BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

Komparasi Kinerja Algoritma Machine Learning dalam Memprediksi Posisi Finish Pembalap Formula 1

Rizkya Nanda Prakoso*, Ade Priyatna, Ridatu Ocanitra

Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Jakarta, Indonesia
Email: [1,*]rizkyaprakoso4college@gmail.com, [2]ade.aeq@bsi.ac.id, [3]ridatu.rdo@bsi.ac.id
Email Penulis Korespondensi: rizkyaprakoso4college@gmail.com

Abstrak−Formula 1 merupakan kompetisi balap mobil dengan kompleksitas tinggi dalam menentukan posisi *finish* pembalap. Penelitian prediksi posisi *finish* Formula 1 menggunakan pendekatan regresi masih terbatas dibandingkan klasifikasi, sehingga performanya belum banyak dieksplorasi. Penelitian ini bertujuan untuk mengomparasikan kinerja algoritma *Ridge Regression, Support Vector Regression* (SVR), dan *Random Forest Regression* dalam memprediksi posisi *finish* pembalap Formula 1, mengidentifikasi faktor-faktor paling berpengaruh, serta memprediksinya pada Miami *Grand Prix* 2026. Dataset bersumber dari *Kaggle,* meliputi data historis Formula 1 periode 2021-2025 serta data tiga balapan awal musim 2026. Penelitian ini menerapkan *preprocessing, feature engineering, feature selection*, pemodelan dengan *hyperparameter tuning*, evaluasi serta interpretasi model terbaik *menggunakan Permutation Feature Importance* dan SHAP. Hasil penelitian menunjukkan bahwa SVR menjadi algoritma terbaik dengan nilai MAE 2,8685, RMSE 4,0919, R^2 0,4946, MAPE 37,2987%, dan *Pearson Correlation Coefficient* 0,7233. Nilai MAE dan MAPE SVR lebih baik dibandingkan *Ridge Regression* (MAE 3,1000; MAPE 56,8081%) dan *Random Forest Regression* (MAE 3,1611; MAPE 57,6259%), meskipun MAPE tersebut tergolong tinggi. Fitur *Position Qualifying* menjadi faktor paling berpengaruh, diikuti oleh *Driver Rank Previous, Constructor Rank Previous*, dan *Grid*. Pada *forecasting* Miami *Grand Prix* 2026, model SVR berhasil memprediksi pemenang balapan dengan tepat, meskipun selisih prediksi pada beberapa pembalap masih cukup besar. Penelitian ini berkontribusi dalam menyediakan kerangka komparatif model regresi linier dan non-linier untuk memprediksi posisi *finish* Formula 1, aspek yang belum banyak dieksplorasi sebelumnya, serta interpretasi hasil prediksi menggunakan PFI dan SHAP.

Kata Kunci: Formula 1; Machine Learning; Ridge Regression; Support Vector Regression; Random Forest Regression; Prediksi Posisi Finish

Abstract−Formula 1 is a motorsport competition with high complexity in determining drivers' finish positions. Research on Formula 1 finish position prediction using regression approaches remains limited compared to classification, leaving its performance underexplored. This study aims to compare the performance of Ridge Regression, Support Vector Regression (SVR), and Random Forest Regression in predicting Formula 1 drivers' finish positions, identify key factors, and predict finish positions at the 2026 Miami Grand Prix. The dataset, sourced from Kaggle, comprises historical Formula 1 data from 2021-2025 and data from the first three 2026 season races. This study applies preprocessing, feature engineering, feature selection, hyperparameter tuning, model evaluation, and interpretation of the best model using Permutation Feature Importance and SHAP. Results show that SVR performed best, with MAE 2.8685, RMSE 4.0919, R^2 0.4946, MAPE 37.2987%, and Pearson Correlation Coefficient 0.7233. SVR's MAE and MAPE outperformed Ridge Regression (MAE 3.1000; MAPE 56.8081%) and Random Forest Regression (MAE 3.1611; MAPE 57.6259%), although this MAPE remains high. Position Qualifying was the most influential feature, followed by Driver Rank Previous, Constructor Rank Previous, and Grid. In the 2026 Miami Grand Prix forecasting, SVR successfully predicted the race winner, although prediction gaps remained considerable for several drivers. This study contributes a comparative framework of linear and non-linear regression models for predicting Formula 1 finish positions, an aspect not widely explored, along with prediction interpretation using PFI and SHAP.

Keywords: Formula 1; Machine Learning; Ridge Regression; Support Vector Regression; Random Forest Regression; Finish Position Prediction

1. PENDAHULUAN

Formula 1 merupakan kompetisi balap mobil paling prestisius yang menggabungkan kemahiran pembalap dan performa mesin unggul di berbagai sirkuit internasional yang diselenggarakan oleh *Fédération Internationale de l'Automobile* (FIA) [1]. Secara umum, Formula 1 terdiri dari 10 tim konstruktor dan 20 pembalap yang berkompetisi dalam 22 hingga 24 *Grand Prix* setiap musimnya. Namun, pada musim 2026 terdapat perubahan regulasi serta ekspansi jumlah peserta menjadi 11 tim konstruktor dan 22 pembalap. Selain itu, terdapat pembatalan dua seri balapan, yaitu Bahrain *Grand Prix* dan Saudi Arabia *Grand Prix* akibat konflik perang antara Iran dengan Amerika Serikat.

Rangkaian balapan Formula 1 umumnya berlangsung selama tiga hari, dengan sesi kualifikasi (*qualifying*) terdiri atas tiga tahap (Q1, Q2, dan Q3) sebagai penentu posisi awal (*starting grid*) sebelum balapan utama (*main race*) dilaksanakan. Posisi kualifikasi memiliki pengaruh yang signifikan terhadap posisi *finish* pembalap, di mana semakin baik posisi yang diperoleh saat kualifikasi, semakin besar peluang pembalap untuk meraih hasil optimal di akhir balapan. Hal ini dibuktikan oleh Weissbock dan Mills (2025) yang menemukan bahwa posisi kualifikasi memiliki hubungan yang lebih kuat dengan posisi *finish* pembalap, dengan nilai koefisien kontingensi (C = 0,741) dan *Spearman's* (ρ = 0,763), dibandingkan dengan nilai koefisien kontingensi (C = 0,734) dan *Spearman's* (ρ = 0,748) pada *starting grid* [2].

Hasil tersebut mengindikasikan bahwa posisi kualifikasi merupakan indikator yang lebih andal terhadap peforma pembalap dan mobil dibandingkan posisi *starting grid*, karena tidak dipengaruhi oleh penalti yang diterapkan setelah sesi kualifikasi berakhir. Meskipun demikian, kompleksitas pada balapan Formula 1 menyebabkan posisi *finish* pembalap tetap bergantung pada faktor lain, seperti strategi *pit stop*, manajemen baterai mobil, pemilihan ban, kondisi lintasan, cuaca, hingga insiden balapan yang tidak terduga. Kondisi tersebut mendorong penerapan dan komparasi berbagai

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

algoritma *machine learning* untuk memprediksi serta mengidentifikasi faktor-faktor yang paling berpengaruh terhadap posisi *finish* pembalap.

Hal ini didukung oleh berbagai penelitian terdahulu yang telah menerapkan *machine learning* dengan beragam model. Krzysztoń dan Smołka (2026) memprediksi hasil balapan Formula 1 menggunakan teknik *ensemble learning* dengan menggabungkan algoritma *Support Vector Machine*, *Gradient Boosting* dan *Random Forest*, serta optimasi *hyperparameter* menggunakan pustaka Optuna [3]. Penelitian tersebut menghasilkan performa yang baik dengan nilai F1-*score* sebesar 77,83% dan akurasi sebesar 82,25%. Shelke et al. (2023) memprediksi pemenang balapan Formula 1 menggunakan algoritma *Support Vector Machine* (SVM) berdasarkan *lap times*, *sector times*, *qualifying times*, serta informasi pembalap dan tim konstruktor [4]. Penelitian tersebut mengklasifikasikan hasil akhir balapan menjadi tiga kelas, yaitu *Podium*, *Finish* dengan poin, dan Tanpa poin. Hasil penelitian menunjukkan bahwa model SVM menghasilkan tingkat akurasi lebih dari 97% serta nilai F1-*score* yang optimal pada setiap kelas.

Sementara itu, Patro et al. (2025) memprediksi waktu putaran balapan (*lap times*) Formula 1 menggunakan beberapa algoritma regresi, seperti *Decision Tree Regression*, *Random Forest Regression*, *Gradient Boosting Regression* (GBR), *K-Nearest Neighbor* (KNN), dan *XGBoost* (*Extreme Gradient Boosting*) [5]. Hasilnya menunjukkan bahwa model *Random Forest Regression* merupakan model terbaik dengan nilai R^2 tertinggi sebesar 0,9985 serta nilai MSE dan MAE terendah masing-masing sebesar 0,1678 dan 0,3057. Selain itu, Sicoie (2022) memprediksi posisi pemenang balapan Formula 1 dan kedudukan kejuaraan pembalap (*Championship Standings*) menggunakan *Random Forest Regression*, *Gradient Boosting Regressor*, dan *Support Vector Regression* (SVR) berbasis data historis Ergast API periode 2014-2022 dengan metode *5-fold cross-validation* [6]. Hasil penelitian menunjukkan bahwa *Gradient Boosting Regression* menghasilkan korelasi tertinggi ($\rho = 0{,}903$), sementara SVR menghasilkan nilai R^2 tertinggi sebesar 0,630 dengan tingkat kesalahan prediksi yang relatif rendah, yaitu sekitar 2-3 posisi dari hasil sebenarnya.

Berdasarkan penelitian terdahulu tersebut, dapat diidentifikasi beberapa kesenjangan penelitian (*research gap*). Pertama, sebagian besar penelitian masih menggunakan pendekatan klasifikasi untuk memprediksi hasil balapan Formula 1 sehingga belum mampu memprediksi urutan posisi *finish* pembalap secara spesifik. Kedua, penelitian yang menggunakan pendekatan regresi masih lebih banyak difokuskan pada prediksi *lap time* dibandingkan prediksi posisi *finish* pembalap. Ketiga, penerapan *Permutation Feature Importance* (PFI) dan SHAP untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap prediksi posisi *finish* pembalap Formula 1 masih terbatas. Selain itu, penerapan *Ridge Regression* dalam komparasi algoritma untuk memprediksi posisi *finish* pembalap Formula 1 juga masih belum banyak dieksplorasi. *Ridge Regression* dipilih sebagai representasi model linier untuk dibandingkan dengan model non-linier (SVR dan *Random Forest Regression*), guna mengevaluasi apakah kompleksitas model non-linier memberikan peningkatan performa dibandingkan pendekatan linier yang lebih sederhana dan interpretatif.

Oleh karena itu, penelitian ini dilakukan untuk mengisi kesenjangan tersebut dengan mengomparasikan kinerja algoritma *Ridge Regression, Support Vector Regression* (SVR), dan *Random Forest Regression* dalam memprediksi posisi *finish* pembalap Formula 1, mengidentifikasi faktor-faktor yang paling berpengaruh terhadap hasil prediksi dari model dengan kinerja terbaik, serta menerapkan model tersebut untuk memprediksi posisi *finish* pembalap pada Miami *Grand Prix* 2026 dan membandingkan hasil prediksi dengan hasil aktual balapan yang diperoleh dari situs resmi Formula 1. Kontribusi penelitian ini terletak pada penyediaan kerangka komparatif antara pendekatan regresi linier dan non-linier untuk prediksi posisi *finish* pembalap Formula 1, sebuah aspek yang belum banyak dieksplorasi karena penelitian terdahulu lebih banyak berfokus pada pendekatan klasifikasi atau prediksi *lap time*. Selain itu, penelitian ini memperkaya interpretabilitas model melalui PFI dan SHAP, serta memberikan bukti empiris atas peforma model pada kondisi balapan nyata melalui *forecasting* Miami *Grand Prix* 2026.

2. METODOLOGI PENELITIAN

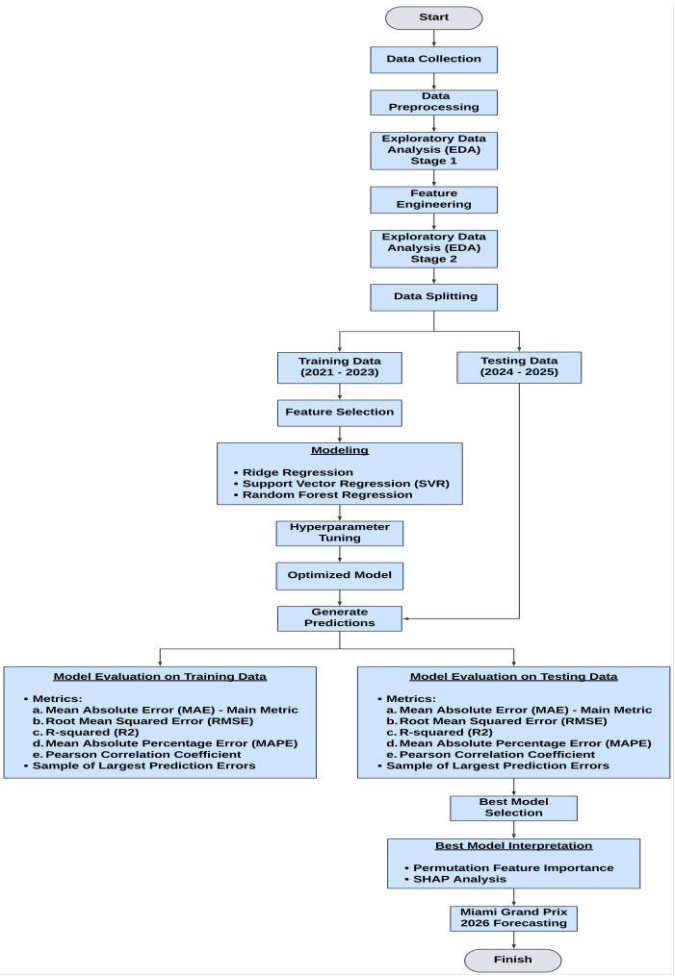
Penelitian ini menggunakan pendekatan kuantitatif berbasis *supervised learning* dengan mengomparasikan tiga algoritma *machine learning* untuk memprediksi posisi *finish* pembalap Formula 1. *Machine learning* merupakan bagian dari kecerdasan buatan yang berfokus pada pembelajaran pola dalam data sehingga sistem mampu melakukan prediksi dan membuat keputusan secara otomatis [7]. Dalam penelitian ini, pendekatan regresi digunakan karena mampu memprediksi nilai numerik berdasarkan hubungan antara variabel dependen dan variabel independen [8]. Meskipun posisi *finish* bersifat ordinal dan diskrit, pendekatan ini tetap digunakan dengan model menghasilkan nilai skor kontinu untuk setiap pembalap, yang selanjutnya diurutkan (*ranking*) terhadap pembalap lain dalam balapan yang sama untuk menghasilkan posisi *finish* prediksi dalam bentuk integer, sesuai dengan jumlah pembalap yang berpartisipasi pada balapan tersebut. Seluruh proses pengolahan dan analisis data dilakukan secara komputasional menggunakan bahasa pemrograman *Python* melalui *Jupyter Notebook* pada *Visual Studio Code* dengan memanfaatkan *library* seperti *Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn* dan SHAP.

2.1 Tahapan Penelitian

Tahapan penelitian ini dilakukan secara terstruktur mengikuti alur pada diagram *flowchart* yang disajikan pada Gambar 1. Berdasarkan Gambar 1, alur penelitian terdiri dari beberapa tahapan yang dilakukan secara sistematis. Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh melalui teknik dokumentasi. Setelah data berhasil dikumpulkan, penelitian dilanjutkan dengan tahapan prapemrosesan data, *exploratory data analysis* (EDA) *stage*

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

1, *feature engineering, exploratory data analysis* (EDA) *stage* 2, *data splitting, feature selection*, pemodelan *machine learning*, evaluasi model, pemilihan dan interpretasi model terbaik, serta *forecasting* Miami *Grand Prix* 2026.



Gambar 1. Diagram Flowchart

2.2 Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari *platform Kaggle*, dipublikasikan oleh James Trotman pada tahun 2026 dan dapat diakses pada https://www.kaggle.com/datasets/jtrotman/formula-1-race-data/data. Dataset tersebut terdiri dari 14 *file* CSV dan 121 kolom yang berisi informasi historis balapan Formula 1 sejak tahun 1950 hingga 2025, serta tiga balapan awal musim 2026 yaitu Australia, China, dan Japan *Grand Prix*. Penelitian ini difokuskan pada variabel yang relevan, sehingga hanya memanfaatkan 9 *file* CSV untuk memprediksi posisi *finish* pembalap Formula 1, yaitu *races, circuits, results, drivers, constructors, qualifying, lap times, driver standings* dan *constructor standings*.

2.3 Data Preprocessing

Data *Preprocessing* adalah proses pembersihan, penyesuaian format, dan persiapan data sehingga proses analisis dapat dilakukan dengan lebih akurat [9]. Tahapan *data preprocessing* dalam penelitian ini meliputi:

a. *Data Filtering*: dataset diseleksi berdasarkan periode tahun 2021-2025 agar sesuai dengan ruang lingkup penelitian.
b. *Data Integration*: sembilan *file* CSV digabungkan menjadi satu dataset utama menggunakan atribut penghubung *raceId, driverId, constructorId,* dan *circuitId*.
c. *Data Cleaning*: dilakukan pembersihan data dengan mengidentifikasi jumlah *missing value*, mengganti nilai tidak valid \N menjadi NaN, menghapus data duplikat, membersihkan spasi berlebih pada kolom teks, serta menyesuaikan tipe data pada beberapa kolom numerik.
d. *Data Transformation*: dilakukan konversi tipe data target *positionFinishOrder* menjadi integer, mengubah beberapa kolom menjadi format numerik, serta imputasi nilai hilang menggunakan nilai median. Median dipilih karena lebih tepat digunakan pada data numerik yang bersifat ordinal/posisi dan berpotensi memiliki *outlier*, dibandingkan mean yang lebih sensitif terhadap nilai ekstrem.

2.4 Exploratory Data Analysis Stage 1

Exploratory Data Analysis (EDA) merupakan tahap awal dalam analisis data yang bertujuan untuk mengidentifikasi pola, mendeteksi outlier, serta menguji dan memvalidasi asumsi terhadap data. EDA berperan penting dalam mengidentifikasi

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

anomali, memahami distribusi data, serta menganalisis hubungan antar variabel sebagai dasar dalam proses pemodelan [10]. Dalam penelitian ini, dilakukan dua tahap EDA, yaitu EDA *Stage* 1 dan *Stage* 2. EDA *Stage* 1 dilakukan sebelum tahap *Feature Engineering* untuk memahami karakteristik dan distribusi data awal. Proses ini mencakup analisis statistik deskriptif pada kolom numerik dan kategorikal untuk memperoleh gambaran umum karakteristik dataset, serta visualisasi *scatter plot* untuk mengkaji hubungan antara *grid position* dan *finish position*.

2.5 Feature Engineering

Feature engineering merupakan proses mengubah data mentah menjadi representasi fitur yang relevan dan dapat diaplikasikan dengan optimal dalam analisis dan pemodelan [11]. Proses ini meliputi pembentukan fitur performa historis pembalap dan konstruktor, penambahan tiga fitur agregasi *lap time*, yaitu *average lap time, best lap time,* dan *lap consistency*, penambahan fitur *best qualifying time* berdasarkan waktu kualifikasi terbaik di antara sesi Q1, Q2, dan Q3, serta penambahan fitur *finish points label* untuk merepresentasikan poin pembalap berdasarkan posisi *finish*.

2.6 Exploratory Data Analysis Stage 2

EDA *Stage* 2 dilakukan untuk memvalidasi fitur yang telah dibuat pada tahap *feature engineering*. Proses ini mencakup analisis korelasi menggunakan visualisasi *heatmap* untuk mengidentifikasi hubungan antar fitur terhadap posisi *finish*, serta visualisasi hubungan antara *best qualifying time* dengan posisi *finish* pembalap.

2.7 Data Splitting dan Feature Selection

Data dibagi menjadi data pelatihan menggunakan data tahun 2021-2023 dan data uji menggunakan data tahun 2024-2025 dengan pendekatan pembagian data berbasis waktu (*time-based split*). Meskipun pembagian data secara temporal ini berpotensi menimbulkan perbedaan karakteristik data antara periode pelatihan dan periode uji akibat perubahan regulasi atau performa tim dan pembalap antar musim, pendekatan ini tetap dipilih karena mempertahankan urutan waktu data, sehingga hubungan antar balapan yang direpresentasikan melalui fitur performa historis pembalap dan konstruktor (seperti *rank, points,* dan *wins* pada balapan-balapan sebelumnya) tetap konsisten antara data pelatihan dan data uji.

Selanjutnya dilakukan *feature selection,* yaitu metode prapemrosesan yang digunakan untuk mempertahankan fitur-fitur yang paling relevan terhadap variabel target sehingga dapat meningkatkan performa model serta mengurangi kompleksitas dan beban komputasi [12]. Proses seleksi fitur dilakukan hanya pada data pelatihan untuk menghindari kebocoran data (*data leakage*) menggunakan metode *filter* berbasis korelasi *Pearson* terhadap posisi *finish* sebagai variabel target. Fitur dengan korelasi tinggi dipertahankan karena menunjukkan hubungan yang kuat terhadap variabel target, sedangkan fitur dengan korelasi rendah dieliminasi karena kontribusinya yang relatif kecil terhadap proses prediksi.

2.8 Modeling dan Hyperparameter Tuning

Penelitian ini menerapkan tiga algoritma *machine learning* berbasis regresi, yaitu *Ridge Regression* yang menggunakan regularisasi untuk mengatasi multikolinearitas [13], *Support Vector Regression* (SVR) yang merupakan pengembangan dari *Support Vector Machine* (SVM) untuk permasalahan regresi [14], dan *Random Forest Regression* yang merupakan metode *ensemble learning* berbasis kumpulan pohon keputusan untuk menghasilkan prediksi numerik [15]. Ketiga algoritma tersebut dibangun menggunakan data pelatihan yang telah melalui tahap *feature selection*.

Sementara itu, *feature scaling* merupakan proses menyeragamkan rentang nilai antar fitur dalam dataset [16]. Dalam penelitian ini, *feature scaling* menggunakan metode *StandardScaler* diterapkan khusus pada algoritma *Ridge Regression* dan *Support Vector Regression* (SVR) untuk menyamakan skala seluruh data numerik dalam dataset agar model tidak bias terhadap fitur-fitur yang memiliki rentang nilai tinggi. Sebaliknya, *Random Forest Regression* tidak menerapkan *feature scaling* karena algoritma berbasis pohon keputusan melakukan pembagian data berdasarkan nilai ambang (*threshold*) pada tiap fitur secara independen, sehingga tidak sensitif terhadap perbedaan skala antar fitur. Dengan demikian, penerapan *feature scaling* secara selektif ini disesuaikan dengan karakteristik masing-masing algoritma dan tidak menciptakan ketimpangan perlakuan dalam proses komparasi model.

Selanjutnya, *hyperparameter tuning* diterapkan pada setiap algoritma menggunakan *GridSearchCV* dengan *5-fold cross-validation* pada data pelatihan untuk memperoleh kombinasi parameter terbaik. Model dengan parameter terbaik kemudian digunakan untuk menghasilkan prediksi pada data pelatihan dan data uji. Seluruh parameter yang diuji untuk setiap algoritma dapat dilihat pada Tabel 1 berikut.

Tabel 1. Hyperparameter Tuning Model

Algoritma	Hyperparameter	Nilai Parameter
Ridge Regression	Alpha	0.0001, 0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10, 25, 50, 100, 250
	Solver	Automatic, Singular Value Decomposition (SVD), Cholesky Decomposition, Least Square QR (LSQR), dan Stochastic Average Gradient (SAG).
Support Vector Regression (SVR)	C	0.5, 1, 3, 5, 10, 20, 50, 100
	Epsilon	0.01, 0.05, 0.1, 0.2, 0.3, 0.5
	Gamma	Scale, Auto, 0.001, 0.01, 0.05, 0.1

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

Algoritma	Hyperparameter	Nilai Parameter
	Kernel	Radial Basis Function (RBF).
Random Forest Regression	n_estimators	100, 200
	max_depth	5, 7, 10
	min_samples_split	5, 10, 15
	min_samples_leaf	2, 4, 6
	max_features	SQRT dan Log2

Berdasarkan Tabel 1, *Random Forest Regression* memiliki jumlah *hyperparameter* yang paling banyak diuji dibandingkan kedua algoritma lainnya, sedangkan *Ridge Regression* memiliki jumlah *hyperparameter* yang paling sedikit.

2.9 Model Evaluation

Evaluasi model diterapkan pada seluruh algoritma menggunakan data pelatihan dan data uji untuk menilai kinerja model serta mengidentifikasi potensi *overfitting* dan *underfitting*. Metrik evaluasi yang digunakan meliputi *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *R-squared* (R^2), *Mean Absolute Percentage Error* (MAPE), dan *Pearson Correlation Coefficient*. MAE digunakan untuk mengukur rata-rata selisih absolut antara nilai aktual dan nilai prediksi, sedangkan RMSE digunakan untuk mengukur rata-rata kesalahan prediksi berdasarkan akar dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. R^2 digunakan untuk mengukur kemampuan model dalam menjelaskan variasi data yang teramati [17], MAPE digunakan untuk mengukur rata-rata persentase kesalahan antara nilai prediksi dan nilai aktual [18], dan *Pearson Correlation Coefficient* digunakan untuk mengukur tingkat hubungan linier antara dua variabel, di mana semakin tinggi nilai koefisien korelasinya, maka hubungan antara kedua variabel tersebut semakin kuat [19].

2.10 Pemilihan Model Terbaik dan Interpretasi Model

Pemilihan model terbaik didasarkan pada hasil evaluasi kinerja model, yaitu rendahnya nilai MAE, RMSE, dan MAPE, serta tingginya nilai R^2 dan *Pearson Correlation Coefficient*. Model dengan kinerja terbaik kemudian diinterpretasikan menggunakan dua metode, yaitu *Permutation Feature Importance* (PFI) dan *Shapley Additive exPlanations* (SHAP) untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap posisi *finish* pembalap. PFI mengukur tingkat pengaruh suatu fitur terhadap performa model dengan cara mengacak nilai pada setiap fitur secara bergantian. Setelah itu, performa model dibandingkan dengan kondisi sebelum pengacakan dilakukan [20]. Sementara itu, SHAP merupakan metode interpretasi model yang memberikan nilai kontribusi pada setiap fitur sehingga pengaruh masing-masing fitur terhadap hasil prediksi dapat dijelaskan secara transparan [21]. Dalam penelitian ini, hasil analisis SHAP divisualisasikan menggunakan *Global Importance Bar Plot* untuk menampilkan rata-rata nilai absolut setiap fitur, serta *Beeswarm Plot* untuk menunjukkan distribusi nilai SHAP dan arah pengaruhnya pada setiap sampel.

2.11 Prediksi Miami Grand Prix 2026

Prediksi pada Miami *Grand Prix* 2026 dilakukan menggunakan algoritma terbaik yang telah melalui tahap *hyperparameter tuning* dan komparasi evaluasi model. Proses prediksi memanfaatkan data performa pembalap dan konstruktor dari tiga balapan pertama Formula 1 musim 2026, yaitu Australia, China, dan Japan *Grand Prix*, termasuk pembalap serta tim konstruktor baru yang bergabung pada musim tersebut. Selain itu, prediksi juga menggunakan fitur hasil *feature selection*, serta nilai rata-rata posisi *qualifying* dan *grid* dari ketiga balapan tersebut untuk merepresentasikan level performa terkini pembalap dan konstruktor menjelang Miami *Grand Prix* 2026. Hasil prediksi selanjutnya dibandingkan dengan hasil aktual balapan Miami *Grand Prix* 2026 yang diperoleh dari situs resmi Formula 1.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data dan Data Preprocessing

Dataset Formula 1 yang diperoleh dari Kaggle terdiri dari 14 file CSV dengan total 121 kolom data. Namun, penelitian ini hanya menggunakan 9 dataset yang dipilih berdasarkan relevansi atribut terhadap proses prediksi posisi finish pembalap. Setelah dilakukan *filtering* pada periode 2021 hingga 2025, diperoleh sebanyak 114 balapan yang digunakan dalam tahap analisis. Perubahan jumlah data pada setiap dataset sebelum dan setelah proses *filtering* disajikan pada Tabel 2 berikut.

Tabel 2. Dimensi Dataset Sebelum dan Setelah Filtering

No	Dataset	Sebelum Filtering	Setelah Filtering
1	Races	(1.171, 18)	(114, 18)
2	Circuits	(78, 9)	(28, 9)
3	Results	(27.304, 18)	(2.278, 18)
4	Drivers	(865, 9)	(35, 9)
5	Constructors	(214, 5)	(12, 5)
6	Qualifying	(11.036, 9)	(2.277, 9)

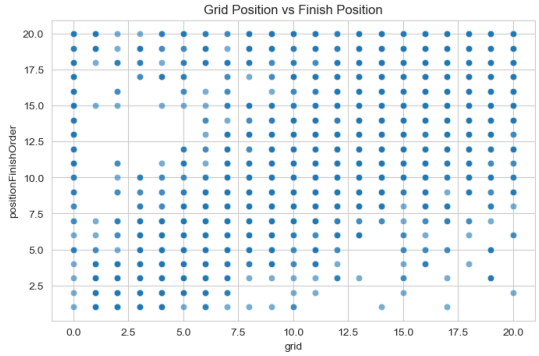
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

No	Dataset	Sebelum Filtering	Setelah Filtering
7	Lap Times	(618.766, 6)	(124.834, 6)
8	Driver Standings	(35.427, 7)	(2.396, 7)
9	Constructor Standings	(13.664, 7)	(1.140, 7)

Selain itu, proses penggabungan (*merge*) sembilan dataset menghasilkan dataset utama yang terdiri dari 2.278 baris data dan 23 atribut asli. Tahap *data cleaning* dan *transformation* berhasil dilakukan sehingga data menjadi bersih, konsisten, dan siap digunakan untuk proses pemodelan.

3.2 Exploratory Data Analysis Stage 1

Analisis statistik deskriptif menunjukkan bahwa pada kolom numerik, nilai rara-rata *grid, qualifying*, dan *finish* berada pada kisaran 10 dengan nilai maksimum mencapai 20, sehingga menunjukkan distribusi posisi balapan yang relatif merata. Sementara itu, analisis pada kolom kategorikal menunjukkan bahwa dataset mencakup 35 pembalap, 12 tim konstruktor, 28 nama sirkuit, dan 26 negara yang merepresentasikan karakteristik balapan Formula 1 selama periode penelitian. Selain itu, hasil visualisasi *scatter plot* disajikan pada Gambar 2 berikut.

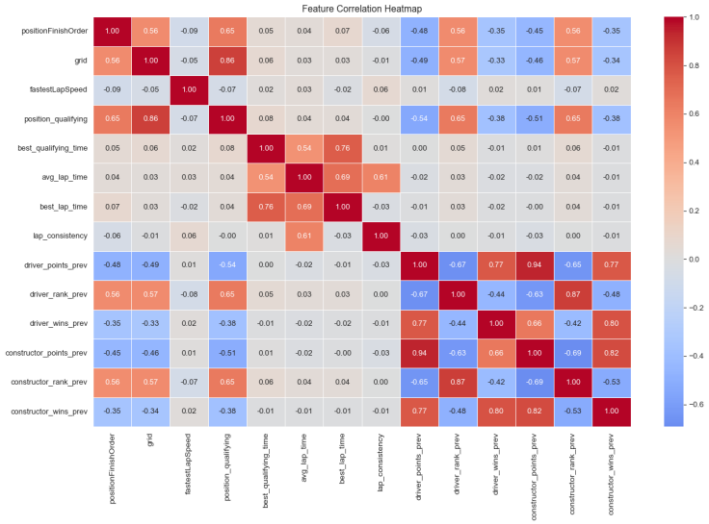


Gambar 2. Hubungan Grid dan Finish Position

Berdasarkan Gambar 2, terdapat kecenderungan bahwa pembalap dengan posisi *grid* yang lebih baik cenderung memperoleh posisi *finish* yang lebih baik, meskipun masih terdapat variasi hasil balapan akibat faktor tertentu selama balapan berlangsung.

3.3 Feature Engineering dan Exploratory Data Analysis Stage 2

Setelah dilakukan proses *feature engineering*, dataset akhir memiliki total 34 kolom yang terdiri dari 23 atribut asli dan 11 fitur tambahan, meliputi performa historis pembalap dan konstruktor, agregasi *lap time*, *best qualifying time*, dan *finish points label*. Selanjutnya, pada tahap EDA *Stage* 2 dilakukan analisis korelasi antar fitur menggunakan *heatmap* sebagaimana ditunjukkan pada Gambar 3 berikut.

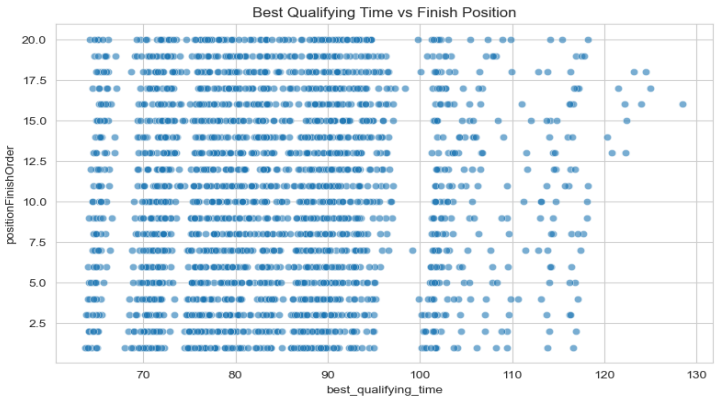


Gambar 3. Feature Correlation Heatmap

Berdasarkan visualisasi pada Gambar 3, terdapat beberapa fitur yang memiliki korelasi kuat, seperti *grid* dengan *position qualifying*, serta *driver points previous* dengan *constructor points previous*. Selain itu, *positionFinishOrder* menunjukkan korelasi tertinggi dengan *position qualifying* dibandingkan fitur lainnya. Hubungan antara fitur *best*

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

qualifying time dan posisi *finish* kemudian divalidasi lebih lanjut menggunakan visualisasi *scatter plot* sebagaimana ditunjukkan pada Gambar 4 berikut.



Gambar 4. Best Qualifying Time dan Posisi Finish

Berdasarkan Gambar 4, pembalap dengan waktu kualifikasi yang lebih cepat cenderung memperoleh posisi *finish* yang lebih baik, mengonfirmasi relevansi fitur kualifikasi sebagai prediktor utama posisi *finish* pembalap.

3.4 Data Splitting dan Feature Selection

Setelah dilakukan *data splitting*, diperoleh 1,320 baris data pelatihan (2021-2023) dan 958 baris data uji (2024-2025). Selanjutnya, sebanyak 13 fitur kandidat dipilih dari 34 kolom dataset dengan mengeliminasi variabel target, identitas, temporal, dan fitur yang berpotensi *data leakage*. Fitur-fitur kandidat tersebut kemudian diseleksi menggunakan metode filter berbasis korelasi *Pearson* terhadap posisi *finish* sebagai variabel target. Hasil seleksi fitur ditunjukkan pada Tabel 3 berikut.

Tabel 3. Hasil Nilai Korelasi Terhadap Posisi Finish

Fitur	Nilai Korelasi	Status
Position Qualifying	0,610	Terpilih
Driver Rank Previous	0,554	Terpilih
Constructor Rank Previous	0,551	Terpilih
Grid	0,518	Terpilih
Driver Points Previous	-0,465	Terpilih
Constructor Points Previous	-0,451	Terpilih
Constructor Wins Previous	-0,344	Terpilih
Driver Wins Previous	-0,329	Terpilih
Fastest Lap Speed	-0,098	Dieliminasi
Lap Consistency	-0,069	Dieliminasi
Best Lap Time	-0,064	Dieliminasi
Best Qualifying Time	-0,056	Dieliminasi
Average Lap Time	-0,029	Dieliminasi

Berdasarkan hasil analisis korelasi pada Tabel 3, fitur dengan nilai absolut korelasi di bawah 0,1 terhadap posisi *finish* dieliminasi karena kontribusinya yang kurang signifikan. Hasil seleksi menghasilkan delapan fitur yang digunakan pada tahap pemodelan, yaitu *position qualifying, driver rank previous, constructor rank previous, grid, driver points previous, constructor points previous, constructor wins previous,* dan *driver wins previous*.

3.5 Modeling dan Hyperparameter Tuning

Sebelum proses pelatihan model, *feature scaling* menggunakan *StandardScaler* diterapkan khusus pada algoritma *Ridge Regression* dan SVR untuk menyeragamkan skala fitur numerik. Selanjutnya, optimasi *hyperparameter* dilakukan menggunakan *GridSearchCV* dengan *5-fold cross-validation* untuk memperoleh konfigurasi parameter terbaik pada masing-masing algoritma, sebagaimana disajikan pada Tabel 4.

Tabel 4. Hasil Hyperparameter Tuning

Algoritma	Parameter Terbaik	Skor
Ridge Regression	Alpha = 0,0001, solver = SAG	3,374
Support Vector Regression (SVR)	C = 5, epsilon = 0,01, gamma = 0,01, kernel = RBF	3,139
Random Forest Regression	max_depth = 7, max_features = log2, min_samples_leaf = 2, min_samples_split = 5, dan n_estimators = 200	3,360

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

Hasil *hyperparameter tuning* pada Tabel 4 menunjukkan bahwa *Support Vector Regression* (SVR) memperoleh skor *cross-validation* terendah sebesar 3,139 yang mengindikasikan performa terbaik pada tahap optimasi parameter.

3.6 Komparasi Kinerja Algoritma Machine Learning

Seluruh konfigurasi parameter optimal yang diperoleh dari proses *hyperparameter tuning* kemudian digunakan untuk melatih dan mengevaluasi ketiga algoritma. Evaluasi dilakukan pada data pelatihan dan data uji menggunakan lima metrik, yaitu MAE, RMSE, R^2, MAPE, dan *Pearson Correlation Coefficient*. Pada metrik MAE, RMSE, dan MAPE, nilai yang rendah menunjukkan performa yang lebih baik, sedangkan pada metrik R^2 dan *Pearson Correlation Coefficient*, nilai yang lebih tinggi menunjukkan performa yang lebih baik. Hasil evaluasi pada data pelatihan disajikan pada Tabel 5, sedangkan hasil evaluasi pada data uji disajikan pada Tabel 6 berikut.

Tabel 5. Hasil Evaluasi Pada Data Pelatihan

Model	Train MAE	Train RMSE	Train R^2	Train MAPE (%)	Train Correlation
Support Vector Regression (SVR)	3,0960	4,5597	0,3747	37,5892	0,6499
Ridge Regression	3,3502	4,3831	0,4222	57,4373	0,6498
Random Forest Regression	2,8288	3,6943	0,5895	48,5766	0,7749

Berdasarkan Tabel 5, *Random Forest Regression* menunjukkan kinerja terbaik pada data pelatihan dengan memperoleh nilai MAE dan RMSE terendah, masing-masing sebesar 2,8288 dan 3,6943, serta nilai R^2 *dan Pearson Correlation tertinggi,* masing-masing sebesar 0,5895 dan 0,7749. Sementara itu, *Ridge Regression* memperoleh nilai MAPE tertinggi sebesar 57,4373%, sedangkan SVR memperoleh nilai RMSE tertinggi sebesar 4,5597 dan nilai R^2 terendah sebesar 0,3747.

Tabel 6. Hasil Evaluasi Pada Data Uji

Model	Test MAE	Test RMSE	Test R^2	Test MAPE (%)	Test Correlation
Support Vector Regression (SVR)	2,8685	4,0919	0,4946	37,2987	0,7233
Ridge Regression	3,1000	4,0104	0,5146	56,8081	0,7214
Random Forest Regression	3,1611	4,0834	0,4967	57,6259	0,7071

Sementara pada Tabel 6, algoritma *Support Vector Regression* (SVR) menghasilkan nilai MAE dan MAPE terendah, masing-masing sebesar 2,8685 dan 37,2987%, serta nilai RMSE sebesar 4,0919 dan R^2 sebesar 0,4946. Selain itu, SVR memperoleh nilai *Pearson Correlation Coefficient* tertinggi sebesar 0,7233. Di sisi lain, *Ridge Regression* unggul pada nilai RMSE terendah sebesar 4,0104 dan nilai R^2 tertinggi sebesar 0,5146. Adapun *Random Forest Regression* menghasilkan nilai RMSE sebesar 4,0834 dan nilai R^2 sebesar 0,4967 yang sedikit lebih baik dibandingkan SVR pada kedua metrik tersebut. Dengan demikian, SVR menunjukkan kinerja terbaik pada tiga dari lima metrik evaluasi, yaitu MAE, MAPE, dan *Pearson Correlation Coefficient*, sedangkan *Ridge Regression* unggul pada RMSE dan R^2. Meskipun SVR menunjukkan kinerja terbaik pada sebagian besar metrik, nilai *error* yang dihasilkan masih menunjukkan bahwa akurasi prediksi belum optimal. Nilai MAE sebesar 2,8685 menunjukkan bahwa hasil prediksi rata-rata menyimpang hampir tiga posisi dari posisi aktual, sedangkan nilai MAPE sebesar 37,2987% menunjukkan bahwa tingkat kesalahan relatif model masih tergolong cukup tinggi. Hal ini menunjukkan bahwa SVR merupakan model dengan kinerja relatif terbaik di antara algoritma yang dibandingkan, tetapi hasil prediksinya masih memiliki keterbatasan dalam mendekati posisi aktual secara konsisten.

Sementara itu, perbandingan hasil evaluasi pada data pelatihan dan data uji menunjukkan bahwa algoritma SVR memiliki kinerja yang paling stabil dengan selisih nilai metrik yang relatif kecil. *Ridge Regression* juga menunjukkan konsistensi kinerja yang baik, sedangkan *Random Forest Regression* mengindikasikan kecenderungan *overfitting* karena menghasilkan performa yang lebih baik pada data pelatihan dibandingkan data uji. Kecenderungan tersebut tetap terlihat meskipun proses *hyperparameter tuning* pada *Random Forest Regression* telah mempertimbangkan parameter pembatas kompleksitas model, yaitu *max_depth, min_samples_split,* dan *min_samples_leaf*, yang digunakan untuk membatasi kompleksitas pohon keputusan. Hal ini menunjukkan bahwa pembatasan kompleksitas yang diterapkan belum sepenuhnya menghilangkan perbedaan performa antara data pelatihan dan data uji. Dengan demikian, model *Random Forest Regression* masih menunjukkan kemampuan yang lebih baik dalam menangkap pola pada data pelatihan dibandingkan pola yang dapat digeneralisasikan pada data uji. Oleh karena itu, penelitian selanjutnya dapat mempertimbangkan penambahan jumlah data pelatihan atau penerapan strategi pengendalian kompleksitas model yang lebih lanjut untuk mengurangi kecenderungan *overfitting* tersebut.

3.7 Model Terbaik dan Interpretasi Model

Berdasarkan hasil komparasi kinerja algoritma, *Support Vector Regression* (SVR) ditetapkan sebagai algoritma terbaik pada penelitian ini. Oleh karena itu, algoritma SVR dipilih untuk tahap interpretasi model menggunakan metode *Permutation Feature Importance* (PFI) dan *Shapley Additive exPlanations* (SHAP). Analisis PFI dilakukan menggunakan 10 iterasi permutasi dengan metrik evaluasi *negative mean absolute error*. Hasil analisis PFI pada algoritma SVR disajikan pada Tabel 7. Berdasarkan Tabel 7, fitur *position qualifying* memiliki nilai *importance* tertinggi sebesar 1,225 sehingga menjadi fitur yang paling berpengaruh terhadap prediksi posisi *finish* pembalap. Selain itu, fitur *driver rank previous, constructor rank previous*, dan *grid* juga menunjukkan pengaruh yang cukup besar terhadap kinerja model,

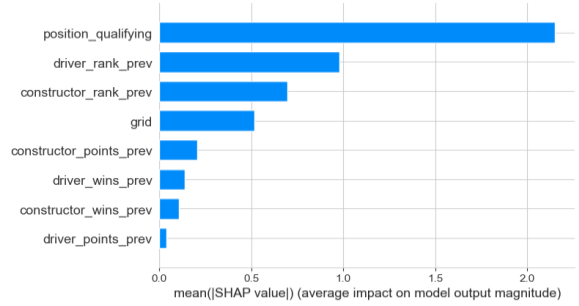
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

sedangkan fitur *constructor points previous, driver wins previous, driver points previous*, dan *constructor wins previous* turut berkontribusi namun dengan tingkat pengaruh yang jauh lebih kecil.

Tabel 7. Hasil Permutation Feature Importance (PFI) pada Algoritma SVR

No	Feature	Nilai Importance
1	Position Qualifying	1,225
2	Driver Rank Previous	0,273
3	Constructor Rank Previous	0,141
4	Grid	0,122
5	Constructor Points Previous	0,048
6	Driver Wins Previous	0,032
7	Driver Points Previous	0,029
8	Constructor Wins Previous	0,014

Sementara itu, interpretasi model menggunakan SHAP dilakukan pada 100 sampel data uji yang dipilih secara acak untuk meningkatkan efisiensi komputasi. Hasil interpretasi kemudian disajikan menggunakan dua representasi visual, yaitu *Global Importance Bar Plot* dan *Beeswarm Plot*. Hasil visualisasi SHAP ditunjukkan pada Gambar 5 dan Gambar 6.



Gambar 5. SHAP Global Feature Importance pada Model SVR

Gambar 5 menunjukkan hasil SHAP *Global Feature Importance* pada model SVR. Hasil tersebut menunjukkan pola yang konsisten dengan *Permutation Feature Importance* (PFI), sehingga memperkuat interpretasi bahwa fitur kualifikasi dan performa historis pembalap maupun konstruktor memiliki kontribusi penting terhadap prediksi posisi *finish* pembalap. Selanjutnya, visualisasi *Beeswarm Plot* disajikan pada Gambar 6 berikut.

Gambar 6. SHAP Beeswarm Plot pada Model SVR

Gambar 6 menampilkan distribusi nilai SHAP pada setiap fitur terhadap hasil prediksi model. Warna merah merepresentasikan nilai fitur yang tinggi, sedangkan warna biru merepresentasikan nilai fitur yang rendah. Berdasarkan visualisasi tersebut, nilai *position qualifying* yang lebih tinggi cenderung menghasilkan nilai SHAP positif, yang menunjukkan bahwa pembalap dengan posisi kualifikasi yang kurang baik cenderung diprediksi *finish* di posisi belakang. Sebaliknya, nilai *driver rank previous* dan *constructor rank previous* yang rendah cenderung menghasilkan nilai SHAP negatif, yang mengindikasikan bahwa pembalap dan konstruktor dengan peringkat historis yang lebih baik cenderung diprediksi *finish* di posisi depan.

3.8 Prediksi Miami Grand Prix 2026

Pada tahap ini, model SVR digunakan untuk memprediksi posisi *finish* pembalap pada Miami *Grand Prix* 2026. Hasil prediksi kemudian dibandingkan dengan hasil aktual balapan untuk mengevaluasi kinerja model. Hasil prediksi posisi *finish* pembalap Miami *Grand Prix* 2026 disajikan pada Gambar 7 berikut.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

	Driver_Name	Constructor_Name	Predicted_Score	Predicted_Finish_Position	Predicted_Points
0	Andrea Kimi Antonelli	Mercedes	1.482050	1	25
1	George Russell	Mercedes	1.925589	2	18
2	Charles Leclerc	Ferrari	3.879576	3	15
3	Lewis Hamilton	Ferrari	4.790485	4	12
4	Oscar Piastri	McLaren	4.992931	5	10
5	Lando Norris	McLaren	5.541516	6	8
6	Isack Hadjar	Red Bull	8.491207	7	6
7	Pierre Gasly	Alpine F1 Team	8.876852	8	4
8	Max Verstappen	Red Bull	9.862593	9	2
9	Oliver Bearman	Haas F1 Team	10.551523	10	1
10	Arvid Lindblad	RB F1 Team	11.143714	11	0
11	Liam Lawson	RB F1 Team	11.301534	12	0
12	Esteban Ocon	Haas F1 Team	11.690911	13	0
13	Gabriel Bortoleto	Audi	11.992415	14	0
14	Nico Hülkenberg	Audi	12.706092	15	0
15	Franco Colapinto	Alpine F1 Team	12.989968	16	0
16	Carlos Sainz	Williams	14.783773	17	0
17	Alexander Albon	Williams	15.430239	18	0
18	Valtteri Bottas	Cadillac	16.745550	19	0
19	Sergio Pérez	Cadillac	16.856374	20	0
20	Fernando Alonso	Aston Martin	16.874548	21	0
21	Lance Stroll	Aston Martin	17.694667	22	0

Gambar 7. Hasil Prediksi Posisi Finish Pembalap Miami Grand Prix 2026

Gambar 7 menampilkan hasil prediksi model SVR terhadap posisi *finish* seluruh pembalap pada Miami *Grand Prix* 2026. Model memprediksi Andrea Kimi Antonelli sebagai pemenang balapan dengan posisi *finish* pertama, diikuti oleh George Russell, dan Charles Leclerc pada posisi tiga besar. Selanjutnya, hasil prediksi tersebut dibandingkan dengan hasil aktual balapan Miami *Grand Prix* 2026 yang diperoleh dari situs resmi Formula 1, sebagaimana disajikan pada Gambar 8 berikut.

MIAMI GP 2026 - ACTUAL VS PREDICTED

	Driver_Name	Constructor_Name	Actual_Finish_Position	Predicted_Finish_Position	Predicted_Points	Actual_Points	Error	Abs_Error
0	Andrea Kimi Antonelli	Mercedes	1	1	25	25	0	0
1	Lando Norris	McLaren	2	6	8	18	4	4
2	Oscar Piastri	McLaren	3	5	10	15	2	2
3	George Russell	Mercedes	4	2	18	12	-2	2
4	Max Verstappen	Red Bull	5	9	2	10	4	4
5	Lewis Hamilton	Ferrari	6	4	12	8	-2	2
6	Franco Colapinto	Alpine F1 Team	7	16	0	6	9	9
7	Charles Leclerc	Ferrari	8	3	15	4	-5	5
8	Carlos Sainz	Williams	9	17	0	2	8	8
9	Alexander Albon	Williams	10	18	0	1	8	8
10	Oliver Bearman	Haas F1 Team	11	10	1	0	-1	1
11	Gabriel Bortoleto	Audi	12	14	0	0	2	2
12	Esteban Ocon	Haas F1 Team	13	13	0	0	0	0
13	Arvid Lindblad	RB F1 Team	14	11	0	0	-3	3
14	Fernando Alonso	Aston Martin	15	21	0	0	6	6
15	Sergio Pérez	Cadillac	16	20	0	0	4	4
16	Lance Stroll	Aston Martin	17	22	0	0	5	5
17	Valtteri Bottas	Cadillac	18	19	0	0	1	1
18	Nico Hülkenberg	Audi	19	15	0	0	-4	4
19	Liam Lawson	RB F1 Team	20	12	0	0	-8	8
20	Pierre Gasly	Alpine F1 Team	21	8	4	0	-13	13
21	Isack Hadjar	Red Bull	22	7	6	0	-15	15

Gambar 8. Perbandingan Posisi Finish Aktual dan Prediksi Miami Grand Prix 2026

Berdasarkan hasil perbandingan pada Gambar 8, SVR mampu menghasilkan beberapa prediksi yang cukup mendekati hasil aktual balapan. Model berhasil memprediksi Andrea Kimi Antonelli sebagai pemenang Miami *Grand Prix* 2026 di posisi pertama dan Esteban Ocon di posisi ketiga belas dengan tepat tanpa selisih. Selain itu, selisih prediksi yang relatif kecil juga terlihat pada beberapa pembalap lain seperti Oscar Piastri, George Russell, Lewis Hamilton, Oliver Bearman, Gabriel Bortoleto, dan Valtteri Bottas. Meskipun demikian, masih terdapat beberapa pembalap yang menunjukkan selisih prediksi cukup besar dibandingkan hasil aktual. Hal ini disebabkan oleh berbagai faktor yang terjadi selama balapan berlangsung dan tidak dapat diantisipasi oleh model, seperti insiden kecelakaan, kendala mekanis, strategi *pit stop*, maupun perubahan performa mobil selama musim berlangsung.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1896-1908
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1190

Sebagai contoh, Nico Hulkenberg, Liam Lawson, Pierre Gasly, dan Isack Hadjar tidak berhasil menyelesaikan balapan (*Did Not Finish*), sehingga posisi aktual mereka jauh berbeda dari prediksi model, dengan selisih prediksi terbesar terjadi pada Isack Hadjar dengan *absolute error* sebesar 15. Secara keseluruhan, model SVR mampu memberikan beberapa prediksi yang mendekati posisi aktual pembalap pada Miami *Grand Prix* 2026. Namun, faktor-faktor tak terduga selama balapan tetap menjadi keterbatasan yang tidak dapat diakomodasi oleh model.

3.9 Pembahasan

Hasil komparasi kinerja ketiga algoritma menunjukkan bahwa *Support Vector Regression* (SVR) unggul pada metrik MAE, MAPE, dan *Pearson Correlation Coefficient*, dengan MAPE sebesar 37,2987% yang jauh lebih rendah dibandingkan *Ridge Regression* dan *Random Forest Regression* (RFR), didukung oleh kernel *Radial Basis Function* (RBF) yang memungkinkan model menangkap hubungan nonlinier antara fitur dan posisi *finish* pembalap. Namun, pada RMSE dan R^2, SVR justru terendah (R^2 = 0,4946) dibandingkan RFR (R^2 = 0,4967) dan *Ridge Regression* (R^2 = 0,5146), serta lebih rendah dari penelitian Sicoie (2022) yang melaporkan R^2 SVR sebesar 0,630 [6]. Perbedaan ini mungkin dipengaruhi oleh perbedaan periode data, fitur yang digunakan, dan variabilitas hasil balapan pada musim yang dianalisis.

Performa model regresi pada penelitian ini perlu dipahami dalam konteks yang berbeda dari penelitian berbasis klasifikasi, seperti Krzysztoń dan Smołka (2026) yang melaporkan F1-*score* 77,83% dan akurasi 82,25% [3], serta Shelke et al. (2023) yang melaporkan akurasi >97% pada klasifikasi tiga kelas (Podium, *Finish* dengan Poin, dan Tanpa Poin) [4]. Kedua pendekatan tersebut tidak dapat dibandingkan secara langsung karena mengevaluasi tugas prediksi yang berbeda: klasifikasi hanya menempatkan pembalap ke dalam beberapa kelas kategori, sedangkan penelitian ini menggunakan regresi untuk memprediksi urutan posisi setiap pembalap *finish* secara spesifik. Hal ini menegaskan pentingnya penelitian ini dalam mengisi kesenjangan yang diidentifikasi sebelumnya, yaitu bahwa pendekatan klasifikasi belum mampu memprediksi urutan posisi *finish* pembalap secara spesifik.

Ridge Regression memperoleh RMSE terendah (4,0104) dan R^2 tertinggi (0,5146) pada data uji, bahkan lebih tinggi dibandingkan data pelatihan (R^2 = 0,4222), yang menunjukkan model ini mampu melakukan generalisasi dengan baik tanpa indikasi *overfitting*. Namun, MAPE *Ridge Regression* mencapai 56,8081% jauh lebih tinggi dibandingkan SVR, sehingga performanya secara keseluruhan tetap dinilai belum melampaui SVR. Temuan ini melengkapi kesenjangan pada penelitian Sicoie (2022), yang tidak menyertakan model linier dalam perbandingannya, sehingga belum dapat menunjukkan apakah kompleksitas model nonlinier benar-benar memberikan keunggulan dibandingkan pendekatan linier yang lebih sederhana [6].

Random Forest Regression menunjukkan performa terbaik pada data pelatihan hampir di semua metrik (R^2 = 0,5895; *Correlation* = 0,7749), namun menurun signifikan pada data uji (R^2 = 0,4967; *Correlation* = 0,7071) dengan MAE 3,1611 dan MAPE 57,6259% tertinggi di antara ketiga algoritma. Pola ini mengindikasikan *overfitting*, berbeda dengan *Ridge* dan SVR yang justru stabil atau meningkat dari *train* ke *test*. Temuan ini juga berbeda dengan Patro et al. (2025), yang melaporkan RFR sebagai model terbaik dengan R^2 sebesar 0,9985 dan MAE sebesar 0,3057 dalam prediksi *lap time* [5]. Perbedaan ini menunjukkan bahwa algoritma *ensemble* sangat bergantung pada karakteristik target: *lap time* bersifat kontinu dan konsisten antar putaran, sedangkan posisi *finish* bersifat ordinal dan rentan terhadap variabilitas tinggi seperti insiden balapan. Ini mendukung kesenjangan bahwa penelitian regresi terdahulu lebih banyak berfokus pada *lap time* dibandingkan posisi *finish*.

Secara keseluruhan, SVR ditetapkan sebagai algoritma dengan kinerja relatif terbaik karena konsisten unggul di MAE, MAPE, dan *Pearson Correlation Coefficient*, serta tidak menunjukkan indikasi *overfitting* seperti *Random Forest Regression,* meskipun keunggulan tersebut bersifat komparatif dan bukan indikasi bahwa performa modelnya sudah memadai secara absolut. Nilai metrik evaluasi ketiga algoritma yang berdekatan menunjukkan bahwa posisi *finish* memang sulit dijelaskan secara memadai, mengingat besarnya faktor dinamis seperti insiden kecelakaan, kegagalan mekanis, dan keputusan strategi tim secara *real-time*.

Hasil interpretasi PFI dan SHAP pada SVR menunjukkan bahwa *position qualifying* adalah fitur paling dominan (*Importance* = 1,225), memperkuat temuan Weissbock dan Mills (2025) bahwa posisi kualifikasi memiliki hubungan lebih kuat dengan posisi *finish* (C = 0,741; ρ = 0,763) dibandingkan posisi *starting grid* (C = 0,734; ρ = 0,748) [2]. Perlu dicatat penelitian tersebut berfokus pada kuantifikasi asosiasi, bukan model prediktif, sehingga penelitian ini mengisi kesenjangan itu dengan model regresi yang memprediksi posisi *finish* secara langsung. Selain itu, fitur *driver rank previous, constructor rank previous,* dan *grid* juga turut memberikan kontribusi terhadap prediksi, meskipun tingkat pengaruhnya lebih kecil dibandingkan *position qualifying,* sebuah aspek yang belum dieksplorasi pada penelitian terdahulu.

Pada tahap *forecasting* Miami *Grand Prix* 2026, model SVR berhasil memprediksi pemenang balapan dengan tepat. Namun, terdapat selisih yang cukup besar pada beberapa prediksi, terutama pada pembalap yang tidak berhasil menyelesaikan balapan (*Did Not Finish*). Keterbatasan ini sejalan dengan yang diungkap oleh Sicoie (2022), yang melaporkan tingkat kesalahan prediksi sekitar 2-3 posisi dari hasil aktual akibat sulitnya memodelkan insiden balapan yang bersifat tidak terduga [6]. Hal ini menegaskan bahwa model yang dibangun berdasarkan data performa historis pembalap dan konstruktor, sebagaimana juga digunakan dalam penelitian ini, memiliki keterbatasan mendasar dalam mengakomodasi faktor-faktor dinamis yang muncul selama balapan berlangsung dan berada di luar pola yang tersedia dalam data historis, sehingga tingkat akurasi model pada kasus-kasus DNF masih perlu ditingkatkan pada penelitian selanjutnya.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa terdapat perbedaan kinerja antara algoritma *Ridge Regression, Support Vector Regression* (SVR), dan *Random Forest Regression* dalam memprediksi posisi *finish* pembalap Formula 1. SVR menunjukkan kinerja terbaik berdasarkan sebagian besar metrik evaluasi yang digunakan, meskipun perbedaan kinerja antara ketiga algoritma relatif kecil. Namun, secara absolut, nilai MAPE sebesar 37,2987% tergolong tinggi, sehingga performa model masih belum sepenuhnya dapat diandalkan untuk penerapan praktis dalam konteks balapan Formula 1. Hasil interpretasi model menggunakan metode *Permutation Feature Importance* (PFI) dan SHAP menunjukkan bahwa *position qualifying* merupakan faktor yang paling berpengaruh terhadap prediksi posisi *finish* pembalap, diikuti oleh *driver rank previous, constructor rank previous*, dan *grid*. Pada tahap *forecasting* Miami *Grand Prix* 2026, model SVR berhasil memprediksi pemenang balapan dengan tepat, meskipun masih terdapat selisih prediksi yang cukup besar pada beberapa pembalap, terutama yang mengalami *Did Not Finish* (DNF). Penelitian ini berkontribusi dalam menyediakan kerangka komparatif antara pendekatan regresi linier dan non-linier untuk memprediksi posisi *finish* pembalap Formula 1, sebuah aspek yang belum banyak dieksplorasi karena penelitian terdahulu lebih banyak berfokus pada pendekatan klasifikasi atau prediksi *lap time*. Selain itu, penelitian ini memperkaya interpretabilitas model melalui PFI dan SHAP, serta memberikan bukti empiris atas performa model pada kondisi balapan nyata melalui *forecasting* Miami *Grand Prix* 2026. Meskipun demikian, penelitian ini tetap memiliki keterbatasan karena belum mempertimbangkan berbagai faktor teknis, strategis, dan eksternal yang dapat memengaruhi hasil balapan, seperti kerusakan mobil, strategi *pit stop*, pemilihan jenis ban, performa baterai mobil, kondisi cuaca, insiden kecelakaan, dan hasil *sprint race*. Oleh karena itu, penelitian selanjutnya dapat mempertimbangkan penambahan faktor-faktor tersebut, melakukan komparasi dengan algoritma *machine learning* lainnya, serta memperluas *forecasting* ke seluruh seri balapan dalam satu musim dengan mempertimbangkan perubahan susunan pembalap dan tim konstruktor pada musim yang diprediksi, guna meningkatkan performa prediksi.

REFERENCES

[1] A. Jafri, "Predicting Formula 1 Race Outcomes : A Machine Learning Approach," New York University Abu Dhabi, Abu Dhabi, 2024. [Online]. Available: https://aliabdullahjafri.com/static/media/Ali_Jafri_CapstoneProject1_Fall2024.c7244022875d46bec5d9.pdf

[2] J. Weissbock and S. Mills, "Re-evaluating the Qualifying / Finish Relationship in Formula One : A Replication and Correction of Prior Findings," Ottawa, Ontario, Canada, 2025. doi: 10.48550/arXiv.2507.10966.

[3] S. A. Krzysztoń and J. Smołka, "Application of machine learning for predicting Formula 1 race results," *J. Comput. Sci. Inst.*, vol. 38, pp. 81–86, 2026, doi: 10.35784/jcsi.8462.

[4] P. Shelke, R. Mirajkar, A. Pande, and K. Yash, "F1 Race Winner Predictor," in *2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, Maharashtra, India: IEEE, 2023, pp. 1–4. doi: 10.1109/ICCUBEA58933.2023.10392224.

[5] A. R. Patro, R. Sahana, S. Gautam, and R. Sapna, "A Predictive Analysis of lap time in Formula 1 using Machine Learning," in *2025 IEEE 6th India Council International Subsections Conference (INDISCON)*, Bengaluru, India: IEEE, 2025, pp. 1–7. doi: 10.1109/INDISCON66021.2025.11252332.

[6] H. Sicoie, "Machine Learning framework for Formula 1 race winner and championship standings predictor," Tilburg University, 2022. [Online]. Available: https://arno.uvt.nl/show.cgi?fid=157635

[7] D. Chopra and R. Khurana, *Introduction to Machine Learning with Python*. New Delhi: Bentham Science Publishers Pte. Ltd. Singapore, 2023. doi: 10.2174/9789815124422123010l.

[8] A. Muzakir, K. Adi, and R. Kusumaningrum, *Penerapan Konsep Machine Learning & Deep Learning*, 1st ed. Semarang: UNDIP Press, 2024. [Online]. Available: https://adsii.or.id/wp-content/uploads/2024/10/Penerapan-Konsep-Machine-Learning-Deep-Learning-Ari-Muzakir-et-al.pdf

[9] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 5, no. 1, pp. 97–100, 2023, doi: 10.26740/jinacs.v5n01.p97-100.

[10] C. Haryanto, N. Rahaningsih, and F. M. Basysyar, "Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 533–539, 2023, doi: 10.36040/jati.v7i1.6343.

[11] J. P. Bharadiya, "The Role of Machine Learning in Transforming Business Intelligence," *Int. J. Comput. Artif. Intell.*, vol. 4, no. 1, pp. 16–24, 2023, doi: 10.33545/27076571.2023.v4.i1a.60.

[12] F. Putra, H. F. Tahiyat, R. M. Ihsan, Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia Febrianda," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[13] M. A. Latief and Y. Karyanti, "Data Mining & Analytic Forecasting Indeks Standar Pencemar Udara Jakarta Menggunakan Metode Linear Regression (Studi Kasus: Dataset Indeks Standar Pencemar Udara Jakarta 2021)," *JOSR J. Soc. Res.*, vol. 1, no. 10, pp. 1164–1174, 2022, doi: 10.55324/josr.v1i10.248.

[14] M. P. Kusuma and A. Kudus, "Penerapan Metode Support Vector Regression (SVR) pada Data Survival KPR PT. Bank ABC, Tbk.," *Bandung Conf. Ser. Stat.*, vol. 2, no. 2, pp. 167–172, 2022, doi: 10.29313/bcss.v2i2.3614.

[15] D. Manurung, B. Zealtiel, and A. H. Lubis, "Prediksi Produksi Tanaman Padi di Indonesia dengan Menggunakan Algoritma Random Forest Regressor," *J. Comput. Informatics Res.*, vol. 4, no. 3, pp. 337–345, 2025, doi: 10.47065/comforch.v4i3.2125.

[16] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, pp. 1–23, 2023, doi: 10.3390/biomedicines11020581.

[17] S. A. N. Cahyo and M. Y. T. Sulistyono, "Comparison of Multiple Linear Regression and Random Forest Methods for Predicting National Rice Production in Indonesia," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3479–3489, 2025, doi:

10.30871/jaic.v9i6.11398.

[18] T. A. D. Kurniawan, A. Setiawan, and F. Tita, "Perbandingan Kinerja Metode Support Vector Regression dan Metode Regresi Linier Berganda dalam Memprediksi BMI pada Dataset ASTHMA," *J. Sains dan Edukasi Sains*, vol. 8, no. 2, pp. 133–142, 2025, doi: 10.24246/juses.v8i2p133-142.

[19] K. Mei, M. Tan, Z. Yang, and S. Shi, "Modeling of Feature Selection Based on Random Forest Algorithm and Pearson Correlation Coefficient," *J. Phys. Conf. Ser.*, vol. 2219, no. 1, pp. 1–9, 2022, doi: 10.1088/1742-6596/2219/1/012046.

[20] A. Khan, A. Ali, J. Khan, F. Ullah, and M. Faheem, "Using Permutation-Based Feature Importance for Improved Machine Learning Model Performance at Reduced Costs," *IEEE Access*, vol. 13, pp. 36421–36435, 2025, doi: 10.1109/ACCESS.2025.3544625.

[21] Y. Huang *et al.*, "Multiparametric MRI model to predict molecular subtypes of breast cancer using Shapley additive explanations interpretability analysis," *Diagn. Interv. Imaging*, vol. 105, no. 5, pp. 191–205, 2024, doi: 10.1016/j.diii.2024.01.004.