MODEL PENYENGGARAAN JANGKAAN BERDASARKAN PENDEKATAN BINARISASI DAN *NAIVE BAYES*

MUHAMMAD SAIFULLAH BIN HJ MUHAMAD JUHARI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

04 Februari 2024

MUHAMMAD SAIFULLAH
BIN HJ MUHAMAD
JUHARI
P118344

PENGHARGAAN

Pertama sekali, penulis ingin mengucapkan syukur ke hadrat Ilahi kerana mengurniakan penulis dengan peluang, kesihatan, dan kemampuan untuk menyambung pengajian dalam program Sarjana Sains Data di Universiti Kebangsaan Malaysia. Penulis juga akur bahawa tanpa izin dan pertolongan daripada Allah, sudah tentu penulis tidak dapat menyempurnakan tesis ini.

Penulis juga ingin mengucapkan terima kasih terutama sekali kepada penyelia utama penulis, Dr. Syaimak Abdul Shukor yang terus menerus membimbing penulis di dalam menyiapkan laporan tesis ini serta semua pensyarah-pensyarah penulis di atas usaha keras dan kesabaran yang mereka tunjukkan dalam mengajar penulis untuk subjek-subjek yang penulis telah ambil sepanjang penulis berada di dalam program ini. Oleh kerana dedikasi yang ditunjukkan oleh penyelia dan pensyarah-pensyarah penulis, pengetahuan, dan kemahiran penulis dalam bidang Sains Data telah meningkat dengan ketara. Penulis berharap dengan minat penulis yang tinggi di dalam bidang Sains Data, penulis dapat menyelami dengan lebih dalam lagi dalam bidang ini dan dapat memberi sumbangan kepada kerjaya penulis.

Akhir sekali, penulis ingin merakamkan ucapan terima kasih yang tidak terhingga kepada keluarga dan rakan-rakan penulis atas dorongan dan kata-kata semangat yang mereka berikan kepada penulis sepanjang penulis berada di dalam program Sarjana Sains Data ini. Penulis bersyukur dikurniakan keluarga dan rakan-rakan yang sentiasa memberikan sokongan kepada penulis agar sentiasa meluangkan masa yang berlebihan untuk mendalami ilmu walaupun sudah berada di alam pekerjaan agar penulis dapat memajukan diri penulis di dalam dunia yang berkembang pesat setiap saat. Penulis sangat menghargai jasa-jasa yang diberikan oleh mereka dan penulis berdoa agar kalian sentiasa dikurniakan umur yang diberkati, kesihatan yang baik serta kebahagiaan di dunia dan akhirat.

ABSTRAK

Pendekatan pembelajaran mesin dalam aplikasi penyenggaraan jangkaan menjadi semakin luas sesuai dengan peredaraan 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 penyelenggaran yang kurang efisyen serta melibatkan kos yang tinggi. Projek ini bertujuan untuk memperluaskan lagi penyelidikan berkaitan dengan pendekatan pembelajaran mesin dalam penyenggaraan jangkaan dengan menggunakan set data sintentik 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] memperolehi nilai Kejituan 0.999 dan Kepersisan 1.0 yang mengatasi hasil kajian yang lepas.

PREDICTIVE MAINTENANCE MODEL BASED ON BINARIZATION AND NAIVE BAYES APPROACH

ABSTRACT

The machine learning approach in the application of Predictive Maintenance has been expanding as time progresses. Various research has been conducted on this topic, and many industries have already implemented machine learning approaches in their Predictive Maintenance applications. However, its use is still not widespread in some industries, and these industries continue to practice less efficient and costly forecasting methods. This project aims to further expand research on machine learning approaches in Predictive Maintenance by using synthetic data sets published for public use. The study conducted in this project focuses on the combination of binarization technique used on the research dataset with the implementation of Naive Bayes algorithm during data minin process, which was not used in previous research papers, as well as the combination of Naive Bayes with Bagging and Boosting techniques. The SMOTE and Undersampling methods were chosen to address the imbalance effects in the data attributes that predict machine failures. Six models consisted of Naive Bayes [SMOTE] model, Naive Bayes with Bagging [SMOTE] model, Naive Bayes with Boosting [SMOTE] model, Naive Bayes [Undersampling] model, Naive Bayes with Bagging [Undersampling] model and Naive Bayes with Boosting [Undersampling] model were developed in this project. The best model which is the Naive Bayes [SMOTE] model achieved an Accuracy of 0.999 and Precision of 1.0, surpassing the results of previous studies.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

ANN	Artificial Neural Network
AUC	Area under the Curve
BNN	Backpropagation Neural Network
CM	Corrective Maintenance
CML	Conventional Machine Learning
CMMS	Computerized Maintenance Management System
CNN	Convolutional Neural Network
ctGAN	Conditional Tabular Generative Adversarial Network
DL	Deep Learning
DT	Decision Tree
ECE	Expected Calibration Error
EFNC	Evolving Fuzzy Neural Classifier
EFNC-Exp	Evolving Fuzzy Neural Classifier with expert rules
ERNN	Elman RNN
FCM	Fuzzy Cognitive Map
FDP	Failure Developing Period
GRU	Gated Recurrent Unit
GS	Global tree Surrogate
HUS-ML	Hybrid Unsupervised and Supervised Machine Learning
k-NN	k-Nearest Neighbour
LGR	Logistic Regression
L-K IF	Liang-Kleeman Information Flow
LR	Linear Regression
LS	Local tree Surrogate
LSTM	Long Short Term Memory

MLP	Multi-layer Perceptron
MRO	Maintenance Repair and Overhaul
NB	Naive Bayes
NMPC	Non-linear Predictive Control
PCA	Principal Component Analysis
PdM	Predictive Maintenance
PM	Predictive Maintenance
RAMS	Reliability, Availability, Maintainability and Safety
RF	Random Forest
RL	Reinforced Learning
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RUL	Remaining Useful Life
SMOTE	Synthetic Minority Oversampling Technique
SODA	Self-Organized Direction Aware data partitioning algorithm
SVM	Support Vector Machine
UKM	Universiti Kebangsaan Malaysia
XGBoost	eXtreme Gradient Boosting

BAB I

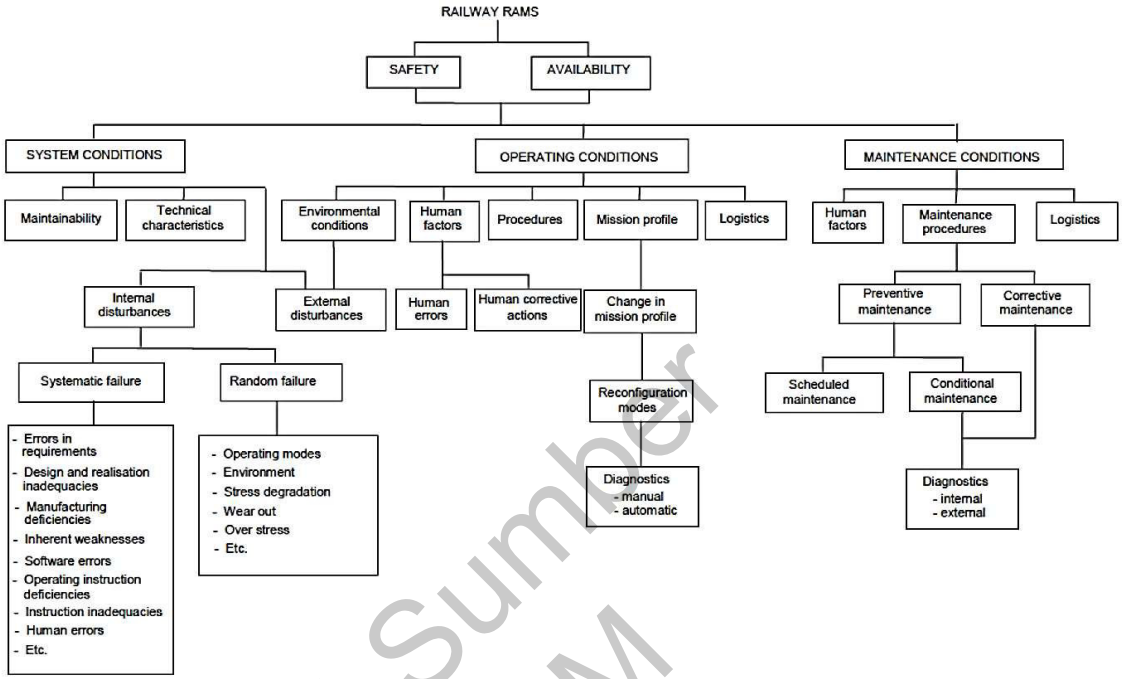
PENGENALAN

1.1 PENDAHULUAN

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 penyenggaraan telah menjadi salah satu syarat kelulusan yang membolehkan sesebuah projek untuk beroperasi, dan piawaian seperti EN50126 mesti dipatuhi untuk memastikan sistem yang akan digunakan untuk orang awam mematuhi garis panduan *Reliability, Availability, Maintainability and Safety* (RAMS) seperti yang ditunjukkan melalui carta alir ringkasan RAMS pada Rajah 1.1.

RAMS merangkumi proses-proses yang memastikan setiap fasa di dalam pembangunan sesebuah produk melalui analisis bagi mengenal pasti potensi kegagalan serta mitigasi untuk menghapuskan atau mengurangkan risiko kegagalan fungsi ke tahap yang boleh diterima mengikut proses yang telah diterangkan di dalam piawaian tersebut. Semasa di fasa operasi dan penyelenggaran, data-data yang diperolehi dari mesin digunakan untuk mengukur sama ada prestasi mesin masih berada di dalam sasaran RAMS yang telah ditentukan dari awal. Pengukuran sasaran RAMS ini membantu kumpulan pengendali untuk mengenal pasti tindakan yang perlu dibuat berdasarkan senarai mitigasi yang telah ditetapkan semasa analisis RAMS ketika fasa

pembangunan produk tersebut. Antara mitigasi-mitigasi yang biasa digunakan adalah penyenggaraan tradisional terdiri daripada dua kaedah yang disebut sebagai *Preventive Maintenance* (PM) dan *Corrective Maintenance* (CM).



Rajah 1.1 Konsep RAMS dalam Sektor Rel

Sumber: (European Committee for Electrotechnical Standardization 2017)

Penyenggaraan PM adalah sejenis penyenggaraan yang dirancang di mana masa untuk melakukan penyenggaraan serta tempoh di antara penyenggaraan semasa dan penyenggaraan 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. Setelah tempoh FDP dikenal pasti, jadual penyenggaraan PM dapat dihasilkan berdasarkan anggaran frekuensi penyenggaraan yang perlu dilakukan sebelum kegagalan mesin berlaku. Contoh penyenggaraan PM dapat dilihat dari servis kereta yang perlu lakukan setiap tiga atau enam bulan berdasarkan buku panduan pengguna yang ditentukan oleh jurutera pereka kenderaan tersebut menggunakan tempoh FDP yang diperolehi dari komponen-komponen kereta itu. Penyenggaraan CM pula adalah jenis penyenggaraan yang tidak dirancang di mana ia dilakukan setelah berlakunya kegagalan pada mana-mana bahagian mesin. Kegagalan mesin yang memerlukan penyenggaraan CM boleh berlaku

pada bila-bila masa dan ia dapat terjadi secara rawak. Mesin yang diselenggara mengikut jadual PM masih perlu menjalani penyenggaraan CM jika kegagalan mesin berlaku di dalam tempoh diantara penyenggaraan PM terkini dan penyenggaraan PM seterusnya. Contoh penyenggaraan CM dapat dilihat dari pembaikan kereta yang perlu dilakukan ketika terdapat kegagalan pada sebarang komponen dalam kereta tersebut seperti kegagalan enjin kereta yang berlaku walaupun masih berada di dalam tempoh penyenggaraan PM.

Sesebuah mesin perlu melalui fasa rekaan, produksi, pembinaan, pengujian, operasi dan penyenggaraan. Fasa operasi dan penyenggaraan biasanya merupakan fasa terpanjang dalam kitaran hidup mesin. Oleh kerana fasa tersebut mengambil tempoh masa yang panjang, banyak firma perniagaan telah sedar mengenai peluang perniagaan yang dapat diperolehi dari aktiviti operasi dan penyenggaraan dengan menambahkan pelaburan di fasa ini untuk meningkatkan margin keuntungan daripada penjualan alat ganti dan servis perkhidmatan penyenggaraan kepada pelanggan. Di samping itu, pada akhir-akhir ini, terdapat kecenderungan di mana pelanggan akan mengeluarkan kontrak bagi skop kerja operasi dan penyenggaraan kepada syarikat swasta untuk mengatasi kesukaran yang dihadapi oleh pelanggan dalam mengurus aktiviti operasi dan penyenggaraan. Bagi sesetengah projek seperti untuk industri rel, ia sudah mula menjadi amalan yang biasa bagi pemilik projek untuk memasukkan bajet bagi kontrak operasi dan penyenggaraan ke dalam keseluruhan kos sesebuah projek sebelum projek tersebut disenaraikan untuk pembidaan. Oleh kerana wujud peluang perniagaan yang luas dalam fasa operasi dan penyenggaraan, bakal pembekal-pembekal komponen untuk mesin telah mula melihat kepada kaedah-kaedah yang boleh diaplikasikan bagi mengurangkan kos pelaksanaan dan dalam masa yang sama dapat mengoptimakan skop kerja bagi menambah margin keuntungan syarikat.

Salah satu cara yang telah mula digunakan untuk menyelesaikan masalah-masalah tersebut adalah dengan menggunakan *Predictive Maintenance* (PdM). Penyenggaraan PdM menggunakan data secara langsung yang dikumpul daripada penderia-penderia yang dipasang pada mesin dan data yang diambil akan menunjukkan keadaan mesin melalui pelbagai nilai ukuran seperti suhu, getaran dan bunyi. Data-data ini akan digunakan dalam menganalisis kebarangkalian bagi kegagalan mesin untuk menentukan waktu yang tepat bagi melakukan penyenggaraan. Secara kebiasaannya,

penyenggaraan 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 penyenggaraan 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 penyenggaraan perlu dilakukan. Penyenggaraan PdM menekankan penyelenggaran pada masa yang betul berbanding penyenggaraan ketika mesin masih mempunyai tempoh yang masih panjang dari masa kegagalan mesin seperti yang sering berlaku dalam penyenggaraan PM atau apabila mesin telah gagal seperti yang sering berlaku di dalam penyenggaraan CM. Hal ini menunjukkan bahawa penggunaan penyenggaraan PdM adalah lebih baik jika dibandingkan dengan teknik penyenggaraan PM dan CM kerana ia dapat mengurangkan kos disebabkan oleh pengurangan kekerapan penyenggaraan dan ini membolehkan pemeliharaan jangka hayat mesin yang lebih panjang.

Pendekatan pembelajaran mesin merupakan salah satu kaedah yang digunakan di dalam penyenggaraan PdM disebabkan oleh ketersediaan pelbagai data dari penderia-penderia yang terdapat pada mesin. Data yang dikumpul dari penderia-penderia pada mesin untuk kegunaan penyenggaraan 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 algoritm-algoritma dengan pendekatan klasifikasi. Kebiasaannya, penyenggaraan 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, penyenggaraan PdM menjadi lebih baik jika dibandingkan dengan penyenggaraan PdM menggunakan kaedah tradisional kerana pendekatan pembelajaran mesin dapat membantu PdM meramal kegagalan dalam kes kegagalan mesin biasa dan juga kegagalan mesin rawak.

Terdapat pelbagai kajian berkaitan pendekatan pembelajaran mesin dalam penyenggaraan PdM yang telah dikeluarkan oleh penyelidik-penyelidik untuk digunakan sebagai rujukan. Namun, set data yang sesuai untuk penyelidikan

pendekatan pembelajaran mesin dalam penyenggaraan 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 bagi menjalankan eksperimen yang berkaitan dengan objektif-objektif projek. Set data buatan tersebut telah dihasilkan berdasarkan set data yang sebenar supaya kajian yang dilakukan dapat mewakili situasi yang nyata. 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 penyenggaraan PdM untuk meramal kegagalan mesin.

1.2 PERMASALAHAN KAJIAN

Pelbagai industri masih menggunakan kaedah PM dan CM dalam proses penyenggaraan. 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 penyenggaraan tradisional adalah kurang cekap kerana jangka masa penyenggaraan berkala ditetapkan berdasarkan ramalan ketika fasa kejuruteraan dan biasanya ia menyebabkan penyelenggaran dilaksanakan jauh lebih awal dari masa kegagalan mesin berlaku. Secara realitinya, setiap mesin berfungsi secara unik bergantung kepada faktor dalaman dan luaran yang berbeza-beza seperti lokasi, cara operasi, masa penggantian komponen, dan banyak lagi. Oleh itu, jika penyelenggaran mesin menggunakan kaedah tradisional, ia akan menyebabkan penyenggaraan dilaksanakan terlalu awal, pembaziran sumber tenaga kerja dan bahan, serta peningkatan jangka hayat yang tidak produktif bagi mesin. Salah satu contoh dapat dilihat daripada penyenggaraan pesawat di mana 29.3% daripada kos penyenggaraan telah dibazirkan dengan penggunaan kaedah tradisional di dalam penyenggaraan pesawat (Lee & Mitici 2023).

Proses penyenggaraan dapat dioptimumkan dengan menggunakan data yang dihasilkan daripada penderia-penderia yang dipasang pada mesin untuk memantau status mesin. Penyenggaraan PdM merupakan salah satu cara penyelesaian yang boleh digunakan untuk mengatasi isu-isu yang terhasil daripada penyenggaraan menggunakan

kaedah tradisional seperti yang disebutkan sebelum ini. Berdasarkan penyelidikan sebelum ini, pemadanan 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). Namun, perlu diketahui bahawa unsur yang paling penting dalam penyelidikan mengenai penyenggaraan PdM adalah penggunaan set data yang betul.

Secara umumnya, set data yang melibatkan penyenggaraan 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 penyenggaraan 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 kesusateraan 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 pendekataan 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 pengunaan NB ke atas set data tersebut. Namun, jumlah forum yang membincangkan tentang penggunaan NB ke atas set data AI4I2020 adalah sedikit. Selain itu, 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. Ini menunjukkan bahawa kajian bagi pendekatan binarisasi dan NB dalam pembangunan

model PdM menggunakan set data AI4I2020 untuk meramal kegagalan masih belum wujud dan ini memberikan peluang kepada projek ini untuk memulakan kajian tersebut.

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. Namun, seperti yang dibincangkan sebelum ini, kaedah NB tidak digunakan untuk kertas penyelidikan rasmi dan hanya beberapa forum sahaja di dalam laman web *Kaggle* yang telah berkongsi kod yang menggunakan teknik NB. 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 menghasil keputusan seperti berikut: kejituan 98.75%, kepersisan 0.9877, kepekaan 0.9875 dan skor-F1 0.9875 (Ghasemkhani, Aktas & Birant 2023).

1.3 SOALAN KAJIAN

Projek ini mempunyai dua soalan kajian seperti yang disenaraikan berikut:

1. Soalan kajian 1: Adakah model PdM dengan pendekatan binarisasi dan NB sesuai digunakan ke atas set data kajian?

2. Soalan kajian 2: Bagaimanakah prestasi model yang dibangunkan di dalam kajian ini serta perbandingannya dengan kajian lepas?

1.4 OBJEKTIF KAJIAN

Projek ini mempunyai dua objektif utama bagi menjawab soalan-soalan kajian yang telah disenaraikan di dalam bab 1.3. Objektif projek adalah seperti berikut:

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

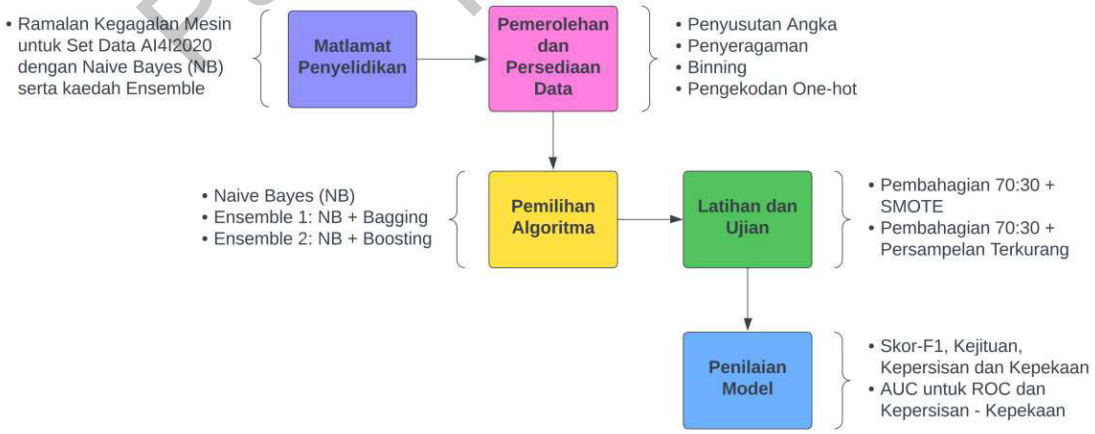
1.5 SKOP KAJIAN

Projek ini menggunakan set data penyenggaraan pada mesin yang telah digunakan dalam kertas kajian "*Explainable Artificial Intelligence for Predictive Maintenance Applications*" yang telah diterbitkan untuk rujukan orang awam bagi pembangunan penyelidikan lanjut (Matzka 2020). Projek ini akan memberikan fokus ke atas pembangunan model berdasarkan pendekatan binarisasi dan menggunakan algoritma NB. Beberapa model lain yang berasaskan teknik binarisasi dan algoritma NB juga akan dibina supaya hasil kajian dapat disahkan dan juga supaya perbandingan antara model dengan binaan yang berbeza dapat dibuat. Selain itu, beberapa teknik yang dapat mengawal prestasi model dari dipengaruhi oleh ciri-ciri set data akan diperkenalkan dan perbandingan antara teknik-teknik ini akan dijalankan. Perbandingan model akan dilakukan berdasarkan jenis ukuran prestasi model yang akan dipilih mengikut kesesuaian ukuran terbabit dengan set data serta fungsi model yang dibina. Sebagai contoh, nilai-nilai untuk *Confusion Matrix* perlu diplotkan untuk model ramalan kegagalan mesin jika ketidakseimbangan dalam set data adalah terlalu ketara. Semua langkah pemprosesan data dalam projek ini akan dilakukan menggunakan *Python*.

1.6 METODOLOGI

Projek ini akan melaksanakan pendekatan yang terdiri daripada lima peringkat berdasarkan proses sains data untuk mencapai objektifnya seperti yang ditunjukkan dalam Rajah 1.2. Pada peringkat pertama, tumpuan akan diberikan untuk mencari matlamat penyelidikan di mana bagi projek ini, matlamat penyelidikan adalah membangunkan model PdM bagi meramal kegagalan mesin. Selepas itu, pada peringkat

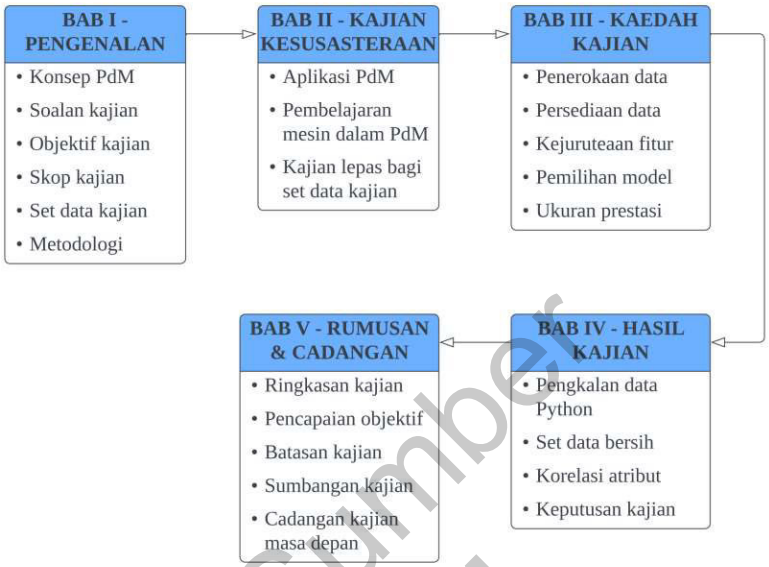
kedua, pemerolehan data akan dilakukan di mana set data yang relevan perlu diperoleh untuk kajian dalam projek ini dan untuk projek ini, set data AI4I2020 dipilih sebagai set data untuk kajian projek. Pra pemprosesan data juga akan dilakukan dalam peringkat ini termasuklah proses binarisasi sebelum set data tersebut digunakan dalam fasa-fasa seterusnya. Peringkat seterusnya adalah pemilihan model di mana algoritma pembelajaran mesin yang jarang digunakan untuk set data AI4I2020 akan dipilih untuk analisis PdM. Kaedah NB akan dipilih sebagai algoritma yang akan digunakan untuk menganalisis data. Seterusnya, kaedah NB akan digabungkan dengan teknik Bagging dan teknik Boosting secara berasingan. Peringkat keempat terdiri daripada latihan dan ujian set data menggunakan model-model yang dipilih. Di dalam peringkat ini, dua kaedah pembahagian data akan digunakan iaitu pembahagian 70:30 dengan kaedah SMOTE serta pembahagian 70:30 dengan kaedah Persampelan Terkurang. Peringkat ini akan menghasilkan keputusan berdasarkan pengukuran yang sesuai yang telah dipilih untuk set data tersebut. Peringkat terakhir terdiri daripada perbandingan prestasi antara model-model yang berbeza, analisis keputusan secara terperinci, dan penilaian model terbaik yang akan menentukan kaedah yang dapat memberikan keputusan yang paling ideal apabila dilaksanakan dalam proses PdM yang menggunakan set data yang serupa.



Rajah 1.2 Gambaran Keseluruhan Carta Alir Metodologi Kajian

1.7 ORGANISASI BAB

Laporan projek ini terdiri daripada lima bab yang utama. Ringkasan bagi setiap bab dalam laporan projek ini dapat dirumuskan seperti di dalam Rajah 1.3 berikut.



Rajah 1.3 Ringkasan Bab Laporan Projek

BAB I: Bab ini menerangkan tentang konsep PdM dan kaitannya dengan pendekatan pembelajaran mesin. Selain itu, dua soalan kajian berserta dua objektif kajian juga disenaraikan di dalam bab ini. Bab ini juga membincangkan tentang skop kajian di mana model yang akan dibangunkan, set data yang akan digunakan serta penilaian model yang akan dijalankan akan diterangkan. Bab ini juga mengandungi metodologi untuk projek ini di mana langkah-langkah untuk menjalankan kajian akan diterangkan secara ringkas. Bab ini akan diakhiri dengan kesimpulan yang akan meringkaskan perkara-perkara yang telah dibincangkan.

BAB II: Bab ini mengandungi kajian kesusateraan mengenai aplikasi PdM secara umum. Bab ini juga akan membincangkan tentang kajian-kajian lepas yang berkaitan dengan pendekatan pembelajaran mesin dalam PdM di dalam pelbagai sektor. Di samping itu, bab ini juga mengandungi kajian kesusateraan mengenai penyelidikan lepas yang berkaitan dengan set data yang akan digunakan di dalam projek ini. Bab ini diakhiri dengan ringkasan yang mengandungi jadual yang menyenaraikan kesemua

ringkasan bagi kesemua kertas kajian yang digunakan dalam kajian kesusasteraan untuk projek ini.

BAB III: Bab ini membincangkan secara teliti tentang kaedah kajian yang digunakan dalam projek ini. Bab ini menerangkan tentang penerokaan data di mana ciri-ciri data akan dikaji secara mendalam. Di samping itu, bab ini juga membincangkan tentang langkah-langkah yang perlu dilalui untuk persediaan data serta senarai pengkalan data *Python* yang digunakan. Kejuruteraan fitur dan aplikasinya dalam projek ini juga akan diterangkan di dalam bab ini. Bab ini juga mengandungi penerangan tentang model-model yang akan dibina serta kaedah-kaedah persediaan data latihan dan ujian yang akan digunakan. Bab ini diakhiri dengan perbincangan mengenai jenis-jenis ukuran yang akan digunakan bagi menilai prestasi model-model yang telah dibina.

BAB IV: Bab ini menerangkan tentang hasil kajian dari projek ini. Bab ini juga akan menunjukkan struktur set data bersih yang terakhir yang digunakan dalam pembelajaran mesin. Kandungan utama dalam bab ini adalah keputusan-keputusan kajian bagi enam model yang dicadangkan dalam projek ini berdasarkan jenis-jenis ukuran yang ditetapkan supaya bersesuaian dengan set data dan matlamat kajian berserta rajah-rajah yang memberi visualisasi hasil kajian. Bab ini akan diakhiri dengan analisis perbandingan prestasi diantara model terbaik dari kajian ini dengan model-model dari kajian lepas.

BAB V: Bab ini merumuskan kajian yang telah dijalankan di dalam projek ini. Ia dimulakan dengan ringkasan kajian yang merangkumi kaedah-kaedah yang telah digunakan serta pencapaian objektif berdasarkan keputusan yang diperoleh melalui hasil kajian terhadap model-model yang berlainan. Di samping itu, bab ini juga membincangkan tentang batasan kajian yang dihadapi di dalam projek ini. Bab ini juga akan menerangkan tentang sumbangan kajian yang dapat dikenal pasti serta cadangan untuk kajian masa depan.

1.8 KESIMPULAN

Bab 1 ini menerangkan secara umum konsep PM, CM dan PdM serta kaitan antara pendekatan pembelajaran mesin dengan aplikasi PdM. Kekurangan set data bersesuaian untuk penyelidikan PdM menjadi salah satu halangan yang umum bagi kajian dalam bidang ini. Walaupun set data untuk kegunaan umum sangat terhad, bab ini telah menerangkan tentang set data yang sepadan dengan keperluan kajian. Dengan membuat kajian terhadap penyelidikan yang lepas, teknik yang belum diguna pakai dapat dikenal pasti di dalam bab ini. Berdasarkan tujuan kajian, set data yang diperolehi serta kaedah yang akan digunakan, soalan-soalan kajian serta objektif kajian dapat disenaraikan di dalam bab ini. Had penyelidikan dalam projek ini ditentukan dengan menyatakan skop kajian. Bab ini juga telah menerangkan tentang metodologi kajian yang perlu diikut sepanjang kajian supaya hasil kajian tidak tersasar jauh dari menjawab soalan-soalan kajian dan mencapai objektif projek ini.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Bab 2 akan membincangkan mengenai kajian kesusateraan ke atas kajian-kajian yang berkaitan dengan PdM, pendekatan pembelajaran mesin dalam PdM serta teknik-teknik yang telah digunakan untuk set data yang digunakan dalam projek ini. Bab 2 terbahagi kepada empat sub topik. Bab 2.2 akan merumuskan tentang teknik-teknik tradisional yang sering digunakan dalam aplikasi PdM. Bab 2.3 akan menerangkan tentang teknik-teknik hibrid yang kerap digunakan untuk aplikasi PdM. Bab 2.4 pula akan menyimpulkan tentang kajian-kajian yang telah dijalankan untuk set data yang akan digunakan dalam kajian ini berdasarkan kertas-kertas kajian yang telah diterbitkan oleh penyelidik-penyelidik lain. Bab ini juga akan menerangkan secara dalam jenis-jenis kaedah pembelajaran mesin yang sering digunakan untuk set data yang serupa. Bab 2 akan diakhiri dengan bab 2.5 yang akan membuat ringkasan terhadap kajian-kajian yang lepas yang dibincangkan dalam bab ini.

2.2 PENYENGGARAAN JANGKAAN (PDM)

PdM adalah salah satu kaedah penyenggaraan 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 mengunakan kaedah statistik merupakan tenik 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 penyenggaraan. 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 penyenggaraan meramal (Theissler et al. 2021). Di samping itu, berdasarkan laporan ulasan mengenai kertas-kertas yang berkaitan dengan pelaksanaan pembelajaran mesin dalam penyenggaraan kaedah DL masih belum diaplikasikan dalam PdM sepenuhnya (Sanzana et al. 2022).

2.3 PENDEKATAN PEMBELAJARAN MESIN DALAM PDM

Terdapat banyak teknik pembelajaran mesin yang sesuai untuk dilaksanakan dalam PdM. Dua kaedah yang biasa digunakan adalah kaedah Ramalan khususnya teknik *Linear Regression* (LR) yang digunakan untuk meramal *Remaining Useful Life* (RUL) mesin dan kaedah Klasifikasi yang digunakan untuk mengklasifikasikan kerosakan mesin kepada pelbagai jenis kegagalan kepada beberapa kelas (Sarvaiya 2021; Gonfalonieri 2019). Selain itu, terdapat juga kertas-kertas kajian yang mengkaji pelaksanaan *Reinforced Leaning* (RL) seperti penyelidikan oleh Rodriguez et al. (2022) yang mengaplikasikan RL yang terdiri dari pelbagai ejen untuk analisis PdM dan Rodriguez et al. (2022) mendapati bahawa teknik ini berjaya memperolehi peningkatan prestasi sebanyak 75% serta berjaya mengatasi teknik CM dan PM (Rodriguez et al. 2022). Lee dan Mitici (2023) juga mengaplikasikan teknik RL dalam analisa PdM untuk pesawat di mana ramalan dilakukan berdasarkan RUL dan kajian inin telah berjaya

membuktikan bahawa kos penyelenggaran pesawat dapat dikurangkan sebanyak 29.3%, 95.6% dari penyenggaraan yang tidak diperlukan dapat dielakkan dan kerugian jangka hayat untuk enjin peswat dapat dikawal supaya berada sekitar 12.81 putaran sahajad (Lee & Mitici 2023). Tambahan pula, terdapat juga kajian-kajian mengenai teknik yang lebih maju seperti *Naive, Varma, Theta, Long Short-term Memory* (LSTM), *Gated Recurrent Unit* (GRU) dan *Elman Recurrent Neural Network* (ERNN) (Tessoni & Amoretti 2022) serta teknik PredMax yang menggabungkan teknik *Deep Convolutional Autoencoder* dengan teknik *Principal Component Analysis* (PCA) (Hajgato et al. 2022) yang mana kedua-dua kajian tersebut menggunakan kombinasi kaedah pembelajaran mesin dengan teknik statistik atau kombinasi lebih daripada satu kaedah pembelajaran mesin.

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 penyelenggaran 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 galas pada roda berdasarkan suhu galas. Kajian ini berjaya menunjukkan bahawa terdapat hubungan yang kuat antara RUL dan suhu galas (Daniyan et al. 2020). Terdapat juga satu lagi kertas kajian mengenai pelaksanaan ANN untuk PdM dalam industri rel yang menggunakan data dari galas 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 penyenggaraan 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. Model hibrid telah menjadi pendekatan yang semakin popular digunakan dalam PdM. 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 eletrik, 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). Model hibrid juga dapat mengenali data tanpa pengenalan terlebih dahulu dan ini telah dibuktikan menggunakan kaedah PredMaX yang menggabungkan *autoencoder Convolutional Neural Network* (CNN) pada langkah pertama dan kemudian menerapkan PCA di mana model tersebut berjaya menentukan masa degradasi tanpa melalui pembelajaran data sebelum itu (Hajgato et al. 2022). Pelaksanaan model hibrid dapat menjimatkan kos penyenggaraan, mengelakkan penyenggaraan yang tidak diperlukan, dan mengurangkan tempoh tidak produktif mesin seperti yang dibuktikan oleh satu kajian tentang penyenggaraan pesawat menggunakan CNN untuk menganggar RUL dan diikuti oleh RL untuk meramal masa kerja penyenggaraan. Hasil yang diperoleh dalam

kertas ini adalah pencegahan 95.6% penyenggaraan yang berlebihan dan pengurangan hayat mesin tidak produktif hanya kepada 12.81 kitaran (Lee & Mitici 2023). Kajian lain untuk industri aeronautikal menggunakan NASA C-MAPSS telah menunjukkan bahawa kombinasi model LSTM dan pemprograman matematik dapat menghasilkan nilai kejituan dan skor F1 yang tinggi dalam tempoh masa yang sesuai (O'Neil, Diallo & Khatab 2022). PdM juga digunakan untuk memantau alat-alat yang digunakan di kilang. Dengan menerapkan ANN dengan teknik SVM untuk menganalisis data daripada bahagian penting seperti alat pemotong dan gelendong motor, kejituan model dapat meningkat hingga 98%. Dalam kajian yang sama, juga ditemui bahawa model ini berfungsi dengan baik dalam meramal status galas (Lee et al. 2019).

2.4 PENYELIDIKAN BERKAITAN SET DATA KAJIAN

Pelbagai kajian telah dijalankan ke atas set data AI4I2020 yang akan digunakan di dalam projek ini. Beberapa kajian menggunakan kaedah pembelajaran mesin yang tradisional untuk analisis PdM ke atas set data AI4I2020. Salah satu kajian tersebut terdiri daripada kertas ulasan seperti yang dikeluarkan oleh Meddaoui , Hain dan Hachmoud (2023) yang telah membandingkan teknik RF, DL serta ANN ke atas set data AI4I2020 dan mendapati bahawa teknik RF menunjukkan kejituan yang lebih tinggi berbanding kaedah ANN di mana teknik RF berjaya menghasilkan kejituan sebanyak 4% lebih tinggi untuk ramalan kegagalan serta kejituan sebanyak 1% lebih tinggi untuk ramalan jenis kegagalan (Meddaoui, Hain & Hachmoud 2023). Terdapat juga kertas kajian yang membandingkan model berdasarkan algoritma-algoritma yang popular seperti yang telah dijalankan oleh Nazara (2022) di mana beliau telah membandingkan teknik DT, *eXtreme Gradient Boosting* (XGBoost), *Gradient Boosting*, RF, LGR serta k-NN dan berjaya membuktikan bahawa teknik XGBoost memperolehi kejituan yang paling tinggi iaitu 97.35 serta *Area under the Curve* (AUC) untuk plot lengkung *Receiver Operating Characteristic* (ROC) paling tinggi dengan nilai 0.972 (Nazara 2022).

Nithin et al. (2022) pula telah menggunakan teknik SVM untuk meramal kegagalan pada set data AI4I2020 dan telah membahagikan set data ke 80% set latihan dan 20% set ujian serta mendapati bahawa teknik SVM menghasilkan kehilangan ralat sebanyak 0.0024 serta ralat klasifikasi lipatan-k sebanyak 0.28% (Nithin et al. 2022).

Sharma et al. (2022) pula telah membandingkan teknik RF, DT, SVM, k-NN dan LGR untuk set data AI4I2020 dan hasil kajian mereka menunjukkan RF memperolehi nilai kejituan paling tinggi berbanding teknik-teknik lain iaitu 0.984 manakala DT memperolehi nilai AUC untuk ROC paling tinggi berbanding teknik-teknik lain iaitu sebanyak 0.837 (Sharma et al. 2022).

Terdapat juga kajian-kajian yang menggunakan teknik-teknik yang lebih kompleks. Salah satu kajian tersebut adalah seperti yang telah dilakukan oleh Chen, Tsung dan Yu (2022) yang mengaplikasikan teknik SmoteCN, *Conditional Tabular Generative Adversarial Network* (ctGAN) dan CatBoost ke atas set data AI4I2020 serta berjaya memperolehi nilai panggilan balik sebanyak 90.68% serta 88.83% bagi kejituan seimbang (Chen, Tsung & Yu 2022). Di samping itu juga, kajian oleh Chen, Tsung dan Yu (2022) juga telah menunjukkan teknik yang dicadangkan memperolehi 1.0 bagi panggilan balik untuk kegagalan kuasa (PWF) dan kegagalan ketegangan (OSF), kejituan sebanyak 68.77% bagi PWF dan kejituan sebanyak 98.05% untuk OSF (Chen, Tsung & Yu 2022). Ghasemkhani, Aktas da Birant (2023) telah mencadangkan teknik yang dipanggil *Balanced K-Star* untuk mengatasi masalah ketidakseimbangan data dan model yang dicadangkan di dalam kajian tersebut menunjukkan kejituan sebanyak 7.01% lebih tinggi berbanding teknik *K-Star* (Ghasemkhani, Aktas & Birant 2023). Selain dari itu, kajian di dalam kertas ini juga telah membuat perbanding untuk kajian-kajian yang menggunakan set data AI4I2020 dan di dalam rumusan kertas tersebut didapati tiada kajian yang menggunakan teknik *Naive Bayes* untuk set data tersebut (Ghasemkhani, Aktas & Birant 2023). Dalam kajian oleh Harichandran , Raphael dan Mukherjee (2023) teknik yang dipanggil *Hybrid Unsupervised and Supervised Machine Learning* (HUS-ML) telah dicadangkan untuk digunakan untuk ramalan kegagalan menggunakan set data AI4I2020 dan kaedah yang telah dicadangkan telah berjaya memperolehi nilai Skor-F1 sebanyak 79.14% dan nilai ini adalah 18.07% lebih tinggi berbanding teknik *Conventional Machine Learning* (CML) (Harichandran, Raphael & Mukherjee 2023). Dalam kajian oleh Iantovics dan Enachescu (2022), teknik *Binary* LGR yang terdiri daripada empat jenis ujian iaitu ujian Wald, ujian Omnibus Pekali Model, ujian Hosmer-Lemeshow dan kemudian klasifikasi menggunakan *Binary* LGR yang mana model tersebut menghasilkan kejituan sebanyak 97.1% (Iantovics & Enachescu 2022).

Johansson, Lofstrom dan Sonstrod (2023) telah mencadangkan penggunaan *Venn-Abers* ke atas algoritma-algoritma asas seperti DT, RF dan XGBoost untuk set data AI4I2020 dan hasil kajian menunjukkan bahawa teknik ini dapat mengurangkan *Expected Calibration Error* (ECE) kepada ketiga-tiga model tersebut (Johansson, Lofstrom & Sonstrod 2023). Dalam kertas kajian yang dihasilkan oleh Mylonas et al. (2022), teknik *LionForest* (LF) telah diaplikasikan ke atas set data AI4I2020 dan dibandingkan dengan teknik *Local tree Surrogate* (LS), *Global tree Surrogate* (GS) dan MARLENA dan teknik LF telah berjaya memperolehi nilai 1 bagi ukuran kepersisan (Mylonas et al. 2022).

Papathanasiou, Demertzis dan Tziritas (2023) telah menjalankan kajian menggunakan kombinasi *Random Survival Forest* dan menapis atribut-atribut yang menyebabkan kegagaln mesin pada set data AI4I2020 serta dapat menunjukkan bahawa teknik yang dicadangkan merupakan teknik yang terbaik dalam kajian tersebut dengan memperolehi kejituan C-index sebanyak 97% bagi kegagalan mesin, 99% bagi kegagalan pelepasan haba (HDF), 97% bagi kegagalan ketegangan (OSF) serta 99% bagi kegagalan kuasa (PWF) (Papathanasiou, Demertzis & Tziritas 2023). Di dalam kajian oleh Souza dan Lughofer (2023), teknik *Envolving Fuzzy Neural Classifier with expert rules* (EFNC-Exp) telah digunakan untuk set data AI4I2020 dan ia telah dibandingkan dengan teknik tradisional yang dipanggil *Self-Organized Direction Aware data partitioning algorithm* (SODA) di mana teknik EFNC-Exp berjaya mengatasi prestasi teknik SODA bagi jumlah sampel diantara 2000 ke 5000 (Souza & Lughofer 2023). Tyrovolas, Liang dan Stylios (2023) pula telah mencadangkan penggunaan teknik *Fuzzy Cognitive Map* (FCM) menggunakan analisis *Liang-Kleeman Information Flow* (L-K IF) dan teknik ini berjaya mengatasi prestasi teknik-teknik FCM yang biasa dengan memperolehi purata kejayaan sebanyak 0.87488 serta agregat kuasa sebanyak 1.69723 (Tyrovolas, Liang & Stylios 2023).

Walaupun tiada kertas kajian rasmi seperti jurnal, laporan projek dan kertas persidangan, disebabkan oleh set data AI4I2020 dapat diakses oleh orang ramai, terdapat kod-kod yang dikongsi di dalam laman web *Kaggle* yang mengaplikasikan teknik NB bagi model PdM untuk meramal kegagalan. Berdasarkan kod yang dikongsi oleh Gaur (2021) di laman web *Kaggle*, dapat dilihat bahawa kod yang dikongsi

menggunakan lapan atribut sahaja dari set data AI4I2020 dan hanya atribut untuk *Type L*, *Type M* dan *Machine failure* yang menggunakan struktur data binari di dalam set data bersih untuk proses perlombongan data dengan algoritma LGR, NB, SVM, k-NN, DT, RF, XGBoost dengan model NB memperolehi kejituan sebanyak 0.78519. (Gaur 2021). Gujarathi (2021) telah menunjukkan kod di dalam laman web *Kaggle* yang hanya menggunakan tujuh atribut dari set data AI4I2020 dan atribut yang dikekalkan dengan jenis binari adalah atribut *Machine failure* sebelum set data yang telah dibersihkan digunakan oleh algoritima-algoritma pembelajaran mesin termasuklah kaedah NB yang memberikan nilai kejituan sebanyak 0.8290 (Gujarathi 2021). Di dalam kod yang dikongsi oleh Kodihalli (2021) di laman web *Kaggle*, hanya tujuh atribut dipilih dari set data AI4I2020 dan cuma atribut *Machine failure* yang menggunakan data jenis binari sebelum set data yang telah menjalani praprosesan dimasukkan ke dalam pelbagai algoritma mesin pembelajaran termasuklah NB yang memperolehi nilai kejituan sebanyak 0.985 (Kodihalli 2021).

Lallahom (2022) pula telah mengongsikan kod di dalam laman web *Kaggle* yang juga menggunakan tujuh atribut dari set data AI4I2020 dan hanya atribut *Machine failure* yang mempunyai struktur data binari di dalam set data bersih sebelum set data tersebut di gunakan untuk membina model PdM berdasarkan beberapa pendekatan pembelajaran mesin termasuklah algoritma NB yang memberikan nilai kejituan sebanyak 0.843 (Lallahom 2022). Kod lain yang dikongi di laman web *Kaggle* juga menggunakan tujuh atribut dari set data AI4I2020 dan hanya atribut *Machine failure* yang mempunyai struktur data binari sebelum digunakan oleh beberapa algoritma pembelajaran mesin termasuk model NB yang memberikan nilai kejituan sebanyak 0.78894 (S. K. 2021). Kod lain pula yang dikongsi oleh Shrimant (2021) di laman web *Kaggle* menggunakan lapan atribut sahaja dari set data AI4I2020 dan pendekatan binari hanya digunakan pada atribut *Machine failure* sebelum set data bersih digunakan untuk membina model PdM bagi meramal kegagalan menggunakan algoritma NB yang memperolehi nilai kejituan sebanyak 0.7615 (Shrimant 2021).

Jadual 2.1 menunjukkan rumusan prestasi model dari kajian lepas untuk model penyenggaraan 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 2.1 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)	*Balanced K-Star*	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)	*Binary* LGR	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)	*Random Survival Forest*	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)	*Random Forest*	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 2.1. 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.

2.5 ANALISIS KAJIAN LEPAS

Jadual 2.2 berikut mengandungi ringkasan bagi kajian-kajian susastera yang telah diterangkan dalam kajian ini. Terdapat pelbagai teknik-teknik yang sesuai untuk aplikasi PdM yang telah digunakan dalam kajian-kajian lepas. Selain itu, dapat juga dilihat bahawa tiada kertas kajian rasmi yang menggunakan pendekatan NB ke atas set data AI4I2020 dan hanya kod-kod yang dikongsi di laman web *Kaggle* yang menunjukkan penggunaan NB.

Jadual 2.2 Ringkasan Kajian Lepas

No.	Tajuk	Penerangan	Jenis Kajian	Kaedah Digunakan	Set Data
1	*Designing a Hybrid Equipment-Failuer Diagnosis Mechanism under Mixed-Type Data with Limited Failure Samples* (Chen, Tsung & Yu 2022)	Kajian ini menggunakan kaedah validasi silang tiga lipatan dengan nisbah data latihan kepada ujian 6700:3300 dan mengaplikasikan SmoteNC + ctGAN + CatBoost sebagai kaedah yang dicadangkan. Kaedah yang dicadangkan mencapai nilai ingatan = 1.0 untuk diagnosis PWF dan OSF, kejituan sebanyak 68.77% (PWF) dan 98.05% (OSF). Selain itu, kaedah yang dicadangkan juga mencapai kadar ingatan sebanyak 90.68% dan kejituan seimbang sebanyak 88.83% untuk analisis kesalahan pelbagai kategori, yang lebih baik daripada CatBoost (tanpa pemanjangan), SmoteNC + CatBoost, dan ctGAN + CatBoost.	Penyelidikan	SmoteNC + ctGAN + CatBoost	AI4I2020
2	*Artificial Intelligence for Predictive Maintenance in the Railcar Learning Factories* (Daniyan et al. 2020)	Kajian ini menggunakan ANN dengan model siri masa dinamik untuk meramalkan keadaan dan kegagalan potensi roda perlahan kereta api. Kajian ini dijalankan menggunakan data masa lalu suhu bantalan roda dan data tersebut dilatih menggunakan algoritma Levenberg Marquardt dalam persekitaran MATLAB 2018a. Kaedah yang dicadangkan berjaya menghubungkan suhu dengan RUL untuk meramalkan penguraian dan kegagalan.	Penyelidikan	ANN	Rel
3	*Data Imbalance+EDA+87% AUC* (Gaur 2021)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma lapan atribut yang digunakan dari set data dan hanya data bagi atribut *Type L*, *Type M* dan *Machine failure* yang	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020

bersambung…

...sambungan

		menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.78519			
4	*Balanced k-Star: An Explainable Machine Learning Method for Internet-of-Things-Enabled Predictive Maintenance in Manufacturing* (Ghasemkhani, Aktas & Birant 2023)	Kajian ini mencadangkan kaedah pembelajaran mesin untuk menangani masalah ketidakseimbangan data dalam PdM. Kaedah yang digunakan ialah K-Star Seimbang. Kaedah yang dicadangkan mencapai kejituan yang lebih tinggi sebanyak 7.01% berbanding kaedah K-Star asas. Di dalam bab 4.2, kertas ini mempersembahkan satu jadual untuk perbandingan dengan kertas-kertas lain dan tidak terdapat kajian yang menggunakan kaedah Naive Bayes atau ensemble dengan kaedah Naive Bayes sejauh ini.	Penyelidikan	Balanced K-Star	AI4I2020
5	*Machine Predictive Maintenance Classification* (Gujarathi 2021)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma tujuh atribut yang digunakan dari set data dan hanya data bagi atribut *Machine failure* yang menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.829	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020
6	*PredMaX: Predictive Maintenance with Explainable Deep Convolutional Autoencoders* (Hajgato et al. 2022)	Kajian ini mencadangkan pengklusteran jangka masa automatik dan pengenalpastian cekap bahagian jentera yang sensitif dengan menggunakan pengetahuan tersembunyi dalam data temporal tak bertanda berdimensi tinggi. Kajian ini menemui bahawa PredMaX mengurangkan dimensi data dalam dua langkah: Satu pemacu otomotif konvolusi mendalam yang dapat dijelaskan digunakan pada data terlebih dahulu, diikuti oleh analisis komponen utama. Selain itu, selang waktu di mana kotak gear	Penyelidikan	PredMax	Kotak Gear

bersambung...

...sambungan

		beroperasi dengan pelincir yang rosak dikenal pasti dengan bantuan PredMaX tanpa pengetahuan khusus mengenai jentera tersebut.			
7	*Equipment Activity Recognition and Early Fault Detection in Automated Construction through a Hybrid Machine Learning Framework* (Harichandran, Raphael & Mukherjee 2023)	Kajian ini mencadangkan satu teknik yang dipanggil HUS-ML. Kaedah yang dicadangkan mempunyai dua langkah pembelajaran utama, yang pertama ialah pembelajaran dengan pengawasan menggunakan model pengelas untuk mengenal pasti aktiviti pembinaan atau operasi yang rosak, dan kemudian diikuti dengan pembelajaran tanpa pengawasan untuk pengesanan kesalahan. Kaedah HUS-ML digunakan pada dataset AI4I2020 untuk mengesahkan prestasinya. Kaedah yang dicadangkan mencapai skor 79.14% dalam F1-Skor, yang lebih tinggi sebanyak 18.07% daripada pendekatan CML.	Penyelidikan	HUS-ML	AI4I2020
8	*An I4.0 Data Intensive Platform Suitable for the Deployment of Machine Learning Models: a Predictive Maintenance Service Case Study* (Herrero & Zorrilla 2022)	Kajian ini menggunakan dataset sensor pam air untuk meramalkan keadaan NORMAL, ROSAK, dan PEMULIHAN. Empat algoritma digunakan termasuk hutan rawak, SVM, k-NN, dan MLP. Algoritma k-NN telah dipilih kerana ia memberikan kejituan tertinggi dengan masa pemprosesan yang paling sedikit.	Penyelidikan	Random Forest, SVM, k-NN, MLP	Pam Air
9	*Method for Data Quality Assessment of Synthetic Industrial Data* (Iantovics & Enachescu 2022)	Kajian ini melaksanakan penilaian kualiti data yang ditekankan secara matematik. Ia menghampiri jenis prediksi multivariat dengan hasil yang boleh digunakan untuk klasifikasi binari. Empat ujian digunakan dalam kaedah yang dicadangkan, iaitu Ujian 1: Ujian Wald, Ujian 2: Ujian Omnibus Koefisien Model, Ujian 3: Ujian Hosmer-Lemeshow, dan Ujian 4: klasifikasi menggunakan regresi logistik binari. Kaedah yang dicadangkan mencapai kejituan akhir sebanyak	Penyelidikan	Regresi Logistik Binari	AI4I2020

bersambung...

...sambungan

		97.1% selepas melaksanakan penyesuaian kepada regresi logistik binari.			
10	*Well-Calibrated Probabilistic Predictive Maintenance using Venn-Abers* (Johansson, Lofstrom & Sonstrod 2023)	Kajian ini mencadangkan peramal Venn-Abers pada pokok keputusan, hutan rawak, dan model XGBoost. Kaedah yang dicadangkan mengaplikasikan Venn-Abers kepada ketiga-tiga model ini agar kesilapan penyesuaian yang dijangka (ECE) dapat dikurangkan. Hasil eksperimen menunjukkan bahawa penggunaan penyesuaian memberikan impak positif apabila mengamati ramalan kelas minoriti. Ini kerana penyesuaian mampu membetulkan hutan rawak yang kurang keyakinan dan teknik pokok keputusan dan XGBoost yang terlalu keyakinan. Selain itu, penyesuaian juga mampu membetulkan kedua-dua hutan rawak yang kurang keyakinan dan terlalu keyakinan dalam teknik pokok keputusan dan XGBoost.	Penyelidikan	Venn-Abers + Decision Tree Venn-Abers + Random Forest Venn-Abers + XGBoost	AI4I2020
11	*Classification models-UpSampled-F1-97%* (Kodihalli 2021)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma tujuh atribut yang digunakan dari set data dan hanya data bagi atribut *Machine failure* yang menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.985	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020
12	*Resampled - AUC: 991* (Lallahom 2022)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma tujuh atribut yang digunakan dari set data dan hanya data bagi atribut *Machine failure* yang menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.843	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020

bersambung...

...sambungan

13	*Deep Reinforcement Learning for Predictive Aircraft Maintenance using Probabilistic Remaining-Useful-Life Prognostics* (Lee & Mitici 2023)	Kajian ini menganggarkan taburan masa RUL menggunakan Rangkaian Neural Konvolusi dengan penurunan Monte Carlo. Prognosis ini dikemaskini dari semasa ke semasa, seiring dengan ketersediaan lebih banyak pengukuran. Selanjutnya, kami menganggap masalah perancangan penyenggaraan sebagai masalah Pembelajaran Pengukuhan Mendalam (DRL) di mana tindakan penyenggaraan dipicu berdasarkan anggaran taburan RUL. Kajian ini berjaya menunjukkan bahawa kos penyenggaraan keseluruhan berkurang sebanyak 29.3% berbanding dengan kes apabila enjin digantikan pada purata anggaran RUL. Selain itu, 95.6% daripada penyenggaraan yang tidak dirancang dapat dielakkan, dan jangka hayat yang terbuang untuk enjin hanya terhad kepada 12.81 kitaran.	Penyelidikan	CNN, RL	Pesawat
14	*Predictive Maintenance of Machine Tool Systems Using Artificial Intelligence Techniques applied to Machine Condition Data* (Lee et al. 2019)	Dalam kajian ini, algoritma berasaskan AI digunakan untuk memantau dua elemen sistem mesin alat yang penting: alat pemotong dan motor puncak. Sokongan vektor mesin dan rangkaian neural tiruan (rangkaian neural ulang dan rangkaian neural konvolusi) dilatih dan diuji untuk memantau dan meramal keadaan alat pemotong dan bantalan, masing-masing. Tiga fungsi kernel yang berbeza diuji untuk mencari fungsi yang paling sesuai, dan kernel polinomial dengan d=3 (SVM Kubik) menunjukkan kejituan purata tertinggi sebanyak 87%. Kejituan purata model dengan isyarat domain masa dan frekuensi adalah 84% dan 98%, secara berturut-turut. Seperti yang dijangkakan, isyarat domain frekuensi menunjukkan prestasi yang lebih baik semasa meramal keadaan bantalan.	Penyelidikan	SVM, ANN	Umum

bersambung...

...sambungan

15	*A Fast Implementation of Coalitional Model Predictive Controllers based on Machine Learning: Application to Solar Power Plants* (Masero et al. 2023)	Kajian ini melaksanakan model ANN untuk meningkatkan kawalan prediktif model bukan linear (NMPC) untuk memaksimumkan tenaga terma yang disediakan oleh medan pengepam solar. Kaedah yang dicadangkan berjaya mengurangkan masa pengkomputeran untuk NMPC berbanding dengan kaedah tradisional.	Penyelidikan	ANN	Solar
16	*The Benefits of Predictive Maintenance in Manufacturing Excellence: A Case Study to Establish Reliable Methods for Predicting Failures* (Meddaoui, Hain & Hachmoud 2023)	Kertas ini mengkaji algoritma-algoritma yang biasanya digunakan untuk penyenggaraan meramal dan mendapati bahawa Hutan Rawak, DL, dan ANN adalah kaedah yang paling biasa digunakan. Satu eksperimen telah dijalankan menggunakan RF dan ANN pada dataset AI4I2020. Model RF menunjukkan kejituan yang lebih tinggi berbanding dengan ANN, dengan perbandingan kejituan yang lebih tinggi sebanyak 4% untuk kegagalan/tiada kegagalan dan 1% lebih tinggi untuk jenis kegagalan yang berbeza.	Ulasan	Random Forest, ANN	AI4I2020
17	*Local Multi-Label Explanations for Random Forest* (Mylonas et al. 2022)	Kajian ini mencadangkan kaedah LionForest (LF). Kaedah yang dicadangkan dibandingkan dengan LS, GS (Gantian Pokok Global), dan MARLENA untuk keseluruhan dataset. Kaedah yang dicadangkan juga dibandingkan dengan kaedah Anchors dan CHIRPS untuk setiap label yang diramalkan secara berasingan. Kaedah LF berjaya mencapai kepersisan sebanyak 1 menggunakan setiap dataset.	Penyelidikan	LionForest	AI4I2020
18	Perancangan Smart Predictive Maintenance untuk Mesin Produksi (Nazara 2022)	Kajian ini membuat perbandingan antara pelbagai teknik Pembelajaran Mesin (ML) pada dataset AI4I2020. XGB mencapai kejituan tertinggi sebanyak 99.067% manakala k-NN mencapai kejituan terendah sebanyak 97.3%. Selain itu, XGB mencapai AUC-ROC tertinggi sebanyak	Penyelidikan	Decision Tree XGBoost Gradient Boosting Random Forest Logistic Regression k-Nearest Neighbour	AI4I2020

bersambung...

...sambungan

		0.972 manakala k-NN mencapai AUC-ROC terendah sebanyak 0.752.			
19	*A Two-Phase Machine Learning Approach for Predictive Maintenance of Low Voltage Industrial Motors* (Nikfar, Bitencourt & Mykoniatis 2022)	Kajian ini mengukur getaran jentera dan menganalisisnya menggunakan SVM, rangkaian neural pembelajaran ulang dan hutan rawak. SVM memberikan prestasi terbaik dengan pencapaian kejituan 100% untuk dua dataset yang mengandungi 15% data yang rosak.	Penyelidikan	NN, SVM, RF	Industri Voltan Rendah
20	*An Approach to Improve Asset Maintenance and Management Priorities using Machine Learning Techniques* (Nithin et al. 2022)	Kajian ini mengkaji tiga kes masalah kejuruteraan dan mencadangkan algoritma pembelajaran mesin untuk setiap masalah. Dataset telah dibahagikan kepada set latihan sebanyak 80% dan set ujian sebanyak 20%. Kaedah yang dicadangkan mencapai kehilangan ralat sebanyak 0.0024 dan kehilangan ralat k-lipatan sebanyak 0.28%.	Penyelidikan	SVM dengan Gaussian Kernel	AI4I2020
21	*A Novel Predictive Selective Maintenance Strategy using Deep Learning and Mathematical Programming* (O'Neil, Diallo & Khatab 2022)	Kajian ini melibatkan pelaksanaan algoritma DL untuk mengenal pasti kebarangkalian setiap komponen menyelesaikan tugas diikuti dengan pelaksanaan model optimasi penyenggaraan selektif yang mengenal pasti tindakan penyenggaraan yang akan memaksimumkan kebolehpercayaan sistem. Dataset NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) digunakan untuk analisis. Terdapat dua objektif utama dalam kertas ini: (i) untuk melanjutkan kerja Hesabi et al. (2021) untuk merangkumi struktur sistem yang lebih kompleks, dan (ii) untuk menyelesaikan masalah optimasi penyenggaraan selektif yang dibangunkan menggunakan pendekatan penyelesaian yang lebih efisien. Rangkaian LSTM bertimbun yang digunakan untuk meramalkan kebolehpercayaan telah terbukti mencapai kejituan dan skor f1 yang tinggi pada	Penyelidikan	LSTM	NASA C-MAPSS

bersambung...

...sambungan

		set pengujian, menunjukkan keupayaannya untuk meramalkan kebolehpercayaan komponen dengan tepat untuk misi yang akan datang. Hasil daripada eksperimen berangka yang diperolehi menunjukkan bahawa kerangka kerja yang dicadangkan dapat menyelesaikan masalah penyenggaraan selektif dalam masa yang munasabah untuk sistem yang kompleks.			
22	*Machine Failure Prediction Using Survival Analysis* (Papathanasiou, Demertzis & Tziritas 2023)	Kajian ini mencadangkan gabungan model hutan rintangan rawak dengan senibina yang menyaring atribut yang akan menyebabkan kegagalan mesin. Dataset telah dibahagikan kepada set latihan sebanyak 75% dan set ujian sebanyak 25%. Model yang dicadangkan mencapai kejituan indeks C sebanyak 97% untuk kegagalan mesin, 99% untuk HDF, 97% untuk OSF, dan 99% untuk PWF.	Penyelidikan	Random Survival Forest	AI4I2020
23	*Multi-agent Deep Reinforcement Learning based Predictive Maintenance on Parallel Machines* (Rodriguez et al. 2022)	Kajian ini bertujuan untuk melaksanakan model RL dalam penyenggaraan meramal. Kaedah yang dicadangkan mengatasi kaedah penyenggaraan tradisional termasuk penyenggaraan korektif dan penyenggaraan pencegahan. Sistem RL mendalam pelbagai ejen telah dikerahkan di mana setiap ejen mengawasi satu jentera. Kaedah yang dicadangkan mencapai peningkatan prestasi sebanyak kira-kira 75%.	Penyelidikan	RL	Mesin
24	*Feature selection, Hyperparameter Tuning* (S. K. 2021)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma tujuh atribut yang digunakan dari set data dan hanya data bagi atribut *Machine failure* yang menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.78894	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020

bersambung...

...sambungan

25	*Model of Aircraft Maintenance Repair and Overhaul Using Artificial Neural Networks* (Safoklov et al. 2022)	Kertas ini mempersembahkan model yang direka bentuk untuk penyenggaraan, pembaikan, dan pemeliharaan MRO sebuah pesawat dengan unit penyenggaraan meramal. ANN dicadangkan sebagai alat penyenggaraan meramal sebagai sebahagian daripada pengesanan sebelum berlakunya kerosakan.	Cadangan	ANN	Pesawat
26	*Application of Deep Learning in Facility Management and Maintenance for Heating, Ventilation, and Air Conditioning* (Sanzana et al. 2022)	Kertas ini mengkaji 100 kertas yang menunjukkan bagaimana rangkaian neural telah berkembang dalam bidang ini dan merangkum aplikasi pembelajaran mendalam dalam pengurusan fasiliti. CNN dan ANN adalah algoritma yang paling biasa digunakan dalam pengurusan fasiliti. Penjadualan penyenggaraan adalah aplikasi kedua yang paling biasa yang menggunakan DL. Kertas ini juga mendapati bahawa pengesanan kegagalan lebih banyak bergantung kepada ANN.	Ulasan	Pelbagai	Kemudahan
27	*Predictive Maintenance: Comparative Study of Machine Learning Algorithms for Fault Diagnosis* (Sharma et al. 2022)	Kajian ini melaksanakan lima teknik pembelajaran mesin iaitu RF, DT, SVM, k-NN dan LGR untuk lima dataset yang berbeza termasuk dataset AI4I2020. Hutan Rawak mencapai kejituan terbaik sebanyak 0.984 dan Pokok Keputusan mencapai AUC terbaik untuk ROC sebanyak 0.837.	Penyelidikan	Random Forest, Decision Tree, SVM, k-NN, Logistic Regression	AI4I2020
28	*Naive Bayes* (Shrimant 2021)	Kod di forum dalam laman web *Kaggle* ini menggunakan pelbagai algoritma pembelajaran mesin termasuklah teknik NB. Cuma lapan atribut yang digunakan dari set data dan hanya data bagi atribut *Machine failure* yang menggunakan data binari sebelum beralih ke fasa perlombongan data. Kejituan yang diperoleh dari model NB adalah 0.7615	Forum laman web	Pelbagai kaedah termasuk NB	AI4I2020

bersambung...

...sambungan

29	*EFNC-Exp: An Evolving Fuzzy Neural Classifier Integrating Expert Rules and Uncertainty* (Souza & Lughofer 2023)	Kajian ini melaksanakan gabungan ketidakpastian dalam respons pakar terhadap atribut sasaran ke dalam algoritma pengelas. Kaedah yang dicadangkan dinamakan ENFC-Exp dengan Ketidakpastian dan telah dibandingkan dengan kaedah rujukan iaitu pendekatan SODA tradisional dan bobot ciri. ENFC-Exp dengan Ketidakpastian telah melampaui kejituan kaedah rujukan untuk bilangan sampel antara 2000 hingga 5000.	Penyelidikan	ENFC-Exp + Uncertainty	AI4I2020
30	*Advanced Statistical and Machine Learning Methods for Multi-step Multivariate Time Series Forecasting in Predictive Maintenance* (Tessoni & Amoretti 2022)	Kajian ini membandingkan kejituan yang terdiri daripada sMAPE, MASE, dan OWA bagi model pembelajaran mesin Naïve, VARMA, Theta, LSTM, GRU, ERNN. Dataset adalah berdasarkan Data Ekonomi Persekutuan (FRED), Kualiti Udara, Jangkaan Tenaga Elektrik Perkakasan, Beijing PM2.5, Turbin Gas CO dan Pemisi Nox. Ditemui bahawa kaedah VARMA adalah model yang paling berprestasi dan Theta adalah model yang paling kurang berprestasi.	Penyelidikan	Naïve, VARMA, Theta, LSTM, GRU, ERNN	Udara
31	*Predictive Maintenance Enabled by Machine Learning: Use Cases and Challenges in the Automotive Industry* (Theissler et al. 2021)	Kajian ini menjalankan tinjauan dan mengkategori kertas-kertas penyelidikan serta menganalisisnya dari sudut aplikasi dan perspektif Pembelajaran Mesin (ML). Ia merangkumi 62 kertas yang telah ditinjau dan dikategorikan berdasarkan (a) kes penggunaan yang berkaitan (dalam hal komponen kenderaan yang dikaji), (b) kaedah ML yang digunakan, (c) tugas ML (contohnya, pengelasan, regresi, dan pengumpulan kumpulan data), dan akhirnya (d) kategori penyenggaraan meramal yang berkaitan, yang telah dikenalpasti terlebih dahulu dari segi manfaat penyenggaraan dan kompleksiti.	Ulasan	Pelbagai	Otomotif

bersambung...

...sambungan

		Kesimpulan berikut telah ditemui: (1) lebih banyak data yang boleh didapati secara awam untuk sistem automotif akan meningkatkan aktiviti penyelidikan, (2) kaedah PdM berasaskan ML adalah berpotensi untuk menemani transformasi sambungan penggerak, (3) penggabungan data dari beberapa sumber boleh meningkatkan kejituan dan membolehkan aplikasi-aplikasi baru, (4) penggunaan kaedah pembelajaran mendalam dalam PdM kemungkinan akan meningkat lagi, tetapi ini memerlukan kaedah yang disesuaikan dari segi kecekapan dan daya tafsiran serta ketersediaan data.			
32	*A Novel Framework for Enhanced Interpretability in Fuzzy Cognitive Maps* (Tyrovolas, Liang & Stylios 2023)	Kajian ini melaksanakan FCM menggunakan analisis L-K IF. Kaedah yang dicadangkan dinamakan IF-FCM. IF-FCM mencatat skor tertinggi dalam kejayaan purata (0.87488) dan kuasa agregat (1.69723) berbanding dengan kaedah FCM yang lain.	Penyelidikan	IF-FCM	AI4I2020
33	*Artificial Intelligence Approaches For The Ladle Predictive Maintenance In Electric Steel Plant* (Vannucci et al. 2022)	Kajian ini mengkaji model pokok keputusan dan hutan rawak yang digunakan untuk anggaran status penuaan secara waktu nyata bagi semua bekas yang beroperasi dalam kilang keluli elektrik. Ia bertujuan untuk pengesanan automatik bagi status akhir hayat bekas (EOL) dan pra-EOL. Kedua-dua pendekatan memberikan hasil yang memuaskan, terutamanya dalam pengenalpastian situasi yang paling kritikal, iaitu apabila bekas hampir mencapai akhir hayatnya atau memerlukan intervensi penyenggaraan segera. Pokok keputusan menunjukkan prestasi sedikit lebih baik daripada hutan rawak untuk penyenggaraan meramal.	Penyelidikan	DT, RF	Pembuatan

Kajian kesusasteraan di dalam projek ini terbahagi kepada dua. Di dalam bahagian pertama, fokus diberikan kepada penyelidikan-penyelidikan yang berkaitan dengan penggunaan pembelajaran mesin dalam PdM secara am. Kertas-kertas kajian yang diterbit mengenai penggunaan pembelajaran mesin dalam PdM terdiri daripada kertas penyelidikan, pengulasan dan cadangan. Bagi kertas-kertas yang tergolong di dalam jenis penyelidikan, penulis-penulis bagi kertas-kertas tersebut membuat cadangan mengenai model-model yang ingin dikaji dan analisis dilakukan menggunakan model-model tersebut untuk mendapatkan prestasi model-model yang dicadangkan. Kertas-kertas berkaitan ulasan penyelidikan yang lepas pula memfokuskan tentang perbandingan antara penyelidikan-penyelidikan yang lepas bagi pendekatan pembelajaran mesin di dalam PdM. Penyelidikan yang tergolong di dalam kategori cadangan pula mencadangkan rangka untuk model yang boleh dibina pada masa hadapan namun kertas-kertas kajian di dalam kumpulan ini tidak menjalankan penyelidikan terhadap model yang dicadangkan. Kertas-kertas kajian yang diperolehi mengenai pendekatan pembelajaran mesin dalam PdM menumpukan pada industri seperti kualiti udara, pesawat, otomotif, pusat kemudahan, kesihatan, pembuatan, angkasa, rel dan pam air. Berdasarkan kertas-kertas kajian yang telah dikumpul, algoritma-algoritma pembelajaran mesin asas yang sering dipilih untuk kegunaan PdM adalah ANN, CNN, DT, k-NN, LR, LSTM, NB, PCA, SVM, RF dan RL.

Di dalam bahagian dua pula, fokus diberikan kepada penyelidikan-penyelidikan yang mengkaji tentang model PdM menggunakan set data sintetik yang digunakan juga untuk projek ini. Kertas-kertas kajian yang diterbit mengenai penggunaan pembelajaran mesin dalam PdM terdiri daripada kertas penyelidikan dan pengulasan. Bagi kertas-kertas yang tergolong di dalam jenis penyelidikan, penulis-penulis bagi kertas-kertas tersebut membuat cadangan mengenai model-model yang ingin dikaji dan analisis dilakukan menggunakan model-model tersebut pada set data sintetik AI4I2020 untuk mendapatkan prestasi model-model yang dicadangkan. Kertas-kertas berkaitan pengulasan pula memfokuskan tentang perbandingan antara penyelidikan-penyelidikan yang lepas bagi pendekatan pembelajaran mesin di dalam PdM yang menggunakan set data kajian. Berdasarkan kertas-kertas kajian yang telah dikumpul, algoritma-algoritma pembelajaran mesin asas yang sering dipilih untuk kegunaan PdM pada set data kajian

adalah ANN, DT, k-NN, LGR, SVM, RF dan RL. Terdapat juga ulasan bagi kod-kod yang dikongsikan di laman web *Kaggle* di mana set data AI4I2020 dapat diperolehi dan fokus diberikan kepada kod-kod yang menggunakan algoritma NB kerana berdasarkan kertas-kertas kajian yang diterbit secara rasmi yang di kumpul untuk projek ini, tiada penyelidikan yang mencadangkan penggunaan algoritma NB.

Berdasarkan kajian kesusateraan bagi projek ini, beberapa algoritma telah dikenal pasti sebagai teknik yang sesuai untuk aplikasi PdM bagi set data kajian tetapi penyelidikan menggunakan sesetengah kaedah untuk set data AI4I2020 masih terhad seperti CNN, LR, LSTM, NB dan PCA. Projek ini memfokuskan pada penggunaan binarisasi untuk mengubah keseluruhan struktur set data kepada jenis binari serta penggunaan algoritma NB kerana set data kajian mempunyai struktur data yang sesuai dengan pendekatan klasifikasi dan NB merupakan salah satu teknik yang popular bagi aplikasi klasifikasi.

2.6 KESIMPULAN

Melalui kertas-kertas kajian yang dikumpul untuk kajian kesusasteraan di dalam projek ini, dapat dilihat bahawa teknik NB tidak dicadangkan di dalam penyelidikan-penyelidikan dalam PdM menggunakan pendekatan pembelajaran mesin bagi set data kajian yang telah diterbit secara rasmi. Namun, melalui kajian-kajian tentang pendekatan pembelajaran mesin dalam PdM secara umum, NB merupakan salah satu algoritma yang penting dan kerap digunakan dalam aplikasi PdM. Hal ini mengukuhkan lagi keperluan untuk menyelidik penggunaan teknik NB dalam aplikasi PdM pada set data AI4I2020 dengan lebih mendalam. Terdapat juga kod-kod yang dikongsi untuk umum di laman web menggunakan set data kajian dan sesetengah kod-kod ini menggunakan algoritma NB untuk pembinaan model PdM bagi meramal kegagalan mesin. Namun, struktur set data yang digunakan di dalam kod-kod yang menggunakan pendekatan NB tidak ditukarkan sepenuhnya kepada jenis binari. Hal ini menunjukkan masih terdapat ruang untuk mengkaji keberkesanan penggunaan teknik binarisasi serta pendekatan NB untuk membina model PdM menggunakan set data AI4I2020 bagi meramal kegagalan mesin.

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

Bab 3 membincangkan tentang proses yang dijalankan di dalam projek ini merangkumi matlamat projek, penerokaan data, persediaan data, pembersihan data, pembahagian data, pembinaan model dan penilaian model. Bab ini terbahagi kepada tujuh bab-bab yang kecil. Bab 3.2 akan mendahului bab-bab yang lain dengan menerangkan secara mendalam kaedah kajian yang akan digunakan di dalam projek ini. Bab 3.3 akan menerangkan tentang penerokaan data di mana sumber data, atribut-atribut data, jenis data sama ada diskret atau berterusan, laporan kualiti data yang merangkumi analisis statistik set data kajian dan visualisasi data melalui graf bar akan dijelaskan secara mendalam. Bab 3.4 pula akan membincangkan tentang persediaan data serta masalah-masalah yang perlu di semak pada sesebuah set data seperti data kosong, data hingar, data terpencil, data tidak seimbang di mana taburan data pada data sasaran tidak serata serta data tidak relevan seperti nombor pengenalan. Kemudian, bab 3.5 pula akan menerangkan mengenai kejuruteraan fitur yang akan digunakan dalam kajian. Bab 3.6 akan mengkhususkan mengenai penerangan tentang model-model yang akan diaplikasikan dalam kajian yang terdiri daripada algoritma NB, teknik Bagging dan teknik Boosting. Bab 3.7 akan menerangkan tentang kaedah penyediaan data latihan dan data ujian untuk diterapkan ke dalam model pembelajaran mesin. Bab 3 akan diakhiri dengan bab 3.8 di mana cara-cara pengukuran prestasi model akan dibincangkan.

3.2 KAEDAH KAJIAN

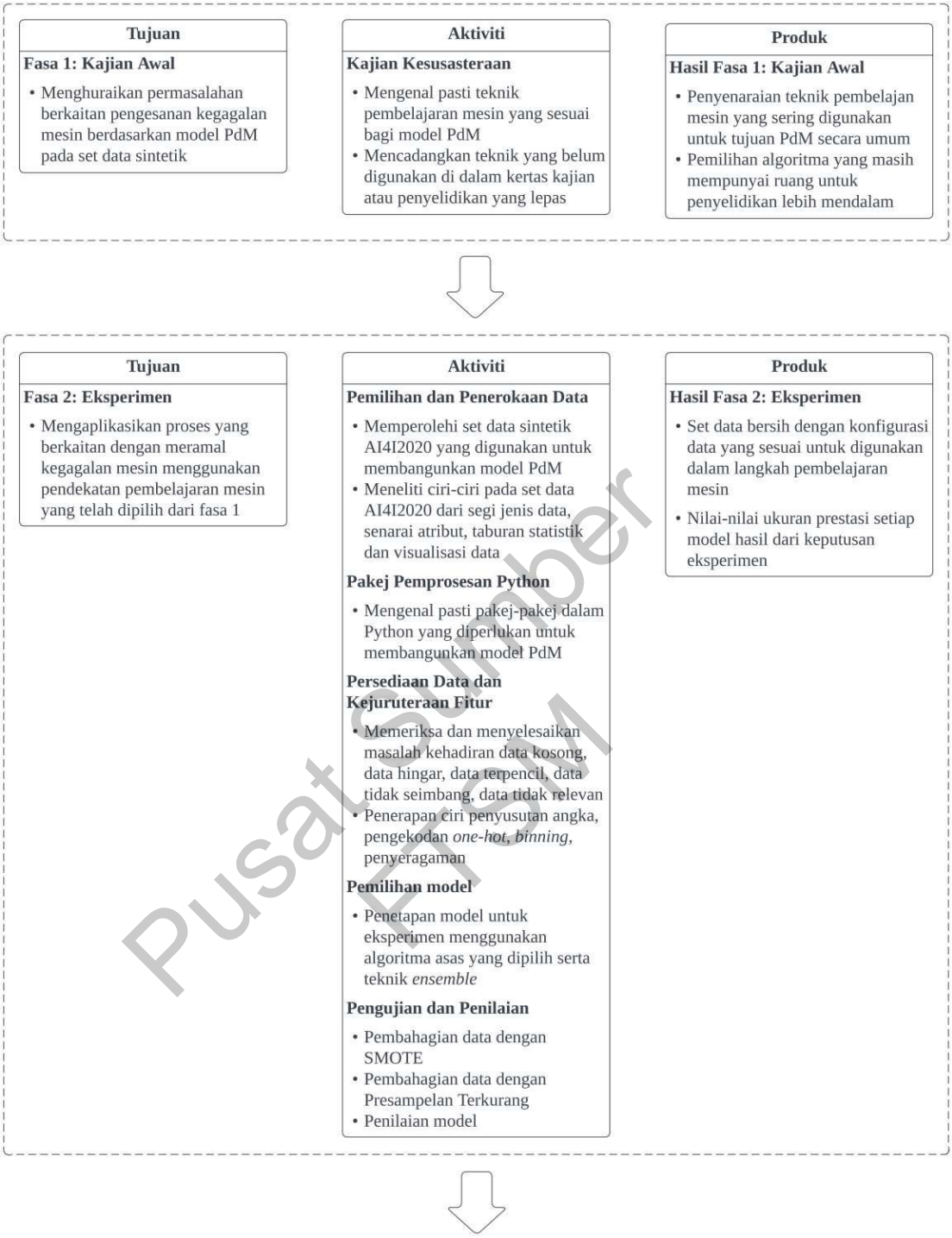
Kaedah kajian di dalam projek ini melibatkan tiga fasa kajian. Fasa 1 adalah kajian awal dengan tujuan untuk mengenal pasti permasalahan kajian iaitu meramal kegagalan mesin menggunakan model PdM ke atas set data sintetik. Di dalam fasa ini, aktiviti yang akan dijalankan termasuklah mencari teknik-teknik 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.

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. Aktiviti-aktivit yang akan dijalankan di dalam fasa ini termasuklah pemilihan dan penerokaan data di mana set data sintetik AI4I2020 dipilih sebagai set data kajian dan ciri-ciri pada set data tersebut akan dianalisis secara teliti bagi mengenal pasti langkah-langkah yang bersesuaian bagi pemprosesan set data pada peringkat seterusnya. Fasa 2 juga melibatkan pemilihan pakej-pakej pemprosesan yang sesuai di dalam aplikasi *Python* yang mengandungi fungsi-fungsi yang diperlukan untuk membina model-model PdM yang telah dicadangkan untuk projek ini. Persediaan data seperti menyelesaikan masalah-masalah yang berkaitan dengan data secara am seperti kehadiran data kosong, data hingar, data terpencil, data tidak seimbang dan data tidak relevan akan dilakukan di dalam fasa ini dan kemudian diikuti dengan pelaksanaan proses binarisasi ke atas set data kajian melalui kejuruteraan fitur menggunakan teknik-teknik seperti penyusutan angkan, pengekodan *one-hot*, *binning* serta penyeragaman. Pemilihan model iaitu menggunakan

algoritma NB termasuklan teknik *ensemble* dengan kombinasi dengan kaedah *boosting* dan *bagging* merupakan salah satu aktiviti di dalam fasa ini. Akhir sekali, aktiviti dalam fasa 2 akan melibatkan peringkat pengujian menggunakan pembahagian data kepada nisbah 70:30 bagi set latihan dan ujian menggunakan kaedah SMOTE dan persampelan terkurang sebelum penilaian model dilaksanakan. Terdapat dua hasil yang akan diperolehi dari fasa ini iaitu set data baru yang bersih dan mempunyai konfigurasi data yang sesuai termasuklah mempunyai data jenis binari bagi semua atribut untuk langkah perlombongan data menggunakan algoritma NB dan setelah eksperimen dijalankan nilai-nilai bagi ukuran prestasi yang berbeza bagi setiap model akan dihasilkan.

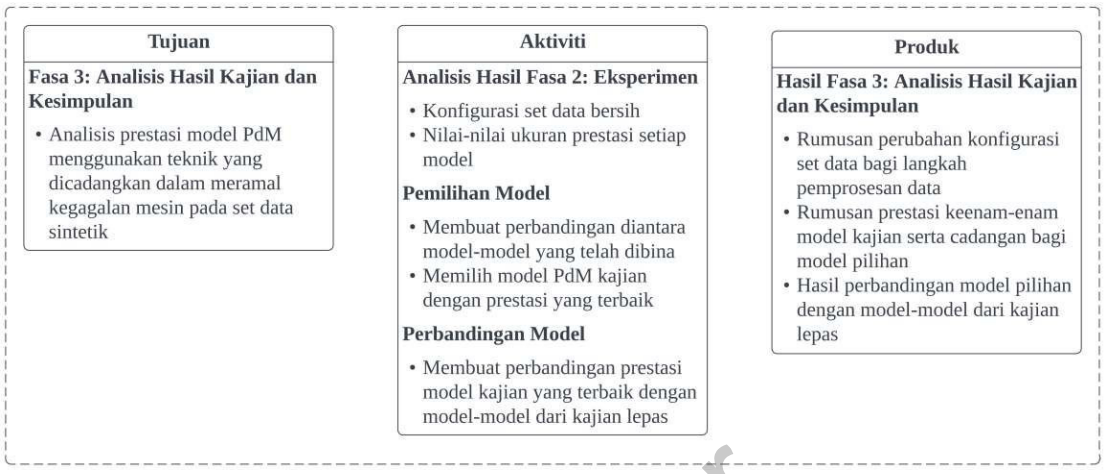
Fasa terakhir bagi kajian dalam projek ini adalah fasa 3 yang melibatkan analisis hasil kajian berserta dengan kesimpulan. Tujuan bagi fasa ini adalah bagi melaksanakan analisis bagi prestasi model-model PdM yang dicadangkan pada awal projek ini. Antara aktiviti-aktiviti yang terlibat di dalam fasa ini adalah analisis bagi hasil dari fasa 2 termasuklah konfigurasi set data bersih yang telah melalui proses binarisasi serta penilaian prestasi berdasarkan hasil eksperiemen 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 kesusateraan. 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 dengn model-model dari kajian lepas.

Kaedah kajian yang terdiri dari fasa 1, fasa 2 dan fasa 3 telah dirumuskan di dalam carta alir yang ditunjukkan di dalam Rajah 3.1 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.



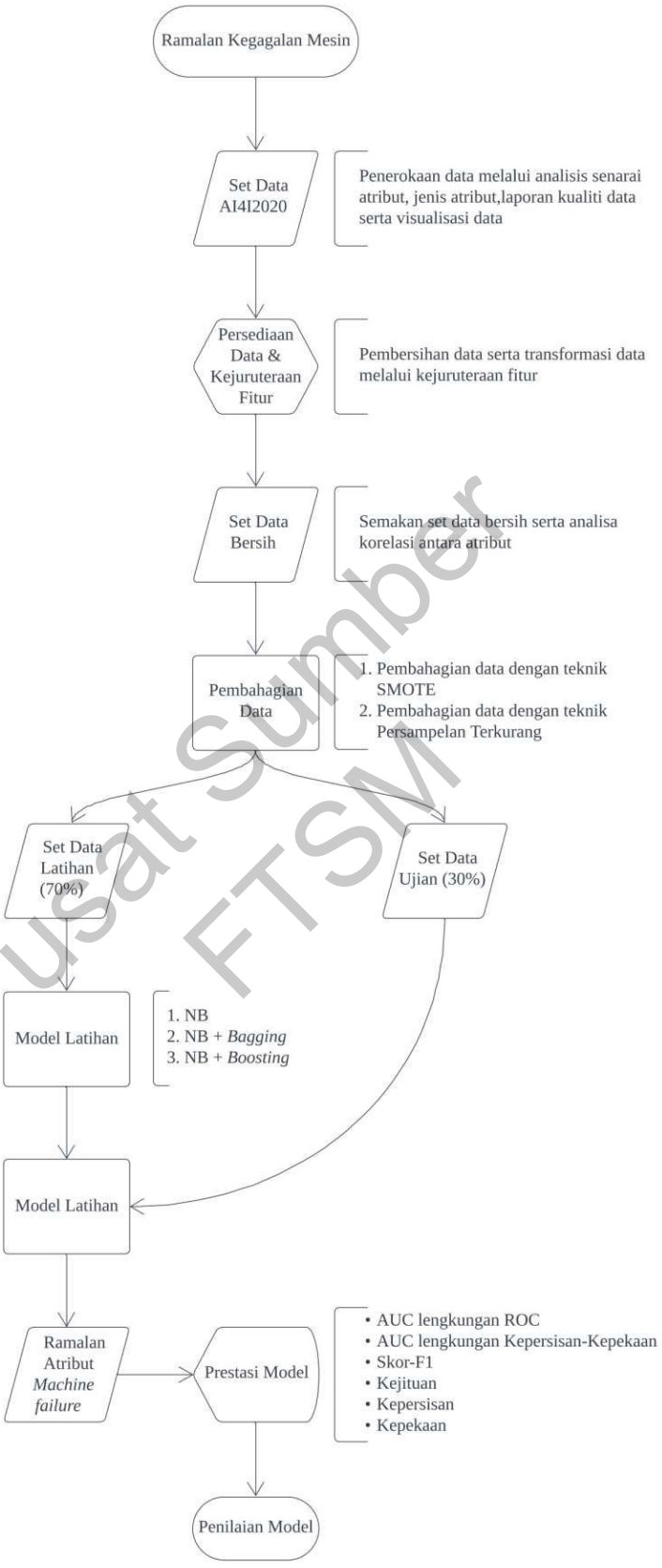
bersambung...

...sambungan



Rajah 3.1 Kaedah Kajian

Ringkasan bagi fasa 2: eksperimen bagi kajian dalam projek ini ditunjukkan dalam bentuk carta alir seperti di Rajah 3.2 berikut.



Rajah 3.2 Carta Alir Fasa 2: Eksperimen

Fasa 2 dimulakan dengan penerokaan data melalui analisis senarai atribut, jenis atribut, laporan kualiti data serta visualisasi data. 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. Setelah itu, semakan set data bersih serta analisa korelasi antara atribut akan dilakukan. 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. 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. 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 Kepersisan-Kepekaan, Skor-F1, Kejituan, Kepersisan dan Kepekaan akan digunakan untuk fasa seterusnya di dalam kajian.

3.3 PENEROKAAN DATA

3.3.1 Sumber Data

Skop penyelidikan di dalam projek ini terhad kepada set data sintetik yang dikenali sebagai AI4I2020 yang dikeluarkan oleh Matzka (2020) dalam kertas penyelidikan mengenai *Explainable Artificial Intelligence for Predictive Maintenance Applications* (Matzka 2020). Sumber mesin bagi set data kajian tidak dinyatakan di dalam kertas

kajian tersebut tetapi penerangan tentang struktur set data tersebut telah dinyatakan di mana set data tersebut telah dibina berdasarkan data asal yang dihasilkan dari mesin yang sebenar.

3.3.2 Senarai Atribut dan Jenis

Set data AI4I2020 terdiri daripada 14 attribut dengan pelbagai ciri. Terdapat empat ciri attribut di dalam set data tersebut iaitu ciri Ordinal yang terdiri daripada dua atribut, ciri Kategori yang terdiri daripada satu atribut, ciri Selang yang terdiri daripada dua atribut, ciri Angka yang terdiri daripada tiga atribut dan ciri Binari yang terdiri daripada enam atribut. Kelas sasaran utama dalam set data tersebut adalah attribut "Machine failure". Senarai atribut berserta ciri dan definisi telah disenaraikan dalam Jadual 3.1 seperti berikut.

Jadual 3.1 Senarai Atribut, Taip dan Definisi

Atribut	**Ciri (Diskret/Berterusan untuk Ciri Angka)**	**Definisi**
UDI	Ordinal	Nombor urutan
Product ID	Ordinal	Nombor unik untuk setiap produk
Type	Kategori	Menunjukkan varian sama ada L=rendah, M=sederhana atau H=tinggi
Air temperature [K]	Selang	Suhu udara
Process temperature [K]	Selang	Suhu operasi
Rotational speed [rpm]	Angka (Berterusan)	Nilai kelajuan putaran
Torque [Nm]	Angka (Berterusan)	Nilai tork
Tool wear [min]	Angka (Berterusan)	Kehausan alatan
Machine failure	Binari	Indikasi status mesin sama ada gagal atau masih berfungsi (0 = berfungsi, 1 = gagal)
TWF	Binari	Kegagalan kehausan alatan (0 = berfungsi, 1 = gagal)
HDF	Binari	Kegagalan pelepasan haba (0 = berfungsi, 1 = gagal)
PWF	Binari	Kegagalan kuasa (0 = berfungsi, 1 = gagal)
OSF	Binari	Kegagalan ketegangan (0 = berfungsi, 1 = gagal)
RNF	Binari	Kegagalan rawak (0 = berfungsi, 1 = gagal)

3.3.3 Laporan Kualiti Data

Laporan Kualiti Data bagi set data AI4I2020 telah dibahagikan kepada dua jadual iaitu jadual dengan ciri Berterusan dan jadual dengan ciri Kategori seperti yang ditunjukkan dalam Jadual 3.2 dan Jadual 3.3. Jumlah data bagi setiap atribut adalah sebanyak 10,000. Selain itu, set data kajian mempunyai nilai bagi setiap atribut dan tiada atribut yang mengandungi atribut kosong.

Jadual 3.2 Laporan Kualiti Data bagi Ciri Berterusan

Atribut	Kiraan	% Data Kosong	Kardinaliti	Minimum	Kuarter Pertama	Purata	Median	Kuarter Ketiga	Maksimum	Sisihan Piawai
Air temperature [K]	10,000	0	1	295.3	298.4	300.005	300.1	301.6	304.5	2
Process temperature [K]	10,000	0	0	305.7	308.9	310.006	310.1	311.1	313.8	1.484
Rotational speed [rpm]	10,000	0	224	1168	1424	1538.776	1504	1612	2886	179.284
Torque [Nm]	10,000	0	71	3.8	33.3	39.987	40.1	46.8	76.6	9.969
Tool wear [min]	10,000	0	6	0	53	107.951	108	163	253	63.654

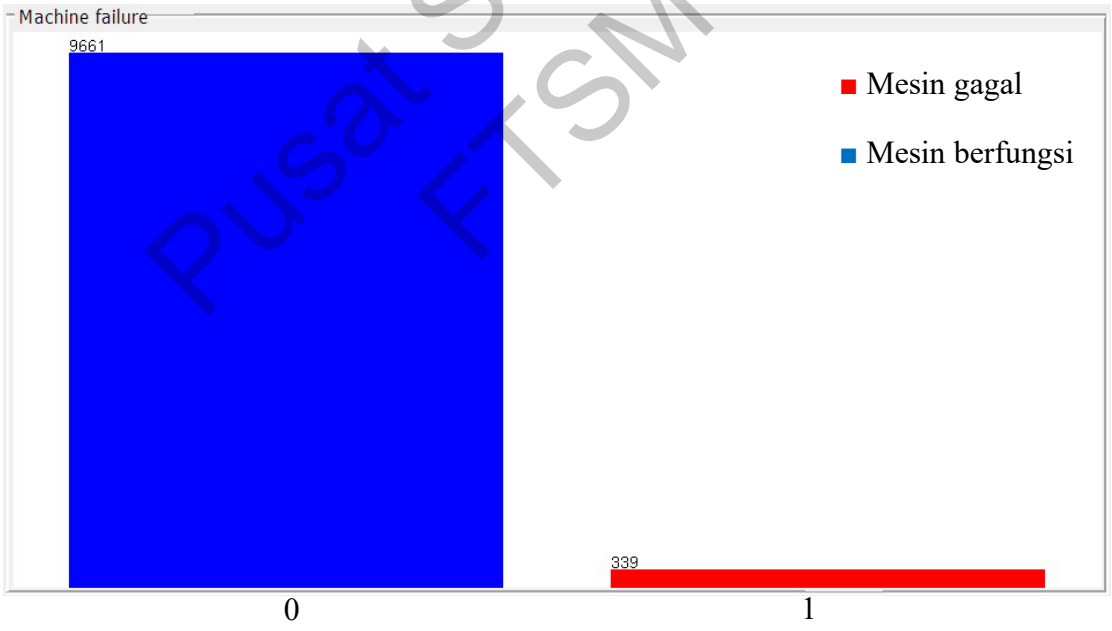
Jadual 3.3 Laporan Kualiti Data bagi Ciri Kategori

Atribut	Kiraan	% Data Kosong	Kardinaliti	Mod	Frekuensi Mod	% Mod	Mod Kedua	Frekuensi Mod Kedua	% Mod Kedua
UDI	10,000	0	10,000	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan
Product ID	10,000	0	10,000	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan	Tidak berkenaan
Type	10,000	0	3	L	6000	60	M	2997	29.97
Machine failure	10,000	0	2	0	9661	96.61	1	339	3.39
TWF	10,000	0	2	0	9954	99.54	1	46	0.46
PWF	10,000	0	2	0	9885	98.85	1	115	1.15
OSF	10,000	0	2	0	9905	99.05	1	95	0.95
RNF	10,000	0	2	0	9981	99.81	1	19	0.19

3.3.4 Visualisasi Data

Visualisasi untuk set data AI4I2020 bagi menganalisis taburan data bagi setiap atribut supaya kaedah yang sesuai dapat dikenal pasti bagi memproses setiap atribut telah dihasilkan menggunakan applikasi Weka. Rajah-rajah bagi visualisasi set data AI4I2020 ditunjukkan dari Rajah 3.3 hingga Rajah 3.14.

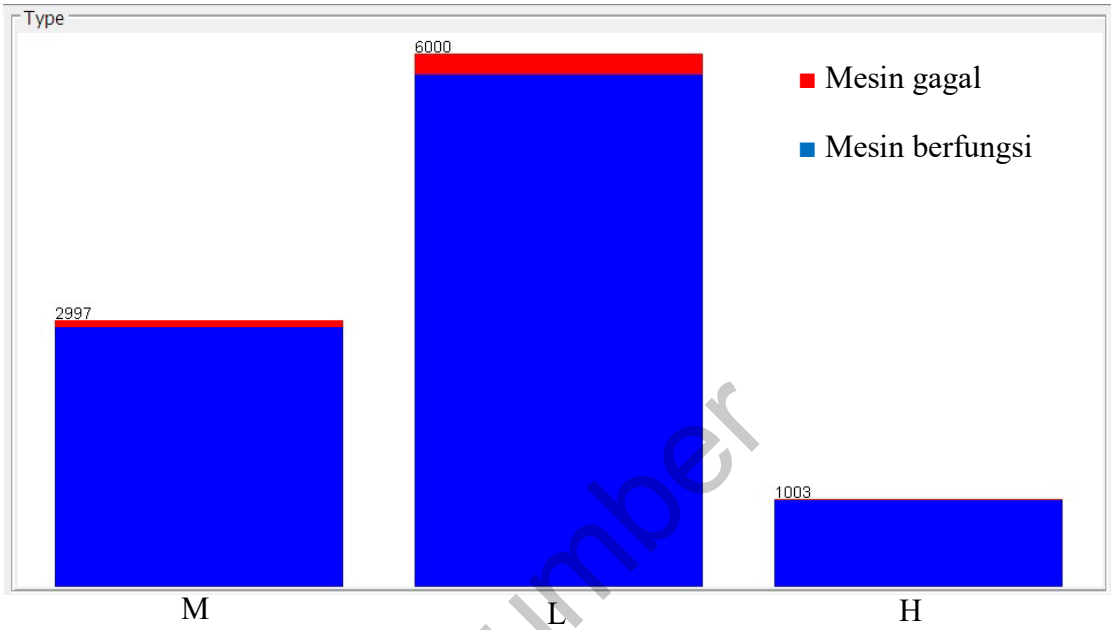
Berdasarkan Jadual 3.3, dapat dilihat bahawa terdapat ketikseimbangan data pada kelas sasaran utama iaitu "Machine failure" di mana hanya 339 data daripada keseluruhan 10,000 data yang mempunyai nilai 1 yang menunjukkan kegagalan mesin manakala data yang selebihnya mempunyai nilai 0 yang tidak menunjukkan kegagalan mesin. Ketidakseimbangan ini dapat dilihat dengan lebih jelas dan ketara berdasarkan graf bar di dalam Rajah 3.3 seperti berikut. Oleh itu, kaedah pemprosesan data yang sesuai perlu digunakan pada set data kajian sebelum langkah perlombongan data untuk mengurangkan kesan pada prestasi model PdM yang akan dibina.



Rajah 3.3 Graf Bar bagi Atribut *Machine Failure*

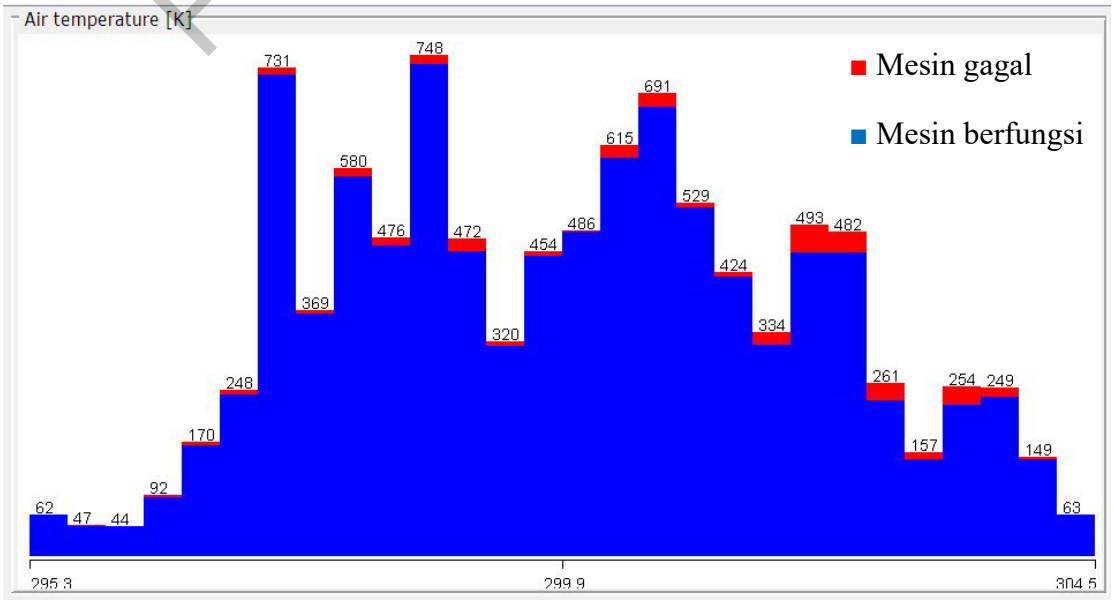
Di dalam rajah-rajah berikut, dapat dilihat bahawa pembahagian data berdasarkan kelas sasaran utama menggunakan penunjuk warna yang berbeza telah digunakan untuk menunjukkan taburan data bagi kes mesin gagal dan kes mesin berfungsi di mana warna biru menandakan kelas 0 iaitu mesin berfungsi manakala

warna merah menandakan kelas 1 iaitu mesin gagal berfungsi bagi atribut-atribut yang berkaitan dengan parameter-parameter pada operasi mesin.



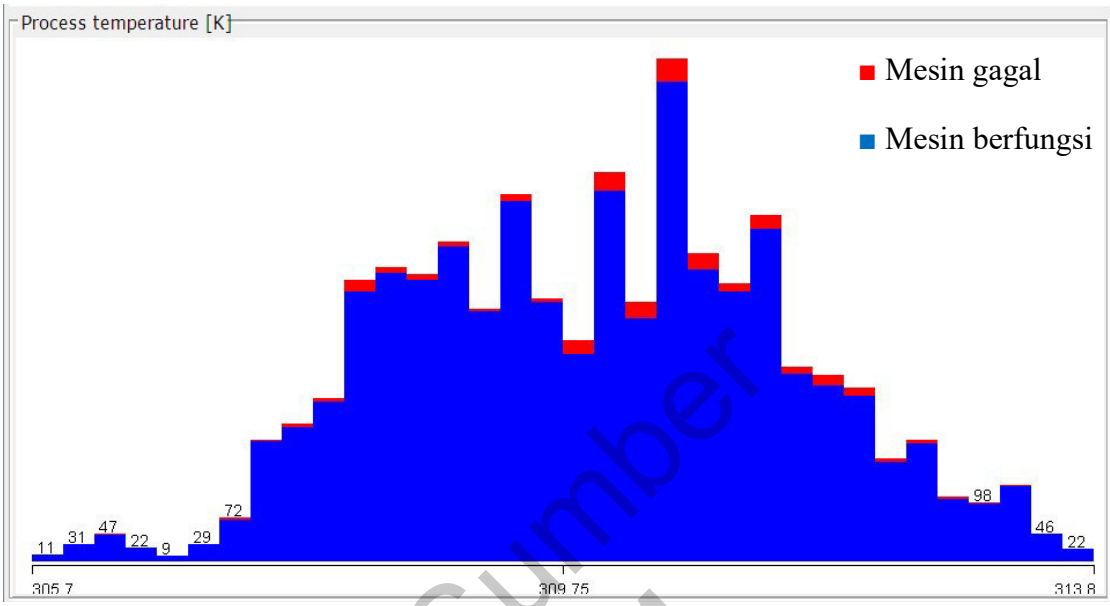
Rajah 3.4 Graf Bar bagi Atribut *Type*

Berdasarkan Rajah 3.4, dapat dilihat bahawa kebanyakan data tergolong di dalam varian L dengan 6000 data manakala data dalam varian H merupakan kumpulan data yang paling sedikit dengan 1003 data.



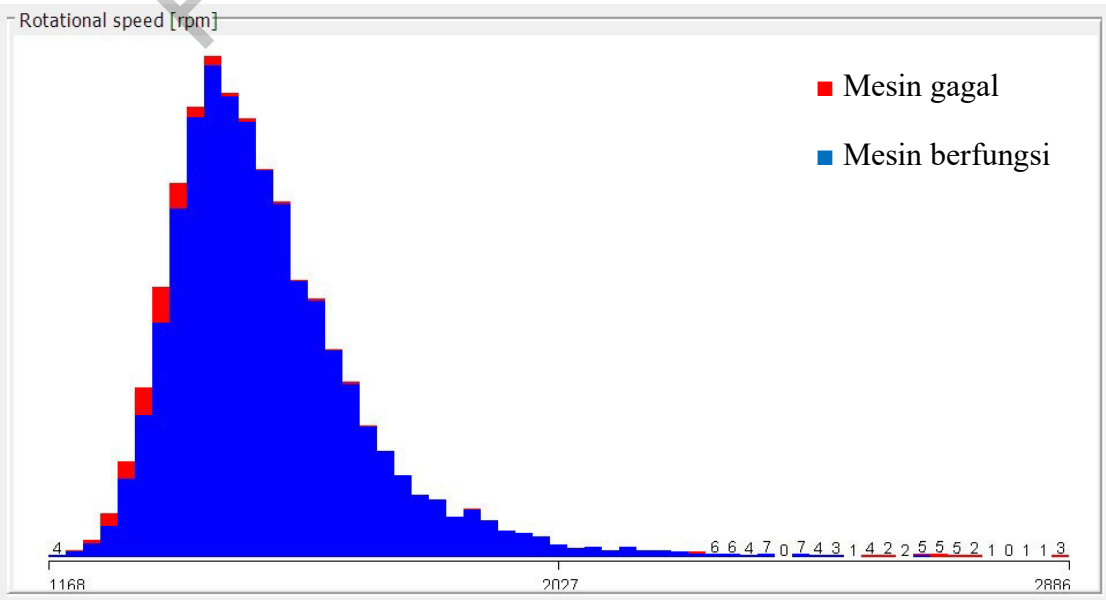
Rajah 3.5 Graf Bar bagi Atribut *Air temperature*

Rajah 3.5 memberi gambaran bahawa set data dalam kajian menunjukkan agihan data dalam bentuk menghampiri lengkungan loceng taburan normal bagi atribut *Air temperature.*



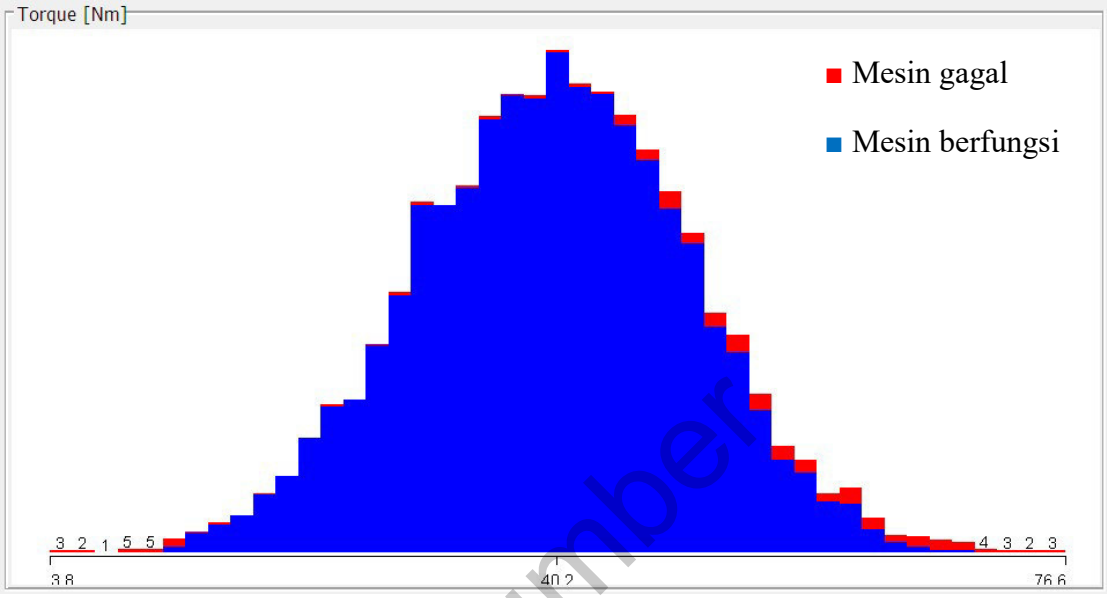
Rajah 3.6 Graf Bar bagi Atribut *Process temperature*

Data bagi atribut *Process temperature* menunjukkan agihan data dalam bentuk menghampiri lengkungan loceng taburan normal seperti ditunjukkan dalam Rajah 3.6.



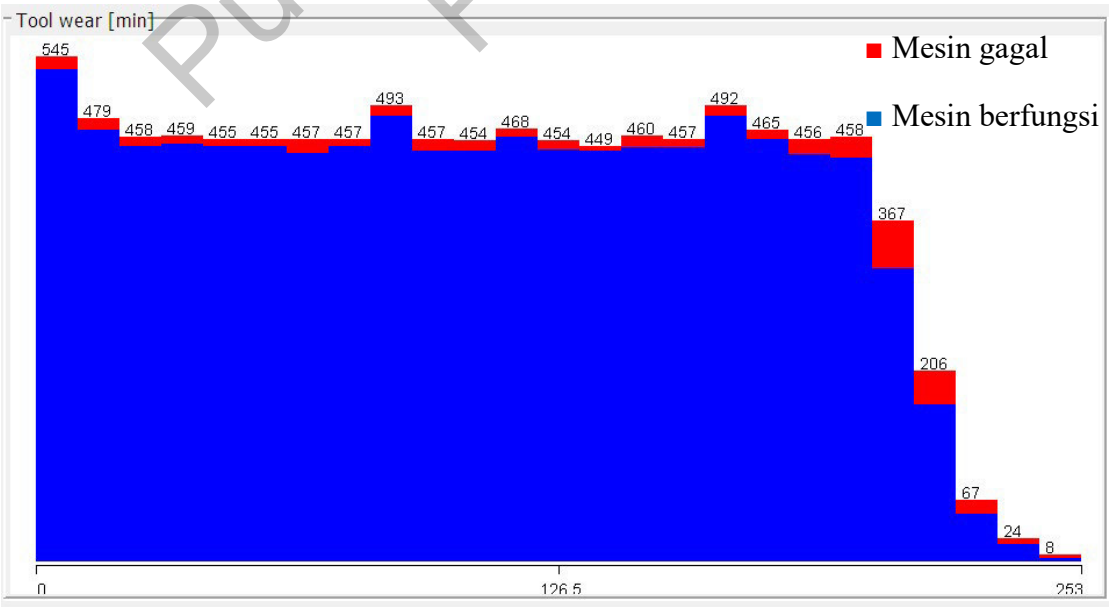
Rajah 3.7 Graf Bar bagi Atribut *Rotational speed*

Rajah 3.7 pula memberi gambaran bahawa set data dalam kajian menunjukkan agihan data dalam bentuk lengkungan condong ke kanan bagi atribut *Rotational speed*.



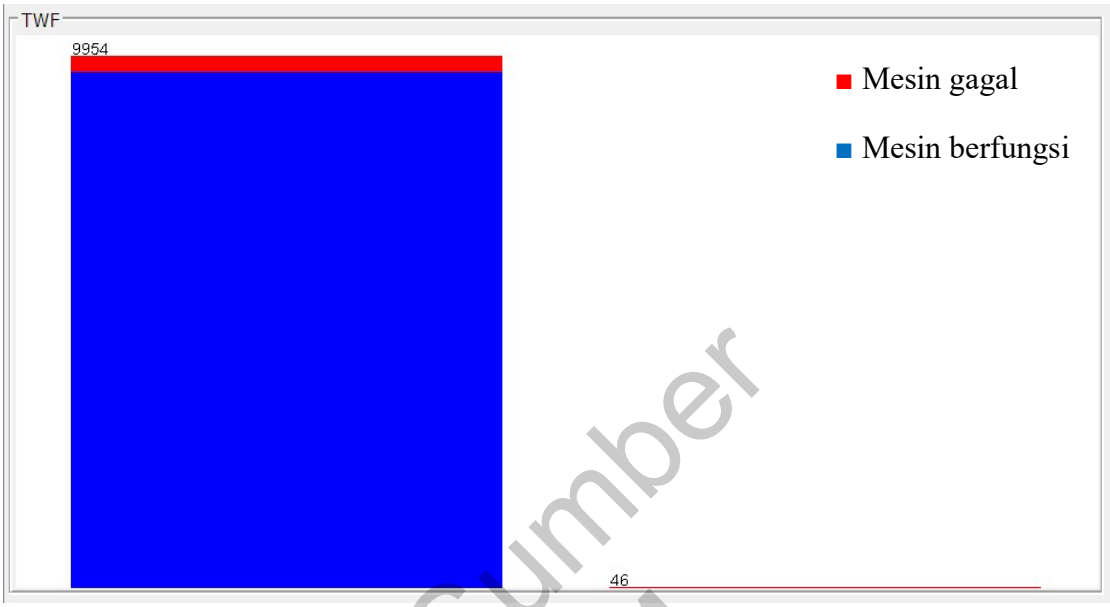
Rajah 3.8 Graf Bar bagi Atribut *Torque*

Untuk agihan data berdasarkan atribut *Torque*, dapat dilihat bahawa data berada dalam bentuk lengkungan loceng taburan normal seperti di Rajah 3.8.



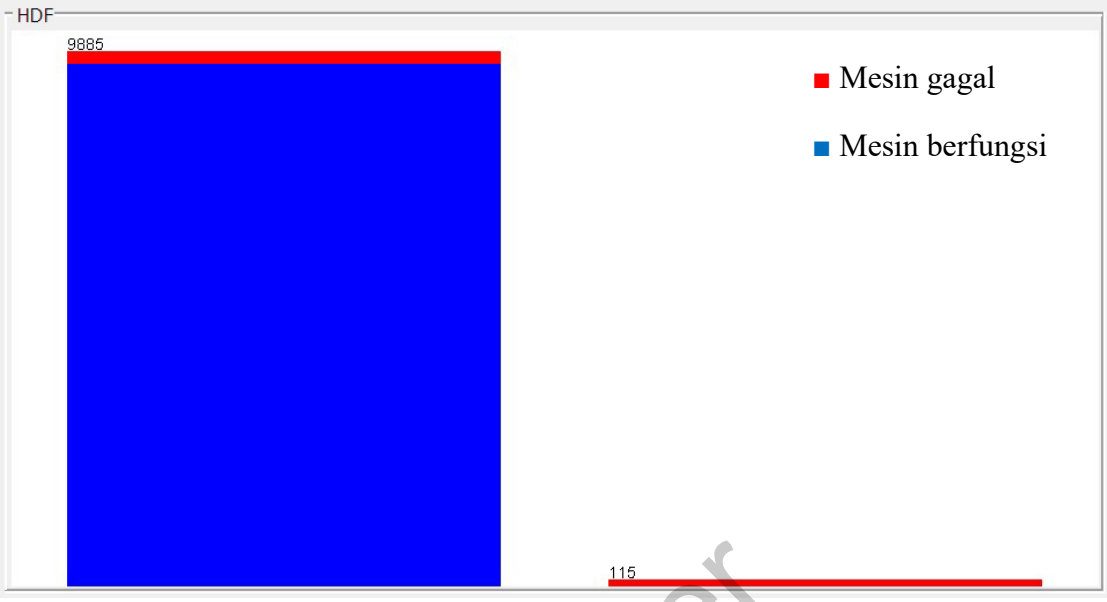
Rajah 3.9 Graf Bar bagi Atribut *Tool wear*

Agihan data bagi attribut *Tool wear* pula menunjukkan bentuk yang mendatar dan mula jatuh mendadak apabila nilai *Tool wear* menghampiri nilai maksimum seperti di Rajah 3.9.



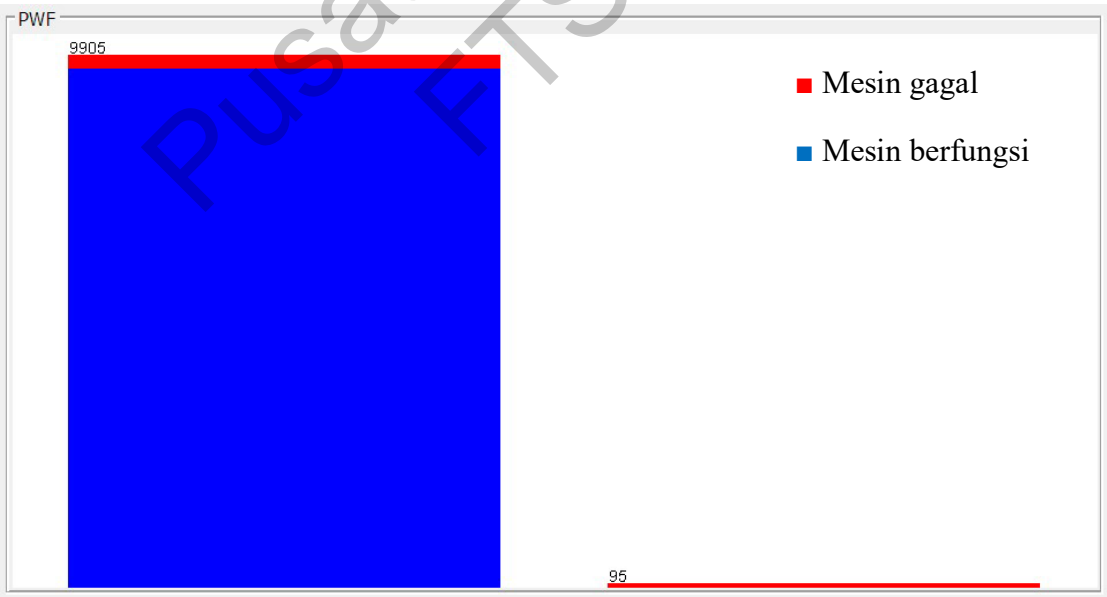
Rajah 3.10 Graf Bar bagi Atribut TWF

Melalui Rajah 3.10, dapat dilihat bahawa berdasarkan artribut TWF, majoriti data berada didalam kelas 0. Selain itu, dapat diperhatikan juga bahawa kesemua data yang berada dalam kelas 1 bagi atribut TWF juga merupakan data dalam kelas 1 bagi atribut *Machine failure*.



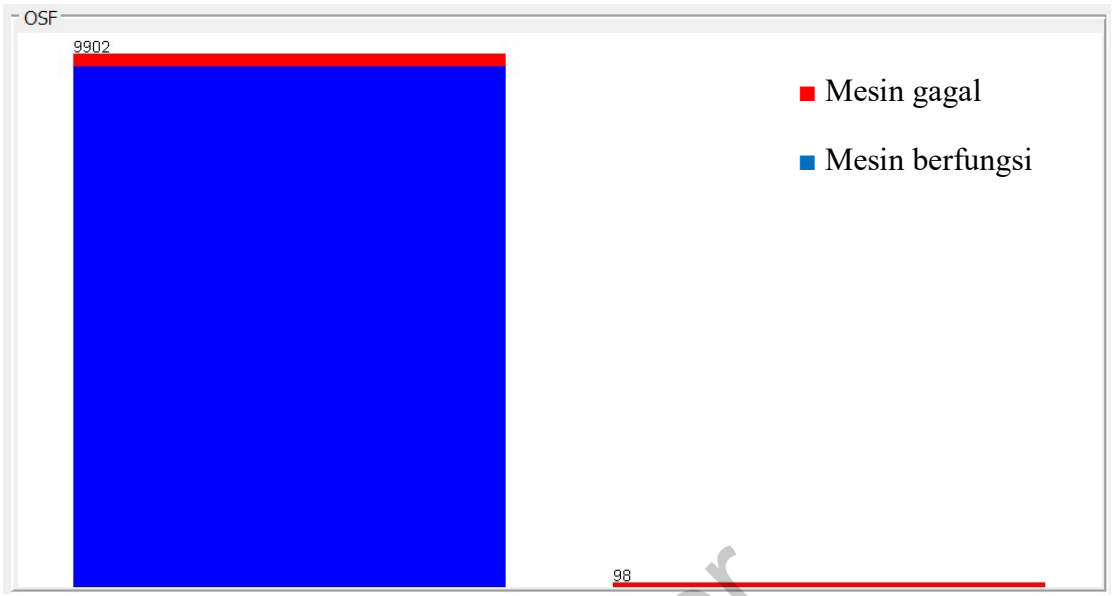
Rajah 3.11 Graf Bar bagi Atribut HDF

Rajah 3.11 menunjukkan bahawa kebanyakan data berada dalam kelas 0 bagi atribut HDF. Selain itu, dapat dilihat bahawa kesemua data yang berada dalam kelas 1 juga merupakan data dalam kelas 1 bagi atribut *Machine failure*.



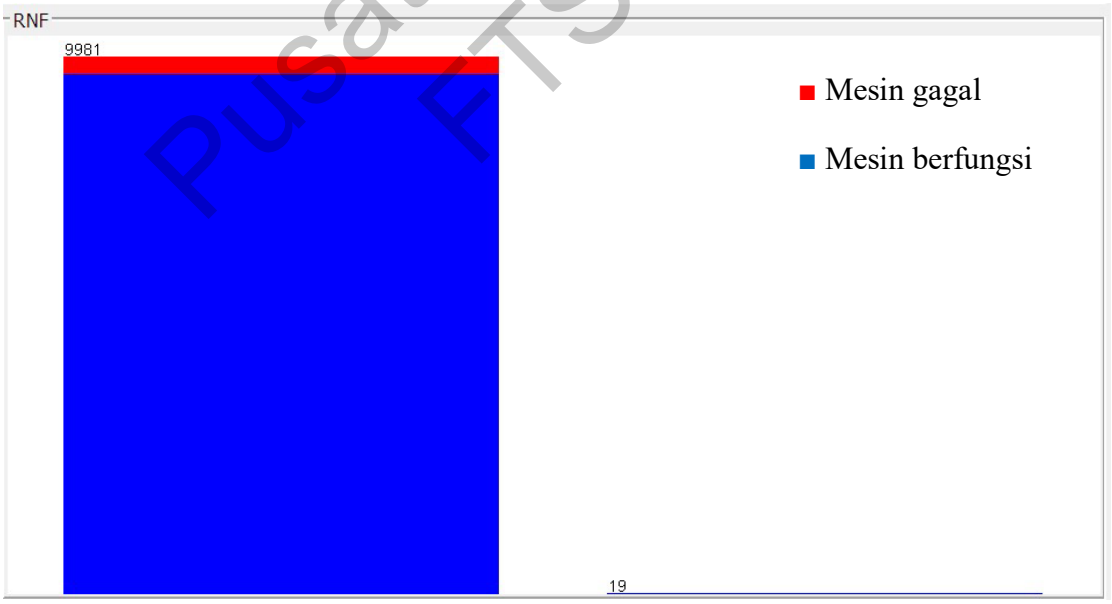
Rajah 3.12 Graf Bar bagi Atribut PWF

Seperti rajah-rajah sebelum ini untuk ciri kategori, majoriti data berada dalam kelas 0 bagi atribut PWF. Selain itu, kesemua data yang berada dalam kelas 1 juga merupakan data dalam kelas 1 bagi atribut *Machine failure* seperti di Rajah 3.12.



Rajah 3.13 Graf Bar bagi Atribut OSF

Rajah 3.13 menunjukkan bahawa majoriti data berada dalam kelas 0 bagi atribut OSF dan kesemua data yang berada dalam kelas 1 juga merupakan data dalam kelas 1 bagi atribut *Machine failure*.

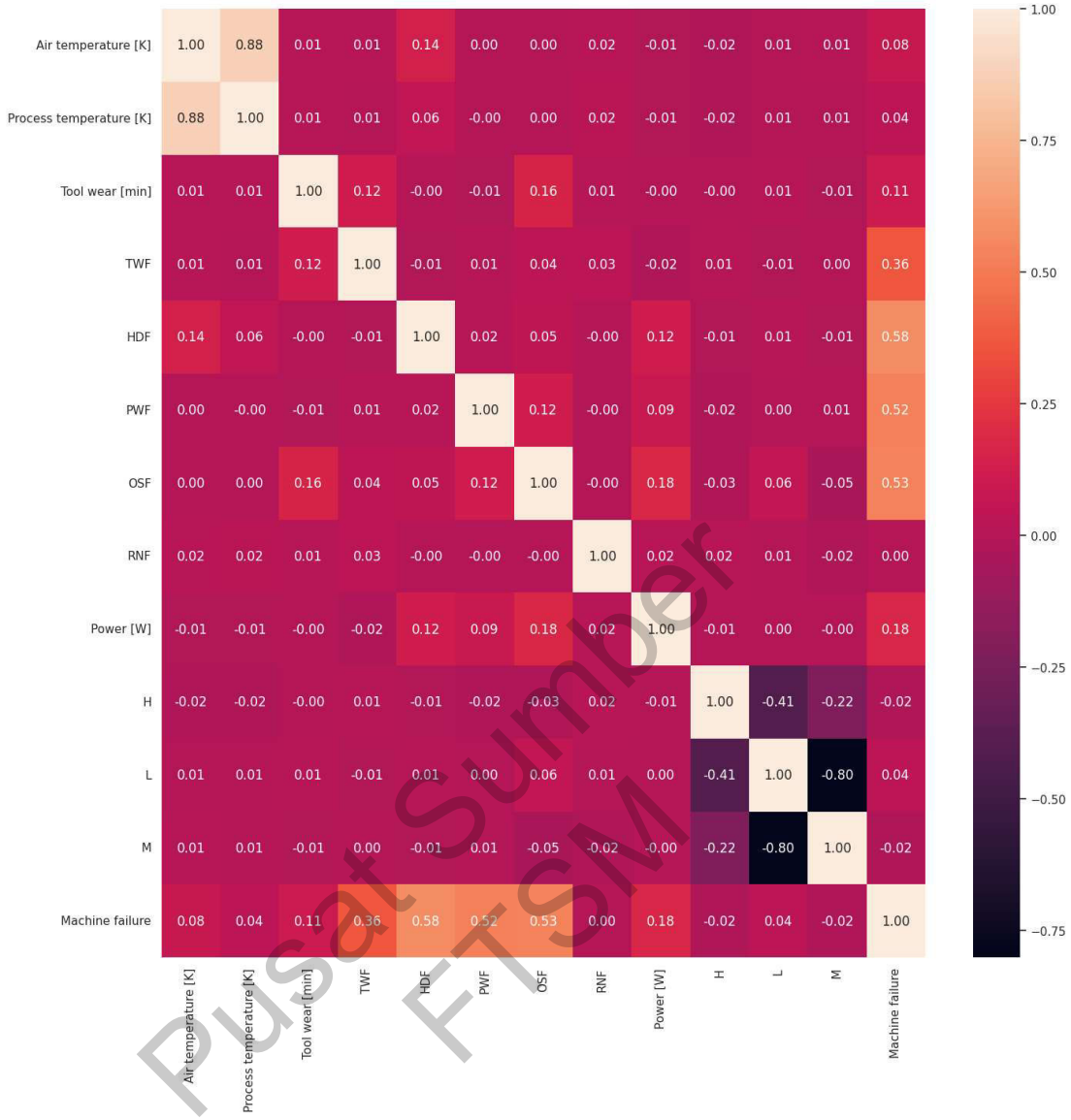


Rajah 3.14 Graf Bar bagi Atribut RNF

Berdasarkan Rajah 3.14, kebanyakan data bagi attribut RNF berada dalam kelas 0 sepertimana yang ditunjukkan oleh rajah-rajah bagi ciri kategori tetapi semua data yang berada dalam kelas 1 merupakan data dari kelas 0 bagi atribut *Machine failure*.

3.3.5 Visualisasi Data

Plot korelasi telah dihasilkan sebelum set data melalui langkah terakhir di dalam persediaan data dan kejuruteraan fitur diaplikasikan. Tujuan plot ini dihasilkan adalah untuk menganalisis korelasi antara atribut-atribut dalam set data kajian supaya fitur-fitur yang mempunyai kecendurangan dalam menyebabkan kegagalan mesin dapat dikenal pasti. Melalui plot korelasi yang ditunjukkan di dalam Rajah 3.15, terdapat empat atribut utama yang menyumbang kepada kegagalan mesin iaitu HDF, OSF, PWF dan TWF. Ini menunjukkan bahawa nilai bagi keempat-empat atribut ini mempunyai pengaruh yang tinggi pada kegagalan mesin berbanding atribut-atribut lain. Atribut RNF pula tidak menunjukkan sebarang korelasi dengan kegagalan mesin. Ini menunjukkan bahawa nilai pada RNF tidak memberi sebarang impak kepada kegagalan mesin. Selain dari itu, plot korelasi juga menunjukkan nilai korelasi yang paling tinggi bagi atribut *Air temperature* dan *Process temperature* menandakan kedua-dua atribut ini mempunyai kaitan yang tinggi antara satu sama lain.



Rajah 3.15 Plot Korelasi Atribut

3.4 PENGKALAN DATA *PYTHON*

Projek ini menggunakan beberapa pakej yang terdapat di dalam pengkalan data *Python*. Senarai pakej yang telah digunakan dirumuskan di dalam Jadual 3.4 berikut.

Jadual 3.4 Pengkalan Data *Python*

Pakej	Penerangan
numpy	Pakej ini membolehkan penggunaan fungsi operasi matematik
imblearn.under_sampling	Pakej ini membolehkan penggunaan fungsi Persampelan Terkurang
imblearn.over_sampling	Pakej ini membolehkan penggunaan fungsi SMOTE
matplotlib	Pakej ini membolehkan penghasilan plot graf
pandas	Pakej ini membolehkan manipulasi data
seaborn	Pakej ini membolehkan penghasilan plot graf
sklearn.ensemble	Pakej ini membolehkan penggunaan fungsi *Bagging* dan *Boosting*
sklearn.metrics	Pakej ini membolehkan penggunaan fungsi skor ukuran prestasi
sklearn.naive_bayes	Pakej ini membolehkan penggunaan fungsi algoritma *Naive Bayes*
sklearn.preprocessing	Pakej ini membolehkan penggunaan fungsi kejuruteraan fitur

3.5 PERSEDIAAN DATA

3.5.1 Pengendalian Data Kosong

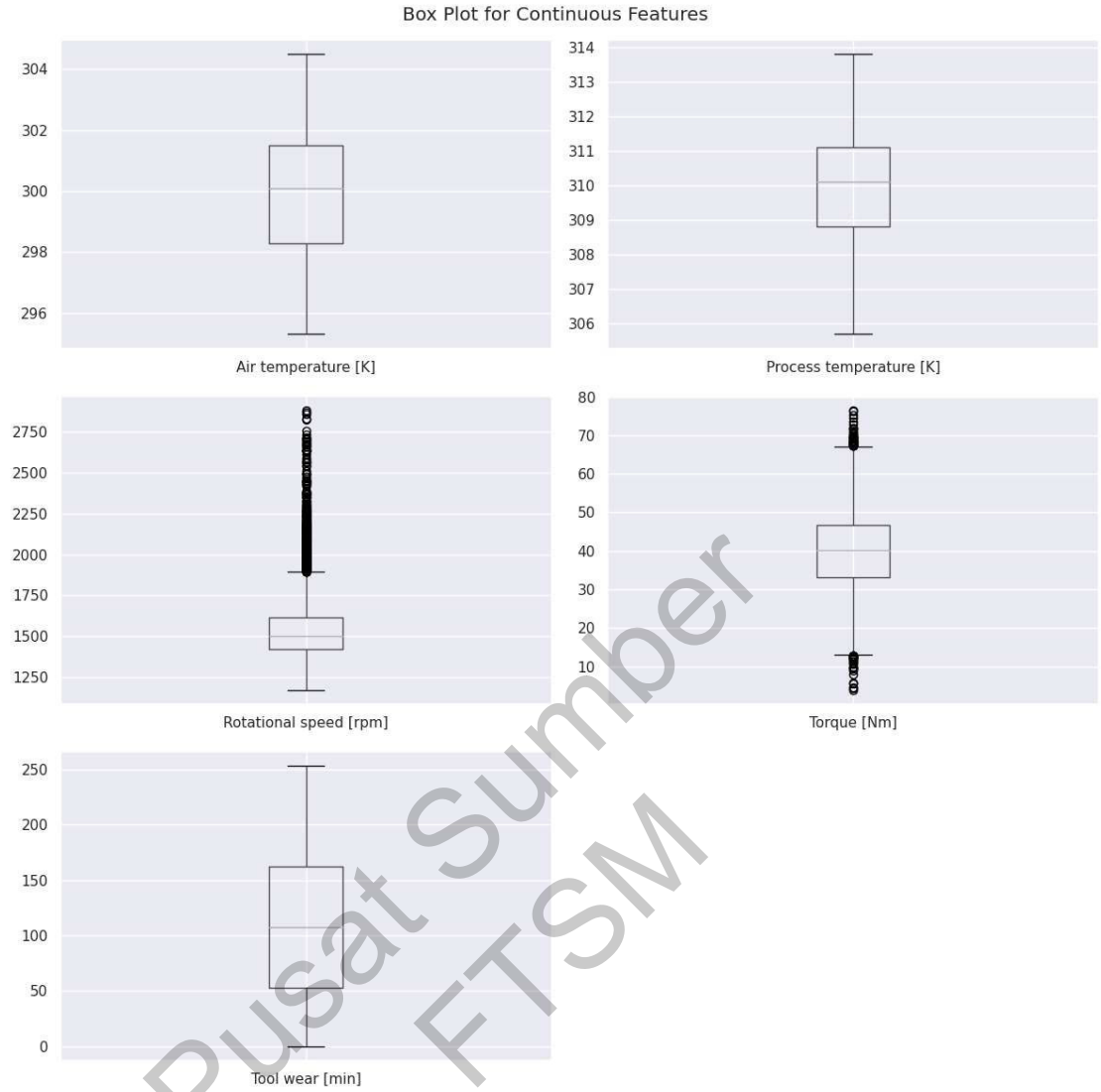
Data kosong merujuk kepada entri data yang tidak mempunyai nilai bagi satu atau lebih atribut yang terdapat pada sesebuah set data. Berdasarkan Jadual 3.2 dan Jadual 3.3, dapat dilihat bahawa setiap data yang terdapat dalam set data AI4I2020 mempunyai nilai bagi setiap atribut. Oleh itu, pengubahsuaian data untuk data kosong tidak perlu dijalankan dalam projek ini.

3.5.2 Pengendalian Data Hingar

Data hingar merujuk kepada data yang tidak logik pada sesebuah set data. Jadual 3.2 dan Jadual 3.3 serta rajah-rajah dalam bab 3.3.4 menunjukkan bahawa tiada data yang tergolong dalam kumpulan data hingar. Oleh itu, projek ini tidak akan melibatkan pengubahsuaian data untuk data hingar

3.5.3 Pengendalian Data Terpencil

Merujuk kepada plot kotak yang terdapat di dalam Rajah 3.16, dapat dilihat bahawa terdapat beberapa data terpencil bagi atribut *Rotational Speed* dan *Torque*. Teknik *Binning* dan binarisasi akan diaplikasikan di dalam projek ini untuk menyelesaikan masalah data terpencil yang terdapat di dalam set data kajian.



Rajah 3.16 Plot Kotak bagi Atribut dengan Ciri Angka (Berterusan)

3.5.4 Pengendalian Data Tidak Seimbang

Berdasarkan Rajah 3.3 yang menunjukkan taburan data bagi kelas sasaran utama iaitu untuk atribut *Machine failure*, dapat dilihat bahawa hanya terdapat 339 data dalam kelas 1. Bilangan ini adalah jauh lebih kecil berbanding data dalam kelas 0 yang mempunyai sebanyak 9661 data. Oleh itu, set data dalam kajian ini menunjukkan ketidakseimbangan taburan data yang ketara. Pengubahsuaian data untuk mengatasi masalah ketidakseimbangan taburan data akan dilaksanakan dalam kajian ini supaya kesan dari masalah ini tidak mempengaruhi prestasi model-model yang akan dibina.

Masalah ini akan diselesaikan menggunakan dua kaedah iaitu kaedah SMOTE serta kaedah Persampelan Terkurang seperti yang diterangkan dalam bab 3.6.

Berdasarkan Rajah 3.5, Rajah 3.6, Rajah 3.7 dan Rajah 3.9, dapat dilihat bahawa atribut *Air temperature*, *Process temperature*, *Rotational speed* dan *Tool wear* menunjukkan taburan data yang tidak seragam. Oleh itu, teknik *Binning* dan binarisasi akan diaplikasikan di dalam projek ini untuk menyelesaikan masalah taburan data yang tidak seragam yang terdapat di dalam set data kajian.

3.5.5 Pengendalian Data Tidak Relevan

Berdasarkan laporan kualiti data yang diterangkan di dalam bab 3.3.3, dapat dilihat bahawa bahawa terdapat dua atribut iaitu UDI dan *Product ID* yang mempunyai nilai yang unik bagi setiap data. Atribut-atribut ini tidak akan memberi sebarang impak terhadap kajian ini malah ia akan melambatkan proses analisis data. Oleh itu, atribut-atribut ini akan dikeluarkan dari set data dalam kajian. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.17 berikut.

```
# Remove 'UDI' and 'Product ID" from train dataset as it is not
useful for data mining
df = df.drop(['UDI'], axis = 1)
df = df.drop(['Product ID'], axis = 1)
```

Rajah 3.17 Kod *Python* bagi Pengendalian Data Tidak Relevan

Jadual 3.5 menunjukkan set data bagi lima barisan teratas setelah melalui proses Pengendalian Data Tidak Relevan. Data set yang baru mempunyai 12 atribut dan jumlah atribut ini menunjukkan pengurangan sebanyak dua atribut berbanding set data yang asal.

Jadual 3.5 Set Data selepas Pengendalian Data Tidak Relevan

Type	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [rpm]	Tool wear [min]	Machine failure	TWF	HDF	PWF	OSF	RNF
M	298.1	308.6	1551	42.8	0	0	0	0	0	0	0
L	298.2	308.7	1408	46.3	3	0	0	0	0	0	0
L	298.1	308.5	1498	49.4	5	0	0	0	0	0	0
L	298.2	308.6	1433	39.5	7	0	0	0	0	0	0
L	298.2	308.7	1408	40.0	9	0	0	0	0	0	0

3.6 KEJURUTERAAN FITUR

3.6.1 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 berdasarkan persamaan 3.1 seperti berikut manakala atribut *Torque* dan *Rotational speed* akan dikeluarkan dari set data yang baru.

$$Power(W) = \frac{2\pi \times \text{Rotational speed(rpm)} \times \text{Torque(Nm)}}{60} \qquad \ldots(3.1)$$

Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.18 berikut.

```
# Combine Rotational speed (rpm) and Torque (Nm) into Power (W)
and drop speed (rpm) and Torque (Nm)

def power_conversion(row):
    return row['Rotational speed [rpm]'] * row['Torque [Nm]'] * 2
* 3.141592653589793/60
```

```
df['Power [W]'] = df.apply(power_conversion, axis=1)

df = df.drop(['Rotational speed [rpm]'], axis = 1)
df = df.drop(['Torque [Nm]'], axis = 1)
```

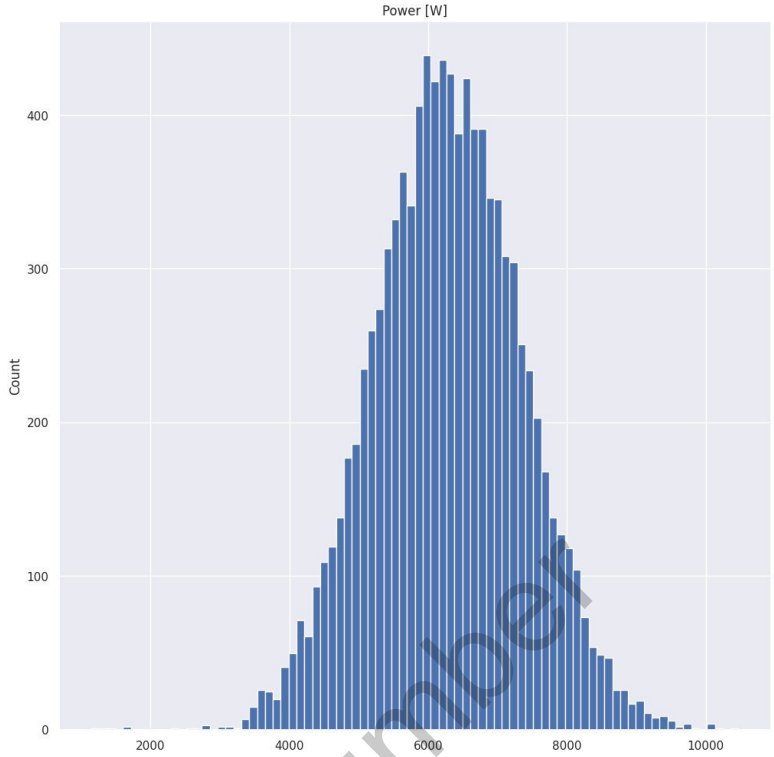
Rajah 3.18 Kod *Python* bagi Penyusutan Angka

Jadual 3.6 menunjukkan set data bagi lima barisan teratas setelah melalui proses Penyusutan Angka. Data set yang baru mempunyai 11 atribut dan jumlah ini menunjukkan pengurangan sebanyak tiga atribut berbanding set data yang asal.

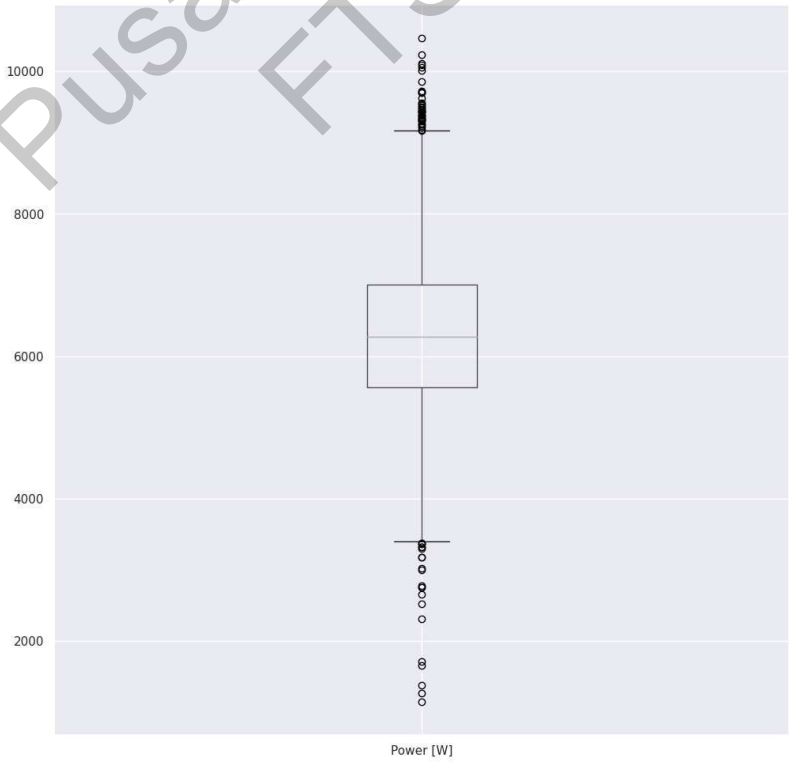
Jadual 3.6 Set Data selepas Penyusutan Angka

Type	Air temperature [K]	Process temperature [K]	Tool wear [min]	Machine failure	TWF	HDF	PWF	OSF	RNF	Power [W]
M	298.1	308.6	0	0	0	0	0	0	0	6951.590560
L	298.2	308.7	3	0	0	0	0	0	0	6826.722724
L	298.1	308.5	5	0	0	0	0	0	0	7749.387543
L	298.2	308.6	7	0	0	0	0	0	0	5927.504659
L	298.2	308.7	9	0	0	0	0	0	0	5897.816608

Rajah 3.19 menunjukkan taburan bagi atribut *Power*. Dapat dilihat bahawa atribut yang baru dikenalkan ini menunjukkan agihan data dalam bentuk lengkungan loceng taburan normal. Melalui Rajah 3.20, plot kotak bagi atribut *Power* menunjukkan masalah data terpencil. Oleh itu, seperti yang telah dibincangkan di dalam bab 3.5.3, teknik *Binning* dan binarisasi akan digunakan ketika proses menghasilkan set data bersih untuk mengatasi masalah data terpencil.



Rajah 3.19 Graf Bar bagi Atribut *Power*



Rajah 3.20 Plot Kotak bagi Atribut *Power*

3.6.2 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 yag asal. Atribut-atribut yang baru ini akan mempunyai nilai sama ada 1 atau 0.

Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.21 berikut.

```
# Converting values of object type to dummy features with binary
values
Type_dum = pd.get_dummies(x.Type)

# Adding the new dummy features to "x" dataset
x_tf = pd.concat([x, Type_dum], axis = 'columns')

# Remove attributes which are still maintaining object type data
x_tf = x_tf.drop(['Type'], axis = 'columns')
```

Rajah 3.21 Kod Python bagi Pengekodan One-hot

Jadual 3.7 menunjukkan set data bagi lima barisan teratas setelah melalui proses Pengekodan *One-hot*. Data set yang baru mempunyai 13 atribut dan jumlah ini menunjukkan pengurangan sebanyak satu atribut berbanding set data yang asal.

Jadual 3.7 Set Data selepas Pengekodan *One-hot*

Air temperature [K]	Process temperature [K]	Tool wear [min]	Machine failure	TWF	HDF	PWF	OSF	RNF	Power [W]	H	L	M
298.1	308.6	0	0	0	0	0	0	0	6951.590560	0	0	1
298.2	308.7	3	0	0	0	0	0	0	6826.722724	0	1	0
298.1	308.5	5	0	0	0	0	0	0	7749.387543	0	1	0
298.2	308.6	7	0	0	0	0	0	0	5927.504659	0	1	0
298.2	308.7	9	0	0	0	0	0	0	5897.816608	0	1	0

3.6.3 *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 menyelasaikan 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 degan ciri berterusan iaitu atribut yang baru dihasilkan iaitu atribut *Power [W]* yang dihasilkan melalui Penyusutan Angka seperti yang diterangkan di dalam bab 3.6.1 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 3.8.

Jadual 3.8 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

Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.22.

```
# Binning Air temperature [K] values

# Specified bin size
bin_size = 2

# Perform binning
x_tf['binned_values_AT'] = pd.cut(x_tf['Air temperature [K]'],
bins=range(int(min(x_tf['Air temperature [K]'])),
int(max(x_tf['Air temperature [K]']) + bin_size), bin_size),
include_lowest=True)

# Convert binned values to integer codes
x_tf['binned_values_AT'] =
pd.Categorical(x_tf['binned_values_AT'])
x_tf['binned_values_AT'] = x_tf['binned_values_AT'].cat.codes

# Use OneHotEncoder from sklearn for one-hot encoding
encoder = OneHotEncoder(sparse=False)
one_hot_encoded =
encoder.fit_transform(x_tf[['binned_values_AT']])

# Create a DataFrame with one-hot encoded columns
one_hot_df = pd.DataFrame(one_hot_encoded, columns=[f'bin_AT{i}'
for i in range(one_hot_encoded.shape[1])])

# Concatenate one-hot encoded DataFrame with the original
DataFrame
x_tf = pd.concat([x_tf, one_hot_df], axis=1)

# Remove non useful attributes
x_tf = x_tf.drop(['Air temperature [K]'], axis = 1)
x_tf = x_tf.drop(['binned_values_AT'], axis = 1)

# Binning Process temperature [K] values

# Specified bin size
bin_size = 2

# Perform binning
x_tf['binned_values_PT'] = pd.cut(x_tf['Process temperature [K]'],
bins=range(int(min(x_tf['Process temperature [K]'])),
int(max(x_tf['Process temperature [K]']) + bin_size), bin_size),
include_lowest=True)
```

```
# Convert binned values to integer codes
x_tf['binned_values_PT'] =
pd.Categorical(x_tf['binned_values_PT'])
x_tf['binned_values_PT'] = x_tf['binned_values_PT'].cat.codes

# Use OneHotEncoder from sklearn for one-hot encoding
encoder = OneHotEncoder(sparse=False)
one_hot_encoded =
encoder.fit_transform(x_tf[['binned_values_PT']])

# Create a DataFrame with one-hot encoded columns
one_hot_df = pd.DataFrame(one_hot_encoded, columns=[f'bin_PT{i}'
for i in range(one_hot_encoded.shape[1])])

# Concatenate one-hot encoded DataFrame with the original
DataFrame
x_tf = pd.concat([x_tf, one_hot_df], axis=1)

# Remove non useful attributes
x_tf = x_tf.drop(['Process temperature [K]'], axis = 1)
x_tf = x_tf.drop(['binned_values_PT'], axis = 1)

# Binning Tool wear [min] values

# Specified bin size
bin_size = 50

# Perform binning
x_tf['binned_values_TW'] = pd.cut(x_tf['Tool wear [min]'],
bins=range(int(min(x_tf['Tool wear [min]'])), int(max(x_tf['Tool
wear [min]']) + bin_size), bin_size), include_lowest=True)

# Convert binned values to integer codes
x_tf['binned_values_TW'] =
pd.Categorical(x_tf['binned_values_TW'])
x_tf['binned_values_TW'] = x_tf['binned_values_TW'].cat.codes

# Use OneHotEncoder from sklearn for one-hot encoding
encoder = OneHotEncoder(sparse=False)
one_hot_encoded =
encoder.fit_transform(x_tf[['binned_values_TW']])

# Create a DataFrame with one-hot encoded columns
one_hot_df = pd.DataFrame(one_hot_encoded, columns=[f'bin_TW{i}'
for i in range(one_hot_encoded.shape[1])])
```

```
# Concatenate one-hot encoded DataFrame with the original
DataFrame
x_tf = pd.concat([x_tf, one_hot_df], axis=1)

# Remove non useful attributes
x_tf = x_tf.drop(['Tool wear [min]'], axis = 1)
x_tf = x_tf.drop(['binned_values_TW'], axis = 1)

# Binning Power [W] values

# Specified bin size
bin_size = 1000

# Perform binning
x_tf['binned_values_PW'] = pd.cut(x_tf['Power [W]'],
bins=range(int(min(x_tf['Power [W]'])), int(max(x_tf['Power [W]'])
+ bin_size), bin_size), include_lowest=True)

# Convert binned values to integer codes
x_tf['binned_values_PW'] =
pd.Categorical(x_tf['binned_values_PW'])
x_tf['binned_values_PW'] = x_tf['binned_values_PW'].cat.codes

# Use OneHotEncoder from sklearn for one-hot encoding
encoder = OneHotEncoder(sparse=False)
one_hot_encoded =
encoder.fit_transform(x_tf[['binned_values_PW']])

# Create a DataFrame with one-hot encoded columns
one_hot_df = pd.DataFrame(one_hot_encoded, columns=[f'bin_PW{i}'
for i in range(one_hot_encoded.shape[1])])

# Concatenate one-hot encoded DataFrame with the original
DataFrame
x_tf = pd.concat([x_tf, one_hot_df], axis=1)

# Remove non useful attributes
x_tf = x_tf.drop(['Power [W]'], axis = 1)
x_tf = x_tf.drop(['binned_values_PW'], axis = 1)
```

Rajah 3.22 Kod *Python* bagi *Binning*

3.6.4 Binarisasi

Binarisasi adalah proses yang sering digunakan dalam pembelajaran mesin yang menjadikan struktur sesebuah set data kepada jenis binari. Kaedah ini dapat mengatasi

masalah data terpencil yang terdapat pada set data kajian. Di samping itu, binarisasi menghasilkan set data yang ringkas. Penggunaan binarisasi juga dapat menyeragamkan set data terutamanya bagi atribut yang menunjukkan masalah agihan data tidak seragam seperti yang terdapat pada set data kajian. Penggunaan set data yang ringkas dan seragam dapat menambahkan prestasi algoritma pembelajaran mesin dalam melakukan perlombongan data. Oleh itu, projek ini akan mengaplikasikan kaedah binarisasi kepada set data kajian.

Seperti yang diterangkan di dalam bab 3.6.2, 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 seperti yang ditunjukkan di dalam Rajah 3.22.

Dengan mengaplikasikan kaedah ini, set data yang baru akan mempunyai lebih banyak atribut berbanding set data yang asal tetapi keseluruhan data akan mempunyai struktur data yang sama sahaja iaitu jenis binari. Struktur set data baru untuk penggunaan perlombongan data bagi projek ini akan diterangkan dengan lebih lanjut lagi di dalam bab 4.2.

3.6.5 Penyeragaman

Penyeragaman adalah teknik yang mengubahsuai skala sesebuah set data dengan menggunakan nilai purata dan sisihan piawai mengunakan persamaan 2.1 seperti berikut. Penyeragaman akan menjadikan menjadikan data purata dalam sesebuah set data sebagai 0 dan data yang lain akan berada di bawah 0 atau di atas 0.

$$X' = \frac{X - \mu}{\sigma} \, dimana \, \mu = purata; \, \sigma = sisihan \, piawai \qquad \ldots(3.2)$$

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. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.23.

```
# Apply standardization
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
```

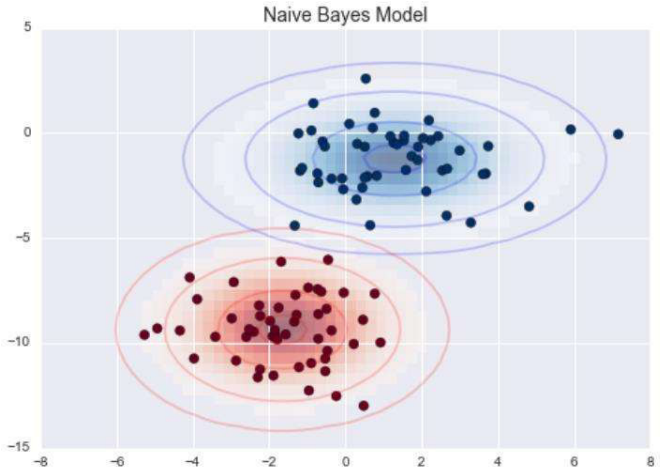
Rajah 3.23 Kod *Python* bagi Penyeragaman

3.7 PEMODELAN DATA

Set data AI4I2020 mengandungi label kelas dengan data jenis binari yang mengandungi nilai 0 dan 1. Oleh kerana hasil sasaran melibatkan nilai diskret, algoritma pembelajaran mesin berasaskan klasifikasi merupakan algoritma yang sesuai untuk meramal kegagalan mesin dalam kajian ini. Terdapat beberapa algoritma pembelajaran mesin klasifikasi popular seperti DT, NB, ANN, SVM, k-NN, dan sebagainya. Dalam projek ini, teknik klasifikasi yang NB dipilih kerana penggunaanya di dalam kajian menggunakan set data AI4I2020 berdasarkan kajian-kajian lepas adalah terhad.

3.7.1 Algoritma *Naive Bayes* (NB)

NB adalah sebuah algoritma yang berdasarkan klasifikasi Bayesian. Ia mengklasifikasikan sasaran dengan menggunakan prinsip hipotesis dan kebarangkalian. Selain itu, NB mengandaikan bahawa ciri-ciri tidak saling berhubungan di antara satu sama lain. Cincin-cincin tersebut mewakili kebarangkalian bagi setiap kelas di mana semakin kecil cincin bagi sekelompok data dalam cincin itu, semakin tinggi kebarangkalian ramalan kelas tersebut. Melalui kajian kesusteraan yang berkaitan dengan projek ini, dapat disimpulkan bahawa teknik NB adalah teknik yang selalu digunakan untuk meramal kegagalan mesin bagi aplikasi PdM namun teknik ini tidak selalu digunakan bagi pembinaan model PdM menggunakan set data kajian. Oleh itu, projek ini akan menggunakan NB bagi menilai kesesuaian algoritma ini dengan set data kajian. Rajah 3.24 menunjukkan visualisasi bagi cara kaedah NB berfungsi.



Rajah 3.24 Model *Naive Bayes*

Sumber: (Vanderplas 2016)

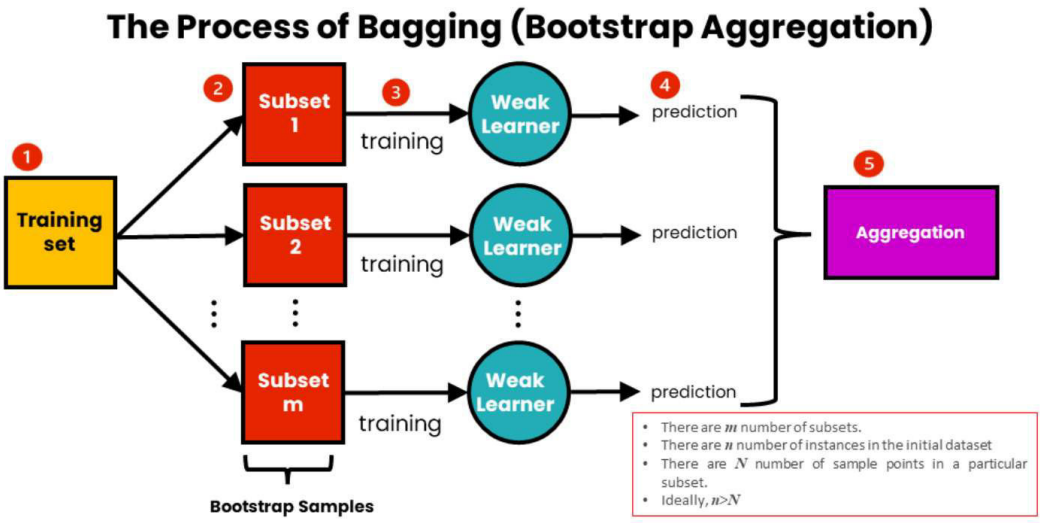
Di dalam projek ini, fitur *GaussianNB* yang terdapat di dalam pengkalan data *Python* akan digunakan untuk mengaplikasikan algoritma NB pada set data dalam kajian. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.25 berikut.

```
# Fit Naive Bayes model
nb.fit(x_train_sm, y_train_sm)
```

Rajah 3.25 Kod *Python* bagi Penggunaan Algoritma NB

3.7.2 Teknik *Bagging*

Teknik *Bagging* atau juga dikenali sebagai *Bootstrap Aggregating* merupakan teknik *ensemble* yang menggunakan algoritma pembelajaran mesin yang asas seperti algoritma klasifikasi dan regresi. *Bagging* berfungsi dengan memilih data secara rawak daripada set data latihan dan kemudian data yang dipilih itu dimasukkan dalam kumpulan-kumpulan yang berlainan. Jumlah data dalam setiap kumpulan adalah kurang berbanding set data latihan. Jika data yang sama terpilih ke dalam kumpulan yang sama, ia akan digantikan dengan data yang lebih baru agar tidak berlaku duplikasi. Data di dalam setiap kumpulan akan dilatih secara berasingan dengan selari dan purata hasil daripada setiap kumpulah akan digunakan sebagai hasil terakhir daripada teknik ini. Teknik *Bagging* digunakan untuk mengurangkan kesan varians tinggi yang mengakibatkan masalah *overfitting*.



Rajah 3.26 Langkah-langkah bagi Teknik Bagging

Sumber: (Kalirane 2023)

Rajah 3.26 menunjukkan proses yang berlaku dalam kaedah *bagging* dimana data dipilih secara rawak daripada set data latihan di langkah 1 dan kemudian data yang dipilih itu dimasukkan dalam kumpulan-kumpulan yang berlainan di langkah 2 dimana data di dalam setiap kumpulan akan digantikan dengan data yang lebih baru jika data yang sama terpilih ke dalam kumpulan yang sama. Data di dalam setiap kumpulan akan dilatih secara berasingan secara selari di langkah 3 dan hasil ramalan dari setiap kumpulan akan dihasilkan di langkah 4. Di langkah 5, purata hasil daripada setiap kumpulah akan digunakan sebagai hasil terakhir daripada teknik ini.

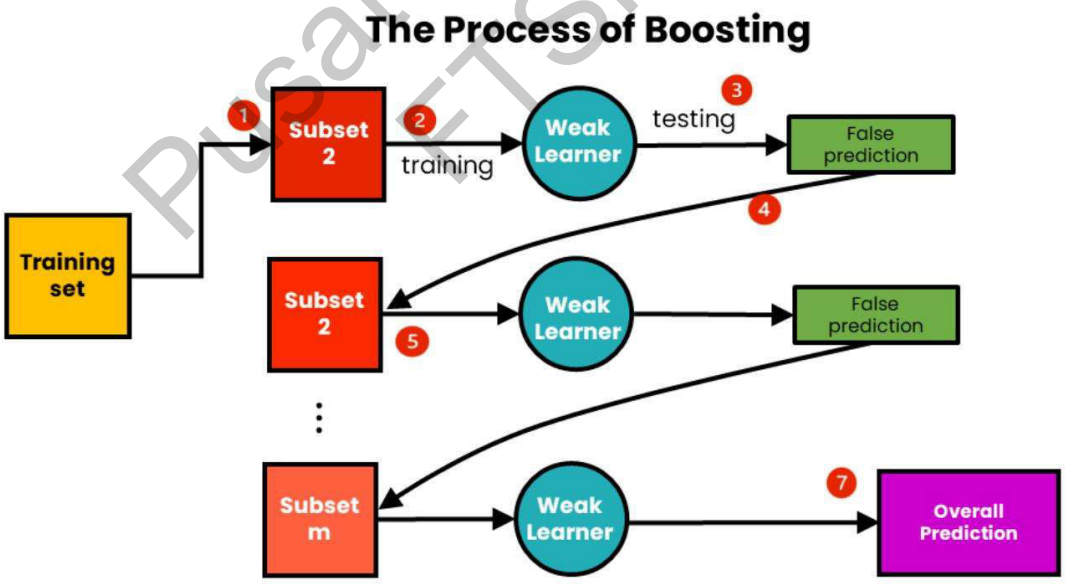
Di dalam projek ini, fitur *BaggingClassifier* yang terdapat di dalam pengkalan data *Python* akan digunakan untuk mengaplikasikan teknik *Bagging* pada set data dalam kajian. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.27 berikut.

```
# Fit Naive Bayes + Bagging model
nbag.fit(x_train_sm, y_train_sm)
```

Rajah 3.27 Kod *Python* bagi Penggunaan Teknik *Bagging*

3.7.3 Teknik *Boosting*

Teknik *Boosting* juga merupakan teknik *ensemble* yang selalu digunakan bersama algoritma pembelajaran mesin yang asas seperti algoritma klasifikasi dan regresi. *Boosting* berfungsi dengan memilih data secara rawak daripada set data latihan dan kemudian data yang dipilih itu dimasukkan dalam kumpulan-kumpulan yang berlainan. Pada kumpulan pertama, data yang telah dipilih akan dilatih dan kemudian model yang dihasilkan daripada latihan ini akan diuji dengan data yang diambil secara rawak dari set data latihan. Dalam kumpulan berikutnya, data masih akan dipilih secara rawak tetapi data yang mempunyai ralat dalam kumpulan sebelumnya akan mempunyai kebarangkalian yang tinggi untuk dipilih juga. Jumlah data dalam setiap kumpulan adalah kurang berbanding set data latihan. Langkah ini akan berterusan sehingga jumlah kumpulan mencapai angka yang telah ditetapkan. Hasil daripada kumpulan yang terakhir akan dijadikan sebagai hasil terakhir untuk teknik ini. Teknik *Boosting* digunakan untuk mengurangkan kesan berat sebelah yang tinggi yang mengakibatkan masalah *underfitting*.



Rajah 3.28 Langkah-langkah bagi Teknik Boosting

Sumber: (Kalirane 2023)

Rajah 3.28 menunjukkan proses yang berlaku dalam kaedah *Boosting* dimana data dipilih secara rawak daripada set data latihan di langkah 1 dan kemudian data yang

dipilih itu dimasukkan dalam kumpulan-kumpulan yang berlainan. Pada kumpulan pertama seperti di langkah 2, data yang telah dipilih akan dilatih dan kemudian model yang dihasilkan daripada latihan ini akan diuji dengan data yang diambil secara rawak dari set data latihan seperti di langkah 3. Data yang mempunyai ralat dari kumpulan pertama seperti yang ditunjukkan di langkah 4 akan dimasukkan ke dalam kumpulan berikutnya dimana data telah dipilih secara rawak seperti di langkah 5. Proses ini akan berterusan sehingga jumlah kumpulan mencapai angka yang telah ditetapkan. Hasil daripada kumpulan yang terakhir seperti yang ditunjukkan di langkah 7 akan dijadikan sebagai hasil sebenar untuk teknik *Boosting*.

Di dalam projek ini, fitur *AdaBoostClassifier* yang terdapat di dalam pengkalan data *Python* akan digunakan untuk mengaplikasikan teknik *Boosting* pada set data dalam kajian. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.29 berikut.

```
# Fit Naive Bayes + boosting model
nbos.fit(x_train_sm, y_train_sm)
```

Rajah 3.29 Kod *Python* bagi Penggunaan Teknik *Boosting*

3.8 PERSEDIAAN DATA LATIHAN DAN DATA UJIAN

Set data AI4I2020 mempunyai masalah data tidak seimbang yang ketara pada atribut sasaran *Machine failure*. Untuk mengawal model pembelajaran mesin yang akan dibina dari mengalami kesan negatif daripada masalah data tidak seimbang ini, dua pendekatan yang berbeza akan digunakan dan kemudian akan dibandingkan antara satu sama lain. Selain dari itu, penggunaan dua pendekatan berbeza dapat memberikan peluang untuk proses validasi keputusan kajian sekiranya terdapat persamaan dalam hasil kajian bagi dua atau lebih model.

3.8.1 Pembahagian 70:30 dengan SMOTE

Bagi membina model pembelajaran mesin, set data perlu dibahagikan kepada set latihan dan set ujian. Set data latihan akan digunakan untuk melatih dan mengoptimumkan model pembelajaran mesin, manakala set data ujian akan digunakan untuk mendapatkan hasil prestasi pengelas. Dalam projek ini, dataset dibahagikan kepada set latihan 70%

dan set ujian 30%. Nisbah 70:30 dipilih untuk pembahagian set data kerana jumlah data di dalam set data kajian iaitu sebanyak 10,000 data dapat dikategorikan sebagai set data yang agak kecil. Seperti yang ditemui semasa penerokaan data, atribut label kelas dipengaruhi oleh masalah data tidak seimbang. Untuk mengatasi masalah ini, SMOTE akan diaplikasikan pada set data latihan. Ini juga membantu mengelakkan masalah *overfitting* pada set data latihan. SMOTE adalah teknik yang menggandakan sampel kelas minoriti untuk menyeimbangkan nisbah antara nilai-nilai dalam label kelas. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.30 berikut.

```
# Apply SMOTE to address dataset imbalance
sm = SMOTE(random_state = 42)
x_train_sm, y_train_sm = sm.fit_resample(x_train, y_train.ravel())
```

Rajah 3.30 Kod *Python* bagi Penggunaan SMOTE

3.8.2 Pembahagian 70:30 dengan Persampelan Terkurang

Kaedah Persampelan Terkurang berfungsi dengan cara yang bertentangan dengan kaedah SMOTE. Dalam kaedah Persampelan Terkurang, data dipilih berdasarkan nisbah yang dipaparkan dalam kelas yang tidak seimbang. Sebahagian data bagi kelas majoriti akan dikeluarkan manakala tiada perubahan akan diaplikasikan ke atas data bagi kelas minoriti. Tujuan pelaksanaan kaedah ini adalah untuk mengatasi isu data tidak seimbang yang terdapat dalam set data AI4I2020. Langkah ini juga diperkenalkan dalam projek ini sebagai cara untuk mengesahkan hasil yang diperoleh menggunakan pembahagian 70:30 dengan kombinasi teknik SMOTE. Terdapat pelbagai teknik dalam Persampelan Terkurang dan untuk projek ini, *RandomUnderSampler* akan digunakan dalam pemprosesan data. Kod yang akan digunakan bagi melaksanakan proses dalam langkah ini ditunjukkan di dalam Rajah 3.31 berikut.

```
# Apply Random Under Sampler to address dataset imbalance
rus = RandomUnderSampler(sampling_strategy='majority')
x_train_rus, y_train_rus = rus.fit_resample(x_train,
y_train.ravel())
```

Rajah 3.31 Kod *Python* bagi Penggunaan Persampelan Terkurang

3.9 PENGUKURAN PRESTASI MODEL

Kajian ini menggunakan keputusan daripada Skor-F1, Kejituan, Kepersisan, Kepekaan AUC untuk plot lengkung ROC dan AUC untuk plot lengkung Kepersisan-Kepekaan untuk membandingkan prestasi bagi model-model yang berbeza.

3.9.1 Skor-F1, Kejituan, Kepersisan dan Panggilan Semula

Kejituan adalah satu parameter yang menunjukkan sekerap mana model dapat meramal instans positif dan instans negatif dengan betul. Ia diwakili oleh persamaan berikut.

$$Kepersisan = \frac{TP + TN}{TN + FP + TP + FN} \quad \text{...(3.3)}$$
$$dimana\ TP = positif\ benar;\ FP = positif\ salah$$

Kepersisan adalah satu parameter yang menunjukkan sekerap mana model dapat meramal instans positif dengan betul dibandingkan dengan semua ramalan instans positif. Ia diwakili oleh persamaan berikut.

$$Kepersisan = \frac{TP}{TP + FP} \quad \text{...(3.4)}$$
$$dimana\ TP = positif\ benar;\ FP = positif\ salah$$

Kepekaan adalah satu parameter yang mengukur sejauh mana model dapat meramalkan instans positif dengan betul dibandingkan dengan instans positif yang betul dan instans negatif yang salah. Ia diwakili oleh persamaan berikut.

$$Kepekaan = \frac{TP}{(TP + FN)} \quad \text{...(3.5)}$$
$$dimana\ TP = positif\ benar;\ FN = negatif\ salah$$

Skor-F1 adalah satu parameter yang mengukur sesebuah model berdasarkan purata harmonik Kepersisan dan Kepekaan. Ia diwakili oleh persamaan berikut.

$$SkorF1 = \frac{2 \times (Kepersisan \times Kepekaan)}{(Kepersisan + Kepekaan)} \quad \ldots(3.6)$$

3.9.2 AUC untuk Lengkungan ROC dan Kepersisan-Kepekaan

Prestasi kesemua model pembelajaran mesin juga dapat dilihat melalui plot lengkungan ROC dan plot lengkungan Kepersisan-Kepekaan. AUC adalah pengukuran prestasi yang sesuai untuk menilai prestasi model-model tersebut. AUC untuk plot lengkungan ROC dapat menunjukkan prestasi keseluruhan model tanpa diganggu oleh isu data tidak seimbang kerana ia mengabaikan klasifikasi negatif, manakala AUC untuk plot lengkungan Kepersisan-Kepekaan akan menentukan prestasi model apabila terdapat isu data tidak seimbang terutamanya apabila label negatif menjadi kelas majoriti seperti di dalam set data AI4I2020.

3.10 KESIMPULAN

Setelah matlamat kajian iaitu ramalan kegagalan mesin telah dikenal pasti, set data yang bersesuaian untuk projek ini iaitu set data AI4I2020 telah diperolehi. Penerokaan data akan dilakukan ke atas set data yang diperolehi untuk menganalisis senarai atribut, jenis atribut, laporan kualiti data serta visualisasi data. Setelah langkah penerokaan data dilaksanakan, langkah persediaan data dan kejuruteraan fitur akan dijalankan di mana pembersihan data serta transformasi data termasuklah proses binarisasi melalui kejuruteraan fitur akan berlaku. Hasil daripada langkah tersebut adalah set data bersih dengan struktur set data yang baru dan berbeza berbading set data asal. Setelah set data bersih diperolehi, pembahagian data dengan nisbah set data latihan: set data ujian sebanyak 70:30 diaplikasikan ke atas set data bersih tersebut dengan dua teknik iaitu teknik SMOTE dan teknik Persampelan Terkurang. Tiga model latihan akan dibina iaitu model NB, model NB + *Bagging* dan model NB + *Boosting* dan kemudian diuji dengan menggunakan set data ujian di mana enam model ramalan atribut *Machine failure* akan diperolehi. Nilai-nilai ukuran prestasi model-model tersebut juga akan diperolehi bagi AUC lengkungan ROC, AUC lengkungan Kepersisan-Kepekaan, Skor-F1, Kejituan, Kepersisan dan Kepekaan. Langkah terakhir merupakan penilaian model yang akan dilaksanakan berdasarkan nilai-nilai ukuran prestasi model yang diperolehi dan model terpilih akan dianalisis dan dibandingkan dengan model-model dari kajian lepas.

BAB IV

HASIL KAJIAN

4.1 PENGENALAN

Bab ini akan membincangkan tentang hasil kajian dalam projek ini. Bab 4.2 akan menerangkan tentang set data bersih yang digunakan di dalam analisis pembelajaran mesin. Bab 4.3 dan bab 4.4 akan merumuskan keputusan bagi analisis menggunakan teknik SMOTE dan teknik Persampelan Terkurang berdasarkan nilai AUC bagi lengkungan ROC, AUC bagi lengkungan Kepersisan-Kepekaan, Skor-F1, Kejituan, Kepersisan dan Kepekaan. Bab 4.5 akan merumuskan hasil kajian bagi setiap model bagi setiap kaedah yang digunakan. Akhir sekali, bab 4.6 akan membuat perbandingan di antara model terbaik dari model-model yang telah dibina di dalam projek ini dengan model-model dari kajian lepas.

4.2 SET DATA BERSIH

Struktur set data yang bersih setelah melalui proses persediaan data serta kejuruteraan fitur mempunyai struktur data yang berbeza berbanding set data yang asal. Set data bersih mempunyai 35 atribut manakala set data yang asal mempunyai 14 atribut. Selain itu, set data bersih hanya mempunyai satu jenis ciri sahaja iaitu binari manakala set data yang asal mempunyai pelbagai jenis ciri. Struktur set data yang bersih yang digunakan untuk projek ini dirumuskan di dalam Jadual 4.1 berikut.

Jadual 4.1 Struktur Data Bersih

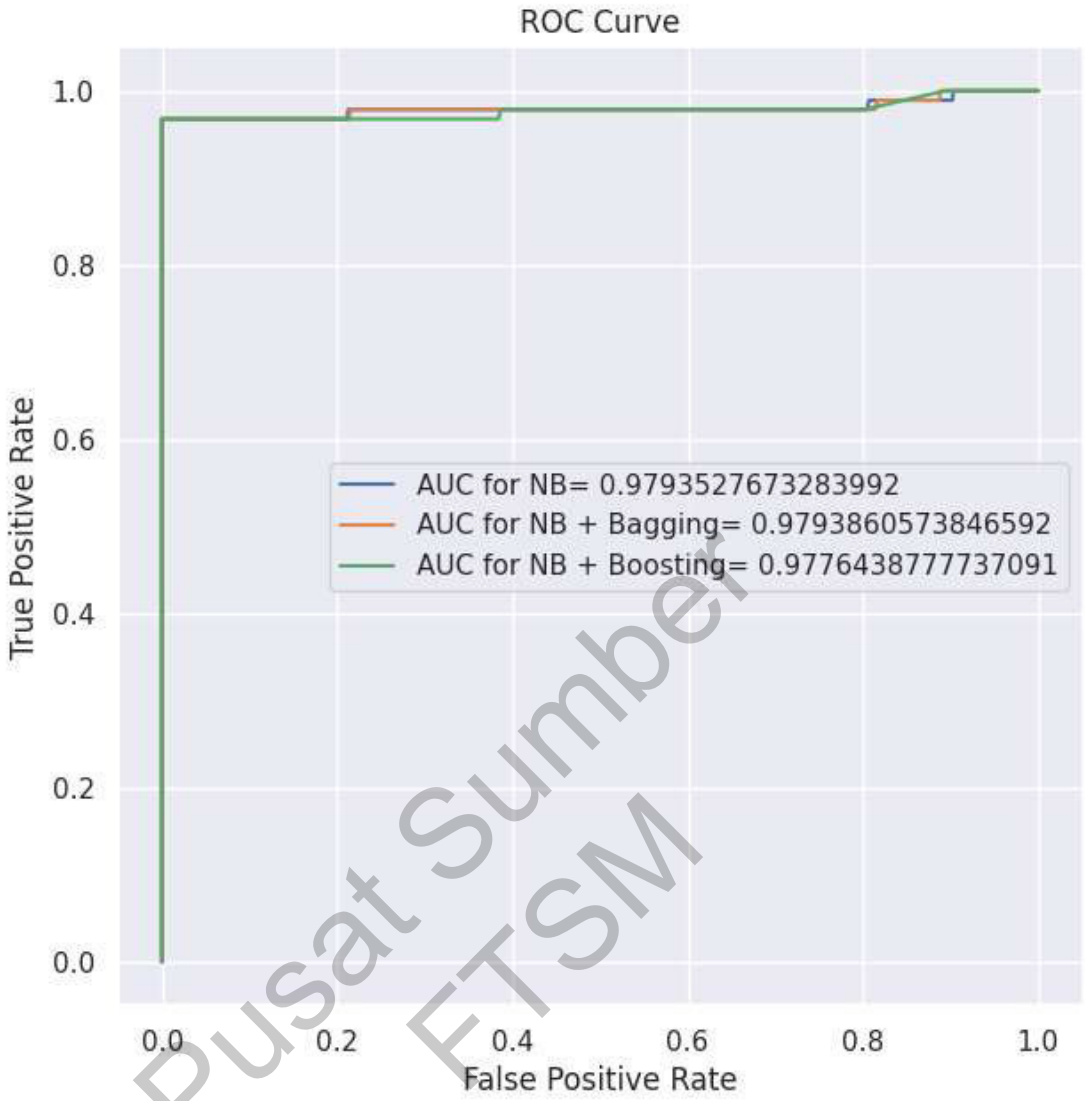
Atribut	Ciri	Penerangan
TWF	Binari	Kegagalan kehausan alatan
HDF	Binari	Kegagalan pelepasan haba
PWF	Binari	Kegagalan kuasa
OSF	Binari	Kegagalan ketegangan
RNF	Binari	Kegagalan rawak
H	Binari	Jenis H = tinggi
L	Binari	Jenis L = rendah
M	Binari	Jenis M = sederhana
bin_AT0	Binari	Bin 1 bagi nilai suhu udara
bin_AT1	Binari	Bin 2 bagi nilai suhu udara
bin_AT2	Binari	Bin 3 bagi nilai suhu udara
bin_AT3	Binari	Bin 4 bagi nilai suhu udara
bin_AT4	Binari	Bin 5 bagi nilai suhu udara
bin_PT0	Binari	Bin 1 bagi nilai suhu operasi
bin_PT1	Binari	Bin 2 bagi nilai suhu operasi
bin_PT2	Binari	Bin 3 bagi nilai suhu operasi
bin_PT3	Binari	Bin 4 bagi nilai suhu operasi
bin_PT4	Binari	Bin 5 bagi nilai suhu operasi
bin_TW0	Binari	Bin 1 bagi nilai kehausan alatan
bin_TW1	Binari	Bin 2 bagi nilai kehausan alatan
bin_TW2	Binari	Bin 3 bagi nilai kehausan alatan
bin_TW3	Binari	Bin 4 bagi nilai kehausan alatan
bin_TW4	Binari	Bin 5 bagi nilai kehausan alatan
bin_TW5	Binari	Bin 6 bagi nilai kehausan alatan
bin_PW0	Binari	Bin 1 bagi nilai kuasa
bin_PW1	Binari	Bin 2 bagi nilai kuasa
bin_PW2	Binari	Bin 3 bagi nilai kuasa
bin_PW3	Binari	Bin 4 bagi nilai kuasa
bin_PW4	Binari	Bin 5 bagi nilai kuasa
bin_PW5	Binari	Bin 6 bagi nilai kuasa
bin_PW6	Binari	Bin 7 bagi nilai kuasa
bin_PW7	Binari	Bin 8 bagi nilai kuasa
bin_PW8	Binari	Bin 9 bagi nilai kuasa
bin_PW9	Binari	Bin 10 bagi nilai kuasa
Machine Failure	Binari	Indikasi status mesin sama ada gagal atau masih

4.3 HASIL KAJIAN UNTUK KAEDAH SMOTE

Bab ini membincangkan tentang hasil kajian dengan penggunaan kaedah SMOTE bagi semua model kajian. Bab 4.3.1 menunjukkan nilai AUC bagi lengkungan ROC berserta dengan plot lengkungan ROC. Bab 4.3.2 pula menunjukkan nilai AUC bagi lengkungan Kepersisan-Kepekaan berserta dengan plot lengkungan Kepersisan-Kepekaan. Bab 4.3.3 membincangkan nilai skor-F1, kejituan, kepersisan dan kepekaan. Di samping itu, bab 4.3.3 juga mengandungi plot Matriks Kekeliruan bagi menggambarkan hasil ramalan kajian.

4.3.1 Keputusan AUC bagi Lengkungan ROC

Rajah 4.1 menunjukkan bahawa nilai AUC bagi lengkungan ROC untuk ketiga-tiga model tidak menunjukkan perbezaan yang ketara di mana nilai yang diperolehi bagi kesemua model kajian adalah sekitar 0.98. Nilai ini boleh dikategorikan sebagai nilai yang tinggi kerana ia hampir kepada had nilai penuh iaitu 1.0. Ini menunjukkan bahawa ketiga-tiga model kajian adalah model klasifikasi yang baik bagi set data sintetik yang digunakan di dalam projek ini.



Rajah 4.1 Plot Lengkungan ROC bagi Kaedah SMOTE

4.3.2 Keputusan AUC bagi Lengkungan Kepersisan-Kepekaan

Rajah 4.2 menunjukkan bahawa nilai AUC bagi lengkungan Kepersisan-Kepekaan untuk ketiga-tiga model tidak menunjukkan perbezaan yang ketara di mana nilai yang diperolehi kesemua model kajian adalah sekitar 0.97. Nilai ini boleh dikategorikan sebagai nilai yang tinggi kerana ia hampir kepada had nilai penuh iaitu 1.0. Ini menunjukkan bahawa ketiga-tiga model kajian adalah model klasifikasi yang baik bagi set data sintetik yang digunakan di dalam projek ini.