors of the previous ensemble. This additive model is optimized using a regularized objective function:

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

Here, $l(y_i, \hat{y}_i)$ denotes the loss function between predicted and actual values, and $\Omega(f_k)$ is the regularization term to control model complexity and prevent overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2)$$

In this formulation, T is the number of leaves in a tree, w_j is the prediction score on the j -th leaf, and γ and λ are regularization parameters.

XGBoost differs from random forests in that it builds trees sequentially, not in parallel [23]. Tree construction in XGBoost is guided by maximizing the information gain at each node. The gain is calculated as:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (3)$$

where G_L , G_R are first-order gradients, and H_L , H_R are second-order derivatives (Hessians) for the left and right child nodes, respectively.

A visual summary of the XGBoost training process is shown in Fig. 2, which demonstrates its iterative and residual-correcting architecture.

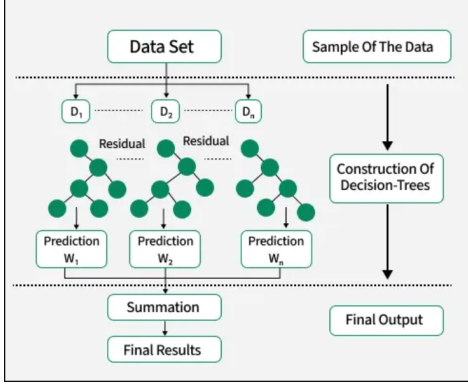


Fig. 2. Illustration of XGBoost model training process [21].

D. Evaluation Metrics

To assess the predictive performance of regression models, several evaluation metrics are commonly used. The three primary metrics are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score) [37], [38].

1) *Mean Absolute Error (MAE)*: MAE measures the average absolute difference between predicted values and actual values. It is robust to outliers and is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

where y_i is the true value, \hat{y}_i is the predicted value, and n is the number of observations.

2) *Root Mean Square Error (RMSE)*: RMSE is the square root of the average squared differences between predictions and actual values. It penalizes larger errors more heavily than MAE [38]:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

3) *R^2 Score (Coefficient of Determination)*: The R^2 score indicates the proportion of the variance in the target variable explained by the model. Its value ranges from 0 to 1 and is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

where \bar{y} is the mean of the observed data. Higher R^2 values indicate better model performance [37].

4) *Cross-Validation*: Cross-validation is a statistical method used to evaluate model generalization on unseen data. The most common approach is k -fold cross-validation, which partitions the dataset into k equal parts. The model is trained on $k - 1$ folds and validated on the remaining fold, repeating the process k times [24]. The final performance is averaged across

all folds, reducing the risk of overfitting and providing a more reliable performance estimate.

III. RESEARCH METHODOLOGY

A. Literature Review

A comprehensive literature review was conducted to establish the theoretical and methodological basis for this research, focusing on the application of machine learning—particularly the XGBoost algorithm—to Formula 1 data. Prior studies on World Driver Champion (WDC) prediction have explored various modeling approaches but show limitations in scope and methodology.

Den Hartog [39] employed machine learning models such as Logistic Regression, Random Forest, AdaBoost, and XGBoost to predict the WDC from 2014 to 2022. XGBoost yielded the best performance with an accuracy of 93% and an F1-score of 0.85. The key features included average finishing position, number of wins, and podiums. However, the research was limited to end-of-season predictions and did not utilize explainable AI (XAI) techniques or contextual race factors such as penalties, pit strategies, or weather conditions.

In contrast, Van Kesteren and Bergkamp [40] analyzed the contribution of drivers versus constructors using a Bayesian multilevel rank-ordered logit regression model. Their results indicated that constructors accounted for 88% of performance variance. Despite offering valuable insights, the research was descriptive in nature and did not develop predictive models for championship outcomes.

This research addresses the research gaps by introducing a mid-season regression-based WDC prediction model using XGBoost with hyperparameter tuning via Optuna. It utilizes a comprehensive dataset from the 1950–2025 seasons, up to round 10 of the 2025 season. The model's performance is evaluated using standard regression metrics including R^2 , Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Unlike previous work, this approach allows for early-season championship forecasts and lays the groundwork for real-time predictive analysis in competitive racing scenarios.

B. Data Collection

Historical Formula 1 data was collected using the Jolpi API, which provides comprehensive access to racing information, including race results, driver standings, team data, and qualifying performance [33]. The data spans from the 1950 to the 2025 season and was retrieved using Python.

The following API endpoints were utilized:

- `/races` – race schedules and locations.
- `/results` – final race outcomes (finish positions).
- `/qualifying` – qualifying session results (start positions).
- `/drivers` – demographic and identity data of drivers.
- `/driverstandings` – seasonal points and rankings of drivers.
- `/constructors` – information about Formula 1 teams.
- `/constructorstandings` – team rankings by accumulated points.

The collected data was stored in `.csv` format to facilitate preprocessing and integration with data analysis tools such as `pandas`. This structured format also supports seamless exploration and visualization in subsequent stages of the analysis.

C. Data Analysis

Data analysis was performed to understand the structure, patterns, and relationships between variables in the historical Formula 1 dataset. Visualization was carried out using the `Plotly` library, enabling interactive exploration of the data.

This step involved exploratory data analysis (EDA) to examine point distributions across seasons, driver performance by team, and the relationship between qualifying positions and final race results. Key objectives of the analysis included:

- Identifying potential predictor variables for the final championship standings.
- Evaluating correlations between numerical and categorical features with the target variable (final driver rank).
- Selecting relevant data to be used as features in the prediction model.

Statistical methods such as *Cramér's V* were used to measure associations between categorical variables. These analyses provided critical insights that guided the feature selection and model development process.

D. Data Pre-processing

The pre-processing stage ensures that the data is clean, consistent, and suitable for machine learning model input. This process involved merging multiple data sources into a unified dataset, removing duplicates, handling missing values, and converting data types.

Feature engineering was applied to derive relevant attributes such as average finishing position, average starting grid position, race completion ratio (DNF rate), and point trends. Categorical features were transformed using `LabelEncoder` and `OrdinalEncoder`, while numerical features were standardized using `StandardScaler` to normalize their distribution.

Feature importance analysis guided the selection of the most predictive attributes. Only race results available before the final round of the 2025 season were included to prevent data leakage in the championship prediction task.

E. Model Design

This research employs the XGBoost (Extreme Gradient Boosting) algorithm, a tree-based ensemble learning method recognized for its speed and performance in handling structured data [?]. The objective is to predict the final championship position of Formula 1 drivers in the 2025 season using various features, including qualifying results, race outcomes, team information, and historical statistics. The model was trained using data from 2014 to 2024 and evaluated on the 2025 season. Several training scopes—10, 20, and 30 seasons—were tested to examine the effect of historical data coverage on model performance.

Initially, a baseline XGBoost model was created using default parameters. To evaluate the model's generalization capability, 5-Fold Cross-Validation was applied on the training data using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) as evaluation metrics [?]. To improve performance, hyperparameter tuning was performed using Optuna, a state-of-the-art Bayesian optimization framework that utilizes Tree-structured Parzen Estimators (TPE) [?]. Optuna's efficiency in exploring complex parameter spaces and support for early stopping made it well-suited for this task. After obtaining the best parameter configuration, the final model was retrained using the entire training set and used to predict the 2025 championship outcomes.

F. Model Evaluation

The performance of the XGBoost model was evaluated to determine its ability to accurately predict the Formula 1 World Driver Champion for the 2025 season. To ensure generalization beyond the training data, three common regression metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). MAE calculates the average magnitude of absolute prediction errors, offering an intuitive measure of prediction accuracy. RMSE, by squaring the errors before averaging, penalizes larger deviations more heavily, making it more sensitive to outliers [?]. The R^2 score evaluates the proportion of variance in the target variable that is predictable from the input features, with values closer to 1 indicating better model performance [?]. These evaluation metrics provide a comprehensive assessment of the model's predictive capability and highlight both its effectiveness and limitations when applied to unseen data.

IV. RESULTS AND DISCUSSION

A. Dataset Description

The dataset used in this research was collected via the Jolpi API using automated scraping implemented in Python. Data from multiple endpoints were retrieved in JSON format and converted into a structured `DataFrame` using the `pandas` library for further processing.

The raw dataset consisted of multiple entities with the following record counts: `driver_standings` (2,097), `drivers` (2,099), `qualifying_results` (10,594), `race_results` (25,385), `race_schedules` (1,149), `constructor_standings` (1,121), and `constructors` (1,121). After merging and cleaning, the final dataset contained 706 rows, covering Formula 1 seasons from 1950 to 2025.

The target variable represents the final driver standings at the end of each season, focusing on predicting the rankings for the 2025 season. The dataset also reflects changes in the F1 point system, particularly the major revision in 2010 and the introduction of sprint races in 2021, which influenced the point distribution system.

B. Import and Merge Dataset

The initial stage involved importing historical Formula 1 data in CSV format, which had been obtained through the Jolpi API. Multiple entities such as drivers, driver standings, qualifying results, race results, race schedules, constructors, and constructor standings were loaded using the pandas library in Python.

These datasets, including the 2025 season data, were merged using the `pd.concat()` function with the `ignore_index=True` parameter to ensure consistent indexing. This resulted in consolidated datasets for each entity, covering the period from 1950 to 2025. The merged datasets formed the basis for further preprocessing and feature construction steps.

C. Preprocessing

1) *Data Merging*: To build a comprehensive dataset for model training, multiple data sources were merged into a single integrated structure. The merging process began by extracting the final driver standings for each season, identified from the last race round of every year. This was followed by merging driver information—such as full name, birthdate, and debut year—to calculate each driver’s age and experience per season. Constructor points from the final race of each season were also added by matching teams with drivers in the corresponding year. Finally, driver performance statistics, including average qualifying position, number of pole positions, average race finish, and Did Not Finish (DNF) rate, were calculated and merged. This unified dataset enabled a broader and interconnected feature space for effective model training.

2) *Data Cleaning and Feature Engineering*: After merging the datasets, data cleaning and feature engineering were performed to ensure the quality and usability of the data. The cleaning process involved standardizing text formats, such as converting constructor names to lowercase and removing unnecessary spaces. Driver birthdates were converted into standard datetime format to calculate driver age per season. Additionally, non-numeric or invalid race results, such as “-” or “D”, were removed, and relevant columns were retained for modeling.

Feature engineering was conducted to enrich the dataset. Driver age was calculated based on the difference between the season year and the driver’s birth year. Another feature, racing experience, was computed as the difference between the current season and the driver’s debut year, with a minimum value of zero to prevent negative values due to inconsistent records.

Unnecessary columns were dropped, and only relevant attributes were preserved, including age, experience, average qualifying position, number of pole positions, average finishing position, DNF ratio, constructor points, and final championship standing. Rows with missing values in these features were removed to ensure robust model training.

3) *Feature Importance and Feature Selection*: To identify the most influential features for predicting the final driver standing, both numerical and categorical correlations were analyzed. Figure 3 presents the Pearson correlation heatmap between numerical variables and the encoded target variable.

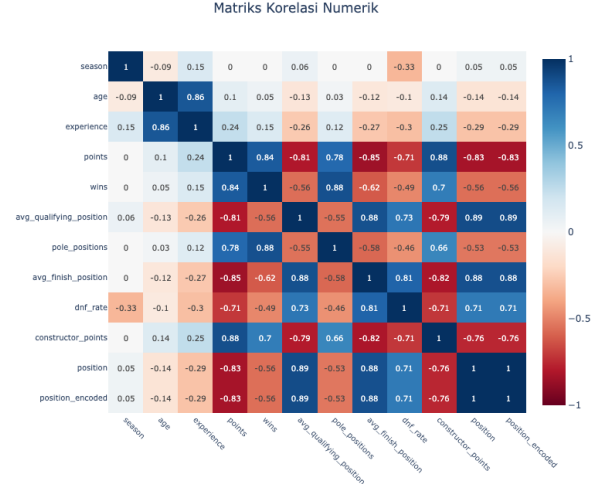


Fig. 3. Pearson Correlation Heatmap

Features such as `points` ($r = -0.83$), `avg_qualifying_position` ($r = 0.89$), and `avg_finish_position` ($r = 0.88$) show strong correlations with the target. These relationships indicate that better qualifying and finishing positions, along with higher points, are strong predictors of better championship outcomes.

Additionally, Cramér’s V was used to assess the relationship between the categorical feature `driver_id` and the target, yielding a moderate association value of 0.53. Although `driver_id` may seem non-generalizable for new drivers, it captures historical performance patterns and improves model accuracy on past data.

Based on correlation analysis and domain knowledge, several features were selected for model training. These include `driver_id`, `age`, `experience`, `points`, `wins`, `avg_qualifying_position`, `pole_positions`, `avg_finish_position`, `dnf_rate`, and `constructor_points`. These features were chosen because they capture key aspects of a driver’s performance, experience, and team competitiveness, which are critical factors in determining final championship standings.

The `season` feature was excluded due to its lack of predictive power for future performance. Proper feature selection ensures the model captures the most relevant information, avoids overfitting, and enhances generalization to new data.

4) *Feature Transformation*: Feature transformation was applied to ensure that the data fed into the model is in a consistent and optimal format. Numerical features such as `points`, `wins`, `pole_positions`, and `constructor_points` were standardized using the z-score method to eliminate scale

bias across different seasons. This transformation enables fair comparisons between driver performances across seasons.

Categorical features, such as the final driver position, were transformed using ordinal encoding, allowing the model to understand the natural order in rankings. Meanwhile, the driver identity (`driver_id`) was encoded using label encoding to convert categorical text data into numeric form without imposing any ordinal relationship.

Additionally, all relevant numeric features were normalized using standard scaling (z-score normalization) to ensure they have a mean of zero and standard deviation of one, improving model convergence and stability.

D. Model Design

To predict the 2025 Formula 1 world champion, an *Extreme Gradient Boosting* (XGBoost) model was designed. The training and evaluation process consisted of multiple phases: train-test data splitting, base model training, cross-validation, and hyperparameter optimization.

1) : *Train-Test Data Split*: The dataset was split temporally, where the 2025 season served as the testing set and past seasons (2014–2024, 2004–2024, and 1994–2024) were used as training data under three scenarios. This allowed an investigation into how the length of historical data affects model performance.

TABLE II
TRAIN-TEST SPLIT SCENARIOS BASED ON HISTORICAL COVERAGE

Scenario	Years	Train Set	Test Set
Scenario 1	10 Years	2014–2024	2025
Scenario 2	20 Years	2004–2024	2025
Scenario 3	30 Years	1994–2024	2025

2) *Base Model Initialization*:: An initial XGBoost regression model was trained using default parameters. This served as a baseline for further optimization.

3) *Cross-Validation*:: Model robustness and generalization were evaluated using 5-Fold Cross-Validation on the training data. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 were used for assessment. This step was essential to detect overfitting and ensure generalizability.

4) *Hyperparameter Tuning*:: To enhance predictive performance, Bayesian Optimization via the Optuna library was employed. It efficiently explored parameter spaces by leveraging past trials. The final optimal configuration is summarized below:

- `n_estimators`: 448
- `max_depth`: 4
- `learning_rate`: 0.0109
- `subsample`: 0.5946
- `colsample_bytree`: 0.9776
- `gamma`: 4.4287
- `reg_alpha`: 2.6161
- `reg_lambda`: 4.8346

This final model configuration was then used for the ultimate training and prediction task, ensuring enhanced predictive power while minimizing overfitting on the training data.

E. Model Evaluation and Results

After determining the optimal hyperparameters through tuning, the final XGBoost model was retrained using the best configuration. The model was then evaluated on the 2025 season test data using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The evaluation results are shown in Table III.

TABLE III
EVALUATION RESULTS USING THE BEST MODEL ON 2025 TEST SET

MAE	RMSE	R^2
1.0610	1.3634	0.9441

The high R^2 score of 0.9441 and low MAE/RMSE values indicate strong predictive performance. The scatter plot in Fig. 4 visualizes the model's continuous predictions compared to actual final driver rankings. Most predictions are close to the perfect diagonal line, suggesting high accuracy.

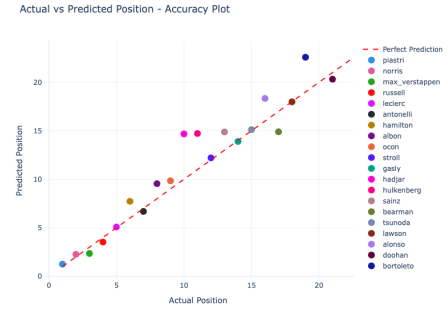


Fig. 4. Model predictions before discretization

Since race rankings are ordinal, the continuous predictions were discretized to obtain final ranking positions. As shown in Fig. 5, the discretized results still maintain alignment with the actual rankings, validating the model's ability to approximate real-world race standings.

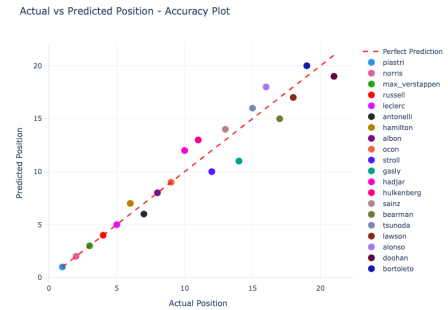


Fig. 5. Model predictions after discretization

The model predicted Oscar Piastri as the 2025 Formula 1 World Champion, based on performance data up to mid-season. While this prediction is robust, it remains subject to real-world uncertainties like strategy changes or unforeseen incidents in the remaining season.

To assess the effect of training data length, three scenarios were tested: using 10, 20, and 30 years of historical data. The results are summarized in Table IV.

TABLE IV
PERFORMANCE COMPARISON ACROSS HISTORICAL DATA SCENARIOS

Scenario	Years	Train Period	MAE	RMSE	R^2
1	10	2014–2024	1.0610	1.3634	0.9441
2	20	2004–2024	1.3288	1.5868	0.9243
3	30	1994–2024	1.1596	1.5234	0.9302

The results show that Scenario 1 (last 10 years) outperformed others, suggesting that longer historical data may introduce outdated patterns that reduce model relevance. Thus, recent data (10 years) provided the most reliable performance for predicting current season outcomes.

V. CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

This research successfully developed a predictive model using the Extreme Gradient Boosting (XGBoost) algorithm to forecast the 2025 Formula 1 World Driver Champion. The model was trained using historical data from the 2014 to 2024 seasons, covering 20 drivers participating in the 2025 season. Nine key features representing both driver and team performance were utilized, including age, experience, total points, average qualifying position, number of pole positions, average race finish position, DNF rate, and constructor performance.

To optimize the model's accuracy, hyperparameter tuning was conducted using Optuna over 255 trials. The best configuration yielded 448 trees, a maximum depth of 4, and a learning rate of 0.01096. The model also applied regularization techniques to prevent overfitting. The final model achieved strong performance metrics on the test set, with an MAE of 1.0610, RMSE of 1.3634, and R^2 score of 0.9441.

Based on the discretized prediction results, the model forecasts Oscar Piastri as the potential 2025 World Champion. These findings demonstrate that XGBoost can deliver accurate and reliable predictions in the highly dynamic and competitive environment of Formula 1.

B. Recommendations

- 1) Addition of external and technical variables. Incorporating external variables such as weather conditions, race incidents, and penalties, as well as technical variables from vehicle telemetry data such as average speed, number of pit stops, and tire usage, can enrich the features in the dataset. The combination of these types of variables is expected to significantly improve the accuracy and quality of the predictive model.

- 2) Implementation of time-series-based approaches. Future research is encouraged to integrate time-series forecasting or sequential learning methods, such as Long Short-Term Memory (LSTM) or Transformer, which are designed to analyze sequential data. This approach allows the model to learn patterns in driver performance changes from one race to another, thereby capturing trends, fluctuations, and performance dynamics more accurately than static regression models. As a result, the prediction can take into account historical context and gradual performance development throughout the season.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Universitas Multimedia Nusantara (UMN) for providing the facilities and academic support necessary for conducting this research. Gratitude is also extended to the supervisors and individuals who offered valuable insights and feedback throughout the development and evaluation of the predictive model. Their contributions were instrumental in improving the accuracy and robustness of this study on forecasting the Formula 1 World Driver Champion using the XGBoost algorithm.

REFERENCES

- [1] Solveig Badillo, Balazs Banfai, Fabian Birzele, Iakov I. Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, Jitao David Zhang, "An Introduction to Machine Learning," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 871-885, 4 2020.
- [2] Christian Janiesch, Patrick Zschech, Kai Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685-695, 9 2021.
- [3] Yiyang Ma, "Challenges and Opportunities in the Business Model of Formula 1," *Advances in Economics, Management and Political Sciences*, vol. 26, no. 1, pp. 108-114, 9 2023.
- [4] Ankur Patil, Nishtha Jain, Rahul Agrahari, Murhaf Hossari, Fabrizio Orlandi, Soumyabrata Dev, "A Data-Driven Analysis of Formula 1 Car Races Outcome," pp. 134-146, 2023.
- [5] Chawin Terawong, Dave Cliff, "XGBoost Learning of Dynamic Wager Placement for In-Play Betting on an Agent-Based Model of a Sports Betting Exchange," 1 2024.
- [6] Shitong Kang, Yunpeng Shi, Yuyao Chen, Haolin Wang, "Enhancing Tennis Match Predictions with AHP, XGBoost, and Genetic Algorithms," *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pp. 314-320, 6 2024.
- [7] Akhil Tyagi, Amandeep Kaur, Aryan Kamboj, Chayandeep Chaulia, Gandharv Mohan, Manpreet Singh, "XGBoosting Cricket: Enhancing Predictive Modeling for Twenty20 Match Results Using Machine Learning and Statistical Techniques," *SN Computer Science*, vol. 5, no. 8, pp. 1036, 11 2024.
- [8] Abdelghani Belgaid, "Statistical Analysis of the Impact of FIA Regulations on Safety, Racing Dynamics, and Spectacle in Formula 1," 10 2024.
- [9] Oliver Budzinski, Arne Feddersen, "Measuring competitive balance in Formula One racing," 6 2020.
- [10] Lili Belkovic, István Takács, "The Digitalization of Formula 1: Innovations from Car Design to Race Strategy," *2023 IEEE 21st Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 000565-000570, 9 2023.
- [11] , "XGBoost," 2025.
- [12] Kavlakoglu, Eda, Russi, Erika, "What is XGBoost?," 2024.
- [13] Kurniawan, Moh. Afaf El, "Sejarah F1, Ajang Balap Mobil Paling Bergengsi di Dunia yang Bermula Setelah Perang," 2024.

- [14] Francis, Anna, "The Four-Time World Champions Verstappen Joins in the All-Time List – and Those Still Left Ahead of Him," , 2024.
- [15] Hardy, Ed, "Who Has Won the Most Consecutive F1 World Drivers' Championships?," , 2024.
- [16] Mitchell, Stewart, "Data Analytics: Managing F1's Digital Gold," , 2022.
- [17] Fédération Internationale de l'Automobile, "FIA Statutes and Internal Regulations," , 2025.
- [18] 1, Formula, "Everything you need to know about F1 – Drivers, teams, cars, circuits and more," , 2025.
- [19] AMG, Mercedes, "Feature: Data and Electronics in F1, Explained!," , 2025.
- [20] Butler, Georgia, "Racing at the Edge: How portable data centers are driving Formula 1," , 2023.
- [21] GeeksForGeeks, "Implementation of XGBoost (eXtreme Gradient Boosting)," , 2025.
- [22] GeeksForGeeks, "Decision Tree," , 2025.
- [23] GeeksForGeeks, "Random Forest Algorithm in Machine Learning," , 2025.
- [24] GeeksForGeeks, "Cross Validation in Machine Learning," , 2025.
- [25] 1, Formula, "The beginner's guide to the F1 Sprint," , 2025.
- [26] 1, Formula, "The beginner's guide to the F1 Drivers' Championship," , 2025.
- [27] 1, Formula, "The beginner's guide to the F1 weekend," , 2025.
- [28] Hardy, Ed, "History of the F1 points system with proposed structure for 2025," , 2024.
- [29] 1, Formula, "From fastest lap to increased rookie running – 7 rule changes you need to know for the 2025 F1 season," , 2025.
- [30] , "How does F1 qualifying work?," , 2025.
- [31] Clark, Amanda, "How Does F1 qualifying work?," , 2025.
- [32] Coursera, "Deep Learning vs. Machine Learning: A Beginner's Guide," , 2023.
- [33] , "jolpi.ca F1 API," .
- [34] Coleman, Madeline, "How F1 sprint races work: New schedule, locations for 2025," , 2025.
- [35] 1, Formula, "Fastest lap point to be scrapped in 2025 after latest FIA World Motor Sport Council meeting," , 2024.
- [36] Fédération Internationale de l'Automobile, "2025 FORMULA ONE SPORTING REGULATIONS," , 4 2025.
- [37] Davide Chicco, Matthijs J. Warrens, Giuseppe Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. e623, 7 2021.
- [38] Timothy O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481-5487, 7 2022.
- [39] I H D Den Hartog, "Data to drive: Personalized visualization in Formula One racing," , 2022.
- [40] Erik-Jan van Kesteren, Tom Bergkamp, "Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage," , 5 2023.