

MODEL PENYENGGARAAN JANGKAAN BERDASARKAN PENDEKATAN BINARISASI DAN *NAIVE BAYES*

Muhammad Saifullah bin Hj Muhamad Juhari, Syaimak binti Abdul Shukor

Fakulti Teknologi Sains dan Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

syaimak@ukm.edu.my

ABSTRAK

Pendekatan pembelajaran mesin dalam aplikasi penyenggaraan jangkaan menjadi semakin luas sesuai dengan peredaran masa. Terdapat pelbagai penyelidikan yang telah dilakukan mengenai topik ini dan sudah terdapat banyak industri yang telah mengaplikasikan pendekatan pembelajaran mesin dalam aplikasi penyenggaraan jangkaan. Namun, penggunaannya masih tidak luas bagi sesetengah industri dan industri tersebut masih lagi mengamalkan cara penyelenggaraan yang kurang efisien serta melibatkan kos yang tinggi. Projek ini bertujuan untuk memperluaskan lagi penyelidikan berkaitan dengan pendekatan pembelajaran mesin dalam penyenggaraan jangkaan dengan menggunakan set data sintetik yang telah diterbitkan untuk kegunaan umum. Kajian yang dilakukan dalam projek ini memfokuskan pada kombinasi teknik binarisasi ke atas set data kajian serta penggunaan algoritma Naive Bayes dalam proses perlombongan data yang tidak digunakan dalam kajian lepas serta gabungan Naive Bayes dengan teknik Bagging dan Boosting. Kaedah SMOTE dan Persampelan Terkurang dipilih untuk mengawal kesan ketidakseimbangan pada data-data dalam atribut yang meramal kegagalan mesin. Enam model dibangunkan di dalam projek ini iaitu model Naive Bayes [SMOTE], model Naive Bayes beserta Bagging [SMOTE], model Naive Bayes beserta Boosting [SMOTE], model Naive Bayes [Persampelan Terkurang], model Naive Bayes beserta Bagging [Persampelan Terkurang] dan model Naive Bayes beserta Boosting [Persampelan Terkurang]. Model yang terbaik iaitu model Naive Bayes [SMOTE] memperoleh nilai Kejituan 0.999 dan Kebersihan 1.0 yang mengatasi hasil kajian yang lepas.

Kata kunci: penyenggaraan jangkaan, pembelajaran mesin, *Naive Bayes*, SMOTE, persampelan terkurang, *boosting*, *bagging*, binarisasi.

I. PENGENALAN

Seiring dengan kemajuan teknologi, kehidupan seharian semakin bergantung pada mesin, termasuklah untuk tugas-tugas yang ringan sekalipun, sehingga jika sesebuah mesin itu mengalami gangguan, aktiviti-aktiviti seharian berkemungkinan besar akan tergendala disebabkan oleh masalah-masalah yang dihadapi oleh mesin tersebut. Bagi mengurangkan kekerapan kegagalan mesin, pelbagai proses dan kaedah telah diperkenalkan pada setiap tahap kitaran hidup mesin bagi mengoptimumkan jangka hayat mesin. Selain itu, untuk sesetengah sektor yang mempunyai sistem yang sangat kritikal dan memberi impak yang besar kepada umum seperti sektor rel, penerapan jaminan fungsi sistem bermula dari fasa kejuruteraan hingga fasa operasi dan penyelenggaraan telah menjadi salah satu syarat kelulusan yang membolehkan sesebuah projek untuk beroperasi.

Penyelenggaraan *Preventive Maintenance* (PM) adalah sejenis penyelenggaraan yang dirancang di mana masa untuk melakukan penyelenggaraan serta tempoh di antara penyelenggaraan semasa dan penyelenggaraan seterusnya ditentukan berdasarkan pengiraan menggunakan *Failure Developing Period* (FDP). FDP adalah tempoh masa dari masa tanda-tanda kegagalan mesin mula muncul sehingga berlakunya kegagalan mesin. Penyelenggaraan *Corrective Maintenance* (CM) pula adalah jenis penyelenggaraan yang tidak dirancang di mana ia dilakukan setelah berlakunya kegagalan pada mana-mana bahagian mesin. Kegagalan

mesin yang memerlukan penyelenggaraan CM boleh berlaku pada bila-bila masa dan ia dapat terjadi secara rawak. Mesin yang diselenggara mengikut jadual PM masih perlu menjalani penyelenggaraan CM jika kegagalan mesin berlaku di dalam tempoh diantara penyelenggaraan PM terkini dan penyelenggaraan PM seterusnya.

Penyelenggaraan *Predictive Maintenance* (PdM) menggunakan data secara langsung yang dikumpul daripada penderia-penderia yang dipasang pada mesin untuk mengawasi keadaan mesin melalui pelbagai nilai ukuran seperti suhu, getaran dan bunyi. Secara kebiasaannya, penyelenggaraan PdM dimulakan dengan menentukan dan menetapkan syarat-syarat yang menunjukkan bahawa sesebuah mesin sedang menghampiri kegagalan berdasarkan rekod data yang lepas yang disimpan di dalam sistem penyelenggaraan yang dipanggil *Computerized Maintenance Management System* (CMMS). Apabila data-data daripada mesin tersebut mencapai atau melebihi nilai-nilai yang menunjukkan mesin tersebut menghampiri kegagalan, CMMS akan mengeluarkan amaran dan mencadangkan supaya penyelenggaraan perlu dilakukan. Penyelenggaraan PdM menekankan penyelenggaraan pada masa yang betul berbanding penyelenggaraan ketika mesin masih mempunyai tempoh yang masih panjang dari masa kegagalan mesin seperti yang sering berlaku dalam penyelenggaraan PM atau apabila mesin telah gagal seperti yang sering berlaku di dalam penyelenggaraan CM.

Pendekatan pembelajaran mesin merupakan salah satu kaedah yang digunakan di dalam penyelenggaraan PdM disebabkan oleh ketersediaan pelbagai data dari penderia-penderia yang terdapat pada mesin. Data yang dikumpul dari penderia-penderia pada mesin untuk kegunaan penyelenggaraan PdM selalunya terdiri dari beberapa atribut yang mewakili keadaan mesin serta sekurang-kurangnya satu atribut yang menunjukkan status kegagalan mesin. Data seperti ini sesuai digunakan dalam pelbagai algoritma pembelajaran mesin terutamanya algoritma-algoritma dengan pendekatan klasifikasi. Kebiasaannya, penyelenggaraan PdM menggunakan kaedah tradisional seperti sistem CMMS dapat meramal kegagalan mesin yang biasa dengan baik tetapi ia tidak dapat meramal kegagalan mesin yang rawak dengan baik. Dengan menggunakan pendekatan pembelajaran mesin, penyelenggaraan PdM dapat meramal kegagalan dalam kes kegagalan mesin biasa dan juga kegagalan mesin rawak.

Terdapat pelbagai kajian berkaitan pendekatan pembelajaran mesin dalam penyelenggaraan PdM yang telah dikeluarkan oleh penyelidik-penyelidik untuk digunakan sebagai rujukan. Namun, set data yang sesuai untuk penyelidikan pendekatan pembelajaran mesin dalam penyelenggaraan PdM yang dapat diakses oleh orang ramai masih terhad kerana banyak syarikat tidak bersedia untuk berkongsi data-data kepada umum. Disebabkan oleh batasan ini, projek ini akan menggunakan set data buatan yang telah diterbitkan untuk umum. Teknik-teknik yang telah digunakan ke atas set data tersebut dalam kajian-kajian lepas akan dibincangkan dan projek ini akan mencadangkan kaedah yang masih belum digunakan untuk set data kajian bagi membangunkan model penyelenggaraan PdM untuk meramal kegagalan mesin.

A. *Penyataan Masalah*

Pelbagai industri masih menggunakan kaedah PM dan CM dalam proses penyelenggaraan. Walaupun PM dan CM telah terbukti menjadi penyelesaian yang berkesan dalam memastikan mesin berfungsi dengan baik dalam jangka masa yang panjang, masih terdapat ruang untuk penambahbaikan bagi mengoptimumkan kos, skop, dan masa sambil mengekalkan kualiti operasi sistem pada sesebuah mesin. Proses penyelenggaraan tradisional adalah kurang cekap kerana jangka masa penyelenggaraan berkala ditetapkan berdasarkan ramalan ketika fasa

kejuruteraan dan biasanya ia menyebabkan penyelenggaraan dilaksanakan jauh lebih awal dari masa kegagalan mesin berlaku. Secara realitinya, penyelenggaraan mesin menggunakan kaedah tradisional menyebabkan penyelenggaraan dilaksanakan terlalu awal, pembaziran sumber tenaga kerja dan bahan, serta peningkatan jangka hayat yang tidak produktif bagi mesin. Salah satu contoh dapat dilihat daripada penyelenggaraan pesawat di mana 29.3% daripada kos penyelenggaraan telah dibazirkan dengan penggunaan kaedah tradisional di dalam penyelenggaraan pesawat (Lee & Mitici 2023).

Proses penyelenggaraan dapat dioptimumkan dengan menggunakan data yang dihasilkan daripada penderia-penderia yang dipasang pada mesin untuk memantau status mesin. Penyelenggaraan PdM merupakan salah satu cara penyelesaian yang boleh digunakan untuk mengatasi isu-isu yang terhasil daripada penyelenggaraan menggunakan kaedah tradisional. Berdasarkan penyelidikan sebelum ini, pepadanan model PdM dengan teknik pembelajaran mesin telah menunjukkan bahawa prestasi sistem dapat meningkat sebanyak 75% (Rodriguez et al. 2022), 95.6% kerja yang tidak diperlukan dapat dielak dari dilaksanakan (Lee & Mitici 2023), dan kejituan untuk pengesanan kerosakan mesin boleh mencapai nilai 98% (Lee et al. 2019).

Secara umumnya, set data yang melibatkan penyelenggaraan PdM mempunyai struktur data yang sesuai untuk digunakan bagi model berasaskan pendekatan klasifikasi iaitu terdiri daripada beberapa atribut yang menunjukkan keadaan mesin bagi nilai ukuran yang berbeza serta sekurang-kurangnya satu atribut yang memberi indikasi sama ada mesin itu gagal atau masih terus berfungsi. Berdasarkan kajian lalu, teknik-teknik asas pembelajaran mesin yang sering digunakan di dalam kajian berkaitan dengan model penyelenggaraan PdM adalah seperti *Artificial Neural Network* (ANN), *Decision Tree* (DT), *Random Forest* (RF), *Support Vector Machine* (SVM), *k-Nearest Neighbour* (k-NN) dan *Logistic Regression* (LGR). Namun begitu, kajian yang menggunakan teknik *Naive Bayes* (NB) dalam aplikasi PdM masih sedikit berdasarkan rujukan penyelidikan yang telah dikumpul untuk kajian kesusasteraan di dalam projek ini.

Matzka (2020) telah menerbitkan set data buatan berdasarkan data dari mesin sebenar yang dikenali sebagai set data AI4I2020 untuk tujuan penyelidikan (Matzka 2020). Terdapat pelbagai kajian berkaitan dengan pendekatan pembelajaran mesin menggunakan set data tersebut tetapi tiada kertas kajian rasmi seperti jurnal, kertas persidangan dan laporan penyelidikan yang mencadangkan pendekatan NB untuk meramal kegagalan menggunakan set data tersebut. Oleh kerana set data ini dapat diakses oleh orang ramai, terdapat juga laman web seperti Kaggle yang mengandungi forum bagi membincangkan analisis penggunaan NB ke atas set data tersebut. Namun, kod-kod menggunakan algoritma NB yang di kongsi di laman web tersebut tidak menggunakan pendekatan binarisasi di mana keseluruhan struktur set data ditukar kepada jenis binari sebelum digunakan untuk perlombongan data.

Kajian-kajian lepas yang berkaitan dengan pembinaan model PdM bagi set data AI4I2020 sering menggunakan pendekatan pembelajaran mesin asas seperti DT, RF, LGR, k-NN, SVM dan ANN. Ghasemkhani, Aktas dan Birant (2023) telah membuat perbandingan di antara pendekatan pembelajaran mesin yang telah digunakan dengan teknik *Balanced K-star* yang dicadangkan di dalam kajian dan merumuskan bahawa purata skor yang diperolehi dari teknik-teknik tersebut adalah seperti berikut: kejituan 91.74%, kepersisan 0.8052, kepekaan 0.6666 dan skor-F1 0.5760 manakala teknik *Balanced K-Star* menghasilkan keputusan seperti berikut: kejituan 98.75%, kepersisan 0.9877, kepekaan 0.9875 dan skor-F1 0.9875 (Ghasemkhani, Aktas & Birant 2023).

B. Objektif Kajian

Projek ini mempunyai dua objektif utama seperti berikut:

- i. Membangunkan model PdM berasaskan kombinasi teknik binarisasi dan algoritma NB bagi meramal kegagalan pada set data kajian.
- ii. Menilai prestasi model yang dibangunkan dan membandingkan model terbaik dengan kajian lepas.

Jurnal ini dibahagikan kepada beberapa bahagian seperti berikut: Bahagian I akan membincangkan latar belakang penyelidikan, pernyataan masalah dan objektif kajian; Bahagian II akan mengupas mengenai kajian kesusasteraan mengenai PdM secara umum serta penyelidikan berkaitan set data kajian; Bahagian III akan menerangkan tentang kaedah kajian yang dilaksanakan di dalam projek ini; Bahagian IV akan menyelusuri hasil kajian dan akhirnya Bahagian V akan membincangkan tentang cadangan dan rumusan kajian.

II. KAJIAN KESUSASTERAAN

A. Penyenggaraan PdM

PdM adalah salah satu kaedah penyelenggaraan yang lebih optimal berbanding PM dan CM kerana PdM boleh mencadangkan penyelesaian pembaikan, mengenal pasti komponen yang perlu diganti, dan menganggarkan kejadian kegagalan yang boleh mengurangkan kos serta memaksimumkan ketersediaan mesin (Sarvaiya 2021). PdM menjadi semakin mudah untuk diaplikasikan disebabkan oleh kewujudan penderia dan pemproses komputer berkemampuan tinggi yang meluas di pasaran dengan harga yang lebih rendah dan lebih mudah untuk dimiliki berbanding sebelum ini. Dengan kemudahan akses kepada perkakasan dengan keupayaan yang tinggi, data dapat dikumpul dan dianalisis untuk membantu proses PdM. Namun, memiliki alat analisis yang berkeupayaan tinggi tidak menjamin PdM untuk memberikan keputusan analisis yang tepat kerana ia juga bergantung kepada beberapa faktor lain seperti ketersediaan data yang bersesuaian, penggunaan kejuruteraan ciri yang betul, dan perbandingan model ramalan yang berkait (Gonfalonieri 2019).

Pendekatan menggunakan kaedah statistik merupakan teknik tradisional yang sering digunakan di dalam PdM. Kertas kajian, cadangan, dan penyelidikan mengenai PdM mula bertambah dalam beberapa tahun kebelakangan ini dan model-model baru juga diperkenalkan dari semasa ke semasa untuk meningkatkan strategi penyelenggaraan. Satu kajian mengenai kertas-kertas yang berkaitan dengan pelaksanaan pembelajaran mesin untuk PdM dalam industri automotif telah menyimpulkan beberapa perkara penting iaitu kebanyakan kajian melaksanakan pembelajaran mesin yang diselia, kajian dalam domain ini mungkin akan meningkat seiring dengan perluasan akses kepada data, prestasi model pembelajaran mesin akan ditingkatkan apabila lebih daripada satu kaedah digunakan dan terdapat trend berterusan di mana kaedah *Deep Learning* (DL) digunakan untuk penyelenggaraan meramal (Theissler et al. 2021). Di samping itu, berdasarkan laporan ulasan mengenai kertas-kertas yang berkaitan dengan pelaksanaan pembelajaran mesin dalam penyelenggaraan kaedah DL masih belum diaplikasikan dalam PdM sepenuhnya (Sanzana et al. 2022).

B. Pendekatan Pembelajaran Mesin dalam PdM

ANN adalah salah satu kaedah yang biasa digunakan dalam PdM. ANN dapat memendekkan tempoh proses PdM seperti yang ditunjukkan oleh kajian ANN untuk PdM pada aplikasi loji kuasa solar di mana ANN dapat mengurangkan masa pemprosesan untuk mendapatkan tenaga terma tertinggi dari pemanas solar berbanding dengan kaedah konvensional *Non-Linear Predictive Control* (NMPC) (Masero et al. 2023). Dalam satu kajian tentang PdM untuk penyelenggaraan pesawat, penggunaan ANN sebagai algoritma yang digunakan sebelum pengesanan kegagalan dicadangkan sebagai sebahagian daripada model *Maintenance Repair and Overhaul* (MRO) (Safoklov et al. 2022). Satu kertas kajian mengenai aplikasi ANN untuk PdM dalam industri rel mengusulkan algoritma ANN sebagai tambahan kepada siri masa dinamik untuk menganggar kegagalan gelas pada roda berdasarkan suhu gelas. Kajian ini berjaya menunjukkan bahawa terdapat hubungan yang kuat antara RUL dan suhu gelas (Daniyan et al. 2020). Terdapat juga satu lagi kertas kajian mengenai pelaksanaan ANN untuk PdM dalam industri rel yang menggunakan data dari gelas pada roda dan kajian ini berjaya menunjukkan bahawa RUL untuk komponen ini adalah 500 jam dalam tempoh 40 hari dan dapat digunakan untuk memberikan had keyakinan dan pengesanan kecerunan (Daniyan et al. 2020).

Selain menggunakan ANN dalam PdM, terdapat juga algoritma pembelajaran mesin lain yang popular dalam aplikasi ini. Satu kajian yang membandingkan empat kaedah pembelajaran mesin termasuk RF, SVM, k-NN dan *Multi-Layer Perceptron* (MLP) untuk meramalkan tiga keadaan pam air (Normal, Rosak atau Pemulihan) dengan menggunakan data yang dikumpul daripada sensor-sensor menunjukkan bahawa model k-NN menghasilkan keputusan kejituan yang terbaik dalam masa yang paling singkat (Herrero & Zorrilla 2022). Dalam kajian lain, model SVM menghasilkan kejituan yang tertinggi (100%) untuk set data dengan atribut kerosakan mesin berbanding dengan RF dan *Backpropagation Neural Network* (BNN) apabila digunakan untuk meramal kegagalan berdasarkan getaran mesin (Nikfar, Bitencourt & Mykoniatis 2022). PdM untuk penyelenggaraan senduk di stesen keluli elektrik adalah berkesan apabila DT dan RF dilaksanakan untuk meramal keadaan penuaan dengan DT menunjukkan prestasi lebih tinggi berbanding dengan RF (Vannucci et al. 2022).

Teknik hibrid merupakan kaedah pembelajaran mesin yang menggabungkan dua atau lebih algoritma pembelajaran mesin. Kaedah ini terdiri daripada pelbagai teknik yang digunakan dalam satu model yang sama dan ia boleh merupakan kombinasi algoritma pembelajaran mesin dengan kaedah bukan pembelajaran mesin atau gabungan pelbagai algoritma dalam pembelajaran mesin. Salah satu contoh adalah penyelidikan yang memberi tumpuan kepada ramalan siri masa berlangkah dengan pelbagai varian dalam PdM di mana model *Naïve* digabungkan dengan kaedah statistik seperti VARMA, *Theta*, LSTM, GRU, dan ERNN untuk menganalisis data daripada *Federal Reserved Economic Data* (FRED), kualiti udara, ramalan untuk perkakas elektrik, Beijing PM2.5, turbin gas CO dan Nox dan didapati dari kajian tersebut bahawa model yang digabungkan dengan VARMA adalah model terbaik manakala hibrid *Naïve* dan *Theta* adalah model yang paling lemah (Tessoni & Amoretti 2022).

C. Penyelidikan Berkaitan Set Data Kajian

Jadual 1 menunjukkan rumusan prestasi model dari kajian lepas untuk model penyelenggaraan prediktif menggunakan set data AI4I2020. Jadual ini juga mengandungi prestasi bagi model NB menggunakan set data yang sama yang dikongsikan untuk rujukan umum di dalam laman web Kaggle seperti yang dibincangkan sebelum ini.

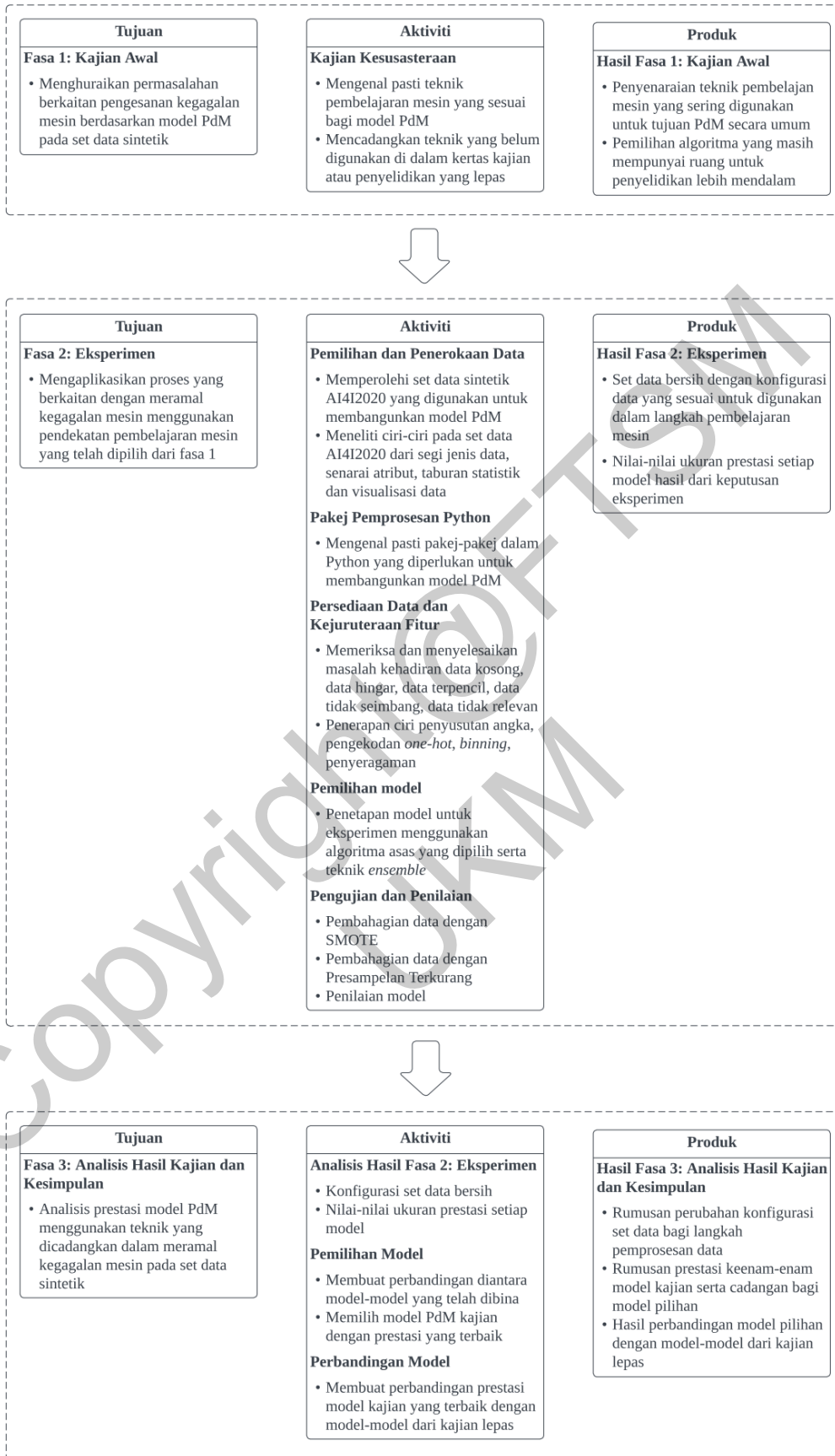
Jadual 1 Rumusan Prestasi Model Kajian Lepas

Kajian	Model	Kejituan	Kepersisan	Kepekaan	Skor-F1	AUC – ROC	AUC – Kepersisan-Kepekaan
(Gaur 2021)	NB	0.785	0.96	0.83	0.88	N/A	N/A
(Ghasemkhani, Aktas & Birant 2023)	<i>Balanced K-Star</i>	0.988	0.988	0.988	0.988	N/A	N/A
(Gujarathi 2021)	NB	0.829	0.836	0.823	0.830	0.901	N/A
(Harichandran, Raphael & Mukherjee 2023)	HUS-ML	0.985	~0.850	~0.750	0.791	N/A	N/A
(Iantovics & Enachescu 2022)	<i>Binary LGR</i>	0.971	N/A	N/A	N/A	N/A	N/A
(Kodihalli 2021)	NB	0.985	0.682	0.968	0.8	N/A	N/A
(Lallahom 2022)	NB	0.843	0.758	0.144	0.242	0.879	N/A
(Nazara 2022)	XGBoost	0.991	N/A	N/A	N/A	0.972	N/A
(Papathanasiou, Demertzis & Tziritas 2023)	<i>Random Survival Forest</i>	0.972	N/A	N/A	N/A	N/A	N/A
(S. K. 2021)	NB	0.789	0.113	0.766	0.196	N/A	N/A
(Sharma et al. 2022)	<i>Random Forest</i>	0.984	N/A	N/A	N/A	0.837	N/A
(Shrimant 2021)	NB	0.762	0.103	0.768	0.182	N/A	N/A

Hasil kajian daripada 12 kertas penyelidikan yang mencadangkan model PdM menggunakan set data AI4I2020 telah dibandingkan di dalam Jadual 1. Model XGBoost yang dicadangkan oleh Nazara (2022) telah memperolehi nilai kejituan yang tertinggi iaitu 0.991. Model *Balanced K-Star* yang digunakan oleh Ghasemkhani, Aktas dan Birant (2023) telah menunjukkan nilai yang paling tinggi bagi ukuran kepersisan iaitu 0.988, kepekaan iaitu 0.988 dan skor-F1 iaitu 0.988. Gujarati (2021) yang menggunakan model NB telah memperolehi nilai AUC-ROC yang tertinggi iaitu 0.901.

III. METODOLOGI KAJIAN

Kaedah kajian yang terdiri dari fasa 1, fasa 2 dan fasa 3 telah dirumuskan di dalam carta alir yang ditunjukkan di dalam Rajah 1 berikut. Rajah ini menunjukkan perkara-perkara penting bagi setiap fasa berdasarkan tujuan fasa tersebut dijalankan, senarai aktiviti-aktiviti yang berlaku bagi setiap fasa serta produk yang akan dihasilkan bagi setiap fasa.



Rajah 1 Kaedah Kajian

A. Fasa 1: Kajian Awal

Fasa 1 adalah kajian awal dengan tujuan untuk mengenal pasti permasalahan kajian iaitu meramal kegagalan mesin menggunakan model PdM ke atas set data sintetik. Di dalam fasa ini, aktiviti yang akan dijalankan termasuklah mencari teknik-teknik yang digunakan untuk membina model PdM serta kaedah pembelajaran mesin yang sering digunakan untuk meramal kegagalan termasuklah model-model yang menggunakan set data sintetik. Setelah bahan rujukan kajian kesusasteraan telah dikumpul dan diteliti, teknik yang belum digunakan untuk membina model PdM menggunakan set data sintetik akan dicadangkan di dalam fasa 1. Hasil dari fasa 1 terdiri daripada senarai algoritma pembelajaran mesin yang digunakan secara umum dan dengan membandingkan teknik-teknik tersebut dengan penyelidikan lepas berkaitan dengan model PdM bagi meramal kegagalan mesin menggunakan set data sintetik, pemilihan algoritma dan teknik yang belum digunakan lagi dapat dicadangkan sebagai kaedah untuk menjalankan eksperimen bagi projek ini.

B. Fasa 2: Eksperimen

Fasa kajian yang seterusnya melibatkan fasa 2 dengan tujuan pelaksanaan eksperimen. Di dalam fasa ini, tujuan yang ditetapkan adalah untuk mengaplikasikan langkah-langkah yang berkaitan dengan pembinaan model PdM menggunakan teknik dan algoritma yang dipilih dari fasa 1 untuk projek ini.

i. Penerokaan Data

Fasa 2 dimulakan dengan penerokaan data melalui analisis senarai atribut, jenis atribut, laporan kualiti data serta visualisasi data bagi set data sintetik AI4I2020 yang terdiri daripada 10,000 data dengan 14 atribut.

ii. Persediaan Data dan Kejuruteraan Fitur

Ia diikuti dengan langkah pembersihan data serta transformasi data melalui kejuruteraan fitur untuk menghasilkan set data bersih yang sesuai untuk kegunaan eksperimen di dalam projek ini.

- *Penyusutan Angka*

Teknik Penyusutan Angka akan diaplikasikan dalam set data kajian untuk mengurangkan jumlah atribut yang akan dimasukkan ke dalam model pembelajaran supaya struktur model tidak menjadi rumit, masa pemprosesan pembelajaran model dapat disingkatkan dan keputusan yang lebih tepat dapat diperolehi. *Torque* dan *Rotational speed* adalah dua ukuran mesin yang saling berkait antara satu sama lain. Dalam kajian ini, atribut baru yang dipanggil *Power [W]* akan dihasilkan manakala atribut *Torque* dan *Rotational speed* akan dikeluarkan dari set data yang baru.

- *Pengekodan One-hot*

Atribut yang mempunyai data dengan ciri kategori kebiasaannya akan membuatkan sesebuah model pembelajaran menjadi rumit dan menambah masa pemprosesan analisis set data. Pengekodan One-hot adalah salah satu teknik untuk mengubah data dengan ciri kategori kepada data dengan ciri angka binari. Teknik ini akan diaplikasikan terhadap atribut Type yang

mempunyai data dengan ciri kategori terdiri daripada L = rendah, M = sederhana dan H = tinggi. Dengan mengaplikasikan teknik ini, atribut *Type* akan dikeluarkan dari set data baru dan ia akan digantikan dengan atribut-atribut baru yang terdiri daripada nilai-nilai di dalam atribut yang asal. Atribut-atribut yang baru ini akan mempunyai nilai sama ada 1 atau 0.

- *Binning*

Binning adalah satu teknik yang sering digunakan dalam pembelajaran mesin untuk data dengan ciri berterusan. *Binning* berfungsi dengan membahagikan data kepada kumpulan-kumpulan berbeza yang berturutan. Teknik ini dipilih untuk diaplikasikan di dalam projek ini untuk menyelesaikan masalah data terpencil dan agihan data yang tidak seragam yang ditunjukkan oleh set data kajian. Selain itu, model pembelajaran mesin dapat memproses data dengan lebih cepat dan tepat menggunakan data dengan struktur yang tidak rumit seperti data yang dihasilkan daripada *Binning*. Teknik *Binning* akan digunakan untuk atribut dengan ciri berterusan iaitu atribut yang baru dihasilkan iaitu atribut *Power* [W] yang dihasilkan melalui Penyusutan Angka serta atribut *Tool wear* [min], *Process temperature* [K] dan *Air temperature* [K]. Saiz bin yang digunakan dalam proses *Binning* dalam kajian ini dirumuskan di dalam Jadual 2.

Jadual 2 Atribut dengan Ciri Berterusan serta Saiz Bin

Atribut	Saiz Bin
Air temperature [K]	2
Process temperature [K]	2
Tool wear [min]	50
Power [W]	1000

- *Binarisasi*

Teknik pengekodan *one-hot* akan digunakan pada data bagi atribut dengan ciri kategori iaitu atribut *Type* untuk menukarkan struktur data tersebut kepada jenis binari. Data dengan ciri berterusan seperti untuk atribut *Air temperature* [K], *Process temperature* [K], *Tool wear* [min] dan *Power* [W] masih akan mempunyai ciri berterusan selepas teknik *Binning*. Untuk memastikan proses binarisasi dilaksanakan ke atas keseluruhan set data, teknik pengekodan *one-hot* akan digunakan ke atas data bagi atribut-atribut tersebut juga untuk menukarkan struktur data kepada jenis binari.

- *Penyeragaman*

Di dalam projek ini, fitur *StandardScaler()* yang terdapat di dalam pangkalan fungsi *Python* akan digunakan untuk menjalankan fungsi penyeragaman pada set data dalam kajian. Langkah ini akan diaplikasikan setelah set data telah dibahagikan kepada set latihan dan set ujian.

iii. Set Data Bersih

Setelah itu, semakan set data bersih serta analisa korelasi antara atribut akan dilakukan.

iv. Pembahagian Data

Langkah seterusnya iaitu pembahagian data yang akan menghasilkan set data latihan merangkumi sebanyak 70% daripada set data bersih dan set data ujian merangkumi sebanyak 30% daripada set data bersih. Set data latihan akan digunakan untuk membina model latihan manakala set data ujian akan digunakan untuk menguji model latihan. Dua teknik pembahagian data akan digunakan bagi setiap model latihan iaitu pembahagian data dengan teknik SMOTE dan pembahagian data dengan teknik Persampelan Terkurang.

v. Model Latihan

Tiga teknik pembelajaran mesin dipilih untuk digunakan pada model latihan iaitu teknik NB, kombinasi teknik NB dan *Bagging*, serta kombinasi teknik NB dan *Boosting*. Enam model yang berlainan iaitu model NB [SMOTE], model NB beserta *Bagging* [SMOTE], model NB beserta *Boosting* [SMOTE], model NB [Persampelan Terkurang], model NB beserta *Bagging* [Persampelan Terkurang] dan model NB beserta *Boosting* [Persampelan Terkurang] akan dihasilkan dari penggunaan dua teknik pembahagian data dan pemilihan tiga kaedah pembelajaran mesin.

vi. Penilaian Model

Selepas itu, pengujian model akan dilakukan dimana model-model latihan akan diuji menggunakan set data ujian untuk meramal atribut *Machine failure*. Hasil eksperimen daripada prestasi model yang terdiri daripada AUC lengkungan ROC, AUC lengkungan Kebersihan-Kepekaan, Skor-F1, Kejituan, Kebersihan dan Kepekaan akan digunakan untuk fasa seterusnya di dalam kajian.

C. Fasa 3: Analisis Hasil Kajian dan Kesimpulan

Fasa terakhir bagi kajian dalam projek ini adalah fasa 3 yang melibatkan analisis hasil kajian berserta dengan kesimpulan. Tujuan bagi fasa ini adalah bagi melaksanakan analisis bagi prestasi model-model PdM yang dicadangkan pada awal projek ini. Antara aktiviti-aktiviti yang terlibat di dalam fasa ini adalah analisis bagi hasil dari fasa 2 termasuklah konfigurasi set data bersih yang telah melalui proses binarisasi serta penilaian prestasi berdasarkan hasil eksperimen menggunakan model-model yang telah dibina. Aktiviti seterusnya di dalam fasa ini adalah pemilihan model terbaik dari enam model yang telah dibina dan model yang terpilih ini akan digunakan untuk aktiviti seterusnya iaitu perbandingan dengan model-model dari kajian yang lepas berdasarkan bahan rujukan yang dikumpul dari fasa 1 untuk kajian kesusasteraan. Produk yang dijangka akan diperolehi dari fasa ini adalah seperti penerangan tentang set data selepas melalui langkah pra-pemprosesan data, rumusan berdasarkan pelbagai nilai ukuran prestasi bagi keenam-enam model yang telah dibina untuk projek ini dan juga hasil perbandingan bagi model terbaik dari model-model cadangan dengan model-model dari kajian lepas.

IV. HASIL KAJIAN

Hasil kajian dari projek ini telah dirumuskan menggunakan skor bagi setiap ukuran untuk setiap model kajian seperti yang ditunjukkan di dalam Jadual 3.

Jadual 3 Rumusan Hasil Kajian

Model	Kejituan	Kebersihan	Kepekaan	Skor-F1	Label Benar bagi Machine Failure = 0	Label Salah bagi Machine Failure = 0	Label Benar bagi Machine Failure = 1	Label Salah bagi Machine Failure = 1	AUC – ROC	AUC – Kebersihan-Kepekaan
NB [SMOTE]	0.999	1.0	0.968	0.984	2907	0	90	3	0.9794	0.9699
NB + <i>Bagging</i> [SMOTE]	0.999	1.0	0.968	0.984	2907	0	90	3	0.9794	0.9699
NB + <i>Boosting</i> [SMOTE]	0.929	0.3	0.968	0.458	2697	210	90	3	0.9776	0.9693
NB [Persampelan Terkurang]	0.997	0.927	0.968	0.947	2900	7	90	3	0.9810	0.9361
NB + <i>Bagging</i> [Persampelan Terkurang]	0.987	0.714	0.968	0.822	2871	36	90	3	0.9813	0.9707
NB + <i>Boosting</i> [Persampelan Terkurang]	0.997	0.927	0.968	0.947	2900	7	90	3	0.9827	0.9343

Model NB [SMOTE] dan model NB + *Bagging* [SMOTE] mendahului model-model lain bagi Kejituan (0.999), Kebersihan (1.0), Skor-F1 (0.984) dan jumlah instans benar bagi label 0 (2907). Kesemua model-model kajian memperolehi Kepekaan yang sama iaitu 0.968. Di samping itu, kesemua model-model kajian dapat meramal kegagalan mesin dengan baik dengan nilai prestasi yang sama di mana 90 instans kegagalan mesin dapat diramal dengan betul daripada 93 instans kegagalan mesin.

Model NB + *Boosting* [Persampelan Terkurang] memperolehi nilai AUC untuk lengkungan ROC yang paling tinggi iaitu 0.9827 diikuti oleh model NB + *Bagging* [Persampelan Terkurang] yang memperolehi nilai 0.9813. Model NB + *Bagging* [Persampelan Terkurang] memperolehi nilai AUC untuk lengkungan Kebersihan-Kepekaan yang paling tinggi iaitu 0.9707 diikuti oleh model NB [SMOTE] dan model NB + *Bagging* [SMOTE] yang memperolehi nilai 0.9699.

Secara keseluruhan, kesemua model-model kajian menunjukkan prestasi yang baik dalam meramal kegagalan mesin. Namun, model NB [SMOTE] dan model NB + *Bagging* [SMOTE] menunjukkan prestasi yang paling baik disebabkan oleh kemampuan kedua-dua model ini dalam meramal kegagalan mesin dan juga ketidakgagalan mesin dengan baik.

Untuk tujuan perbandingan model terbaik dari projek ini dengan model-model dari penyelidikan lepas, model NB [SMOTE] dipilih sebagai model terbaik untuk perbandingan dengan kajian lepas kerana ia adalah model asas yang tidak memerlukan masa yang lebih untuk pemrosesan data dibandingkan dengan gabungan dengan teknik *Bagging* yang tidak menunjukkan sebarang perubahan prestasi model asal.

Jadual 4 menunjukkan perbandingan antara model terbaik daripada projek ini dengan model-model dari kajian lepas.

Jadual 4 Perbandingan Prestasi Model Cadangan dengan Kajian Lepas

Kajian	Model	Kejituan	Kepersisan	Kepekaan	Skor-F1	AUC – ROC	AUC – Kepersisan-Kepekaan
Model Cadangan Projek	NB + SMOTE	0.999	1.0	0.968	0.984	0.9794	0.9699
(Gaur 2021)	NB	0.785	0.96	0.83	0.88	N/A	N/A
(Ghasemkhani, Aktas & Birant 2023)	<i>Balanced K-Star</i>	0.988	0.988	0.988	0.988	N/A	N/A
(Gujarathi 2021)	NB	0.829	0.836	0.823	0.830	0.901	N/A
(Harichandran, Raphael & Mukherjee 2023)	HUS-ML	0.985	~0.850	~0.750	0.791	N/A	N/A
(Iantovics & Enachescu 2022)	<i>Binary LGR</i>	0.971	N/A	N/A	N/A	N/A	N/A
(Kodihalli 2021)	NB	0.985	0.682	0.968	0.8	N/A	N/A
(Lallahom 2022)	NB	0.843	0.758	0.144	0.242	0.879	N/A
(Nazara 2022)	XGBoost	0.991	N/A	N/A	N/A	0.972	N/A
(Papathanasiou, Demertzis & Tziritas 2023)	<i>Random Survival Forest</i>	0.972	N/A	N/A	N/A	N/A	N/A
(S. K. 2021)	NB	0.789	0.113	0.766	0.196	N/A	N/A
(Sharma et al. 2022)	<i>Random Forest</i>	0.984	N/A	N/A	N/A	0.776	N/A
(Shrimant 2021)	NB	0.762	0.103	0.768	0.182	N/A	N/A

Berdasarkan ukuran nilai Kejituan, model cadangan dari projek ini iaitu model NB berserta SMOTE memperoleh nilai tertinggi iaitu 0.999 diikuti oleh model XGBoost oleh Nazara (2022) dengan nilai 0.991 dan seterusnya model *Balanced K-Star* oleh Ghasemkhani, Aktas dan Birant (2023) dengan nilai 0.988. Bagi ukuran Kepersisan pula, model cadangan dari projek ini masih memperoleh nilai tertinggi iaitu 1.0 diikuti oleh model *Balanced K-Star* oleh Ghasemkhani, Aktas dan Birant (2023) dengan nilai 0.988 dan seterusnya model NB oleh Gaur (2021) dengan nilai 0.96. Model *Balanced K-Star* oleh Ghasemkhani, Aktas dan Birant (2023) memperoleh nilai Kepekaan paling tinggi iaitu 0.988 diikuti oleh model cadangan dari projek ini serta model NB oleh Kodihalli (2021) yang mendapat nilai Kepekaan 0.968 bagi kedua-dua model tersebut. Model *Balanced K-Star* oleh Ghasemkhani, Aktas dan Birant (2023) juga memperoleh nilai Skor-F1 paling tinggi iaitu 0.988 diikuti oleh model cadangan dari projek ini dengan nilai 0.984 dan seterusnya model NB oleh Gaur (2021) yang mendapat nilai 0.88. Bagi ukuran AUC untuk lengkungan ROC pula, model cadangan dari projek ini memperoleh nilai tertinggi iaitu 0.9794 diikuti oleh model XGBoost oleh Nazara (2022) dengan nilai 0.972 dan seterusnya model NB oleh Gujarathi (2021) dengan nilai 0.901. Model cadangan dari projek ini memperoleh 0.9699 bagi ukuran AUC untuk lengkungan Kepersisan-Kepekaan manakala model-model dari kertas-kertas penyelidikan lepas tidak mengandungi nilai bagi ukuran ini.

Melalui Jadual 4, dapat dilihat bahawa prestasi model NB yang telah dibina dalam kajian ini mengatasi prestasi model-model dari kajian yang lepas bagi ukuran Kejituan, Keperisian, AUC untuk lengkungan ROC dan AUC untuk lengkungan Keperisian-Kepekaan dalam meramal kegagalan mesin. Berdasarkan jadual yang sama juga, dapat dilihat bahawa model NB yang dibina di dalam projek ini mengatasi prestasi model yang dikongsi di dalam laman web Kaggle dari segi Kejituan, Keperisian dan Skor-F1.

V. CADANGAN DAN RUMUSAN

Kajian yang telah dilakukan dalam projek ini dapat membantu dalam tugas utama PdM iaitu meramal kegagalan. Model yang dipilih telah berjaya meramal hampir kesemua instans kegagalan mesin di dalam set data sintetik. Walaupun projek ini memfokuskan pada pendekatan pembelajaran mesin menggunakan gabungan teknik binarisasi dan algoritma NB dalam PdM, model-model yang dibangunkan dalam projek ini dapat juga digunakan untuk aplikasi selain PdM.

Terdapat beberapa cadangan yang dapat dikenal pasti untuk kajian masa depan seperti berikut.

1. Kajian ini dapat diperkembangkan lagi dengan penggunaan set data yang sebenar yang diperolehi terus dari mesin. Ini dapat mengesahkan prestasi model-model yang telah dicadangkan di dalam projek ini berdasarkan aplikasi yang sebenar.
2. Kajian ini memfokuskan pada ramalan tunggal atribut *Machine failure*. Kajian ini dapat diluaskan untuk membuat ramalan atribut kegagalan yang lain seperti atribut TWF, HDF, PWF, OSF dan RNF untuk melihat prestasi model dalam meramal kegagalan sistem yang lebih kecil pada mesin.
3. Kajian ini hanya melibatkan penggunaan kaedah binarisasi dengan algoritma NB. Kajian ini dapat diluaskan dengan penggunaan kaedah binarisasi dengan algoritma mesin pembelajaran yang lain. Ini dapat mengesahkan sama ada teknik binarisasi yang digunakan dalam projek ini memberi prestasi lebih baik atau sebaliknya walaupun menggunakan algoritma lain yang pernah digunakan dalam kajian-kajian lepas.
4. Kajian ini hanya melibatkan pembahagian data yang mudah seperti kaedah SMOTE dan Persampelan Terkurang. Penggunaan teknik Pengesahan Bersilang dapat digunakan untuk kajian seterusnya supaya perbandingan dengan kaedah SMOTE dan Persampelan Terkurang dapat diperolehi.
5. Kajian ini cuma menggabungkan model asas NB dengan kaedah *Bagging* dan kaedah *Boosting*. Kajian ini dapat diluaskan lagi melalui kombinasi dengan kaedah *Stacking* supaya perbandingan dengan dua teknik yang digunakan dalam projek ini dapat diperolehi.
6. Projek ini telah menunjukkan bahawa kaedah binarisasi adalah sesuai dengan set data kajian berdasarkan prestasi yang baik yang ditunjukkan oleh model-model yang telah dibina di dalam kajian ini. Namun, secara umumnya, kaedah binarisasi menyebabkan kehilangan maklumat pada sesebut set data. Untuk mengesahkan bahawa tiada kehilangan maklumat yang penting, projek mencadangkan supaya kaedah ini digunakan pada set data dari mesin sama yang berfungsi di dalam persekitaran yang

berlainnan dan perbandingan ramalan kegagalan dilakukan diantara model tersebut dengan model kajian.

CADANGAN DAN RUMUSAN

Penulis berterima kasih kepada Universiti Kebangsaan Malaysia khususnya Fakulti Teknologi dan Sains Maklumat di atas peluang menjalankan projek ini.

RUJUKAN

- Daniyan, Ilesanmi, Khumbulani Mpofo, Moses Oyesola, Boitumelo Ramatsetse & Adefemi Adeodu. 2020. "Artificial Intelligence for Predictive Maintenance in the Railcar Learning Factories." *Procedia Manufacturing* (45): 13-18.
- Gaur, Durgance. 2021. *Data Imbalance+EDA+87% AUC*. Accessed October 14, 2023. <https://www.kaggle.com/code/durgancegaur/data-imbalance-eda-87-auc>.
- Ghasemkhani, Bitan, Ozlem Aktas & Derya Birant. 2023. "Balanced K-Star: An Explainable Machine Learning Method for Internet-of-Things-Enabled Predictive Maintenance in Manufacturing." *Machines* 11 (322): 1-20.
- Gonfalonieri, Alexandre. 2019. *Towards Data Science*. Accessed June 9, 2023. <https://towardsdatascience.com/how-to-implement-machine-learning-for-predictive-maintenance-4633cdbe4860>.
- Gujarathi, Rudra. 2021. *Machine Predictive Maintenance Classification*. Accessed October 14, 2023. <https://www.kaggle.com/code/rudragujarathi/machine-predictive-maintenance-classification>.
- Harichandran, Aparna, Benny Raphael & Abhijit Mukherjee. 2023. "Equipment Activity Recognition and Early Fault Detection in Automated Construction through a Hybrid Machine Learning Framework." *Computer-Aided Civil and Infrastructure Engineering* (38): 253-268.
- Herrero, Ricardo D & Marta Zorrilla. 2022. "An I4.0 Data Intensive Platform Suitable for the Deployment of Machine Learning Models: a Predictive Maintenance Service Case Study." *Procedia Computer Science* (200): 1014-1023.
- Iantovics, Laszlo Barna & Calin Enachescu. 2022. "Method for Data Quality Assessment of Synthetic Industrial Data." *Sensors* 22 (1608): 1-21.
- Kodihalli, Sarayu Pramod. 2021. *9 Classification models-UpSampled-F1-97%*. Accessed October 14, 2023. <https://www.kaggle.com/code/sarayukodihalli/9-classification-models-upsampled-f1-97>.
- Lallahom, Omar Ben. 2022. *Resampled - AUC: 991*. Accessed October 14, 2023. <https://www.kaggle.com/code/omarbenlallahom/resampled-auc-991>.

- Lee, Juseong, & Mihaela Mitici. 2023. "Deep Reinforcement Learning for Predictive Aircraft Maintenance using Probabilistic Remaining-Useful-Life Prognostics." *Reliability Engineering and System Safety* (230): 1-14.
- Lee, Wo J, Haiyue Wu, Huitaek Yun, Hanjun Kim, Martin B. G. Jun & John W. Sutherland. 2019. "A Novel Predictive Selective Maintenance Tool Systems using Artificial Intelligence Techniques applied to Machine Condition Data." *Procedia CIRP* (80): 506-511.
- Masero, Eva, Sara Ruiz-Moreno, Jose R D Frejo, Jose M. Maestre & Eduardo F Camacho. 2023. "A Fast Implementation of Coalitional Model Predictive Controllers based on Machine Learning: Application to Solar Power Plants." *Engineering Applications of Artificial Intelligence* (118): 1-10.
- Matzka, Stephan. 2020. "Explainable Artificial Intelligence for Predictive Maintenance Applications." IEEE.
- Nazara, Krisman Yusuf. 2022. "Perancangan Smart Predictive Maintenance untuk Mesin Produksi." Jakarta Timur: Tim Publikasi & TIK.
- Nikfar, Mohsen, Julia Bitencourt & Konstantinos Mykoniatis. 2022. "A Two-Phase Machine Learning Approach for Predictive Maintenance of Low Voltage Industrial Motors." *Procedia Computer Science* (200): 111-120.
- Papathanasiou, Dimitris, Konstantinos Demertzis & Nikos Tziritas. 2023. "Machine Failure Prediction Using Survival Analysis." *Future Internet* 15 (153): 1-26.
- Rodriguez, Marcelo L R, Sylvain Kubler, Andrea de Giorgio, Maxime Cordy, Jeremy Robert & Yves Le Traon. 2022. "Multi-agent Deep Reinforcement Learning based Predictive Maintenance on Parallel Machines." *Robotics and Computer-Integrated Manufacturing* (78): 1-12.
- S. K., Vijayaragavan. 2021. *Feature selection, Hyperparameter Tuning*. Accessed October 14, 2023. <https://www.kaggle.com/code/vijayaragavansk/feature-selection-hyperparameter-tuning>.
- Safoklov, Boris, Denis Prokopenko, Yury Deniskin & Mikhail Kostyshak. 2022. "Model of Aircraft Maintenance Repair and Overhaul Using Artificial Neural Networks." *Transportation Research Procedia* (63): 1534-1543.
- Sanzana, Mirza R, Tomas Maul, Jing Y Wong, Mostafa O M Abdulrazie & Chun-Chieh Yip. 2022. "Application of Deep Learning in Facility Management and Maintenance for Heating, Ventilation, and Air Conditioning." *Automation in Construction* (141): 1-13.
- Sarvaiya, Dhaval. 2021. *AIM*. Accessed June 9, 2023. <https://analyticsindiamag.com/machine-learning-for-predictive-maintenance-key-approaches-techniques-to-consider/>.
- Sharma, Neelam, Tejasv Singh Sidana, Saransh Singhal & Sachin Jindal. 2022. "Predictive Maintenance: Comparative Study of Machine Learning Algorithms for Fault Diagnosis." Delhi: Elsevier.

- Shrimant, Shubham. 2021. *Naive Bayes*. Accessed October 14, 2023. <https://www.kaggle.com/code/shubhamshrimant/naive-bayes>.
- Tessoni, Valentina & Michele Amoretti. 2022. "Advanced Statistical and Machine Learning Methods for Multi-step Multivariate Time Series Forecasting in Predictive Maintenance." *Procedia Computer Science* (200): 748-757.
- Theissler, Andreas, Judith Perez-Velazquez, Marcel Kettelgerdes & Gordon Elger. 2021. "Predictive Maintenance Enabled by Machine Learning: Use Cases and Challenges in the Automotive Industry." *Reliability Engineering and System Safety* (215): 1-21.
- Vannucci, Marco, Valentina Colla, Matteo Chini, Daniele Gaspardo & Birgit Palm. 2022. "Artificial Intelligence Approaches For The Ladle Predictive Maintenance In Electric Steel Plant." *IFAC PapersOnLine II* (55): 331-336.

Copyright@FTSM
UKM