

SISTEM RAMALAN PENYENGGARAAN TRANSIT ALIRAN RINGAN (LRT) MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN

¹Nur Dyana Binti Nasir Ahmad, ¹Zalinda Othman

**¹Fakulti Teknologi & Sains Maklumat
43600 Universiti Kebangsaan Malaysia**

Abstrak

Transit Aliran Ringan (LRT) di Malaysia merupakan salah satu bentuk pengangkutan awam utama yang membantu mengurangkan kesesakan lalu lintas serta menyediakan kemudahan pengangkutan yang efisien. Walaupun sistem ini semakin berkembang dengan penggunaan teknologi moden, masih terdapat isu yang perlu diberi perhatian, khususnya berkaitan penyenggaraan. Masalah utama yang dikenalpasti dalam kajian ini ialah penjadualan penyenggaraan yang tidak efisien, yang boleh menjejaskan kelancaran operasi dan keselamatan pengguna. Oleh itu, fokus kajian ini adalah untuk membangunkan satu sistem ramalan penyenggaraan LRT menggunakan teknik pembelajaran mesin seperti Logistic Regression (LR), K-Nearest Neighbours (KNN), dan Random Forest (RF) dengan pendekatan CRISP-DM. Data untuk sistem ini diperoleh daripada Kaggle. Set Data Klasifikasi Penyenggaraan Ramalan Mesin (Dataset Machine Predictive Maintenance Classification) yang digunakan adalah sintetik, mengandungi 10,000 baris data dengan 5 atribut. Penilaian prestasi model dilakukan menggunakan metrik seperti ketepatan (accuracy), kejituan (precision), dapatan semula (recall), skor F1 (F1-score), dan AUC (Area Under Curve), bagi memastikan kebolehpercayaan dan keberkesanan model ramalan yang dibangunkan. Hasil daripada kajian ini mendapati bahawa Random Forest adalah model terbaik dengan nilai ketetapan 96%. Sistem ramalan penyenggaraan ini juga dilengkapi dengan papan pemuka interaktif yang memaparkan analisis ramalan penyenggaraan dalam format yang mudah difahami. Walau bagaimanapun, kekangan seperti kualiti dan kuantiti data, dapat mempengaruhi ketepatan ramalan dan keberkesanan sistem.

Abstract

The Light Rail Transit (LRT) system in Malaysia is one of the main forms of public transportation that helps reduce traffic congestion while providing efficient transport services. Although the system has been expanding with the adoption of modern technologies, there are still issues that need to be addressed, particularly in relation to maintenance. The main problem identified in this study is the inefficiency of maintenance scheduling, which can affect operational smoothness and user safety. Therefore, the focus of this study is to develop a predictive maintenance system for the LRT using machine learning techniques such as Logistic Regression (LR), K-Nearest Neighbours (KNN), and Random Forest (RF), following the CRISP-DM approach. The data for this system was obtained from Kaggle. The dataset used, titled "Machine Predictive Maintenance Classification Dataset," is synthetic and consists of 10,000 rows of data with 5 attributes. The performance of the models was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC (Area Under Curve), to ensure the reliability and effectiveness of the predictive models developed. The results of the study show that Random Forest is the best-performing model with an accuracy of 96%. The predictive maintenance system is also equipped with an interactive dashboard that displays maintenance prediction analysis in an easily understandable format. However, limitations such as the quality and quantity of data may affect the accuracy of predictions and the overall effectiveness of the system.

1.0 PENGENALAN

LRT adalah salah satu bentuk pengangkutan awam yang penting di Malaysia. Sistem LRT yang beroperasi dengan baik membantu mengurangkan kesesakan lalu lintas dan memberikan kemudahan pengangkutan yang efisien kepada orang ramai. Penyenggaraan yang berkesan adalah amat penting untuk memastikan keselamatan dan kecekapan sistem tersebut. Penyenggaraan bukan sahaja memastikan operasi yang optimum, tetapi juga memanjangkan jangka hayat aset dan mengurangkan risiko kerosakan mengejut. Bagi mengurangkan kejadian kerosakan besar atau kegagalan sistem, penyenggaraan perlu dilakukan secara berkala dan sistematik. Namun begitu, banyak sistem penyenggaraan masih menggunakan pendekatan berjadual secara manual yang kurang efisien. Keadaan ini sering membawa kepada peningkatan kos operasi, masa henti yang lebih lama, dan risiko keselamatan yang lebih tinggi kepada pengguna serta pengendali sistem. Sebagai alternatif, penyenggaraan prediktif yang menggunakan analisis data telah menjadi pilihan utama dalam industri moden. Kaedah ini membolehkan sistem mengenal pasti corak dan tanda awal kerosakan berdasarkan data sejarah, dan menjadualkan penyenggaraan sebelum kegagalan berlaku. Walau bagaimanapun, cabaran utama bagi pendekatan ini adalah bagaimana untuk memanfaatkan data yang tersedia secara berkesan bagi meramalkan kerosakan.

Objektif projek ini adalah untuk membina model ramalan dan membangunkan sistem penyenggaraan yang berasaskan data. Sistem ini dibangunkan menggunakan data sintetik daripada Kaggle, yang akan dipra-proses sebelum digunakan dalam pembinaan model ramalan. Antara algoritma yang digunakan termasuk Logistik Regresi, K-Jiran Terdekat dan Hutan Rawak, yang kemudiannya dinilai menggunakan metrik prestasi seperti ketepatan, kejituan, dapatan semula, skor F1 dan AUC.

Sistem ini turut dilengkapi dengan papan pemuka interaktif yang memaparkan hasil analisis dalam format visual yang mudah difahami. Skop projek hanya melibatkan penggunaan data terbuka dan tidak menyentuh sistem sebenar LRT. Kekangan utama ialah kesukaran mendapatkan data sebenar atas faktor kerahsiaan. Projek ini juga menggunakan pendekatan CRISP-DM yang merangkumi enam fasa utama iaitu pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian dan pembangunan. Metodologi ini menyediakan rangka kerja yang sistematik dan fleksibel bagi memastikan keberkesanan pembangunan sistem.

2.0 KAJIAN LITERATUR

Pelbagai penyelidikan telah dijalankan berkaitan penggunaan teknik pembelajaran mesin dalam sistem ramalan penyenggaraan, khususnya dalam bidang pengangkutan awam. Kajian oleh Guven & Sahin (2022) membangunkan sistem ramalan penyenggaraan kenderaan awam menggunakan algoritma seperti Logistic Regression, SVM, Naive Bayes, KNN dan Random Forest yang digabungkan dengan logik fuzzy. Sistem ini mampu memantau keadaan kenderaan secara masa nyata dan memberikan amaran awal kepada pusat kawalan, sekali gus membolehkan tindakan pencegahan diambil sebelum berlaku kerosakan kritikal.

Kajian oleh Bianchi et al. (2023) pula memberi fokus kepada infrastruktur rel, terutamanya peralihan ke arah penyenggaraan ramalan berdasarkan data dan AI. Penyelidikan ini menyesuaikan pendekatan sistem Prognostics and Health Management (PHM) bagi mengoptimumkan jadual penyenggaraan serta meningkatkan keselamatan operasi. Komponen kritikal seperti sambungan rel dan struktur sokongan dikenal pasti sebagai elemen penting dalam pelaksanaan PdM (Predictive Maintenance) untuk rel.

Zhou et al. (2022) meneliti ramalan kos operasi sistem LRT di Amerika Syarikat menggunakan pelbagai algoritma seperti ANN, KNN, Random Forest dan Support Vector Regression. Kajian ini mendapati bahawa KNN memberikan ketepatan ramalan yang tinggi berbanding kaedah tradisional seperti Multiple Linear Regression, dan dapat membantu perancangan bajet operasi yang lebih efisien.

Dalam konteks lain, Chazhoor et al. (2020) menekankan kepentingan proses prapemprosesan data dalam meningkatkan ketepatan model klasifikasi. Teknik seperti SMOTE, PCA dan LDA digunakan untuk menangani ketidakseimbangan data dan mengurangkan dimensi, manakala pelbagai model seperti Random Forest, Logistic Regression, MLP dan AdaBoost diuji. Walaupun kajian ini tertumpu pada industri semikonduktor, teknik yang digunakan adalah relevan untuk aplikasi dalam sistem pengangkutan. Secara keseluruhan, penyelidikan-penyelidikan ini menegaskan bahawa pemilihan algoritma, kualiti data dan integrasi teknologi merupakan faktor penting dalam keberkesanan sistem ramalan penyenggaraan berasaskan pembelajaran mesin.

3.0 METODOLOGI

Metodologi CRISP-DM digunakan dalam proyek ini sebagai pendekatan sistematis yang merangkumi enam fasa utama: pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian dan pembangunan.

3.1 Pemahaman Perniagaan

Masalah utama yang dikenalpasti ialah penjadualan penyenggaraan LRT yang tidak konsisten, mengakibatkan gangguan operasi, penurunan prestasi dan ketidakselesaian kepada pengguna. Oleh itu, projek ini membangunkan sistem ramalan penyenggaraan menggunakan teknik pembelajaran mesin.

3.2 Pemahaman Data

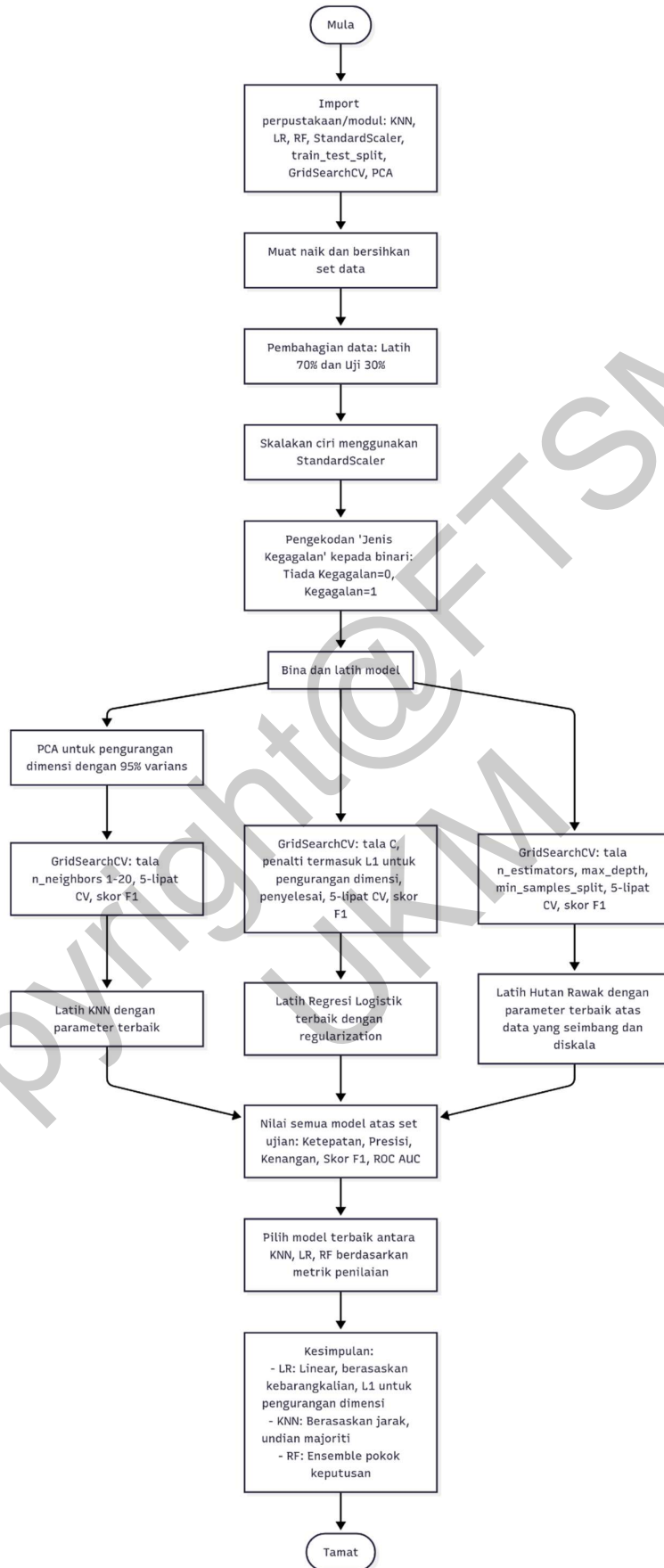
Data yang digunakan ialah data sintetik dari Kaggle iaitu *Dataset Predictive Maintenance Classification*, yang mengandungi 10,000 baris dengan lima atribut penting iaitu Air Temperature [K], Process Temperature [K], Rotational Speed [rpm], Torque [Nm], Tool Wear [min], dan label kelas sasaran 'Failure Type'.

3.3 Penyediaan Data

Pra-pemprosesan melibatkan beberapa langkah: menukar label Failure Type kepada bentuk binari (0 untuk No Failure, 1 untuk Failure), serta penskalaan data menggunakan `StandardScaler` untuk menormalisasi nilai-nilai numerik. Semua proses dijalankan dalam Google Colab.

3.4 Pemodelan

Tiga algoritma digunakan iaitu Logistic Regression, K-Nearest Neighbour dan Random Forest. KNN menggunakan pendekatan berdasarkan jarak, Logistic Regression berdasarkan model linear kebarangkalian, manakala Random Forest menggunakan gabungan pokok keputusan secara rawak. Semua model menerima input yang sama dan diuji untuk mengenal pasti prestasi terbaik.



Rajah 0.1 Carta Alir Pembinaan Model

Ketiga-tiga model memulakan proses dengan input data yang sama, merangkumi 5 atribut iaitu Air Temperature [K], Process Temperature [K], Rotational Speed [rpm], Torque [Nm], dan Tool Wear [min]. Sebelum digunakan dalam pemodelan, semua model memerlukan pra-pemprosesan data yang melibatkan dua langkah utama iaitu, penskalaan ciri yang dilakukan menggunakan StandardScaler bagi memastikan semua atribut berada pada julat yang sama dan mengelakkan dominasi satu atribut ke atas yang lain dan pengekodan label kelas di mana nilai Failure Type ditukar kepada bentuk binari, iaitu No Failure = 0 dan Failure = 1. Langkah ini adalah sama dan penting untuk semua model kerana ia memastikan input bersedia untuk dimasukkan ke dalam algoritma pembelajaran mesin masing-masing.

Model KNN menggunakan pendekatan berasaskan jarak dan tidak melibatkan proses pembelajaran. Selepas data dipra-proses, KNN menyimpan keseluruhan set latihan. Apabila menerima data baharu, model mengira jarak Euclidean antara input dan setiap titik data dalam set latihan. Kemudian, ia memilih K titik terdekat dan meramalkan kelas berdasarkan majoriti undian jiran tersebut. Berbeza dengan model lain, Logistik Regresi ialah model berasaskan persamaan linear. Ia membina model dengan menganggarkan pemberat bagi setiap atribut semasa proses latihan. Input digabungkan secara linear menggunakan pemberat dan dimasukkan ke dalam fungsi logistik untuk menghasilkan kebarangkalian ramalan bagi setiap kelas. Ramalan akhir diberikan berdasarkan kebarangkalian tertinggi, menjadikannya model paling mudah dan cepat, tetapi terhad jika hubungan antara pembolehubah adalah tidak linear. Hutan Rawak berbeza dari kedua-dua model sebelumnya kerana ia menggunakan kaedah ensemble learning, dengan membina berbilang pokok keputusan. Setiap pokok dilatih pada subset rawak data menggunakan teknik bagging, dan pada setiap nod, hanya subset atribut digunakan untuk pemisahan, menjadikan pokok-pokok itu berbeza antara satu sama lain. Semasa ramalan, setiap pokok membuat keputusan sendiri, dan hasil akhir ditentukan melalui undian majoriti.

3.5 Penilaian

Penilaian model adalah langkah penting untuk memastikan ketepatan, kebolehpercayaan, dan keberkesanan model dalam meramal keperluan penyenggaraan. Penilaian ini dilakukan menggunakan metrik prestasi seperti ketepatan, kejituan, dapatan semula, skor F1 dan ROC AUC. Fasa ini membantu memastikan bahawa model tidak hanya baik pada data latihan, tetapi juga boleh diaplikasikan kepada data baharu dengan hasil yang konsisten dan boleh dipercayai. Model yang menunjukkan prestasi terbaik akan dipilih untuk integrasi dalam sistem akhir.

3.6 Pembangunan

Sistem akhir dibangunkan sebagai antara muka ramalan interaktif. Pengguna akan memasukkan nilai atribut dan sistem akan memaparkan hasil ramalan sama ada berlaku “Failure” atau “No Failure”. Pembangunan menggunakan bahasa Python dan platform Google Colab. Sistem juga memerlukan keperluan perkakasan minimum seperti Windows 10, Intel Core i5 dan RAM 8GB. Antara keperluan fungsian termasuk penggunaan model pembelajaran mesin dan paparan papan pemuka ramalan. Keperluan bukan fungsian pula menekankan kebolehgunaan sistem agar mesra pengguna.

4.0 HASIL

4.1 Implementasi Sistem

Sistem dibangun berdasarkan keputusan fasa pemodelan dan penilaian yang dijalankan sebelum ini. Proses pembangunan dilakukan dalam persekitaran Google Colab menggunakan bahasa pengaturcaraan Python, serta pakej berkaitan pembelajaran mesin seperti scikit-learn, pandas, matplotlib, seaborn dan SHAP. Model ramalan yang dibangun turut diintegrasikan ke dalam sistem antara muka pengguna menggunakan Streamlit.

4.2 Model KNN

Model KNN menggunakan pendekatan berasaskan jarak untuk mengklasifikasikan data. Sebelum penalaan parameter, prestasi model menunjukkan overfitting dengan ketepatan 1.0 pada set latihan tetapi prestasi lemah pada set ujian. Selepas penalaan menggunakan GridSearchCV, nilai `n_neighbors = 1` memberikan keputusan terbaik dengan ketepatan set latihan 1.0 dan set ujian 0.95. Model KNN sesuai kerana keupayaannya mengendalikan data kecil dan menghasilkan keputusan klasifikasi yang baik selepas dioptimumkan.

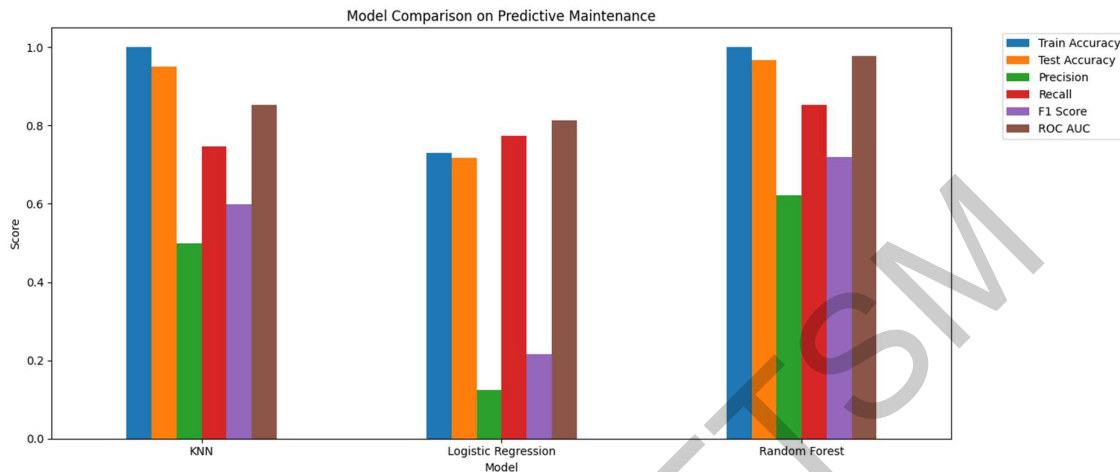
4.3 Model Logistic Regression

Model Logistic Regression digunakan untuk meramalkan kebarangkalian kejadian berdasarkan input. Prestasi awal model menunjukkan nilai ketepatan sederhana (sekitar 0.70), mencerminkan kemungkinan masalah ketidakseimbangan data. Selepas penalaan parameter menggunakan GridSearchCV (dengan `C = 10`, `penalty = 'l1'`, `solver = 'liblinear'`), ketepatan meningkat kepada 0.72 untuk data latihan dan 0.71 untuk data ujian. AUC model meningkat kepada 0.81, menunjukkan kebolehan pemisahan yang lebih baik.

4.4 Model Random Forest

Model Random Forest menggunakan gabungan pokok keputusan untuk meningkatkan kestabilan dan ketepatan. Sebelum penalaan, prestasi kurang konsisten. Selepas penalaan menggunakan GridSearchCV (dengan `n_estimators = 50`, `max_depth = 20`, `min_samples_split = 2`), model menunjukkan ketepatan 0.99 (latihan) dan 0.96 (ujian). AUC model ialah 0.97, menjadikannya model terbaik dalam kajian ini dari segi prestasi keseluruhan.

4.5 Penilaian dan Perbandingan Model



Rajah 4.1 Graf Perbandingan Model

Jadual 4.1 Perbandingan Metrik Prestasi bagi setiap Model

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	ROC AUC
KNN	1.000000	0.950000	0.500000	0.746667	0.598930	0.853684
Logistic Regression	0.729549	0.718333	0.125135	0.773333	0.215413	0.812461
Random Forest	0.999699	0.966667	0.621359	0.853333	0.719101	0.978543

Dalam kajian ini, tiga model pembelajaran mesin telah dibangunkan dan dibandingkan untuk tugas pengelasan binari. Penilaian dibuat berdasarkan beberapa metrik penting seperti ketepatan latihan dan ujian, kejitian, dapatan semula, skor F1 dan skor ROC AUC.

Model KNN menunjukkan ketepatan latihan yang sempurna (1.000) dan ketepatan ujian yang tinggi iaitu 0.950. Ini menunjukkan model terlalu sesuai dengan data latihan (*overfitting*), namun masih mampu memberi prestasi baik pada data ujian. Kejitian model ini adalah 0.500 dan kepekaan sebanyak 0.7467, manakala skor F1 ialah 0.5989. Skor ROC AUC yang diperoleh ialah 0.8537, menunjukkan keupayaan pembezaan model terhadap kelas yang baik.

Model Logistik Regresi pula mempunyai prestasi yang paling rendah antara ketiga-tiga model. Ketepatan latihan dan ujian masing-masing adalah 0.7295 dan 0.7183. Walaupun model ini menunjukkan kejitian yang tinggi (0.7733), ia mencatat

dapatan semula yang sangat rendah (0.1251), yang menunjukkan banyak ramalan positif palsu telah dilakukan. Skor F1 hanya 0.2154, manakala ROC AUC ialah 0.8125, membuktikan bahawa model ini masih boleh membezakan kelas, namun tidak cekap dari segi ketepatan ramalan.

Model Hutan Rawak mencatat prestasi terbaik secara keseluruhan. Ketepatan latihan hampir sempurna (0.9997) dan ketepatan ujian adalah yang tertinggi (0.9667), menandakan model ini mampu menyesuaikan diri dengan baik tanpa terlalu *overfit*. Model ini mencatat kejituan sebanyak 0.6214 dan dapatan semula tertinggi iaitu 0.8533, menunjukkan keseimbangan antara mengenal pasti kes sebenar dan mengurangkan ramalan salah. Skor F1 adalah 0.7191, dan skor ROC AUC adalah yang tertinggi iaitu 0.9785, menandakan keupayaan klasifikasi yang sangat tinggi.

Berdasarkan perbandingan ini, Hutan Rawak merupakan model terbaik untuk digunakan dalam sistem ramalan ini kerana ia menunjukkan keseimbangan yang baik antara semua metrik prestasi. KNN juga memberi prestasi yang baik tetapi menunjukkan kemungkinan berlakunya *overfitting*. Sementara itu, Logistik Regresi kurang sesuai untuk data ini kerana prestasinya yang rendah dari segi kejituan dan skor F1.

4.6 Antara Muka Pengguna

Antara muka pengguna yang dibina membolehkan pengguna memasukkan input ciri-ciri yang relevan untuk proses ramalan seperti Air temperature [K], Process temperature [K], Rotational speed [rpm], Torque [Nm], dan Tool wear [min]. Setelah pengguna memasukkan input, butang 'Submit' boleh ditekan untuk menghantar data tersebut ke backend sistem. Kod Python akan memproses input ini dan menggunakan model pembelajaran mesin Hutan Rawak untuk menjana ramalan kegagalan. Model ini digunakan adalah kerana ia mencapai penilaian metrik yang paling tinggi berbanding dua model pembelajaran mesin yang dilatih.

Hasil ramalan akan dipaparkan secara langsung di panel keputusan menggunakan fungsi `st.write()` dan `st.success()` Streamlit. Rajah menunjukkan contoh antara muka menggunakan Streamlit dan paparan ramalan yang terhasil berdasarkan input pengguna.

PREDICTIVE MAINTENANCE OF LIGHT RAIL TRANSIT (LRT)

Enter Data

Air temperature [K]: 298.9

Process temperature [K]: 309.1

Rotational speed [rpm]: 2861

Torque [Nm]: 4.8

Tool wear [min]: 143

PREDICTION

FAILURE DETECTED

Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Failure
298.9	309.1	2861.0	4.8	143.0	X
298.1	308.6	2851.0	42.8	7.0	0

Download Prediction History

SUBMIT RESET

Rajah 4.2 Antaramuka Sistem Ramalan Penyenggaraan LRT

5.0 KESIMPULAN

Kajian ini telah berjaya merealisasikan pembangunan sebuah sistem ramalan penyenggaraan LRT dengan menggunakan teknik pembelajaran mesin berasaskan pendekatan CRISP-DM. Objektif utama kajian adalah untuk membina sistem ramalan penyenggaraan yang mampu meramal kegagalan LRT.

Sistem yang dibangunkan menggunakan dataset sintetik berstruktur yang diperoleh daripada platform Kaggle, yang merangkumi atribut seperti Air temperature [K], Process temperature [K], Rotational speed [rpm], Torque [Nm], dan Tool wear [min]. Atribut ini dipilih berdasarkan kebolehannya untuk mencerminkan keadaan sebenar komponen sistem mekanikal dalam LRT. Kajian ini menjalankan proses eksplorasi data secara menyeluruh bagi memahami hubungan antara ciri-ciri tersebut, di mana analisis korelasi mendapati bahawa terdapat hubungan linear yang kuat antara Air Temperature [K] dan Process Temperature [K] serta antara Rotational Speed [rpm] dan Torque [Nm].

Dalam fasa permodelan, tiga algoritma pembelajaran mesin telah digunakan dan dibandingkan iaitu Logistik Regresi, KNN dan Hutan Rawak. Setiap model dilatih dan diuji dengan menggunakan pecahan data 70:30 untuk mendapatkan hasil yang optimum. Untuk meningkatkan prestasi model, teknik tambahan seperti PCA digunakan dalam model KNN bagi menangani isu multikolineariti, manakala regularisasi digunakan untuk model LR. Berdasarkan hasil penilaian menggunakan metrik prestasi seperti ketepatan, ketelitian, kepekaan, skor F1, dan skor roc auc, didapati bahawa model Hutan Rawak memberikan prestasi terbaik dari segi kebolehan mengklasifikasikan label kelas dengan lebih tepat. Ini menunjukkan keupayaan model tersebut dalam menangani data bersaiz besar dan pelbagai atribut dengan baik.

Keseluruhan projek telah menggunakan pendekatan CRISP-DM sebagai kerangka metodologi utama. Proses ini bermula dengan fasa pemahaman perniagaan, diikuti pemahaman data, penyediaan data, permodelan, penilaian dan pembangunan. Setiap fasa telah dijalankan secara sistematik untuk memastikan pembangunan sistem memenuhi keperluan perniagaan dan teknikal. Pendekatan ini membuktikan bahawa penggunaan metodologi perlombongan data yang berstruktur mampu menghasilkan sistem ramalan yang terancang dan boleh digunakan secara praktikal.

Secara keseluruhan, kajian ini telah membuktikan bahawa penggunaan pembelajaran mesin dalam bidang penyenggaraan prediktif adalah berpotensi untuk diimplementasikan dalam sektor pengangkutan awam seperti LRT.

6.0 RUJUKAN

- Bianchi G, Fanelli C, Freddi F, Giuliani F, La Placa A. 2025. Systematic review railway infrastructure monitoring: From classic techniques to predictive maintenance. *Advances in Mechanical Engineering*. 17(1).
- Chazhoor, A., Mounika, Y., Vergin Raja Sarobin, M., Sanjana, M. V., & Yasashvini, R. 2020. Predictive maintenance using machine learning based classification models. *IOP Conference Series: Materials Science and Engineering*, 954(1), 012001.
- Craig, L., & Tucci, L. 2024. What is machine learning? guide, definition and examples. Search Enterprise AI.
- Erbiyik, H. 2023. Definition of Maintenance and Maintenance Types with Due Care on Preventive Maintenance. IntechOpen.
- Fikri, M. N. 2020. The impact of fuzzy discretization's output on classification accuracy of random forest classifier. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3), 3950–3956.
- Framework, B. D. 2024. Data types: Structured vs. Unstructured Data: Enterprise Big Data... Enterprise Big Data Framework.
- Gawde, S., Patil, S., Kumar, S., Kamat, P., Kotecha, K., & Alfarhood, S. 2024. Explainable predictive maintenance of rotating machines using lime, shap, PDP, ice. *IEEE Access*, 12, 29345–29361.
- GÜVEN, Ö., & ŞAHİN, H. 2022. Predictive maintenance based on machine learning in public transportation vehicles. *Mühendislik Bilimleri ve Araştırmaları Dergisi*, 4(1), 89–98.
- Hassan, N. A. 2022. Pengguna LRT Mengeluh Perkhidmatan kerap “bermasalah.” *Harian Metro*.
- Jacquelyne AB, Azlis-Sani J, Nor Azali A, Nur-Annuar M, Shahrul Azhar S, & Mifzal-Nazhan M. (2020). TASK ANALYSIS ON MAINTENANCE WORKER (RAIL GRINDER) OF LIGHT RAIL TRANSIT (LRT). *Malaysian Journal of Public Health Medicine*, 20(Special1), 223–230.

- Jing Zhu, J. Z., Jing Zhu, S. H., Song Huang, Y. S., Yaqing Shi, K. W., & Kaishun Wu, Y. W. 2022. An algorithm combining random forest classification and Fuzzy Comprehensive Evaluation. 網際網路技術學刊, 23(4), 735–747.
- Kargar-Sharif-Abad, M., Kharazian, Z., Miliou, I., & Lindgren, T. 2024. Shap-driven explainability in survival analysis for Predictive Maintenance Applications.
- Leung, C. K., Elias, J. D., Minuk, S. M., Roy R. de Jesus, A., & Cuzzocrea, A. 2020. An innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on Big Transportation Data. 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–8.
- Masirin, M. I., Salin, A. M., Zainorabidin, A., Martin, D., & Samsuddin, N. (2017). Review on Malaysian Rail Transit Operation and management system: Issues and solution in integration. IOP Conference Series: Materials Science and Engineering, 226, 012029.
- Noursalehi, P., Koutsopoulos, H. N., & Zhao, J. 2020. Machine-learning-augmented analysis of textual data: Application in transit disruption management. IEEE Open Journal of Intelligent Transportation Systems, 1, 227–236.
- Prayogo, R. H., Kurnianto, B. A., Nababan, N. P., & Suharjito, S. 2024b. Predictive maintenance using linear regression. Syntax Literate ; Jurnal Ilmiah Indonesia, 9(3), 1958–1970.
- Rajabinasab, M., Pakdaman, F., Zimek, A., & Gabbouj, M. 2025. Randomized PCA forest for approximate K-nearest neighbor search. Expert Systems with Applications, 281, 126254.
- Ren, Y., Zhu, X., Bai, K., & Zhang, R. 2023. A new random forest ensemble of intuitionistic fuzzy decision trees. IEEE Transactions on Fuzzy Systems, 31(5), 1729–1741.
- Ricketts, J., Barry, D., Guo, W., & Pelham, J. 2023. A scoping literature review of natural language processing application to safety occurrence reports. Safety, 9(2), 22. <https://doi.org/10.3390/safety9020022> [
- Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. 2019. A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101.

Trihandaru, S., Parhusip, H. A., Heriadi, A. H., Santosa, P. P., Sardjono, Y., & Lea, L. 2024. Selection dominant features using principal component analysis for predictive maintenance of heave engines. JTAM (Jurnal Teori Dan Aplikasi Matematika), 8(4), 1026.

Yu, Y., Yu, F., He, Y., & Yang, W. 2021. A hybrid fuzzy random forest algorithm using harmonic search algorithm for parameter tuning. 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 936–939.

Zhou, G., Etemadi, A., & Mardon, A. 2022. Machine learning-based cost predictive model for better operating expenditure estimations of U.S. Light Rail Transit Projects. Journal of Public Transportation, 24, 100031.

Nur Dyana Binti Nasir Ahmad (A193348)

Prof Madya Dr. Zalinda Othman

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia