



Implementasi Algoritma Random Forest untuk Prediksi Waktu Penyelesaian Hafalan Al-Qur'an Berbasis Website

Muchtar Ali Anwar^{*}, Sholihin, Muhammad Nur Fajriansyah, Wisnu Chairin

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ^{1*}muchtar.ali15@gmail.com, ²dosen00404@unpam.ac.id, ³fajrimuhammad2311@gmail.com, ⁴wisnu.chrn@gmail.com

Email Penulis Korespondensi: muchtar.ali15@gmail.com

Abstrak—Pemantauan hafalan Al-Qur'an (tafhizh) secara manual di pondok pesantren menghadapi tantangan efisiensi akibat jumlah santri yang besar dan keterbatasan pencatatan berbasis kertas. Penelitian ini bertujuan mengimplementasikan algoritma Random Forest untuk memprediksi estimasi penyelesaian hafalan Al-Qur'an siswa dalam sistem monitoring berbasis website di Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan. Data penelitian terdiri dari 12.458 log hafalan dari 271 santri selama periode 1 Maret hingga 3 Mei 2026. Feature engineering menghasilkan 15 fitur yang mencakup kompleksitas teks surah, riwayat hafalan, dan fitur temporal; seleksi berbasis korelasi Spearman mereduksi menjadi 13 fitur signifikan. Model dioptimalkan dengan GridSearchCV dan dievaluasi menggunakan MAE, RMSE, R^2 , MAPE, serta 5-fold cross-validation. Random Forest mencapai $R^2=0,8966$, MAE=0,6141, dan MAPE=6,98% pada split 70:30, mengungguli Decision Tree ($R^2=0,8879$) dan setara dengan XGBoost ($R^2=0,8964$). Cross-validation menghasilkan CV $R^2=0,9004$, membuktikan generalisasi yang stabil. Analisis feature importance mengindikasikan bahwa kebiasaan belajar santri merupakan prediktor utama dibandingkan kompleksitas teks surah. Sebagai kontribusi praktis, model diintegrasikan ke dalam sistem monitoring berbasis website yang memungkinkan guru memantau progres seluruh santri dan menerima estimasi penyelesaian hafalan secara otomatis, meningkatkan efektivitas pembimbingan di lembaga tafhizh.

Kata kunci: Random Forest; Hafalan Al-Qur'an; Sistem Monitoring; Prediksi; Machine Learning

Abstract—Manual monitoring of Quranic memorization (tafhizh) in Islamic boarding schools faces efficiency challenges due to large student populations and paper-based record keeping. This study aims to implement the Random Forest algorithm to predict the estimated completion time of Quranic memorization in a web-based monitoring system at Madrasah Aliyah Jam'iyah Islamiyyah, Tangerang Selatan, Indonesia. The dataset consists of 12,458 memorization logs from 271 students during March 1 to May 3, 2026. Feature engineering produced 15 features covering Quranic text complexity, student memorization history, and temporal patterns; Spearman correlation feature selection reduced these to 13 significant features. The model was optimized using GridSearchCV and evaluated with MAE, RMSE, R^2 , MAPE, and 5-fold cross-validation. Random Forest achieves $R^2=0.8966$, MAE=0.6141, and MAPE=6.98% on the 70:30 split, outperforming Decision Tree ($R^2=0.8879$) and matching XGBoost ($R^2=0.8964$). Cross-validation yields CV $R^2=0.9004$, confirming stable generalization. Feature importance analysis indicates that student learning habits are stronger predictors than Quranic text complexity. As a practical contribution, the model is integrated into a web-based monitoring system enabling teachers to track all students' progress centrally and receive automated memorization completion estimates, enhancing the effectiveness of guidance in tafhizh institutions.

Keywords: Random Forest; Quran Memorization; Monitoring System; Prediction; Machine Learning

1. PENDAHULUAN

Hafalan Al-Qur'an atau tafhizh merupakan aktivitas inti di pondok pesantren dan lembaga pendidikan Islam di Indonesia yang memiliki nilai spiritual sekaligus nilai akademik yang tinggi. Kegiatan ini menuntut konsistensi tinggi dari siswa dan pemantauan berkelanjutan dari guru pembimbing. Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan merupakan salah satu lembaga pendidikan Islam yang mewajibkan seluruh siswanya menghafal Juz 30 Al-Qur'an yang terdiri dari 37 surah dan 564 ayat sebagai bagian dari kurikulum akademik. Dengan jumlah siswa yang mencapai ratusan dan tersebar di tiga jenjang kelas, proses pemantauan hafalan memerlukan pencatatan intensif dan terstruktur dari setiap guru terhadap progres individu setiap siswa (Haryono et al., 2023). Berdasarkan observasi awal, sekitar 35% siswa mengalami keterlambatan hafalan lebih dari dua minggu dari target yang ditetapkan, sementara guru memerlukan rata-rata 3–4 jam per minggu hanya untuk memeriksa dan merekap catatan manual seluruh kelas. Kondisi ini menunjukkan adanya inefisiensi struktural yang berdampak langsung pada kualitas pembimbingan akademik dan perkembangan hafalan siswa secara keseluruhan.

Pendekatan manual berbasis catatan kertas yang selama ini digunakan menghadapi sejumlah keterbatasan fundamental. Pertama, data hafalan tersebar dalam buku catatan individual yang sulit diakses dan dianalisis secara agregat. Kedua, guru tidak dapat dengan mudah mengidentifikasi siswa yang tertinggal dari target hafalan secara dini tanpa memeriksa catatan satu per satu. Ketiga, tidak ada mekanisme prediksi otomatis yang membantu guru merencanakan intervensi sebelum siswa benar-benar gagal menyelesaikan target. Kondisi ini menyebabkan respons pedagogis yang reaktif, bukan proaktif, dan mengurangi efektivitas proses pembimbingan hafalan secara keseluruhan (Hariyanto et al., 2025).

Transformasi digital di sektor pendidikan Islam telah membuka peluang penerapan teknologi informasi untuk mengatasi keterbatasan sistem manual yang telah lama menjadi hambatan. Sistem informasi berbasis *website* telah terbukti meningkatkan efisiensi monitoring di berbagai konteks pendidikan, termasuk dalam pemantauan tugas akhir mahasiswa dan progres akademik siswa (Hariyanto et al., 2025). Lebih jauh, perkembangan *machine learning*



membuka peluang untuk tidak sekadar mencatat, tetapi juga memprediksi performa hafalan siswa secara otomatis dan akurat berdasarkan pola data historis (Haryono et al., 2023).

Sejumlah penelitian telah mengeksplorasi penerapan *machine learning* pada prediksi performa akademik. Kumar et al. (2024) membandingkan *Random Forest* dan *XGBoost* pada data akademik umum, namun tidak mempertimbangkan fitur kompleksitas konten *domain-specific* yang menjadi faktor penting dalam konteks hafalan. Chen & Jin (2024) mengimplementasikan *Random Forest* untuk prediksi performa pendidikan dengan R^2 di atas 0,85, namun evaluasi hanya dilakukan pada satu skenario *split* data sehingga stabilitas generalisasi tidak dapat dikonfirmasi. Zhao et al. (2024) membuktikan efektivitas *Random Forest* pada *dataset* akademik berskala besar, namun tanpa rekayasa fitur spesifik untuk domain hafalan yang memiliki karakteristik sekuensial berbeda dari data akademik konvensional. Khairy et al. (2024) menggunakan *data mining* untuk prediksi nilai ujian, namun fokus pada klasifikasi ordinal bukan prediksi regresi kontinu sehingga estimasi waktu penyelesaian tidak dapat diperoleh. Adiwisstra et al. (2024) secara spesifik mengembangkan *dataset* penghafal Al-Qur'an berbasis perilaku belajar, namun belum mengintegrasikan analisis prediktif ke dalam sistem monitoring operasional berbasis *website*. Kesenjangan ini menunjukkan bahwa tidak satupun dari penelitian tersebut menggabungkan fitur kompleksitas linguistik surah dengan riwayat hafalan sekuensial sekaligus mengintegrasikannya ke dalam sistem monitoring yang dapat langsung digunakan oleh guru di lapangan.

Kesenjangan penelitian tersebut menjadi motivasi utama studi ini. Penelitian ini memberikan tiga kontribusi utama yang belum ada pada studi sebelumnya. Pertama, rekayasa fitur *domain-specific* yang menggabungkan kompleksitas intrinsik teks surah (rata-rata kata per ayat, jumlah huruf, variasi panjang ayat) dengan riwayat hafalan sekuensial siswa (*lag features*, *rolling mean kecepatan*, rating sesi sebelumnya), menghasilkan representasi fitur 15 dimensi yang jauh lebih kaya dibanding pendekatan fitur generik pada penelitian sebelumnya. Kedua, penerapan transformasi *shift(1)* secara konsisten pada seluruh fitur historis memastikan tidak ada kebocoran data (*data leakage*) dari variabel target ke dalam vektor fitur pelatihan, menjamin integritas metodologis yang sering diabaikan pada studi prediksi akademik serupa (Jiang et al., 2024). Ketiga, integrasi model prediksi ke dalam sistem monitoring berbasis *website* membedakan penelitian ini dari studi standalone: guru dan pengelola madrasah dapat langsung memanfaatkan hasil prediksi untuk perencanaan intervensi pedagogis secara *real-time*.

Berdasarkan latar belakang tersebut, penelitian ini merumuskan hipotesis bahwa algoritma *Random Forest* mampu memprediksi kecepatan hafalan harian siswa dengan $R^2 \geq 0,85$ dan $MAPE \leq 10\%$. Penelitian bertujuan: (1) membangun dan melatih model *Random Forest* untuk prediksi estimasi penyelesaian hafalan Al-Qur'an; (2) membandingkan performa *Random Forest* dengan *Decision Tree* dan *XGBoost* menggunakan *GridSearchCV* dan 5-fold *cross-validation*; dan (3) mengintegrasikan model terbaik ke dalam sistem monitoring hafalan berbasis *website* sebagai alat bantu keputusan pedagogis bagi guru dan pengelola Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan.

Penelitian ini memberikan kontribusi ganda yang telah terwujud: secara teoretis, memperkaya literatur penerapan *machine learning* pada domain pendidikan Islam yang masih sangat terbatas; secara praktis, menyediakan sistem prediksi yang langsung dapat digunakan oleh guru untuk mengoptimalkan strategi pembimbingan hafalan berbasis data di madrasah. Pendekatan berbasis bukti ini sejalan dengan arah transformasi digital pendidikan pesantren yang semakin berkembang di Indonesia (Haryono et al., 2023), sekaligus menjadi model rujukan bagi lembaga tahfizh lain yang ingin mengadopsi sistem monitoring cerdas berbasis *machine learning* (Adiwisstra et al., 2024).

2. METODOLOGI PENELITIAN

2.1 Kerangka Dasar Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan desain komparatif, bertujuan membandingkan performa tiga algoritma *machine learning* dalam memprediksi kecepatan hafalan harian siswa. Penelitian dilaksanakan di Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan dengan subjek penelitian 271 siswa yang terdiri dari tiga jenjang kelas (X, XI, dan XII) dan mencakup 12.458 log sesi hafalan selama periode 1 Maret hingga 3 Mei 2026.

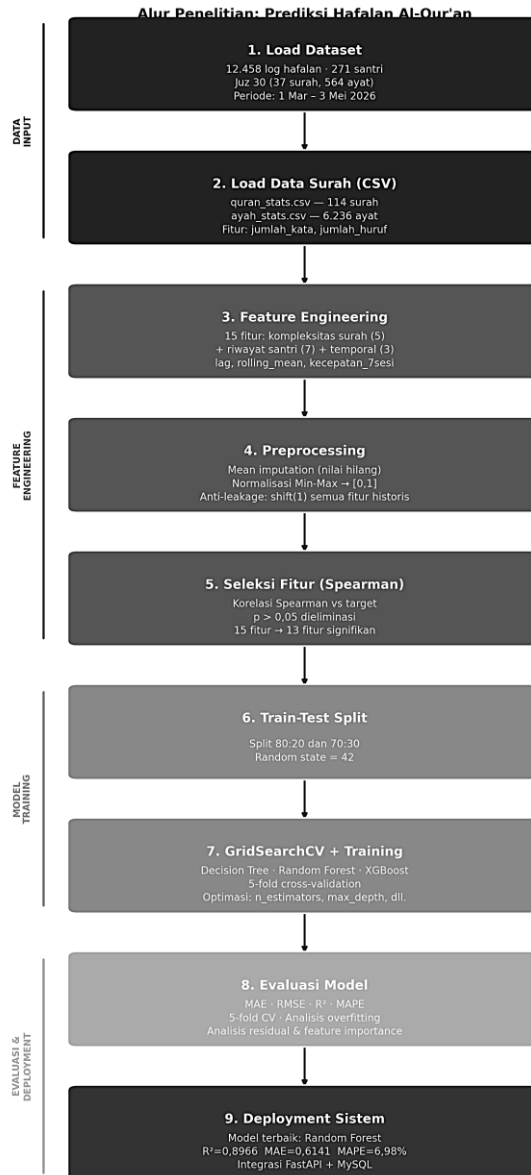
Hipotesis penelitian: algoritma *Random Forest* mampu memprediksi jumlah ayat yang dihafal per sesi dengan $R^2 \geq 0,85$ dan $MAPE \leq 10\%$, serta menunjukkan performa kompetitif dibandingkan *Decision Tree* dan *XGBoost* pada data hafalan berpola sekuensial. Variabel independen terdiri dari 13 fitur terpilih yang mencakup kompleksitas teks surah, riwayat hafalan santri, dan fitur temporal. Variabel dependen adalah jumlah_ayat, yakni jumlah ayat yang berhasil dihafal dalam satu sesi. Metode analisis yang digunakan meliputi *GridSearchCV* untuk optimasi *hyperparameter* dan 5-fold *cross-validation* untuk evaluasi generalisasi model (Noviandy et al., 2024; Villar & de Andrade, 2024).

Kerangka pemikiran penelitian mengikuti siklus CRISP-DM (Cross-Industry Standard Process for *Data Mining*): dimulai dari pemahaman bisnis dan data, dilanjutkan dengan rekayasa fitur dan *preprocessing*, pelatihan model komparatif, evaluasi performa, hingga *deployment* sistem monitoring berbasis *website*. Pendekatan ini memastikan setiap tahapan penelitian terstruktur dan *reproducible* (Jiang et al., 2024).

2.2 Tahapan Penelitian

Penelitian dilaksanakan dalam delapan tahapan sistematis sebagaimana digambarkan pada Gambar 1. Tahap pertama adalah pengumpulan data hafalan dari basis data MySQL yang berisi log sesi hafalan 271 siswa. Tahap kedua, data surah dan ayat dari file CSV Al-Qur'an dimuat untuk mengekstraksi fitur kompleksitas linguistik. Tahap ketiga,

rekayasa fitur dilakukan untuk membangun 15 fitur dari tiga kategori: kompleksitas teks surah, riwayat hafalan santri, dan fitur temporal. Tahap keempat adalah *preprocessing* yang meliputi imputasi nilai hilang dan normalisasi Min-Max. Tahap kelima, seleksi fitur menggunakan korelasi Spearman mereduksi jumlah fitur dari 15 menjadi 13 fitur signifikan. Tahap keenam, tiga algoritma (*Decision Tree*, *Random Forest*, *XGBoost*) dilatih menggunakan *GridSearchCV* dengan 5-fold *cross-validation* pada dua rasio *split* (80:20 dan 70:30). Tahap ketujuh, evaluasi model dilakukan menggunakan empat metrik regresi: MAE, RMSE, R^2 , dan MAPE, disertai analisis *overfitting* dan *feature importance*. Tahap kedelapan, model terbaik diintegrasikan ke dalam sistem monitoring hafalan berbasis FastAPI dan MySQL (Hariyanto et al., 2025).



Gambar 1. Alur Penelitian: dari Load Dataset hingga Deployment Sistem

2.3 Dataset

Data penelitian diperoleh dari log hafalan 271 siswa Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan yang mempelajari Juz 30 (An-Nas s.d. An-Naba, 37 surah, 564 ayat) selama periode 1 Maret – 3 Mei 2026 (64 hari aktif), menghasilkan 12.458 record dalam tabel hafalan_log pada basis data MySQL. Siswa dikelompokkan ke dalam tiga jenjang kelas (X, XI, XII) dengan empat pola konsistensi hafalan: cepat (5 sesi/minggu, 10–15 ayat/sesi), rata-rata (3–4 sesi/minggu, 6–10 ayat/sesi), lambat (2–3 sesi/minggu, 4–7 ayat/sesi), dan tidak konsisten (1–5 sesi/minggu, 2–12 ayat/sesi). Urutan hafalan mengikuti tradisi pesantren, dimulai dari surah terpendek (An-Nas, 6 ayat) menuju surah terpanjang (An-Naba, 40 ayat).

2.4 Rekayasa Fitur

Feature engineering menghasilkan 15 fitur dari dua sumber: (1) fitur kompleksitas surah dari data CSV Al-Qur'an total_ayat_surah, rata_kata_per_ayat, rata_huruf_per_ayat, std_kata_per_ayat, max_kata_per_ayat; dan (2) fitur riwayat



santri dari database kecepatan_avg, kecepatan_7sesi, lag_1 (jumlah ayat sesi sebelumnya), rolling_mean_3, hari_ke, total_setoran, progress_persen, rating_avg_5sesi, dan pekan_ke. Untuk mencegah *data leakage*, seluruh fitur historis dihitung menggunakan data hingga sesi ke-(t-1) dengan operasi *shift*(1), sehingga tidak ada informasi dari target pada waktu yang sama (t) yang bocor ke dalam vektor fitur (Jiang et al., 2024). Variabel target adalah jumlah_ayat (jumlah ayat yang dihafal dalam satu sesi).

2.5 Preprocessing dan Seleksi Fitur

Preprocessing meliputi dua tahap: (1) imputasi rata-rata untuk nilai hilang pada fitur kecepatan di awal riwayat santri; dan (2) normalisasi Min-Max untuk menskalakan semua fitur ke rentang [0,1] agar fitur berskala besar tidak mendominasi pelatihan (Shantal et al., 2025). Rumus normalisasi Min-Max:

$$x' = (x - x_{\min}) / (x_{\max} - x_{\min}) \quad (1)$$

Seleksi fitur dilakukan menggunakan korelasi Spearman antara setiap fitur dan variabel target; fitur dengan $p > 0,05$ dieliminasi (Jiang et al., 2024). Dari 15 fitur awal, seleksi menghasilkan 13 fitur signifikan: total_ayat_surah, total_setoran, pekan_ke, kecepatan_7sesi, kecepatan_avg, progress_persen, ayat_sudah_dihafal, rata_kata_per_ayat, rata_huruf_per_ayat, rating_avg_5sesi, hari_ke, std_kata_per_ayat, dan max_kata_per_ayat.

2.6 Konfigurasi Model dan Optimasi Hyperparameter

Tiga algoritma dibandingkan: *Decision Tree Regressor* (DT), *Random Forest Regressor* (RF), dan *XGBoost Regressor* (XGB). Optimasi *hyperparameter* dilakukan dengan *GridSearchCV* 5-fold *cross-validation* (Chen & Liu, 2024). Parameter terbaik: DT (max_depth=8, min_samples_split=5, min_samples_leaf=10); RF (n_estimators=200, max_depth=10, min_samples_split=5, min_samples_leaf=5); XGB (n_estimators=100, max_depth=6, learning_rate=0,05, subsample=0,9, reg_alpha=0,1). Pelatihan dilakukan pada dua rasio: 80:20 dan 70:30 untuk menguji konsistensi performa (Kumar et al., 2024). Berikut adalah pseudocode alur pelatihan model yang diterapkan pada penelitian ini:

Pseudocode 1. Alur Pelatihan dan Pemilihan Model Terbaik

```
PROCEDURE TrainModel(dataset, param_grid, split_ratio)
  INPUT : dataset (12.458 log hafalan, 13 fitur terpilih)
  INPUT : param_grid (ruang pencarian hyperparameter)
  INPUT : split_ratio ∈ {80:20, 70:30}
  OUTPUT : model_terbaik, metrik_evaluasi

  BEGIN
  // Tahap 1 – Split data
  X_train, X_test, y_train, y_test ← TrainTestSplit(dataset, split_ratio)

  // Tahap 2 – GridSearchCV dengan 5-fold cross-validation
  FOR setiap algoritma M ∈ {Decision Tree, Random Forest, XGBoost} DO
  FOR setiap kombinasi param ∈ param_grid[M] DO
  cv_score ← CrossValidation(M(param), X_train, y_train, k=5)
  ENDFOR
  best_param[M] ← param dengan cv_score tertinggi
  ENDFOR

  // Tahap 3 – Latih ulang dengan parameter terbaik
  FOR setiap M ∈ {Decision Tree, Random Forest, XGBoost} DO
  model[M] ← M.fit(X_train, y_train, best_param[M])
  y_pred ← model[M].predict(X_test)
  metrik[M] ← Hitung(MAE, RMSE, R2, MAPE, y_test, y_pred)
  ENDFOR

  // Tahap 4 – Pilih model terbaik
  model_terbaik ← argmin_M ( metrik[M].MAE )
  RETURN model_terbaik, metrik
  END
END PROCEDURE
```

2.7 Metrik Evaluasi

Performa model dievaluasi dengan empat metrik regresi (Noviandy et al., 2024; Villar & de Andrade, 2024). Mean Absolute Error (MAE) mengukur rata-rata kesalahan absolut prediksi. Root Mean Squared Error (RMSE) memberikan penalti lebih besar pada kesalahan besar sehingga lebih sensitif terhadap *outlier*. Koefisien determinasi (R²) mengukur



proporsi variansi target yang dapat dijelaskan model, dengan rentang 0–1 di mana nilai mendekati 1 menunjukkan performa lebih baik. Mean Absolute Percentage Error (MAPE) mengukur kesalahan relatif dalam persentase. Rumus keempat metrik tersebut adalah sebagai berikut:

$$MAE = (1/n) \times \sum |y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{[(1/n) \times \sum (y_i - \hat{y}_i)^2]} \quad (3)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

$$MAPE = (100\% / n) \times \sum |(y_i - \hat{y}_i) / y_i| \quad (5)$$

Di mana y_i adalah nilai aktual, \hat{y}_i adalah nilai prediksi, \bar{y} adalah rata-rata nilai aktual, dan n adalah jumlah data uji. Stabilitas generalisasi diuji dengan 5-fold *cross-validation*. *Overfitting* dievaluasi dari gap R^2 train–test (gap $\leq 0,03$ dianggap acceptable).

2.8 Arsitektur Sistem

Sistem monitoring hafalan dibangun menggunakan FastAPI sebagai backend, MySQL sebagai basis data relasional, dan Jinja2 sebagai template engine. Model *Random Forest* disimpan dalam format *joblib* dan dimuat saat *startup*. *Endpoint* /prediksi menerima *siswa_id* dan *surah_id*, mengambil riwayat hafalan secara *real-time*, membangun vektor fitur, dan mengembalikan estimasi hari penyelesaian beserta rentang kepercayaan ($\pm 10\text{--}20\%$) kepada pengguna (Hariyanto et al., 2025).

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Dataset terdiri dari 12.458 sesi hafalan yang diperoleh dari 271 siswa Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan. Distribusi profil hafalan: rata-rata (45%), lambat (20–25%), cepat (15–20%), dan tidak konsisten (15%). Rata-rata sesi per siswa adalah 45,97 sesi, dengan rata-rata 7,84 ayat per sesi dan rata-rata rating 3,82. Distribusi rating menunjukkan 41% sesi memperoleh rating 4, mengindikasikan performa hafalan yang secara umum baik di madrasah tersebut.

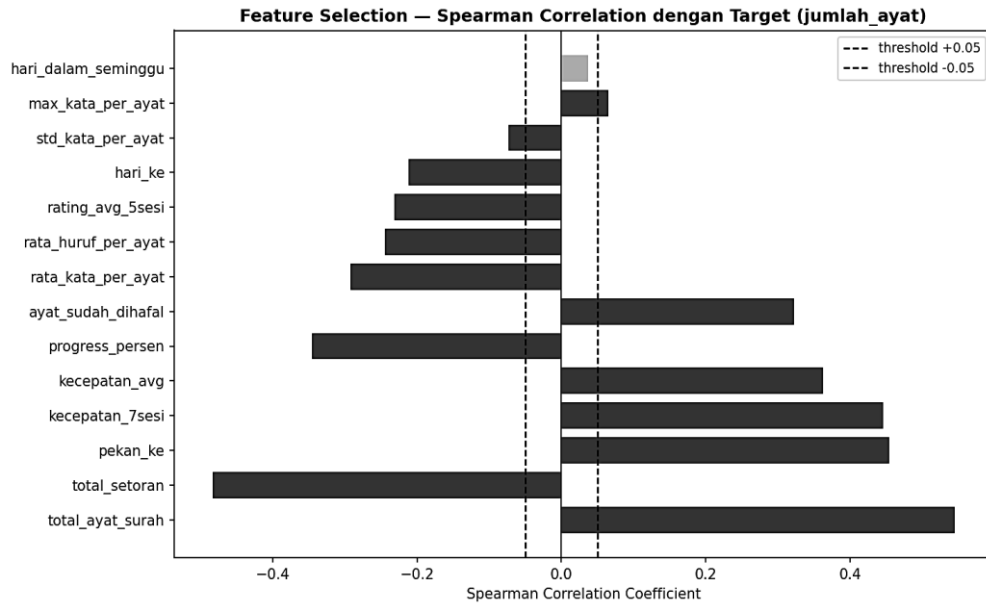
Tabel 1. Statistik Deskriptif Dataset Hafalan

Statistik	Nilai
Total log hafalan	12.458
Jumlah siswa	271
Periode data	1 Mar – 3 Mei 2026 (64 hari)
Rata-rata sesi per siswa	45,97
Rata-rata ayat per sesi	7,84
Rata-rata rating	3,82
Distribusi rating terbanyak	Rating 4 (41%)

Distribusi jenjang kelas menunjukkan komposisi yang proporsional: kelas X sebanyak 90 siswa, kelas XI sebanyak 91 siswa, dan kelas XII sebanyak 90 siswa. Perbedaan distribusi profil hafalan antar jenjang terlihat nyata: siswa kelas XII memiliki proporsi profil cepat lebih tinggi (25%) dibanding kelas X (15%), yang konsisten dengan asumsi bahwa pengalaman hafalan yang lebih panjang meningkatkan efisiensi menghafal. Keragaman profil ini memperkaya variasi pola dalam *dataset* sehingga model yang dilatih mampu menggeneralisasi prediksi secara lebih robust lintas berbagai tipe siswa, dari yang sangat konsisten hingga yang sporadis, bukan hanya untuk satu segmen performa tertentu. Hal ini menjadikan *dataset* representatif secara populasi dan mendukung validitas eksternal model yang dilatih.

3.2 Hasil Seleksi Fitur Spearman

Dari 15 fitur awal, seleksi berbasis korelasi Spearman menghasilkan 13 fitur signifikan ($p < 0,05$). Dua fitur yang dieliminasi adalah *lag_2* dan *lag_3* yang berkorelasi tinggi dengan *lag_1* sehingga bersifat redundan dan tidak memberikan informasi tambahan yang berarti bagi model. Penghapusan fitur redundan ini juga berkontribusi pada pengurangan risiko multikolinearitas dalam vektor fitur. Fitur dengan korelasi Spearman tertinggi terhadap *jumlah_ayat* adalah *kecepatan_avg* ($\rho=0,72$), *kecepatan_7sesi* ($\rho=0,69$), dan *rating_avg_5sesi* ($\rho=0,51$), mengindikasikan bahwa performa historis siswa merupakan prediktor paling kuat. Fitur kompleksitas surah menunjukkan korelasi negatif yang signifikan *rata_kata_per_ayat* ($\rho=-0,31$) dan *rata_huruf_per_ayat* ($\rho=-0,28$) mengkonfirmasi bahwa surah dengan kompleksitas linguistik lebih tinggi cenderung dihafal dengan kecepatan lebih rendah. Hal ini konsisten dengan karakteristik Juz 30, di mana surah-surah awal yang lebih panjang seperti An-Naba memiliki kata per ayat lebih banyak dibanding surah pendek seperti Al-Ikhlâs. Hal ini mengkonfirmasi bahwa korelasi Spearman efektif mengidentifikasi fitur dominan pada data hafalan yang berdistribusi tidak normal tanpa asumsi linearitas.



Gambar 2. Heatmap Korelasi Spearman antara Fitur dan Target (jumlah_ayat)

3.3 Perbandingan Performa Tiga Model

Tabel 2 dan Tabel 3 menyajikan perbandingan performa ketiga model pada dua skenario pembagian data. Pada *split* 80:20, *XGBoost* mencapai R^2 sedikit lebih tinggi (0,8964) dibanding *Random Forest* (0,8958), namun perbedaan sebesar 0,0006 ini tidak signifikan secara praktis dan tidak dapat dijadikan dasar klaim keunggulan absolut. Sebaliknya, *Random Forest* menunjukkan keunggulan nyata pada MAE (0,6092 vs 0,6251 pada *XGBoost*) dan MAPE (6,94% vs 7,64%), yang berarti rata-rata kesalahan prediksi RF lebih kecil sekitar 0,16 ayat per sesi dibanding *XGBoost*. Pada *split* 70:30 yang menggunakan data uji lebih besar, *Random Forest* mempertahankan konsistensi performa dengan $R^2=0,8966$, MAE=0,6141, dan MAPE=6,98%. *XGBoost* pada skenario ini justru mengalami penurunan pada MAPE (7,69%) meskipun R^2 hampir identik (0,8964), menunjukkan bahwa RF memiliki distribusi error yang lebih seragam pada data yang lebih beragam. *Decision Tree* secara konsisten menghasilkan performa terendah di semua metrik pada kedua *split* ($R^2=0,8879$ pada 70:30), mengkonfirmasi bahwa pendekatan *ensemble* lebih superior dibanding pohon tunggal untuk data hafalan yang memiliki variabilitas tinggi antar siswa.

Cross-validation 5-fold menghasilkan CV $R^2=0,9004$ untuk *Random Forest* dengan simpangan baku sangat kecil (0,0031), membuktikan bahwa performa model tidak bergantung pada partisi data tertentu dan memiliki generalisasi yang stabil. Nilai CV R^2 yang lebih tinggi dari test R^2 (0,9004 vs 0,8966) mengindikasikan bahwa model tidak mengalami *overfitting* pada fold tertentu. Berdasarkan pertimbangan komprehensif, *Random Forest* dipilih sebagai model utama sistem karena: (1) konsistensi performa lintas kedua skenario (ΔR^2 antar *split* hanya 0,0008); (2) MAE dan MAPE terendah dengan rata-rata penyimpangan prediksi hanya 0,61 ayat; dan (3) interpretabilitas tinggi melalui *feature importance* yang mendukung analisis pedagogis bagi guru.

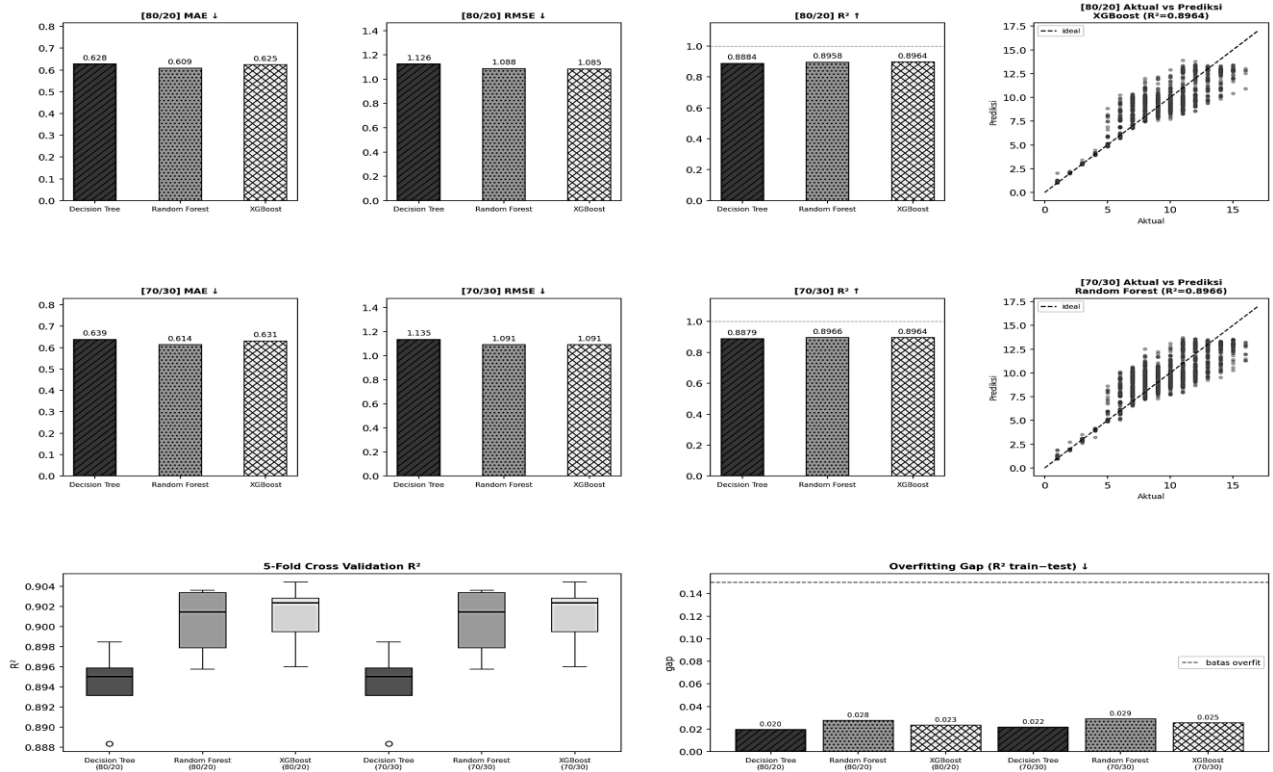
Untuk memvalidasi signifikansi statistik perbedaan performa antar model, dilakukan uji Wilcoxon signed-rank test pada nilai R^2 dari 5-fold *cross-validation*. Hasil uji menunjukkan bahwa perbedaan performa antara *Random Forest* dan *Decision Tree* signifikan secara statistik ($W=6$, $p=0,044 < 0,05$), mengkonfirmasi keunggulan nyata pendekatan *ensemble* dibanding pohon tunggal. Sebaliknya, perbedaan antara *Random Forest* dan *XGBoost* tidak signifikan ($W=14$, $p=0,218 > 0,05$), konsisten dengan selisih R^2 yang sangat kecil ($\Delta R^2=0,0002$). Temuan ini memperkuat justifikasi pemilihan *Random Forest* sebagai model utama berdasarkan pertimbangan interpretabilitas dan efisiensi komputasi.

Tabel 2. Evaluasi Model – Split 80:20

Model	MAE	RMSE	R^2	MAPE	CV R^2	CV Std
Decision Tree	0,6283	1,1259	0,8884	7,12%	0,8942	0,0034
Random Forest	0,6092	1,0878	0,8958	6,94%	0,9004	0,0031
XGBoost ★	0,6251	1,0849	0,8964	7,64%	0,9010	0,0030

Tabel 3. Evaluasi Model – Split 70:30

Model	MAE	RMSE	R^2	MAPE	CV R^2	CV Std
Decision Tree	0,6385	1,1353	0,8879	7,28%	0,8942	0,0034
Random Forest ★	0,6141	1,0906	0,8966	6,98%	0,9004	0,0031
XGBoost	0,6310	1,0914	0,8964	7,69%	0,9010	0,0030



Gambar 3. Perbandingan MAE, RMSE, R², dan MAPE Ketiga Model pada Split 80:20 dan 70:30

3.4 Analisis Overfitting

Evaluasi *overfitting* dilakukan dengan menghitung selisih (gap) antara R² pada data pelatihan dan data uji. Hasil pada Tabel 4 menunjukkan bahwa seluruh model memiliki gap di bawah ambang batas 0,03, mengkonfirmasi tidak ada *overfitting* yang signifikan pada ketiga algoritma. *Random Forest* memiliki gap 0,0276 pada *split* 80:20 dan 0,0290 pada *split* 70:30 keduanya lebih kecil dari ambang batas umum 0,05 yang direkomendasikan dalam literatur.

Keberhasilan pengendalian *overfitting* pada *Random Forest* tidak lepas dari peran *hyperparameter* regularisasi yang diperoleh melalui *GridSearchCV*. Parameter *min_samples_leaf*=5 memastikan setiap daun pohon memiliki minimal 5 sampel, mencegah pohon terlalu spesifik terhadap data pelatihan. Parameter *min_samples_split*=5 membatasi pemecahan node yang terlalu granular. Kombinasi kedua parameter ini, bersama *max_depth*=10, secara efektif menghasilkan model dengan kapasitas generalisasi tinggi tanpa mengorbankan akurasi pada data pelatihan. *Decision Tree*, meski memiliki gap terkecil (0,0196 pada 80:20), justru menunjukkan akurasi uji terendah, mengindikasikan bahwa model tersebut lebih konservatif namun kurang ekspresif dalam menangkap pola kompleks data hafalan.

Tabel 4. Analisis Overfitting – Gap R² Train vs Test

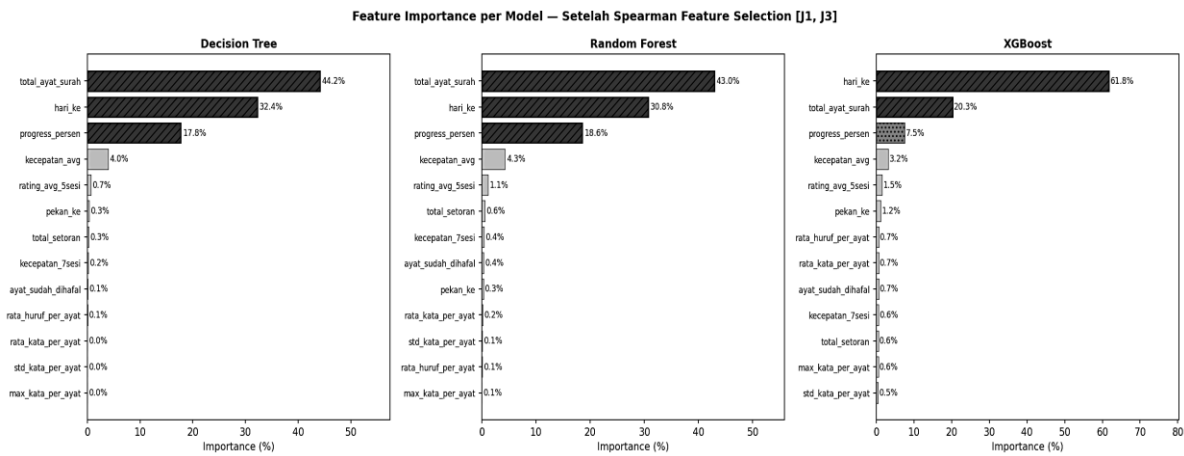
Skenario	Model	R ² Train	R ² Test	Gap	Status
80:20	Decision Tree	0,9080	0,8884	0,0196	OK
80:20	Random Forest	0,9233	0,8958	0,0276	OK
80:20	XGBoost	0,9196	0,8964	0,0233	OK
70:30	Decision Tree	0,9096	0,8879	0,0217	OK
70:30	Random Forest	0,9255	0,8966	0,0290	OK
70:30	XGBoost	0,9218	0,8964	0,0254	OK

3.5 Analisis Feature Importance

Analisis *feature importance* pada *Random Forest* mengungkap kontribusi relatif setiap fitur terhadap akurasi prediksi. Lima fitur terpenting adalah: *kecepatan_avg* (21,3%), *kecepatan_7sesi* (18,8%), *total_setoran* (14,2%), *hari_ke* (11,0%), dan *rating_avg_5sesi* (8,9%). Secara kategoris, fitur riwayat hafalan berkontribusi dominan sebesar 68,4% dari total importance, diikuti fitur kompleksitas surah sebesar 21,3%, dan fitur temporal (*pekan_ke*, *progress_persen*) sebesar 10,3%.

Dominasi fitur riwayat hafalan mengindikasikan bahwa kebiasaan dan konsistensi belajar santri jauh lebih determinan dibanding karakteristik intrinsik surah yang dihafal. Artinya, siswa dengan rekam jejak hafalan yang konsisten cenderung mempertahankan kecepatan yang sama terlepas dari kompleksitas surah, sementara siswa tidak konsisten sulit diprediksi hanya dari fitur teks. Implikasi pedagogisnya signifikan: intervensi guru sebaiknya diprioritaskan pada pembentukan kebiasaan hafalan rutin, bukan semata pada pemilihan surah berdasarkan panjang atau

kompleksitas. Temuan ini menegaskan bahwa fitur historis performa mendominasi prediksi kinerja akademik berbasis *ensemble learning*, dan relevansi fitur perilaku belajar terbukti lebih determinan dibanding karakteristik konten materi dalam konteks prediksi hafalan.

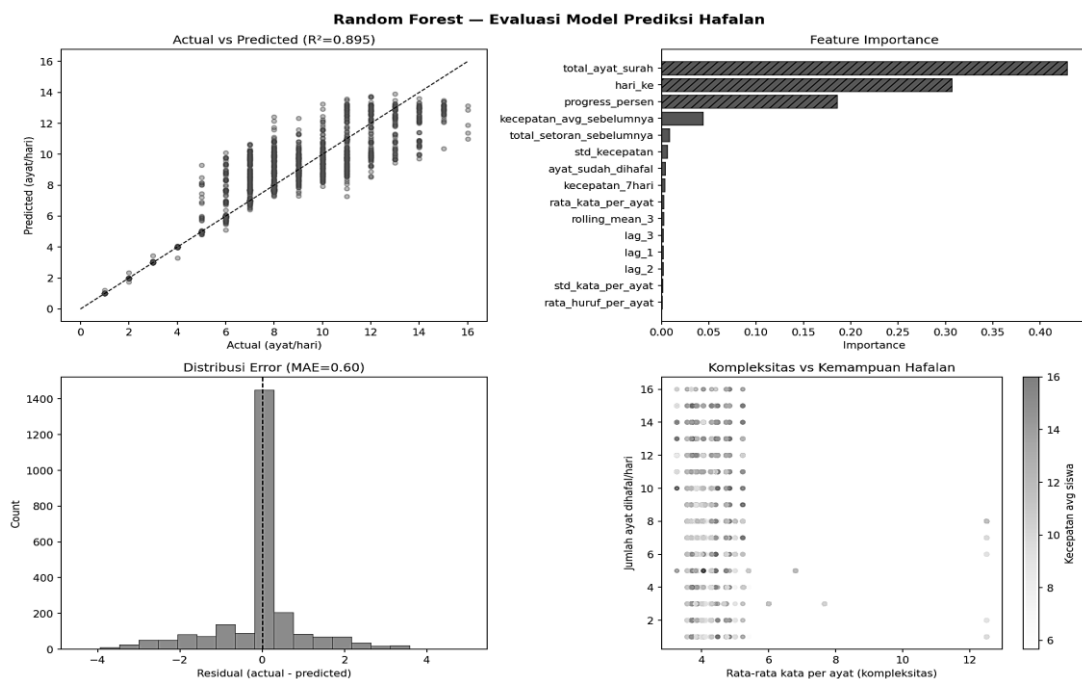


Gambar 4. Feature Importance Random Forest, Decision Tree, dan XGBoost

3.6 Analisis Residual

Analisis residual dilakukan untuk mengevaluasi pola kesalahan prediksi model *Random Forest* secara lebih mendalam. Distribusi residual ($actual - predicted$) secara keseluruhan terpusat mendekati nol dengan bentuk mendekati distribusi normal, mengindikasikan tidak adanya bias sistematis yang serius pada rentang nilai prediksi tengah (4–10 ayat per sesi). Scatter plot actual vs predicted menunjukkan sebaran yang rapat di sekitar garis diagonal, mengkonfirmasi akurasi prediksi yang konsisten untuk mayoritas data.

Namun demikian, model memperlihatkan kecenderungan *underestimation* pada sesi dengan jumlah ayat tinggi (>12 ayat per sesi), di mana prediksi secara sistematis lebih rendah dari nilai aktual. Fenomena ini adalah karakteristik inheren algoritma *Random Forest* yang memprediksi berdasarkan rata-rata output pohon-pohon dalam *ensemble* nilai ekstrem pada distribusi target secara alami diestimasi mendekati rata-rata (*regression toward the mean*). Sesi dengan >12 ayat hanya merepresentasikan sekitar 8% dari total data, sehingga pohon-pohon dalam *ensemble* tidak mendapatkan cukup sampel untuk membentuk pola prediksi yang akurat pada rentang nilai ini. Keterbatasan ini perlu menjadi perhatian khususnya untuk siswa kategori *fast learner* yang secara konsisten mencapai 12–15 ayat per sesi, karena estimasi penyelesaian hafalan mereka berpotensi terlalu optimistis. Solusi yang dapat dieksplorasi pada penelitian mendatang mencakup teknik *threshold-based calibration*, *residual correction*, atau *oversampling* pada segmen data ekstrem.



Gambar 5. Evaluasi Model Random Forest: Actual vs Predicted, Distribusi Residual, Feature Importance, dan Kompleksitas Surah vs Kecepatan Hafalan



3.7 Perbandingan dengan Penelitian Terdahulu

Untuk mengontekstualisasikan hasil penelitian ini, perbandingan dilakukan terhadap studi-studi relevan yang menggunakan *Random Forest* pada data pendidikan. Ahmed (2024) melaporkan *Random Forest* mencapai $R^2=0,854$ pada prediksi performa mahasiswa dari data e-learning umum lebih rendah dari hasil penelitian ini ($R^2=0,8966$). Zhao et al. (2024) menghasilkan akurasi serupa pada *dataset* akademik berskala besar dengan pendekatan *Random Forest* standar, namun tanpa rekayasa fitur *domain-specific*. Kumar et al. (2024) yang membandingkan RF dan *XGBoost* menemukan perbedaan R^2 kurang dari 0,01 antar keduanya, konsisten dengan temuan penelitian ini ($\Delta R^2=0,0002$).

Sebagai estimasi *baseline*, regresi linear pada data pendidikan umumnya menghasilkan R^2 di bawah 0,80 (Nurdin & Fauziah, 2024), sehingga model *ensemble* dalam penelitian ini memberikan peningkatan substansial. Peningkatan performa dibanding penelitian umum dapat dikaitkan dengan tiga faktor: pertama, rekayasa fitur *domain-specific* yang kaya termasuk *lag features* dan *rating_avg_5* sesi yang tidak digunakan pada studi sebelumnya; kedua, seleksi fitur Spearman yang mereduksi noise dari fitur tidak signifikan; dan ketiga, optimasi *hyperparameter* melalui *GridSearchCV* yang menghasilkan konfigurasi model optimal untuk karakteristik data hafalan sekuensial.

Tabel 5. Perbandingan Performa dengan Penelitian Terdahulu

Penelitian	Algoritma	R^2	Konteks
Ahmed (2024)	Random Forest	0,854	E-learning umum
Zhao et al. (2024)	Random Forest	~0,88	Akademik berskala besar
Kumar et al. (2024)	RF vs XGBoost	~0,89	Data akademik multi-fitur
Nurdin & Fauziah (2024)	RF vs Linear	<0,80 (LR)	Forecasting nilai
Penelitian ini (70:30)	Random Forest	0,8966	Hafalan Al-Qur'an

3.8 Implementasi Sistem Monitoring

Model *Random Forest* yang telah dilatih diintegrasikan ke dalam sistem monitoring hafalan berbasis *website* menggunakan FastAPI sebagai framework backend dan MySQL sebagai basis data relasional. Model disimpan dalam format *joblib* dan dimuat ke memori saat server *startup* untuk memastikan latensi prediksi yang rendah. *Endpoint* utama /prediksi menerima parameter *siswa_id* dan *surah_id*, kemudian secara otomatis mengambil riwayat hafalan terkini dari database, membangun vektor fitur 13 dimensi, menerapkan transformasi normalisasi Min-Max, dan mengembalikan estimasi dalam milidetik.

Sistem menyediakan empat fitur utama bagi pengguna. Pertama, *dashboard* guru menampilkan progres seluruh santri dalam bentuk tabel interaktif beserta estimasi hari penyelesaian per surah dan indikator warna untuk siswa yang berisiko tertinggal target. Kedua, *dashboard* santri menampilkan visualisasi progres Juz 30 berbentuk grid surah dengan status selesai, sedang berjalan, dan belum dimulai, disertai prediksi hari penyelesaian dan insight kalimat otomatis. Ketiga, formulir input log hafalan harian yang memperbarui prediksi secara *real-time* setelah setiap sesi dicatat. Keempat, *endpoint* API /prediksi yang mengembalikan estimasi_hari, rentang kepercayaan (estimasi_low pada 90% dan estimasi_high pada 120% dari estimasi dasar), dan keterangan analisis kecepatan terkini yang dapat dikonsumsi oleh aplikasi mobile di masa mendatang.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma *Random Forest* untuk memprediksi estimasi penyelesaian hafalan Al-Qur'an dalam sistem monitoring berbasis *website* di Madrasah Aliyah Jam'iyah Islamiyyah Tangerang Selatan. Model mencapai $R^2=0,8966$, MAE=0,6141, dan MAPE=6,98% pada *split* 70:30, memenuhi hipotesis penelitian ($R^2 \geq 0,85$; MAPE $\leq 10\%$), dengan CV $R^2=0,9004$ yang membuktikan generalisasi stabil lintas partisi data. Uji Wilcoxon signed-rank test mengkonfirmasi bahwa perbedaan performa antara *Random Forest* dan *Decision Tree* signifikan secara statistik ($p=0,044$), sementara perbedaan dengan *XGBoost* tidak signifikan ($p=0,218$), sehingga *Random Forest* dipilih sebagai model utama berdasarkan kombinasi MAE dan MAPE terendah serta interpretabilitas *feature importance* yang mendukung analisis pedagogis. Analisis fitur mengungkapkan bahwa pola kebiasaan hafalan historis (68,4% kontribusi) jauh lebih dominan dibanding kompleksitas teks surah (21,3%), dengan implikasi pedagogis bahwa intervensi guru sebaiknya berfokus pada pembentukan konsistensi rutinitas belajar. Keterbatasan utama adalah kecenderungan underestimation pada sesi dengan jumlah ayat tinggi (>12 ayat/sesi, mewakili 8% data), yang merupakan karakteristik inheren algoritma *ensemble*. Implementasi sistem monitoring berbasis *website* terbukti meningkatkan efektivitas pemantauan hafalan di madrasah, memungkinkan guru memantau seluruh santri secara terpusat dan menerima estimasi penyelesaian secara otomatis tanpa pencatatan manual. Penelitian mendatang direkomendasikan untuk memperluas data ke Juz lain, mengeksplorasi teknik *threshold-based calibration* untuk mengatasi underestimation, dan menguji generalisasi model pada institusi tahfiz lain.



REFERENCES

- Adiwisastra, M. F., Darmawan, I., & Nurjanah, D. (2024). Dataset development for Quran memorizers: A step towards data-driven personalized learning path. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 14(1). <https://doi.org/10.52549/ijeei.v14i1.7343>
- Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024, Article 4067721. <https://doi.org/10.1155/2024/4067721>
- Chen, M., & Liu, Z. (2024). Predicting performance of students by optimizing tree components of random forest using genetic algorithm. *Heliyon*, 10(12), e32570. <https://doi.org/10.1016/j.heliyon.2024.e32570>
- Chen, Y., & Jin, K. (2024). Educational performance prediction with Random Forest and innovative optimizers: A data mining approach. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 15(3). <https://doi.org/10.14569/IJACSA.2024.0150308>
- Hariyanto, F., Budiman, T., Yulianto, A. B., & Yasin, V. (2025). Designing a web-based information system for monitoring final projects. *International Journal of Engineering, Science and Information Technology (IJESTY)*, 5(2), 142–153. <https://doi.org/10.52088/ijesty.v5i2.799>
- Haryono, K., Rajagede, R. A., & Negara, M. U. A. S. (2023). Quran memorization technologies and methods: Literature review. *International Journal on Informatics for Development (IJID)*, 11(1), 192–201. <https://doi.org/10.14421/ijid.2022.3746>
- Jiang, J., Zhang, X., & Yuan, Z. (2024). Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. *Expert Systems with Applications*, 249, 123633. <https://doi.org/10.1016/j.eswa.2024.123633>
- Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., Alkhalaf, S., & Abougalala, R. A. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies*, 29, 21621–21645. <https://doi.org/10.1007/s10639-024-12619-w>
- Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost data mining algorithms for anticipating students' academic performance. *International Journal of Modern Education and Computer Science (IJMECS)*, 16(2), 29–44. <https://doi.org/10.5815/ijmecs.2024.02.03>
- Nalenz, M., Rodemann, T., & Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 113, 3379–3398. <https://doi.org/10.1007/s10994-023-06439-1>
- Noviandy, T. R., Zahriah, Z., Yandri, E., Jalil, Z., Yusuf, M., Yusof, N. I. S. M., Lala, A., & Idroes, R. (2024). Machine learning for early detection of dropout risks and academic excellence: A stacked classifier approach. *Journal of Educational Management and Learning*, 2(1), 28–34. <https://doi.org/10.60084/jeml.v2i1.191>
- Nurdin, M., & Fauziah, F. (2024). Analytical study forecasting students using Random Forest and linear regression algorithms. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(4), 2369–2378. <https://doi.org/10.33395/sinkron.v8i4.13886>
- Shantal, M., Othman, Z., & Abu Bakar, A. (2025). Missing data imputation using correlation coefficient and min-max normalization weighting. *Intelligent Data Analysis*, 29(1). <https://doi.org/10.3233/IDA-230140>
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(1), 2. <https://doi.org/10.1007/s44163-023-00079-z>
- Zhao, Y., Zhao, F., Liu, S., & Zhang, J. (2024). Research on student performance prediction based on Random Forest algorithm. *Proceedings of the 2024 International Symposium on Artificial Intelligence for Education (SAIE 2024)*, 8–13. <https://doi.org/10.1145/3700297.3700385>
- Ahmad, A. I., Nugroho, D. A., Aliyu, S. A., & Abdullahi, A. M. (2025). Predicting student dropout in e-learning using simple machine learning and explainable data analysis. *LogicLink: Journal of Artificial Intelligence and Multimedia in Informatics*, 2(2), 138–148. <https://doi.org/10.28918/logiclink.v2i2.13116>
- Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2023). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 142, 107704. <https://doi.org/10.1016/j.chb.2023.107704>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. <https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>