JOURNAL
COMPUTER SCIENCES INSTITUTE

JCSI 38 (2026) 81-86
Received: 24 September 2025
Accepted: 12 October 2025

Application of machine learning for predicting Formula 1 race results

Sylwia Agata Krzysztoń*, Jakub Smołka

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

With the growing popularity of motorsports and the increasing availability of large telemetry datasets, machine learning techniques to predict Formula 1 race results have become particularly well-justified. The purpose of this study is to build an ensemble learning model using Support Vector Machine, Gradient Boosting and Random Forest for race result classification. The Optuna library was used for hyperparameter optimisation. The models are based on historical data on races and drivers. The model achieved an F1 score of 77.83% ± 4.18% (macro-averaged) and an accuracy of 80.22% ± 4.69% on the validation set. The results confirm the effectiveness of the applied methods and highlight the significant impact of telemetry data on prediction quality. Ensemble learning can serve as a valuable tool to support Formula 1 race strategies.

Keywords: machine learning; classification; ensemble learning; formula 1

*Corresponding author

Email address: sylwia.krzy64@gmail.com (S. A. Krzysztoń)

1. Introduction

In recent years, Formula 1 has been experiencing a true renaissance. Although its history dates to the 1950s, it is only now that its popularity has been on the rise. The Formula 1 World Championship (Grand Prix) is the most prestigious series of car races. The 2024 season includes a record number of 24 Formula 1 races, which is the highest number in the history of the championship. Formula 1 World Championship circuits cannot be shorter than 305 km (the exception is the Monaco Grand Prix, where the track length is 260 km) and the races cannot last longer than 2 hours. A total of 10 teams will compete in the 2025 season, each with two drivers, giving a total of 20 competitors in the stake. The races take place on Sundays and are preceded by practice sessions (Fridays and Saturdays), qualifying sessions and on some race weekends, a sprint (a short, intense race over approximately 1/3 of the circuit). In the case of a weekend with sprint, there is also a qualifying session for the sprint. Every race week is extremely exciting for Formula 1 fans, always sparking questions and predictions about who will take pole position and who will win the Grand Prix.

In an era of extensive and rapid access to data, machine learning is also experiencing one of its most dynamic periods of development, allowing for the creation of high-quality predictive models. Predicting the results of sporting events, including Formula 1 races, is a popular topic in machine learning model development due to the large and high-quality dataset of telemetry data. The domain of machine learning offers a large selection of algorithms that can be used to create advanced models, and their selection is closely related to the data structure and analysis purposes. In the context of predicting the results of sporting events, such as Formula 1 races, the use of machine learning algorithms becomes particularly relevant. They allow for effective analysis of complex nonlinear data, which results in highly accurate forecasts. Research in this domain could be of highly significant importance in the planning of racing strategies.

2. Literature overview

Currently, several studies are being conducted that focus on forecasting processes or events based on historical data and factors that influence their development. The result of the popularity of this type of research is the development of new, advanced models using machine learning techniques. Due to the different specifications of the data used and the purpose of the prediction, developers implement different approaches. Previously published articles focusing on predicting the progress of Formula 1 races and the impact of individual features on their result have been analysed. Presenting existing achievements in this area will allow for an understanding of key aspects in research on predicting race results and progress, while also identifying areas that require further study.

Publication [1] presents the use of Support Vector Machine (SVM) based on two types of kernels (linear and polynomial). Grand Prix winners were predicted based on historical data. The output data in the model refers to three metrics: podium position, points scored, no points scored, or failure to finish the race. The choice of kernel affected the model's performance in prediction tasks. The developed model achieved an accuracy of 97% for the linear kernel and 96% for the polynomial kernel. For the other selected metrics, i.e. sensitivity, precision and F1 score, the model with the linear kernel proved to be slightly better (differences in the order of parts per hundred). The presented method showed the potential for a significant impact on the field of Formula 1 racing and data analysis using machine learning.

In his work [2], Franssen compared the effectiveness of three different machine learning models: a base Artificial Neural Network model, a Deep Neural Network (DNN), and a neural network with a Radial Basis Function (RBF) in predicting Formula 1 race results. Both DNN and RBF network performed better than the baseline model in the F1 metric. Also, both networks achieved higher F1 results in the training set (RBF 69%, DNN 67%), but significantly lower in the test set (RBF

55%, DNN 58%). The RBF network also had a lower validation result (58%), while the DNN achieved a result closer to the test result. Method showed that the DNN outperforms the RBF network in performance, scoring a higher F1 result and lower loss.

Sicoie studied supervised machine learning algorithms for predicting the winner of a Formula 1 race [3]. The developed structure includes Random Forest (RF), Gradient Boosting (GB) and Support Vector Regression (SVR) models. All three finely tuned models (using grid and random search) generated high correlation results (with the actual race position) of 0.883 - 0.903. The study showed that the method is effective, but it has limitations that can be overcome by using more detailed data and optimising the algorithms used.

Stoppels uses an artificial neural network to predict the results of Formula 1 races [4]. The dataset contains information about 21 races from the 2016/2017 season for four drivers and additional data about 21 races generated manually. The neural network created for the set of 42 races correctly predicted 77% of cases in the training set and 69% in the validation set. An analysis was also carried out in which one of the drivers was considered a bad driver in the rain. The results were 75% for the validation set. Reducing the amount of input data (only 21 races) did not cause significant deviations (75% on the validation set). The research showed that artificial neural networks can effectively predict race results, but their effectiveness depends on the structure of the network and the quality of the data.

Article [5] discusses the use of linear regression combined with Monte Carlo simulations to analyse Formula 1 race results. The analysis takes into consideration the contribution of individual drivers and constructors to the overall performance of the race. The Monte Carlo algorithm allowed for the simulation of thousands of possible race scenarios. The linear regression model used in this study proved to be highly effective, with over 88% of the differences in race results attributable to constructor performance. The analysis conducted may be useful in evaluating results where competition depends on multiple factors.

Publication [6] describes the development of the RankNet machine learning model, which allows for the modelling of events related to pit stops and changes in ranking positions, taking into consideration cause-and-effect relationships. The RankNet model (79% accuracy) proved to be the best solution, combining a coder and decoder network with a separate multi-layer perceptron (MLP) network capable of learning the relationships between pit stops and changes in position in the ranking. The method showed that model decomposition based on cause-and-effect relationships is crucial for improving ranking positions. Further work should focus on transfer learning to overcome the limitations associated with the lack of data on extreme racing events.

Zhao's article [7] is about the use of deep neural networks (DNN) to predict the fastest lap time in Formula 1 qualifying. The input dataset contains the fastest lap times from qualifying sessions (Q1, Q2, Q3) from 2014 to 2023, with individual data processing for each track. The regression model achieves high accuracy with a deviation from the actual result of less than 5%. The limitation of the presented approach is that it does not consider variable weather conditions or driver status. Research showed that the deep neural network model can be successfully used to predict the results of Formula 1 drivers.

The impact of starting in pole position on the final race result was also analysed using logit and Poisson regression models [8]. Two cases were examined: the probability of winning the race and the final position in the race. Starting in pole position gives a significant advantage over other drivers in the field (the advantage is approximately two places at the finish line or approximately 10 percentage points higher probability of winning the race). The use of econometric modelling methods is effective in studying the impact of pole position on Formula 1 race results. The advantage associated with pole position varies over time and is influenced by various factors, such as rules and regulations, technological developments and unobserved factors.

Telemetry data from Formula 1 races contain a huge number of features. Research [9] was conducted to identify the most important factors that influence the number of points scored by drivers during the season. Principal component analysis was used to reduce the dimensionality. The regression model explains 99% of the variability in points. It turned out that the biggest impact is made by completed races, tyre type, positions on the track and average starting position. The original features space can be significantly reduced to a lower-dimensional subspace without significant loss of information, which will also reduce the required computing power.

Marupaka and Rangineni focused on integrating machine learning techniques to predict and evaluate data quality within the ETL (Extract-Transform-Load) process [10]. They used the developed data as input for machine learning algorithms. The accuracy, precision and sensitivity achieved by these models highlight the potential for improving data quality in real time and reducing the risks associated with the low quality of data. The technique developed is flexible and can be successfully used in other domains.

The presented overview of scientific articles reflects current research achievements in predicting Formula 1 race results. It has shown that an approach using various machine learning techniques gives promising results. In most studies, researchers used support vector machines, neural networks and regression models. Their choice was based on their effectiveness and ability to model complex relationships in the data. The best results were achieved by a model built from a SVM with a linear kernel. However, the literature does not clearly indicate a single, best technique for predicting the results of Formula 1 races. The most frequently cited problems include excessive model fitting to data, insufficiently detailed and qualitative dataset, high computational complexity, and the need to optimise algorithms.

A review of the literature showed that various machine learning techniques are used to predict Formula

1 race results, but none of them is clearly dominant. This shows that the problem under study is more complex and requires further research to develop more effective methods. The presented research confirms the effectiveness and potential of machine learning to predict the race results of Formula 1 with high accuracy. The achievements to date provide a solid basis for further research, which will allow for the supplementation of scientific achievements with new approaches and innovative applications.

3. Purpose of the study and research hypotheses

The purpose of this work is to develop a new model using advanced machine learning techniques that will enable the prediction of Formula 1 race results. The work focuses on the model development and analysis of telemetry data, which will be provided as input data. The model development involves selecting the appropriate machine learning algorithm, creating its structure and fine-tuning its parameters to achieve the best possible results. The finished model is trained and tested to evaluate it and check whether it has predictive capabilities on new, unknown data. The following research hypotheses have been defined as part of the work:

1. Advanced machine learning techniques significantly improve the accuracy of predicting Formula 1 race results.
2. Ensemble learning ensures high predictive performance and reliability of the model for predicting Formula 1 race results.
3. Telemetry data has a significant impact on the accuracy of predicting Formula 1 race results.

4. Research methods

The research methods developed are crucial for the repeatability of the studies conducted. It was designed to ensure the highest possible data quality and reliable model results. The research methods can be divided into several key stages. A series of steps were taken, including data processing, building and optimising a machine learning model, and compiling the results.

4.1. Programming environment and libraries

For research, open-source tools were used. Due to the specific nature of the research and the domain of machine learning, Python version 3.13.2 was chosen as the programming language. To enhance the implementation process, the libraries listed in Table 1 were used.

The implementation work was done in the Visual Studio Code programming environment, version 1.104.0. In addition, the Jupyter plugin version 2025.8.0 was used to create interactive notebooks. This configuration of the environment made it possible to divide the code into fragments conveniently run individual parts and increase the efficiency of the entire process. Additionally, this simplifies the repeatability of the test.

Table 1: Libraries used

Library	Version	Application
requests	2.32.5	HTTP queries
pandas	2.3.2	Data analysis
numpy	2.3.3	Numerical calculations
scikit-learn	2.3.3	Machine learning
optuna	4.5.0	Hyperparameter optimisation

4.2. Dataset

The first step is to prepare a dataset from the open-source OpenF1 API [11]. The dataset contains races taking place in 2023-2024. Since 2014, there has been the hybrid era in Formula 1, due to a change in regulations and the introduction of a new generation of engines (1.6-litre turbocharged V6) supported by an electrical system. For this reason, analysing races prior to this change is unreliable due to the change in a key component, the engine. The dataset consists of information about racing drivers and their results in Formula 1 racing sessions. The compiled dataset consists of 1,635 records and 15 input features describing Formula 1 racing sessions, such as: driver number, broadcast name, country code, year, session name, start date, start time, end date, end time, GMT offset, starting position, wins before, pit stops count, average pit stop duration, qualifying position. The target variable is the driver's final position in the race, which has been divided into four key categories (Table 2).

Table 2: Description of classes in multi-class classification

Class	Description
winner	1st place
top3	2nd and 3rd place
points	4th to 10th place
no points	11th to 20th place

Based on the collected data, empty records and missing data were removed. New variables were created to store data on the number of wins, pit stops, and average pit stop times. Categorical variables were appropriately encoded using Label Encoder, while numerical features were scaled using Standard Scaler. The dataset was divided into a training set and a validation set in a ratio of 70% to 30%. To assist with data acquisition and processing, a simplified ETL process was created, consisting of three stages:

- extract - retrieving data from an external source,
- transform - cleaning, scaling, encoding, defining new variables,
- load - save to a variable and export to a text file.

Thanks to this approach, adding data about new seasons and analyzing it will be more effective. In addition, the dataset created in this way can be successfully used for other approaches.

4.3. Machine learning model

Considering the characteristics of telemetry data, a complex machine learning model was developed. Several basic models were combined to create a single optimal classification model that would enable the prediction of Formula 1 race results. Through the ensemble learning

approach, the classification ability is increased by using the strengths of different algorithms. An ensemble learning model was built consisting of a Support Vector Machine (SVM), a Random Forest (RF) and Gradient Boosting (GB). The SVM allows for the categorisation of nonlinear relationships by mapping data onto a hyperplane. The class is assigned based on the sign of the decision function value for a given sample (formula 1):

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \quad (1)$$

where: x is feature vector of the classified point, N is number of training samples, $\alpha_i \geq 0$ is Lagrange coefficients determined during training, x_i is training vectors, $y_i \epsilon \{-1,1\}$ is class labels, $K(x_i, x)$ is kernel function (e.g. linear, RBF), b is bias (offset). Random Forest uses a base model of a decision tree on random samples of data and features, with the result obtained by aggregation in that case majority voting (formula 2):

$$\hat{y} = \text{mode}\{\text{h}_1(\text{x}),\ \text{h}_2(\text{x}),\ \ldots,\ \text{h}_\text{T}(\text{x})\} \quad (2)$$

where: x is feature vector, $\hat{y}$ is predicted class for sample x, T is number of trees in the random forest, h_T is predicted class by tree t. Gradient Boosting assists the classifier when working with highly variable data by combining multiple weak predictive models into a single ensemble (formula 3):

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \quad (3)$$

where: m is number of iterations/trees, x is feature vector, ν is learning rate, $h_m(x)$ is tree fitted to the gradient of the loss function relative to current predictions.

The models were integrated using the Voting Classifier, considering prediction probabilities. The hyperparameters for the models were adjusted using the Optuna to maximise the F1-score. The optimisation process was carried out on 50 trials. Table 3 presents the best hyperparameter values obtained during model optimisation.

Table 3: Best hyperparameter values

Hyperparameter	Value	Explanation
rf_n_estimators	131	Number of trees in RF
rf max depth	14	Max. depth of RF tree
gb n estimators	92	Number of trees in GB
gb learning rate	0.22	GB learning rate
gb max depth	10	Max. depth of GB trees
svm C	1.75	SVM regularisation parameter
svm kernel	rbf	SVM kernel type

The developed ensemble learning model was trained on the training set, and its effectiveness was verified on the validation set. Cross-validation was used in the model training process to minimise the risk of type III error, reduce overfitting and to reliably assess the accuracy of the predictive model's forecasts. In this process, the training data was divided into 5 equal subsets, with 4 being used as training sets and one as a validation set in each iteration.

4.4. Evaluation of the model

Several basic metrics were used to evaluate the quality of the developed model. Since a team learning model consisting of several basic models was used, the metrics were selected in such a way as to fully reflect their reliability. For this purpose, confusion matrix was created. Due to the multi-class classification (Table 2), the confusion matrix takes the form of a 4×4 square table. Correctly classified cases are on the diagonal of the matrix (Table 4). In the case of multi-class classification, metrics for each class are calculated separately using one-vs-rest strategies. The statistics calculated in this way are then averaged for all classes using macro averaging. The F1 score averaged in this way due to class unbalance. The accuracy of the model is calculated globally. The following symbols are used in formulas 4-8: TP is true positive/hit, TN is true negative/correct rejection, FP is false positive/false alarm, FN is false negative/miss, n is number of classes, i is class index.

The accuracy of the model calculated globally is the ratio of the sum of correct recognitions to the sum of all recognitions (formula 4).

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i + TN_i + FN_i)} \quad (4)$$

The macro-average precision of the model is the ratio of the number of true positive recognitions to the sum of true positive and false positive recognitions (formula 5).

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \quad (5)$$

The macro-average recall indicates the ratio of the sum of true positive recognitions to the sum of true positive and false negative recognitions (formula 6).

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \quad (6)$$

The macro-average specificity is the arithmetic mean of the sum of true negative recognitions to true negative and false negative recognitions (formula 7).

$$\text{Specificity} = \frac{1}{n}\sum_{i=1}^{n} \frac{TN_i}{TN_i + FP_i} \quad (7)$$

The macro-average of F1 is the harmonic mean of precision and recall (formula 8).

$$\text{F1} = \frac{1}{n}\sum_{i=1}^{n} \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \quad (8)$$

The root mean square error (RMSE) and logarithmic error (LogLoss) were also used to evaluate the effectiveness of the model. The use of multiple metrics provides a more complete picture of the performance of the constructed model. Both classification ability and the ability to avoid false positive results are tested.

RMSE is the square root of the mean of the squares of the differences between the forecast values and the actual values (formula 9):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (9)$$

where: m is number of i N is number of classes in the validation set, P_i is predicted value for class i, O_i is actual value for class i. Logarithmic error measures the degree of divergence between the predicted probability and the actual recognition (formula 10):

$$\mathrm{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} O_{i,c}\log\left(P_{i,c}\right) \quad (10)$$

where: N is number of classes in the validation set, C is total number of classes in the problem, $O_{i,c}$ is actual value for class c in class i (1 if class c is true for class i, 0 otherwise), $P_{i,c}$ is predicted probability of assigning class c to class i.

5. Results

All metrics were calculated based on the summed (5-fold cross-validation) confusion matrix presented in Table 4. The marked diagonal of the confusion matrix shows the number of correctly classified cases for each class.

Table 4: Summed confusion matrix for multi-class classification

Predicted class	Actual class			
	no_ponts	points	top3	winner
no_points	**503**	64	2	1
points	55	**326**	19	2
top3	6	35	**69**	5
winner	0	6	8	**43**

The developed model with optimal parameters aimed at maximising the F1-score achieved a value of 77.83% ± 4.18%. The high F1-score metric indicates good classifier performance in multi-class conditions. The remaining performance metrics are presented in Table 5. The developed model performs is highly effective in classifying the no points class (503 correct predictions), while the winner class is the most difficult to predict (43 correct predictions), which is a result of it being naturally the least numerous class. Note that the model often confuses the points class with the no points class (64 cases). The metrics show that the model correctly classified 82.25% ± 3.83% of cases. In addition, it has a high ability to correctly reject negative samples (92.91%). The model performs well in terms of balanced evaluation between classes (precision and recall at 80.22% ± 4.69% and 76.20% ± 3.75%, sequentially). The values presented in Table 6 reflect the discrepancies between the model predictions and the actual class labels. The error values obtained in the context of multi-class classification are not exceptionally high and are within an acceptable range.

Table 5: Classification results with standard deviation

Metric	Average value [%]
F1-score	77.83 ± 4.18
Accuracy	82.25 ± 3.83
Precision	80.22 ± 4.69
Recall	76.20 ± 3.75
Specificity	92.91 ± 1.57

Table 6: Classification error metrics

Metric	Average value
RMSE	0.4719 ± 0.0609
LogLoss	0.5216 ± 0.0513

6. Conclusion

Based on the experiment conducted and the results achieved, the hypotheses can be interpreted. Hypothesis 1, stating that advanced machine learning techniques significantly improve the accuracy of Formula 1 race result predictions, has been confirmed. Metrics such as F1-score at 77.83% ± 4.18% and accuracy at 82.25% ± 3.83% indicate the high effectiveness and accuracy of the model used. Hypothesis 2 concerning the impact of ensemble learning on the predictive performance of the model and its reliability has also been confirmed. The use of three different SVM classifiers, RF and BG, allowed for stable and correct predictions in most cases and relatively low error rates in the context of multi-class classification. Hypothesis 3, indicating that telemetry data has a significant impact on the accuracy of Formula 1 race result predictions, was also confirmed. Based on the confusion matrix, the models effectively utilise the provided data. The integration of various sources of information related to Formula 1 races, including telemetry data, contributes to a significant improvement in the quality of predictions, as confirmed by the classification results obtained. The main limitations include the size and quality of the available telemetry data. Work on developing the dataset and model should include integrating additional data sources, such as detailed information on team strategies or weather conditions, and applying class weighting techniques, which may further improve the quality of predictions.

References

[1] P. Shelke, A. Pande, S. Kale, Y. Paralikar, A. Kulkarni, F1 Race Winner Predictor, In 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA) (2023) 1–4, https://doi.org/10.1109/ICCUBEA58933.2023.10392224.

[2] K. Franssen, Comparison of neural network architectures in race prediction: Predicting the racing outcomes of the 2021 Formula 1 season, Master thesis, Tilburg University, Tilburg, 2022.

[3] H. Sicoie, Machine learning framework for Formula 1 race winner and championship standings predictor, Bachelor thesis, Tilburg University, Tilburg, 2022.

[4] E. Stoppels, Predicting race results using artificial neural networks, Master thesis, University of Twente, Enschede, 2017.

[5] S. A. Menon, M. K. Ranjan, A. Kumar, B. Gopalsamy, F1 lap analysis and result prediction, International Research Journal of Modernization in Engineering, Technology and Science 6(11) (2024) 1848-1861.

[6] B. Peng, J. Li, S. Akkas, T. Araki, O. Yoshiyuki, J. Qiu, Rank Position Forecasting in Car Racing, In 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS) (2021) 724–733, https://doi.org/10.1109/IPDPS49936.2021.00082.

[7] Z. Zhao, Deep neural network-based lap time forecasting of Formula 1 racing, Applied and Computational Engineering 47 (2024) 61–66, https://doi.org/10.54254/2755-2721/47/20241191.

[8] D. Wesselbaum, P. D. Owen, The value of pole position in Formula 1 history, Australian Economic Review 54(1) (2021) 164–173, https://doi.org/10.1111/1467-8462.1240.

[9] A. Patil, N. Jain, R. Agrahari, M. Hossari, F. Orlandi, S. Dev, A Data-Driven Analysis of Formula 1 Car Races Outcome, In: L. Longo, R. O'Reilly (eds) Artificial Intelligence and Cognitive Science. Communications in Computer and Information Science, Springer 1662 (2023) 137–151, https://doi.org/10.1007/978-3-031-26438-2_11.

[10] D. Marupaka, S. Rangineni, Machine learning-driven predictive data quality assessment in ETL frameworks, International Journal of Computer Trends and Technology 72 (2024) 53–60, https://doi.org/10.14445/22312803/IJCTT-V72I3P108.

[11] OpenF1 API, https://openf1.org, [16.09.2025].