RAMALAN DAN KLASIFIKASI MENGGUNA MESIN VEKTOR SOKONGAN TERHADAP SISTEM PEMBACAAN METER KAWALAN JAUH

SITI HAJAR BINTI IBRAHIM

UNIVERSITI KEBANGSAAN MALAYSIA

RAMALAN DAN KLASIFIKASI MENGGUNA MESIN VEKTOR SOKONGAN TERHADAP SISTEM PEMBACAAN METER KAWALAN JAUH

SITI HAJAR BINTI IBRAHIM

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2023

PENGAKUAN

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

18 Ogos 2023

SITI HAJAR BINTI IBRAHIM
P112478

PENGHARGAAN

Alhamdulillah, setinggi-tinggi kesyukuran dipanjatkan kehadrat Ilahi dengan izin dan kurnianNya dapat saya menyempurnakan tugasan kajian bagi kod projek TTTU607C untuk memenuhi syarat Ijazah Sarjana Modul (Data Sains) dalam tempoh masa yang ditetapkan.

Dengan kesempatan yang ada ini saya amat berbesar hati untuk mengucapkan jutaan terima kasih kepada Prof. Madya Dr. Shahnorbanun Sahran, selaku pensyarah saya kerana telah meletakkan sepenuh kepercayaan beliau kepada saya untuk menyiapkan tugasan kajian ini. Ucapan terima kasih juga kepada beliau atas budi bicara dalam memberi tunjuk ajar sepanjang tugasan kajian ini dijalankan. Selain itu, sekalung penghargaan buat semua pensyarah saya di Fakulti Teknologi & Sains Maklumat dan pihak Universiti Kebangsaan Malaysia (UKM) sepanjang pengajian saya di sini.

Saya juga ingin mengucapkan ribuan terima kasih terutamanya kepada pihak majikan saya, kerana menyokong dan membenarkan kajian dilakukan dengan sebaiknya. Selain itu, saya turut berterima kasih kepada rakan-rakan seperjuangan saya kerana telah banyak menghulurkan bantuan dan kerjasama bagi merealisasikan usaha menyempurnakan tugasan ini dengan jayanya.

Ucapan terima kasih ini juga ditujukan khas buat kedua ibubapa saya yang telah banyak membantu saya, juga buat suami tercinta yang menyokong serta ahli keluarga saya yang sentiasa menjadi tulang belakang kejayaan saya selama ini. Saya dedikasikan penyelidikan kajian saya ini kepada suami Abdul Hakim Bin Adzman, dan anak-anak tersayang Umair Khairi, Zayd Khairi, Hanaa Yusra dan Siddiq Khairi.

Akhir sekalian, ucapan ini ditujukan kepada semua pihak yang telah terlibat dalam menjayakan tugasan ini sama ada secara langsung atau tidak langsung. Segala bantuan yang telah dihulurkan, amatlah saya hargai dan akan dikenang jasa baik kalian semua. Terima kasih.

ABSTRAK

Pendekatan kepada inovasi global hari ini, menyediakan beberapa peningkatan dan analisis yang lebih memberi tumpuan kepada perkembangan penggunaan pembacaan meter jauh (RMR) dalam sistem grid tenaga. Terdapat beberapa perkembangan dalam tahun-tahun kebelakangan ini yang telah memberi tumpuan kepada ramalan prediksi kegagalan dalam sistem meter RMR ini. Ia adalah penting untuk mendapatkan tahap ketahanan dan kebolehpercayaan yang dikehendaki supaya sistem berfungsi dengan sebaiknya. Walau bagaimanapun, proses analisa yang dilakukan, mendapati pengendalian pemantauan masalah secara tradisional telah menunjukkan kekurangan dari segi pengurusan masa, kos dan penggunaan kaedah yang tidak komprehensif yang menyebabkan solusi permasalahan kurang dapat ditangani. Kepentingan kepada penambahbaikan kajian adalah perlu dengan melihat penggunaan teknologi pembacaan meter ini semakin berkembang yang mana secara tidak langsung menyumbang kepada peningkatan kegagalan jumlah operasi meter kerana rangkaian data yang semakin kompleks. Oleh itu, sains data digunakan untuk mengekstrak maklumat bersih daripada data mentah untuk perolehan pandangan yang boleh dilakukan. Kajian ini telah mengesyorkan bahawa kaedah baru untuk pemeriksaan ketersediaan profil beban data dalam sistem meter RMR perlu dianalisis lebih dalam bagi mendapatkan prestasi yang lebih baik. Daripada metodologi asas di dalam kajian ini, mengetengahkan tiga perspektif peringkat analisis utama yang merupakan analisis deskriptif, prediktif, dan preskriptif. Dalam kajian ini, model pembelajaran pengawasan dengan mesin vektor sokongan (SVM) utama dijalankan bersama-sama dengan pelbagai model perbandingan regresi dan klasifikasi untuk mencadangkan model terbaik yang berguna dalam menjejaki tugas ketersediaan data di masa depan. Sebahagian daripada analisis siri masa juga dijalankan dengan bantuan model SARIMA. Hasil eksperimen yang dijalankan pada data daripada simulasi menunjukkan bahawa strategi yang dibincangkan adalah tepat dan boleh dipercayai dalam hal meningkatkan prestasi kajian ini. Keberkesanan hasil kajian ini secara tidak langsung memberikan suatu sumbangan kepada sektor utiliti tenaga sekaligus membangunkan potensi analisa berkesan kepada perkembangan penyelidikan yang dilakukan hari ini.

PREDICTIVE AND CLASSIFICATION METHOD USING SUPPORT VECTOR MACHINE FOR REMOTE METER READING SYSTEM

ABSTRACT

The approach to today's global innovation, provides some improvements and analyses that are more focused on the development of the use of remote meter readings (RMR) in power grid systems. There have been several developments in recent years that have focused on predicting failure forecasts in this RMR meter system. It is important to obtain the level of durability and reliability required so that the system works optimally. However, the analysis process carried out, found that the operation of problem monitoring traditionally has shown shortcomings in terms of time management, costly and the use of incomprehensive methods that cause problem solutions to be less manageable. Interest in improving the study is necessary by seeing the increasing use of this meter reading technology that conjunction with the increasing of failure numbers in the operation meters due to an increasingly complex data network. Therefore, data science is used to extract clean information from raw data for achievable insights. The study has suggested that new methods for checking the availability of data load profiles in RMR meters should be analyzed more deeply in order to better performance. From the basic methodology in this study, it outlines three main level analysis perspectives that are descriptive, predictive, and prescriptive analysis. In this study, a supervised learning model with a primary support vector machine (SVM) was executed together with a variety of regression and classification comparison models to suggest the best models useful in tracking future data availability tasks. Part of the time series analysis is also carried out with the help of the SARIMA model. The results of experiments conducted on data from the simulation showed that the strategies discussed were accurate and reliable in terms of improving the performance of this study. The effectiveness of the results of this study indirectly contributes to the energy utility sector while developing the potential for effective analysis of today's research.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

AMRA	Automatic Meter Reading Association
AUC	Area Under the Curve
CART	Classification and Regression Trees
CFNN	Cascade-Forward Backpropagation Neural Network
CFSVR	Clifford Fuzzy Support Vector Regression
CNN	Convolution Neutral Network
CSVR	Clifford Support Vector Regression
DAG	Directed acyclic graph
EDA	Exploratory data analysis
EUROAMRA	Europe Automation Meter Reading Association
FDIA	False Data Injection Attack
FFNN	Feedforward Backpropagation Neural Network
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GGSN	Gateway GPRS support node
GIS	Geographic Information System
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
ID	Identification
IoT	Internet of Things
KDD	Knowledge discovery in databases
KNN	K-Nearest Neighbour
KNNR	K-Nearest Neighbour Regression
LR	Linear Regression

LSTM	Long Short-Term Memory Network
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MRU	Meter reading units
MSE	Mean Squared Error
NB	Naive Bayes
PLC	Power-line communication
R^2	R-Squared
RF	Random Forest
RFR	Random Forest Regression
RMR	Remote Meter Reading
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
RR	Ridge Regression
RSSI	Received Signal Strength Indication
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMA-LSTM	Seasonal Autoregressive Integrated Moving Average Long-Short Term Memory neural network
SGSN	Serving GPRS Support Node
SMOGN	Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-IPF	Synthetic Minority Oversampling Technique Iterative-Partitioning Filter
SMOTE-LOF	Synthetic Minority Oversampling Technique Local Outlier Factor

SMOTER	Synthetic Minority Over-sampling with Enhanced Regularization
SSA-SARIMA -LSTM	SARIMA-LSTM based on Singular Spectrum Analysis
SVC	Support Vector Classification
SVC LINEAR	Support Vector Classification Linear
SVC POLY	Support Vector Classification Polynomial
SVC RBF	Support Vector Classification Radial Basic Frequency
SVC SIGMOID	Support Vector Classification Sigmoid
SVM	Support Vector Machine
SVM-GAB	Support Vector Machine Gentle-Adaboost
SVM-SMOTE	Support Vector Machine Synthetic Minority Oversampling Technique
SVR	Support Vector Regression
SVR LINEAR	Support Vector Regression Linear
SVR POLY	Support Vector Regression Polynomial
SVR RBF	Support Vector Regression Radial Basic Frequency
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
TTP	Trusted third party
UKAMRA	United Kingdom Automation Meter Reading Association
VC	Vector Classification
WIFI	Wireless fidelity
2G	Second-generation of wireless mobile telecommunications technology

3.5G	Third-generation of wireless mobile telecommunications technology
4G	Fourth generation of wireless mobile telecommunications technology
5G	Fifth-generation of wireless mobile telecommunications technology

BAB I

PENGENALAN

1.1 LATAR BELAKANG KAJIAN

Kelestarian tenaga dan perlindungan alam sekitar telah menjadi kebimbangan global, yang membangkitkan pembangunan pelaksanaan dasar baru untuk menggalakkan kecekapan tenaga dalam sektor elektrik dengan salah satu keutamaan pembacaan kawalan meter jauh iaitu *Remote Meter Reading* (RMR) seperti yang dinyatakan oleh Stoyanov et al (2021). Pendedahan teknologi global hari ini menjadikan pelbagai transformasi digital yang semakin berkembang pesat sejajar dengan keupayaan dan fungsinya tersendiri. Metodologi baru yang telah mendapat daya tarikan yang signifikan adalah penggunaan pembacaan meter kawalan jauh yang dikendalikan oleh pelbagai sektor hari ini. Dengan pendekatan transformasi ini, menghasilkan banyak faedah, termasuk cara yang lebih sistematik, berdaya tahan, dan relevan terhadap keupayaan dalam meningkatkan hasil dinamik ke atas pangkalan pengetahuan proses kerja. Salah satu sektor industri yang telah dipengaruhi oleh transformasi digital ialah rangkaian elektrik pintar. Ianya telah dicapai melalui pemasangan pembacaan meter kawalan jauh, yang disambungkan melalui peranti komunikasi dalam sistem untuk memperoleh data mengenai maklumat penggunaan. Penyebaran sistem grid elektrik telah membawa kepada pembangunan rangkaian pintar, yang kini meluas di kebanyakan negara berkembang di seluruh dunia. Implementasi rangkaian pintar telah membawa kepada perubahan yang signifikan dalam kaedah pengeluaran, pengedaran dan pemantauan elektrik, yang membawa kepada peningkatan ketersediaan data. Sehubungan itu, aplikasi meter RMR boleh dikenali sebagai sistem yang melibatkan integrasi meter

pintar, komunikasi rangkaian, dan sistem pengurusan data dalam pelaksanaan rantaian pintar. Penggunaan meter RMR menghasilkan ketersediaan data masa nyata, yang memudahkan penubuhan proses yang berlanjutan yang mempromosikan kecekapan tenaga. Kemajuan teknologi memudahkan pertukaran maklumat antara utiliti korporat dan pengguna akhir dapat dilakukan.

Meter RMR adalah peranti elektronik yang berfungsi sebagai komponen utama dalam sistem RMR, bertanggungjawab untuk mengukur kuantiti penggunaan kuasa dan maklumat sistem peranti. RMR meter mampu menghantar data kepada syarikat utiliti secara autonomi, tanpa memerlukan sebarang campur tangan manual atau pengawasan secara mikro-tingkat. Kelebihan RMR meter melampaui keperluan operasi mereka, kerana data yang melimpah yang dihasilkan oleh peranti pintar ini boleh digunakan oleh pelbagai keperluan penyelidikan hari ini. Pendekatan ini bertujuan untuk menentukan penggabungan optimal teknologi baru yang boleh secara berkesan dalam memudahkan pelaksanaan strategi perniagaan, menilai atribut yang mempengaruhi penggunaan tenaga domestik, dan menawarkan penyelesaian pemeliharaan ramalan dalam konteks automasi industri. Baru-baru ini, tetapan *Internet of Things* (IoT) telah menyaksikan penggunaan sistem elektrik. Peranti ini menunjukkan keupayaan untuk melaksanakan fungsi bacaan, penangkapan maklumat, dan rekod data, yang kemudian dihantar ke pusat data kawalan pada tempoh yang telah ditentukan, mengikut jadual rutin yang ditetapkan. Sehubungan dengan itu, pelbagai kajian dilakukan dalam menganalisis keperluan fungsi meter RMR ini di mana boleh dikaitkan dengan kajian Y. Wang et al. (2019). Konsep yang diusulkan untuk sistem bacaan meter yang dikendalikan jauh melibatkan penyediaan peranti komunikasi untuk mengautomasikan proses bacaan setiap peranti meter elektrik yang digunakan. Integrasi metodologi konvensional dengan teknologi digital kontemporari mempunyai potensi untuk meningkatkan prosedur dalaman dan mengurangkan ketidaktepatan semasa memperoleh atau mendokumentasikan maklumat penting. Implementasi sistem ini membolehkan pengenalan ciri-ciri penting dalam data mentah yang diperolehi daripada pelbagai peranti, yang sebelum ini tidak boleh dicapai. Dalam konteks tertentu ini, penggunaan istilah pengambilan data terbukti relevan kerana ia berfungsi sebagai metodologi yang berharga untuk mengekstrak wawasan penting daripada data yang tidak diproses. Maklumat ini kemudian boleh digunakan untuk memaklumkan proses pengambilan

keputusan dan memudahkan analisis pengetahuan berkaitan dengan isu-isu semasa. Sistem pengesanan jarak jauh digunakan bukan sahaja untuk merekodkan bacaan data penting tetapi juga untuk menjalankan analisis yang komprehensif. Analisis ini boleh dikategorikan kepada tiga peringkat iaitu analisis deskriptif, yang bertujuan untuk menyediakan analisis maklumat yang hadir dalam data mentah, analisis prediktif yang menjelaskan ramalan hasil data, dan analisis preskriptif yang menyediakan keputusan berdasarkan hasil akhir dan menawarkan cadangan untuk meningkatkan proses yang sedia ada. Kajian ini bertujuan untuk mengkaji kecekapan operasi dan melaksanakan teknik baru untuk mengesan potensi masalah dalam sistem peranti di dalam memenuhi keperluan masa nyata. Daripada mengintegrasikan data baru melalui sistem peranti, teknik prediktif yang sedia ada boleh digunakan untuk memanfaatkan maklumat semasa. Topik utama perbincangan di kalangan kajian ramalan dan pembelajaran berkaitan dengan penggunaan pembelajaran *Support Vector Machine* (SVM) bersama dengan model algoritma yang lain sebagai model analisis. Teknik-teknik yang telah dicadangkan dianggap berkesan dalam kedua-dua senario prediktif dan klasifikasi, dengan itu menjadikan ianya sesuai untuk usaha penyelidikan pada masa hadapan. Dengan menggunakan kaedah analisis yang dinyatakan di atas, ianya membolehkan proses menganalisis data sistem secara automatik untuk tujuan menjalankan penilaian yang lebih menyeluruh ke atas sistem rangkaian yang sedia ada. Pengurusan yang lebih baik dan peningkatan konsistensi dalam tindak balas segera akan dicapai. Pendekatan baru yang dipaparkan dalam kajian ini membolehkan pengesanan kaedah-kaedah baru yang memudahkan pengenalan segera isu-isu dalam penyesuaian dan mengenal pasti variasi masalah yang ada.

1.2 PERMASALAHAN KAJIAN

Penggunaan sistem pembacaan meter jauh, terutamanya dalam industri elektrik, telah menjadi subjek di dalam kajian. Walaubagaimanapun, kebanyakan penganalisis hanya mencatat dan menganalisis hasil dengan cara yang umum, tanpa menggunakan ramalan analitik yang menggunakan kedua-dua pendekatan pengelasan dan kaedah ramalan biasa. Penyelidikan terdahulu telah memberi tumpuan terutamanya kepada fungsi-fungsi tunggal, seperti klasifikasi atau ramalan, dan tidak menyeluruh mengkaji penggunaan langkah-langkah analitik yang berkaitan dengan metodologi deskriptif, prediktif, dan preskriptif untuk memeriksa masalah secara komprehensif bagi

meningkatkan usaha pemantauan. Kajian oleh Yang, Zhou (2020) ini, menunjukkan sebuah kajian tunggal klasifikasi dengan menjalankan kajian kaedah terhadap bacaan meter selamat berdasarkan pembangunan sistem mekanisme perkongsian kunci dan disokong oleh algoritma *Support Vector Machine* (SVM). Dalam kajian ini, pendekatan baharu dicadangkan iaitu pemindahan data daripada meter pintar ke pusat kawalan tanpa melibatkan pihak ketiga kerana *Trusted third party* (TTP) tidak lagi selamat untuk menghantar dan menyimpan rekod data. Bagi pelaksanaan kaedah penghantaran terus data daripada meter pintar, pengelasan daripada kedua pihak data pengguna dan data di pusat kawalan perlu dilakukan untuk memastikan ketepatan makluman data yang diperolehi. Hasil daripada kajian ini, menunjukkan sistem bacaan meter pintar tanpa pihak ketiga adalah boleh dipercayai apabila melakukan kaedah analisis di kedua bahagian data. Pertamanya data dihantar ke data di pusat kawalan dengan menggunakan analisis algoritma Shamir dan algoritma Laplace tanpa pihak ketiga yang melakukannya. Di bahagian pengguna, algoritma mesin vektor sokongan digunakan untuk mengklasifikasikan ketulenan data, dan mencari kegagalan meter serta rompakan elektrik sekiranya ada. Keputusan menunjukkan bahawa algoritma yang dicadangkan tersebut mempunyai ketepatan klasifikasi yang lebih tinggi dan kecekapan pelaksanaan yang lebih tinggi. Kajian ini jelas menunjukkan algoritma mesin vektor sokongan dapat membantu dalam aktiviti klasifikasi dengan baik namun kajian tidak menggunakan pendekatan ramalan menggunakan algoritma yang sama. Sekiranya kaedah-kaedah tersebut tidak hadir atau tidak mencukupi, penyelidikan di dalam kajian ini berusaha untuk menggunakan rangka kerja analisis perniagaan, yang menggunakan teknik analisis yang menggabungkan pengurusan pengetahuan, untuk membezakan dan menilai faktor dalaman dan luaran. Salah satu daripada tugas utama adalah untuk menilai proses pemantauan untuk kegagalan pembacaan data meter kawalan jauh, dengan tumpuan kepada kedua-dua pemasangan di tapak dan kemas kini data dalam sistem. Penggunaan data yang didaftarkan membolehkan penggunaan teknik analisis statistik untuk menghasilkan ramalan dan wawasan yang boleh digunakan untuk meningkatkan sistem analitik statistik semasa. Pemilihan teknik yang sesuai bergantung kepada sifat penyelidikan yang akan dijalankan. Kajian oleh Doaa A.Bashawyah (2021) menerangkan kajian beban jangka pendek berdasarkan pembelajaran mesin ramalan untuk penggunaan tenaga meter pintar data dalam isi rumah London. Konsep kajian adalah berdasarkan pembangunan model pembelajaran mesin yang digunakan ialah

jiran terdekat *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM). Pendekatan kajian ini adalah bertujuan untuk menyediakan aplikasi jangka masa untuk membuat anggaran profil beban pelanggan yang dipengaruhi oleh beberapa faktor seperti tingkah laku pengguna, aplikasi operasi, masa dalam sehari, tempoh cuti, keadaan cuaca dan lain-lain lagi. Istilah ramalan yang digunakan dikategorikan dari segi ramalan jangka pendek, ramalan beban jangka sederhana dan ramalan beban jangka panjang. Oleh itu, beberapa teknik diaplikasikan dalam kajian ini dengan menggunakan teknik statistik bagi menentukan hubungan antara atribut beban dan input, analisis siri masa dan sistem kabur menggunakan istilah pembelajaran mesin. Dapatan kajian ini menyebut bahawa aplikasi pengelas SVM menunjukkan prestasi yang baik berbanding dengan pengelas KNN berdasarkan prestasi *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE) di mana SVM memberikan pengurangan MAPE. Hasil kajian menunjukkan SVM memberikan prestasi yang lebih baik dalam ramalan profil beban tenaga dalam kajian ini. Oleh itu, kajian menyeluruh kaedah statistik terdahulu dijalankan untuk menentukan pendekatan yang paling sesuai untuk setiap contoh analisis data yang sediada. Selepas itu, tahap pemprosesan data melibatkan pengklasifikasian keadaan data melalui penciptaan visualisasi data. Laporan dan visualisasi data digunakan sebagai alat komunikasi.

Proses analisis data melibatkan kajian hipotesis prediktif menggunakan kaedah kuantitatif dan mengenal pasti potensi untuk pembangunan pembelajaran mesin melalui model pengelasan dan ramalan. Penyebab kebiasaan kegagalan ketidakupayaan pembacaan meter jarak jauh perlu dipelajari untuk memastikan kesilapan dalam operasi meter tersebut dapat diambil tindakan yang sesuai dan faktor kegagalan boleh dikenal pasti. Beberapa faktor telah dilihat sebagai punca potensi kegagalan ketersediaan data dalam meter ini, termasuk perbezaan dalam konfigurasi antara pemasangan baru dan sedia ada, perbezaan rangkaian komunikasi, dan kesilapan dalam tarikh rekod bacaan akhir. Laporan ini merangkumi kajian analisis untuk mengkaji maklumat data melalui pernyataan kegagalan dan tempoh kesilapan ketersediaan data, di samping mengklasifikasi isu-isu tambahan yang berlaku dalam proses analisis semasa. Teknik yang digunakan dalam kajian ini termasuk analisis kuantitatif, pemodelan dinamik, dan penciptaan model pembelajaran mesin yang pelbagai untuk dibandingkan dengan *Support Vector Classification* (SVC) dan *Support Vector Regression* (SVR). Hasil

analisis digunakan untuk mencadangkan kaedah alternatif yang meningkatkan penemuan awal, sambil juga menggabungkan maklumat baru untuk memenuhi keperluan semasa.

Dengan kepelbagaian bidang analisis dan ramalan yang semakin maju, ia telah membawa kepada haluan pelbagai cabaran baru. Terutamanya, terdapat penekanan yang semakin meningkat kepada menjalankan analisis komprehensif menggunakan teknik pengukuran data terkini. Kajian ini bertujuan untuk meneroka pendekatan pelbagai analisis yang boleh digunakan dalam konteks teknologi meter. Kajian ini mengkaji pelbagai faktor yang berkaitan dengan isu-isu asas, seperti pengenalan aktiviti yang mengakibatkan kegagalan dalam ketersediaan data lengkap. Selain itu, penyelidikan mengenai korelasi antara setiap atribut dan hubungan antara mereka boleh diperluaskan. Penyelidikan ditambah dengan kajian metodologi yang sesuai untuk mengkategorikan dan meramalkan kajian, untuk menilai sejauh mana teknik konvensional yang sedia ada digunakan digantikan dengan teknik-teknik baru yang boleh meningkatkan usaha penyelidikannya.

1.3 OBJEKTIF KAJIAN

Penyelidikan di dalam kajian ini adalah membincangkan objektif utama kajian yang membincangkan berkenaan matlamat dan tujuan kajian ini dilakukan. Objektif utama kajian ini menggariskan empat elemen utama kajian seperti dinyatakan di bawah:

1. Merangka sebuah kerangka kerja diagnostik yang bersepadu pada setiap fasa kajian dengan mengikut kepada pendekatan model yang dibangunkan.

2. Membina model analisa kajian awal ke atas data mentah untuk memahami dan mengenal pasti keseluruhan set data dalam meningkatkan kefahaman isu kajian.

3. Melakukan penilaian hasil kajian ramalan regresi dan klasifikasi dalam membentuk sebuah keputusan hasil akhir yang komprehensif.

4. Menjalankan analisis yang boleh meramalkan corak masa ke hadapan yang dicirikan oleh had masa dalam pembangunan kaedah yang lebih optimum.

1.4 SKOP KAJIAN

Untuk menangani permasalahan yang dinyatakan di atas, kajian ini memfokuskan penyelidikan berkenaan isu ketersediaan data profil beban pengguna yang diperolehi oleh pembacaan meter RMR berdasarkan jumlah maklumat bilangan selang masa (*interval*) dengan selang masa penuh sebanyak 48 *interval* sehari. Perolehan data profil pengguna yang dicapai oleh meter RMR ini akan dihantar kepada pusat pengumpulan data meter melalui rangkaian komunikasi tanpa wayar pada setiap selang masa iaitu 30minit. Pembacaan meter RMR yang berjaya memperoleh data profil beban pengguna dengan bilangan selang masa penuh ini, akan dilakukan proses pengebilan dengan lebih tepat. Sebaliknya, kegagalan pembacaan meter RMR memperoleh data profil beban yang lengkap menyebabkan kehilangan jumlah selang masa sebenar berkurang di mana menjadi antara tanda ciri kegagalan operasi meter RMR dalam memperoleh data penuh dan menghantar data profil beban pengguna ke pusat data.

Oleh hal yang sedemikian, skop kajian ini merangkumi data perolehan mengikut status ketersediaan data yang telah dikumpulkan dari 2006 sehingga 2022 dengan saiz sampel iaitu 122,053. Pembinaan rangka kerja diagnostik di dalam kajian ini adalah merangkumi pelbagai dimensi tindakan, termasuk pemprosesan maklumat data, mengenal pasti kegagalan operasi menggunakan pendekatan model yang sesuai, dan mengantisipasi masalah masa depan yang berpotensi melalui ramalan. Apabila tugasan mengidentifikasi kerosakan dalam operasi dilakukan, model ramalan sistem dibangunkan untuk membolehkan pelaksanaan tindakan yang bertujuan untuk mengenal pasti ciri-ciri tambahan peranti yang berkaitan dengan klasifikasi hasil sasaran. Kajian dilakukan dengan merangkumi tiga peringkat yang berbeza iaitu analisis deskriptif, analisis prediktif, dan analisis preskriptif. Analisis deskriptif melibatkan pengumpulan maklumat data dan penerangan penerokaan data. Analisis prediktif melibatkan proses pembinaan model ramalan semasa mengenai berdasarkan senario sejarah yang diperoleh. Akhirnya, analisis preskriptif melibatkan menilai kesimpulan yang boleh diambil daripada maklumat hasil yang sedia ada berdasarkan ramalan hasil pada masa hadapan dan tindakan yang sewajar yang patut diambil.

Jadual 1.1 Eleman skop kajian dan penerangannya

Bilangan	*Skop kajian*	*Penerangan*
1	Penetapan Sumber data	Perolehan sumber data mentah dilakukan melalui integrasi hasil perbincangan dan penelitian isu oleh pakar domain.
2	Reka kerangka kerja berperingkat & spesifikasinya	Penyusunan dan perangkaan isu dengan gabungan aliran kerja berperingkat iaitu desktiptif, prediktif dan preskriptif bersama dengan spesifikasinya.
3	Pembangunan model analisa pada setiap fasa	Pembinaan model algoritma dilakukan mengikut peranan setiap fasa analisa yang diperlukan.
4	Penghasilan keputusan akhir kajian	Dapatan hasil dan penilaian keputusan bagi menyediakan keputusan prestasi yang terbaik.

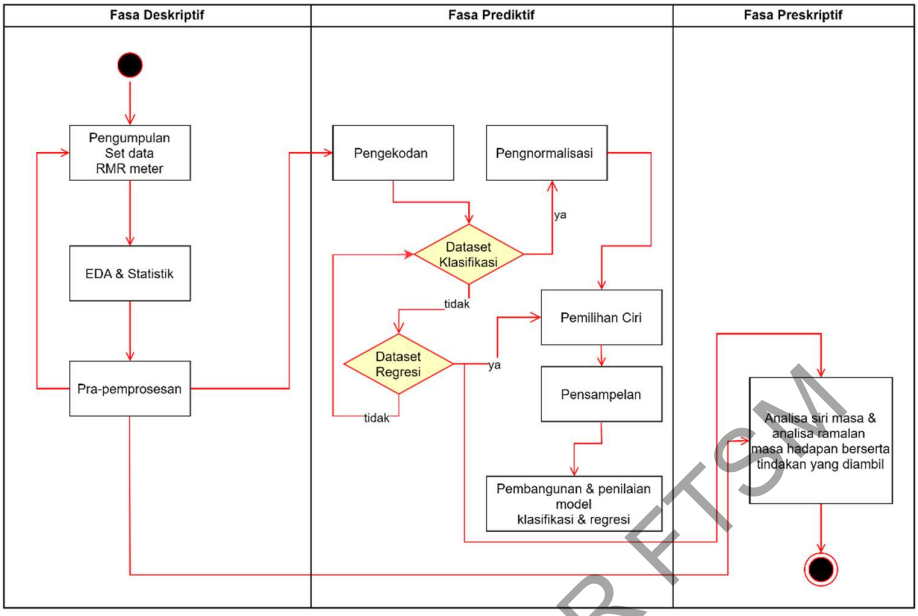
1.5 PERSOALAN KAJIAN

Di dalam menjayakan matlamat kajian, beberapa persoalan penting dikenalpasti dengan melihat kepada persoalan bagaimana kerangka kerja kajian ini dapat dijalankan. Berikut ialah pertanyaan-pertanyaan yang dikenalpasti bagi pembangunan kajian ini.

Jadual 1.2 Objektif dan persoalan kajian

Bilangan	*Objektif kajian*	*Persoalan Kajian*
1	Merangka sebuah kerangka kerja diagnostik yang bersepadu pada setiap fasa kajian dengan mengikut kepada pendekatan model yang dibangunkan.	Di awal perancangan projek, apakah peringkat kajian yang perlu ditetapkan? Apakah matlamat menjalankan kajian mengenai analisis ketersediaan data?
2	Membina model analisa kajian awal ke atas data mentah untuk memahami dan mengenal pasti keseluruhan set data dalam meningkatkan kefahaman isu kajian.	Apakah analisa yang sesuai dilakukan pada awal kajian dan bagaimana analisa ini dapat membantu memahami keseluruhan set data ?
3	Melakukan penilaian hasil kajian ramalan regresi dan klasifikasi dalam membentuk sebuah keputusan hasil akhir yang komprehensif.	Bagaimana membentuk model yang sesuai untuk dilakukan analisis? Adakah perbandingan beberapa pendekatan model lain perlu dilakukan bersama?
4	Menjalankan analisis yang boleh meramalkan corak masa ke hadapan yang dicirikan oleh had masa dalam pembangunan kaedah yang lebih optimum.	Apakah analisis akhir yang boleh dilakukan melalui keputusan yang dihasilkan dalam menguruskan penyediaan data bacaan meter kawalan jauh ini untuk beroperasi dengan lebih baik pada masa hadapan?

1.6 METODOLOGI KAJIAN



Rajah 1.1 Kerangka kerja bagi fasa kajian ketersediaan data meter RMR

Berdasarkan penerangan matlamat dan halatuju kajian ini, berikut adalah keterangan proses bagi setiap fasa kajian diterangkan dalam Rajah 1.1. Ringkasan metodologi kajian ditunjukkan mengikut fasa analisa yang dicadangkan bagi melakukan kajian mengikut sasaran objektif yang ditetapkan. Fasa pertama adalah fasa analisa deskriptif yang mempunyai skop dengan menyiasat melalui analisis deskriptif dengan mencari kebarangkalian kegagalan operasi kesediaan meter yang berlaku pada setiap kategori lokasi negeri, jenis rangkaian komunikasi, dan status kehilangan *interval* yang dimiliki berserta pemeringkatan ketersediaan data. Tindakan menjalankan aktiviti pembersihan data semasa pra-pemprosesan data dilakukan dalam fasa ini bagi memastikan kejituan analisis dapat dilakukan. Fasa kedua adalah analisa prediktif yang merangkumi kajian dalam menentukan klasifikasi dan regresi sasaran bagi isu kegagalan meter untuk menganalisis berapa banyak isu yang berkaitan dengan memilih dan menjelaskan jenis ciri yang ada. Di dalam fasa ini turut melakukan pembahagian set data bagi penetapan set latihan dan ujian dengan penambahan tugasan pensampelan semula sekiranya terdapat ciri sasaran data yang tidak mempunyai keseimbangan pada analisa kajian. Pada peringkat akhirnya, tugasan meramalkan hasil sasaran bagi melakukan tindakan awal dan memperoleh makluman awal untuk ramalan sasaran isu semasa. Penilaian dan perbandingan model yang dibangunkan dianalisa dalam mengkaji keberkesanan model

yang dipilih oleh kajian. Fasa ketiga adalah analisa preskriptif yang menerangkan skop kajian dalam mencari hubungan hasil melalui analisa berkenaan siri masa dalam mengidentifikasi tahap ketersediaan data mengikut tahap prestasi sejarah. Dalam fasa ini, dilakukan juga tindakan merangka jangkaan dengan penglibatan paparan hasil analisa siri masa dan tindakan penambahbaikan sistem pada masa hadapan.

1.7 KEPENTINGAN KAJIAN

Integrasi pembacaan meter kawalan jauh, rangkaian komunikasi, dan sistem pengurusan data telah menjadi komponen penting dalam infrastruktur sistem grid elektrik kontemporari di samping menyokong kelestarian inovasi semakin pesat membangun secara global pada hari ini. Oleh itu, beberapa kepentingan kajian dinyatakan di dalam kajian bagi menghuraikan setiap perspektif isu kajian yang dilakukan. Implementasi sistem ini memerlukan kajian dan analisis yang menyeluruh, kerana ianya penting untuk meningkatkan ketahanan sistem perantian ini. Pendekatan ini perlu merangkumi penilaian komprehensif terhadap aspek analitik bersejarah, semasa dan masa hadapan. Peranti ini juga memainkan peranan penting dalam merekodkan data profil beban pengguna sebenar dengan tepat di mana keupayaan keberkesanan kajian ini dapat mengurangkan masalah yang berpotensi berkaitan dengan ketidaktepatan, kesilapan dan tindakan mengenal pasti kegagalan dalam proses ketersediaan data dapat dihasilkan. Selain fungsi utama pengebilan di dalam operasi meter RMR ini, pembacaan meter RMR ini juga menawarkan pelbagai keupayaan analisis yang boleh dihasilkan daripada data yang dikumpulkan, dengan itu membolehkan pelbagai inisiatif analisis di peringkat yang selanjutnya dapat dihasilkan. Oleh itu, pentingnya menjalankan penyelidikan mengenai isu ramalan ketersediaan data yang menjadi elemen utama kajian ini, kerana ia membolehkan kajian yang lebih komprehensif mengenai kepelbagaian senario yang signifikan berlaku pada masa kini dan akan datang atas faktor penggunaan global rangkaian inovasi meter pada hari ini. Identifikasi masalah pada masanya adalah penting untuk memulihkan segera pelan pembangunan ramalan dalam meningkatkan kecekapan dan kesinambungan sektor permintaan yang telah mendapat perhatian di dalam perbincangan kontemporari.

1.8 ORGANISASI TESIS KAJIAN

Menerusi tesis kajian yang dijalankan, ianya merangkumi lima bab utama seperti yang dinyatakan di bawah:

Bab I: Bahagian bab kajian yang menyatakan secara ringkas berkenaan latar belakang kajian, pernyataan masalah kajian, objektif utama kajian , skop kajian dan kepentingan kajian yang dijalankan.

Bab II: Bahagian bab kajian yang membincangkan hasil kajian terdahulu berkenaan kajian literasi bagi melihat sejauh mana penyelidikan dan metodologi yang digunakan dalam mengadaptasi pendekatan yang dapat diambil ke dalam kajian ini.

Bab III: Bahagian bab kajian yang menerangkan metodologi bagi kerangka kerja utama dan keseluruhan dengan penerangan pelaksanaan terperinci proses kerja yang berlaku di dalam kajian yang merangkumi fasa pemeringkatan analisa kajian.

Bab IV: Bahagian bab kajian yang menjalankan pengumpulan hasil dapatan kajian menerusi analisa pada setiap fasa kajian. Penilaian ke atas prestasi setiap model yang dibangunkan dilakukan dalam melakukan analisa perbandingan hasil dan pendekatan model yang terbaik sebagai hasil akhir kajian.

Bab V: Bahagian bab kajian ini adalah kesimpulan dan rumusan akhir kajian dengan menyatakan sumbangan kajian dan juga cadangan pelaksanaan penambahbaikan yang boleh dilakukan.

BAB II

KAJIAN LITERASI

2.1 TEKNOLOGI *REMOTE METER READING* (RMR)

Kemajuan teknologi pembacaan meter automatik berkait rapat dengan kemajuan yang dibuat dalam bidang kuasa elektrik, elektronik, sains komputer, dan teknologi komunikasi tanpa wayar. Kajian berkenaan teknologi automatik pembacaan meter ini telah dilakukan oleh Wenbin Zheng et al. (2017) dengan kenyataan berkenaan kemunculan awal teknologi ini telah ditemui terutamanya di negara-negara Eropah dan Amerika. Gabungan penyelidikan berkenaan *Automatic Meter Reading Association* (AMRA) ini telah ditubuhkan pada tahun 1986 di Amerika Syarikat dengan mengalami pertumbuhan yang signifikan dalam skala pembacaan meter automatik selama hampir tiga dekad pembangunannya. Negara-negara Eropah menunjukkan komitmen mereka untuk kekal di barisan hadapan dalam bidang ini, seperti yang ditunjukkan oleh penubuhan *Europe Automation Meter Reading Association* (EUROAMRA) dan *United Kingdom Automation Meter Reading Association* (UKAMRA). Penggunaan teknologi pembacaan meter automatik telah mengalami pertumbuhan yang ketara dalam industri kuasa. Bidang sistem pembacaan dan pemantauan meter mempunyai komponen penting berkaitan dengan penggunaan teknologi komunikasi tanpa wayar dan rangkaian komputer. Ia juga telah dilakukan kajian oleh Feham (2012) dan Meng et al. (2019), berkenaan teknologi komunikasi meter RMR dengan protokol modem telah dicadangkan untuk tujuan operasi. Medium utama yang digunakan untuk penghantaran data melalui sistem meter tanpa wayar termasuk *Global System for Mobile Communication* (GSM), *General Packet Radio Service* (GPRS), *Bluetooth, Zigbee*,

wireless fidelity (WIFI), talian telefon, internet, dan *power-line communication* (PLC). Kajian tersebut menerangkan penggunaan komunikasi data GPRS untuk meningkatkan penghantaran rangkaian tanpa wayar adalah kerana kekurangan sistem pada data GSM. Teknologi rangkaian GPRS adalah sub-rangkaian GSM menggunakan teknologi pertukaran paket dan penghantaran dengan memiliki *Serving GPRS Support Node* (SGSN) , *Gateway GPRS Support Node* (GGSN) dan nod jaringan lain dalam membantu sistem badan utama meningkatkan kebolehpercayaan komunikasi. Secara lebih ringkas, pembacaan meter jauh berasaskan GPRS mudah dikendalikan, tidak memerlukan pembinaan semula rangkaian, dan tidak mempunyai had geografi. Kapasiti penghantaran sistem bacaan meter jauh mencukupi untuk penangkapan data dengan hasil keberkesanan penghantaran meningkat secara ketara apabila pembangunan dan perubahan semasa penggunaan dari *Second-generation* (2G) kepada *Fifth-generation* (5G) telah memenuhi matlamat kecekapan, kelajuan dan kebolehpercayaan pemindahan data telah meningkat.

Perkembangan teknologi meter pembacaan jauh ini juga telah diterangkan di dalam kajian oleh Dunuweera et al. (2017) berkenaan kajian perkembangan rangkaian komunikasi menggunakan pendekatan *Zigbee*. Kajian beliau menjelaskan tentang sejarah terdahulu, berkenaan meter elektromekanikal mendominasi pembacaan tenaga elektrik pada awal peringkat utiliti elektrik dibangunkan. Perilaku medan magnetik adalah berkadar dengan voltan dan arus tiada medan komunikasi yang wujud. Teknologi komunikasi membawa meter pintar baru diwujudkan terhadap pembacaan tenaga moden dalam pengukuran lebih banyak parameter bacaan seperti voltan, arus, kuasa aktif, kuasa reaktif, kuasa yang kelihatan, profil beban, dan log kesilapan. Dengan perisian modul GPRS, setiap meter ini mempunyai *Subscriber Identity Module* (SIM) dengan *Identification* (ID) unik dalam sebahagian komponen pemancar komunikasi tanpa wayar, penerima, dan unit penyimpanan data, serta sensor pengukuran data dan elektronik pemeteran bacaan sistem secara jauh hari ini.

2.2 PENYELIDIKAN PEMBELAJARAN MESIN KE ATAS TEKNOLOGI METER

Penggunaan data pembacaan meter kawalan jauh atau sebahagian juga daripada meter pintar yang dikendalikan untuk meningkatkan kecekapan dan kesinambungan segmen permintaan telah mendapat perhatian yang ketara di peringkat global, dan

dianggap sebagai topik penting di zaman kontemporari ini. Perantian meter ini memainkan peranan penting dalam menangkap data profil beban pengguna dengan tepat untuk mengelakkan ketidaktepatan, ketidaksamaan dan kegagalan pengiraan. Walau bagaimanapun, perlu dilihat bahawa pembacaan meter kawalan jauh berkhidmat kepada pelbagai tujuan di luar aktiviti kelancaran pengebilan, kerana data yang dikumpulkan boleh digunakan untuk pelbagai usaha analisis. Oleh itu, pentingnya menjalankan penyelidikan meter automatik pintar dalam perspektif yang berbeza kerana penggunakan meter jenis ini adalah juga berfungsi dalam sistem kuasa, pasaran elektrik, sistem maklumat meteorologi, media sosial, dan sebagainya. Setiap aspek grid pintar, termasuk operasi, penyelenggaraan, ramalan beban hasil, perlindungan serta pengesanan kesalahan dan lokasi telah mempunyai usaha penyelidikan ke atasnya.

Kerja berkaitan penyelidikan pembelajaran tingkah laku pengguna elektrik tambahan oleh Y. Wang et al., (2019) merangkumi aplikasi analisis data meter pintar dalam ramalan beban pengguna, pengesanan anomali, segmentasi pengguna, dan respons permintaan. Perkembangan terkini telah disimpulkan dan disemak semula. Kajian ini mencadangkan kajian lanjut mengenai data besar, pembelajaran mesin, model perniagaan kreatif, transformasi sistem tenaga, dan kepelbagaian inisiatif pengurusan data. Analisis data meter pintar menjanjikan menyediakan gambaran wawasan yang mendalam dengan mengubah data menjadi penilaian yang lebih baik. Ia biasanya dibahagikan kepada tiga perspektif analisis iaitu deskriptif, prediktif dan preskriptif. Kaedah statistik untuk analisis tahunan dan keadaan jumlah kehilangan selang masa yang hilang pada kesediaan data dikaji oleh Stoyanov et al. (2021). Tujuan kajian ini dibuat adalah untuk mendapatkan tabiat penggunaan elektrik isi rumah pada musim bunga di Ruse, Bulgaria. Alat Meter Pintar telah dipasang pada bangunan berkenaan untuk memantau penggunaan elektrik dan air. Setiap isi rumah diberikan akses untuk melihat penggunaan harian mereka. Perisian Matlab digunakan untuk pengesahan undang-undang pengedaran. Penggunaan Meter Pintar ini memudahkan pengguna memantau penggunaan harian mereka dan membantu mereka mengubah sikap dalam tatacara penggunaan sumber tenaga. Bagi kajian oleh Farrugia et al. (2022), proses statistik dan analisis memerlukan set data yang lengkap tanpa kehilangan nilai data. Dalam operasi meter pintar, beberapa profil beban tidak dibaca selama hari-hari, yang mengakibatkan data yang hilang. Utiliti ini boleh menyediakan pengguna dengan

profil beban yang komprehensif dengan menggantikan data tersebut. Untuk memenuhi matlamat ini, pendekatan imputasi telah dilakukan dengan menggantikan nilai menggunakan purata berat bahagian-bahagian yang hilang daripada pengiraan arus maksimum dalam nod dan mencari tapak kegagalan yang mungkin ada pada profil beban terperinci. Kaedah imputasi berasaskan *K-Nearest Neighbour* (KNN) mengkaji sejarah penggunaan pengguna untuk tingkah laku yang sebanding dengan sebelum data yang hilang. Asas penyelidikan ini turut dilakukan oleh Benítez & Díez (2022) dengan matlamat utama kajian ini adalah untuk menyediakan ulasan terperinci tentang pengesanan automatik corak profil beban penggunaan tenaga elektrik melalui teknik terkini untuk pengelompokan. Prestasi algoritma telah diukur dengan menggunakan indeks kuantitatif yang biasa digunakan dalam pengelompokan. Keputusan menunjukkan bahawa berasaskan *K-means* algoritma pengelompokan dinamik dengan jarak berasaskan *Euclidean* atau *Hausdorff* menyediakan hasil terbaik. Juga daripada Rai & De (2022) yang memperkenalkan *ensemble* rata-rata dan *ensemble* bertumpuk. Teknik purata ditimbang menggabungkan teknologi prediksi beban pintar *Feedforward Backpropagation Neural Network* (FFNN), *Cascade-Forward Backpropagation Neural Network* (CFNN) dan SVR. *Ensembles* mengumpulkan menggabungkan pendekatan semasa dengan melonggarkan eksponensial. Bahagian berikut menerangkan model ini. Akhirnya, banyak kajian kes mengenai set data beban perumahan dan campuran menunjukkan kegunaan kaedah ini. Penyelidikan ini menggunakan data beban masa nyata untuk melaksanakan beberapa algoritma ramalan beban di dua nod yang berbeza dalam rangkaian kampus akademik perumahan. Kaedah ramalan beban *ensemble* rata-rata dan mengumpulkan gabungan ramalan dibandingkan dengan pendekatan ramalan individu.

Selain itu, perkembangan meter RMR ini juga telah menghasilkan kajian berkenaan implementasi *Geographic Information System* (GIS) oleh Zulkflee et al. (2019). Kajian ini menerangkan keberhasilan syarikat bekalan elektrik, Tenaga Nasional Berhad telah berjaya melancarkan teknologi pembacaan meter jauh kepada lebih daripada 70% pelanggan yang banyak meningkatkan operasi membaca meter dan mengurangkan kerugian teknikal. Oleh itu, seiring dengan penggunaan elektrik yang pesat telah mesti memenuhi populasi yang semakin meningkat di dalam aktiviti pengedaran bil di bandar-bandar dan kampung-kampung terpencil. Oleh itu,

kepentingan kajian ini dalam mengetengahkan pembinaan model kajian menggunakan GIS untuk menilai semula sempadan *meter reading units* (MRU) menggunakan analisis spasial untuk meminimakan kos, masa perjalanan, dan keperluan logistik lain bagi pengedaran bil kepada pengguna. Analisis rangkaian mengoptimumkan laluan dan mengurangkan masa kerja pembaca meter telah dilakukan dengan menggunakan konsep pengkelompokan dan analisis rangkaian. Analisis tetangga terdekat digunakan untuk membahagikan MRU kepada sub-kluster premis pengguna. Kajian lain oleh Pilo et al. (2021) mencadangkan teknik baharu berdasarkan analisis data lanjutan untuk menganalisis data pelanggan profil beban yang didaftarkan oleh meter pintar. Pelbagai algoritma pengelompokan telah dicadangkan dalam literatur beliau seperti kaedah deterministik, statistik, dan kecerdasan buatan. Di antara semua teknik ini, teknik pengelompokan telah digunakan secara meluas untuk mengukur corak pengguna dalam mengkategorikan pelanggan dan menentukan profil beban khusus untuk setiap kategori. Dengan peningkatan pengguna di dalam penggunaan meter pintar ini, analisa berkenaan profil pengguna turut dilakukan menerusi ketersediaan data pengguna yang diperolehi daripada pemasangan meter RMR ini. Bagi kajian oleh Hurst (2020), pembelajaran mesin dilakukan dengan mengenalpasti kadar pengangguran di kalangan isi rumah bujang. Hasil dapatan menunjukkan ia mampu untuk mendapatkan status kebolehpasaran menggunakan kaedah rangkaian *neural perceptron* berbilang lapisan. Ketepatan profil pengguna dapat diperoleh melalui corak tingkahlaku penduduk isi rumah. Kaedah regresi logistik adalah model parametrik yang paling biasa digunakan untuk analisis pembolehubah hasil binari.

2.3 PENYELIDIKAN PENDEKATAN MESIN VEKTOR SOKONGAN(SVM)

Sehubungan dengan hasil kajian menerusi skop persamaan kajian yang diterangkan, ianya membawa kepada kajian analisa dengan penglibatan model analisa vektor sokongan SVM. Di kajian lain Hassani et al. (2022), membentangkan sokongan keputusan pintar dan model gabungan untuk pengesanan ralat dan lokasi dalam grid kuasa. Hasil daripada kajian ini menunjukkan kepintaran pengiraan, sistem sokongan keputusan, pengesanan kesalahan, lokasi kerosakan, dan pembelajaran mesin di dalam sistem grid kuasa dengan kajian menyeluruh dapat dilakukan dengan hasil yang mencapai sasaran yang dikehendaki. Bagi mengkaji lagi sejauhmana kajian penggunaan

mesin vektor sokongan berdasarkan kajian lepas, satu kajian diperoleh yang mana dibuat oleh R. Wang et al. (2021) melalui kajian berkenaan mesin vektor sokongan *Clifford Fuzzy* untuk regresi dan penggunaannya dalam ramalan beban elektrik tenaga sistem diketengahkan. Konsep daripada kajian ini adalah berdasarkan regresi vektor sokongan *Clifford Support Vector Regression* (CSVR) oleh algebra geometri *Clifford* dengan tujuan dalam membuat ramalan profil beban pengguna yang akan digunakan pada masa hadapan. Untuk melakukan ramalan beban pengguna ini, penggunaan *fuzzy* ditetapkan untuk memberi berat dalam set latihan yang boleh memberikan sumbangan kepada fungsi regresi ini. Perlaksanaan penetapan *fuzzy* dapat dijalankan apabila nilai *outlier* dikurangkan. Kelebihan lain daripada konsep metod ini ialah ramalan menjadi lebih tepat apabila titik yang diramalkan mempunyai lebih sumbangan kepada nilai yang diramalkan. Oleh itu, algebra *fuzzy* ini diadaptasi kepada CSVR dan kemudian kepada ke dalam *Clifford Fuzzy Support Vector Regression* (CFSVR). Hasil daripada pendekatan kaedah CFSVR ini, meningkatkan ketepatan ramalan profil beban yang diperlukan dengan berkesan terhadap SVR algoritma lain. Berkenaan kajian lain yang berkaitan, kajian tentang pengesanan serangan suntikan data palsu dalam sistem fizikal maklumat kuasa berdasarkan algoritma *Support Vector Machine Gentle-Adaboost* (SVM-GAB) oleh Xiaoping Xiong et al. (2021). Di dalam kajian ini, tindakan mengenalpasti lampiran suntikan data palsu iaitu *False Data Injection Attack* (FDIA) dengan menggunakan syarat pengurangan dan pengelasan dimensi sepanjang data dengan tujuan utama adalah untuk melakukan penyiasatan ke atas serangan suntikan data palsu FDIA menggunakan kaedah ini dengan membandingkan algoritma pengesanan arus perdana dalam sistem *standard* IEEE-14 dan IEEE-39. Dengan pendekatan ini, algoritma SVM dan *Gentle-Adaboost* berjaya merealisasikan pengesanan masa nyata FDIA dalam kajian ini berbanding dengan algoritma pengesanan tradisional.

Untuk mengkaji algoritma mesin vektor sokongan dalam bentuk kedua-dua aktiviti klasifikasi dan ramalan, satu kajian ditemui melalui kajian berkenaan kaedah tafsiran log kerintangan rendah - kontras minyak dibayar di Chang 8 Standstone, kawasan Huanxian, Basin Ordos dengan mesin vektor sokongan oleh Bai et al. (2022) dengan model pengelasan dan model regresi menggunakan kaedah SVM. Asas kepada idea adalah untuk ramalan parameter takungan dalam memetakan ruang input kepada

dimensi tinggi untuk tujuan memisahkan secara *linear* info data yang ada dengan menggunakan fungsi *kernel* dan kemudian menyelesaikan *hyperplane* atau fungsi yang boleh dipisahkan secara *linear.* Sekiranya jarak pemisahan lebih baik kesannya boleh dilihat semasa proses pengelasannya. Akhir sekali, keupayaan diskriminasi bukan *linear* bagi data asal dikenalpasti. Dapat disimpulkan disini, vektor sokongan teknologi mesin (SVM) telah digunakan untuk mentafsir kerintangan minyak kontras rendah yang dibayar di Chang dengan urutan data input keluk pembalakan telah dipilih analisis hubungan diantara jenis minyak takungan dan data pembalakan dilakukan. Kemudian, model pengelasan SVM dijalankan untuk pengecaman minyak dan model regresi SVR dilakukan bagi meramal parameter takungan tersebut. Dengan aktiviti mesin vektor bagi kedua-dua aktiviti ini dalam sebahagian kajian, menjadikan satu pendekatan baru yang perlu dilaksanakan di dalam pengurusan infrastruktur kesediaan data di dalam pembacaan meter kawalan jauh yang bukan hanya menjalankan satu bidang tugasan tetapi kepada dua kaedah seperti kajian yang telah dilakukan oleh penyelidikan sebelum ini.

2.4 PENYELIDIKAN PENGGUNAAN PERLOMBONGAN DATA & PEMBELAJARAN MESIN

Secara amnya, kajian oleh H et al. (2011) telah menjelaskan tugas pra-pemprosesan data adalah bahagian yang sangat penting dalam prosedur perlombongan data. Sebilangan besar data perlu dirawat, diubah suai dan diubah untuk mempersembahkan prestasi yang lebih baik dalam ketepatan data yang diuji. Tugas pembelajaran ini akan membawa kepada perlombongan data untuk mendapatkan corak pengetahuan yang berkualiti. Perlombongan data melibatkan penyepaduan teknik daripada pelbagai disiplin seperti pangkalan data dan teknologi pergudangan data, statistik, pembelajaran mesin, pengkomputeran berprestasi tinggi, pengecaman corak, rangkaian saraf, visualisasi data, pengambilan maklumat, pemprosesan imej dan isyarat, dan data sampel analisis. Penglibatan model pembelajaran mesin untuk meramal kegagalan instrumen meter dan sebarang anomali yang boleh dikenal pasti telah dibincangkan. Antara kajian, Ming Liu et al. (2020) menggunakan set data elektrik meter yang sama, yang menerangkan tentang pengesanan pembelajaran mendalam bagi meter elektrik pintar yang tidak tepat. Dalam penyelidikan ini, pengesanan pada meter pintar yang tidak tepat dan menyasarkannya untuk diganti dengan melaksanakan kaedah pembelajaran

mendalam berdasarkan *Long Short-Term Memory Network* (LSTM) dan *Convolution Neutral Network* (CNN) yang diubah suai untuk meramal trajektori penggunaan elektrik pada data sejarah.

2.4.1 Konsep Kajian Awal

Kajian ini mencadangkan rangka kerja dari penyelidikan oleh Lo & Pachamanova (2023) untuk beralih daripada sains data tradisional. Dengan pembaharuan sebelum ini iaitu tumpuan analisis merangkumi daripada mengekstrak nilai daripada data yang tersedia kepada membuat keputusan analisis yang dipacu matlamat. Untuk melakukannya, objektif perniagaan dikenalpasti terlebih dahulu melalui penyepaduan pelbagai elemen teknik analisis dalam keadaan biasa. Setiap analisis dibincangkan hubungan antara ramalan analisis dan analisis preskriptif dalam perumusan dan konteks masalah. Setiap analisis preskriptif diterangkan dengan mengandaikan pautan antara keputusan dan hasil. Ketepatan analisis dan penjajaran dengan matlamat perniagaan muktamad menjadikan kesepaduan di dalam kaedah ini. Di bahagian ini, menyentuh struktur kajian dijalankan iaitu dibahagikan kajian empirikal kuantitatif dan kualitatif kepada empat asas kajian yang utama dengan peranan kajian yang berbeza setiapnya. Peringkat fasa awal kajian bermula dengan analisis kuantitatif iaitu analisis deskriptif dan seterusnya pada fasa selanjutnya iaitu analisis kualitatif pada fasa analisis ramalan, pengelasan dan preskriptif. Setiap analisis ini merangkumi sub-proses yang lain seperti pengelompokan, pra-pemprosesan, pengekodkan, pemilihan ciri, pembelajaran mesin, analisis siri masa dan proses selanjutnya.

2.4.2 Fasa Analisis Deskriptif

Analisis kuantitatif adalah teknik awal yang dilakukan pada peringkat fasa ini. Tujuan fasa ini dibina adalah bertujuan untuk menyediakan set data yang telah dibersihkan dan telah dikaji corak data mentah yang disediakan. Menurut Benítez & Díez (2022), tugas deskriptif adalah berkaitan dengan perbincangan tentang corak data, laporan perhubungan antara atribut dan jenis pangkalan data diterangkan secara terperinci di dalam fasa ini. Fungsi fasa analisis deskriptif ini disebut juga oleh Li et al. (2021) dengan menjelaskan bahawa fasa ini, kebiasaannya merupakan langkah pertama dan prasyarat untuk analisis selanjutnya. Aktiviti yang dilakukan adalah sering melakukan

pra-pemprosesan data terlebih dahulu, membahagikan keseluruhan data kepada beberapa bahagian dan mencipta kluster data dan analisis sifat setiap kluster yang sama coraknya atau sebaliknya. Pengelompokan biasanya merupakan langkah pertama dalam fasa analisis ini. Ia memberi tumpuan kepada mengekstrak maklumat berharga yang boleh dikumpulkan daripada data mentah. Aspek penting ditunjukkan dalam membangunkan keupayaan untuk mengurus, menganalisis dan mentafsir data penyelidikan kuantitatif mempunyai pelbagai implikasi dalam memahami, menilai dan menggunakan bukti kuantitatif ini. Menurut kajian yang dibuat, penyelidik menerangkan bahawa analisis deskriptif meringkaskan data untuk menerangkan corak sampel dan ia adalah cara paling mudah untuk melaporkan maklumat ini dengan hanya menggunakan kiraan iaitu, jumlah dan peratusan kekerapan sahaja bagi corak data yang dihasilkan. Dalam artikel yang diterbitkan menjelaskan frekuensi boleh dipaparkan dalam jadual atau graf adalah untuk memekatkan maklumat menjadikan ianya menarik secara visual dan mudah difahami.

2.4.3 Fasa Analisis Prediktif

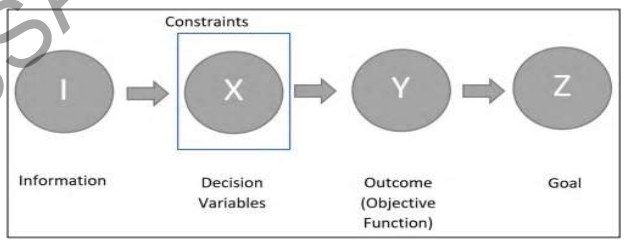
Kajian analisis ramalan merupakan fasa selepas analisis fasa deskriptif awal dilakukan. Objektif pembinaan fasa ini adalah untuk melakukan analisis kualitatif yang diterangkan konsep dalam melihat kepada hasil jangkaan yang terjadi. Untuk mencipta model ramalan dan pengelasan, kaedah asas kualitatif digunakan sebagai mod perjalanan di fasa ini. Penyelidik oleh Ma (2020) menggunakan data berstruktur yang dikumpul, disimpan dan disusun dalam pangkalan data hubungan yang ditubuhkan, dan seterusnya membuat perbezaan dua jenis data berstruktur iaitu data berangka dan data kategori. Di dalam huraiannya, menyatakan berkenaan data berangka biasanya data sebenar yang membawa kepada sesuatu kesan berterusan terhadap matlamat pembelajaran. Bagi data kategori mendapat set nilai tetap yang mewakili set kategori ialah binari yang hanya menerima dua nilai yang boleh memberikannya nilai kualitatifnya.

Dalam metodologi ini, Alocén et al.(2022) menunjukkan domain yang digunakan untuk membina meta-pembelajar adalah dengan menggunakan pembelajaran mesin yang mana dipilih berdasarkan kriteria kualiti prestasi dan sifatnya. Model yang cukup tepat dipilih untuk menampung sebanyak mungkin maklumat vektor ramalan.

Oleh itu, di dalam kajian ini akan menggunakan kumpulan pokok dari dua model algorithm utama iaitu model regresi dan model pengelasan. Mengenai integrasi kaedah, menyusun dan menggabungkan meta-pembelajaran kedua-duanya dicipta oleh vektor ramalan yang dilatih untuk menggunakan faktor luaran dan masa sebagai penjelasan atribut. Hasil pembelajaran mesin ini akan memberikan keputusan ketepatan setiap model yang dijalankan.

2.4.4 Fasa Analisis Preskriptif

Analisa kajian di peringkat ini menunjukkan penyelidikan berkenaan siri masa dan analisanya terhadap tindakan yang mana banyak dilakukan oleh penyelidik. Salah satunya adalah dari penyelidikan oleh Li et al. (2021) dalam menggunakan analisis domain frekuensi dan masa yang mempunyai tumpuan yang berbeza serta kekuatan analisis domain masa mempunyai kapasiti yang lebih besar untuk menangkap corak umum profil harian seperti perubahan beban atau isu-isu lain yang diperlukan analisis domain frekuensi membantu menangkap beban kebolehubahan yang berterusan. Walaupun penyelidikan sebelumnya hanya menggunakan pendekatan domain masa atau frekuensi-domain, namun menerusi analisa siri masa, pelbagai aliran umum di dalam penyelidikan dapat diambil sekaligus membantu analisa preskriptif dikaji dengan sokongan keputusan yang lebih baik.



Rajah 2.1 *Directed acyclic graph* (DAG) sebagai rangka kerja analisis Li et al. (2023)

Pemahaman lanjut berkenaan pendekatan di dalam analisa ini dirujuk menerusi sebuah kajian yang dilakukan oleh Lo & Pachamanova (2023). Rajah 2.1 di atas mencadangkan bagaimana analisa preskriptif dapat dijalankan. Aliran proses yang ditunjukkan adalah digunakan untuk menunjukkan rangka kerja analitik preskriptif kausal yang dicadangkan. Daripada gambar rajah yang ditunjukkan, penemuan X yang dikelilingi oleh kotak adalah berkemungkinan sebagai makluman kekangan diwakili

oleh kotak di sekeliling keputusan nilai X yang boleh dicapai. Dengan mencadangkan satu atau lebih pilihan, analisis ini melangkaui analisis deskriptif dan ramalan. Pada asasnya, analisa ini lebih meramalkan pelbagai hasil hadapan dan membolehkan organisasi menilai pelbagai hasil yang mungkin berdasarkan sejarah dapatan yang lepas. Analisa preskriptif menggunakan algoritma, peraturan perniagaan, *Machine Learning* (ML) dan prosedur pemodelan pengiraan. Data sejarah atau suapan data masa nyata adalah antara pelbagai set data yang boleh dimasukkan menggunakan teknik ini. Apabila dilaksanakan dengan betul, ia boleh memberi kesan yang ketara kepada proses keputusan perniagaan untuk mengoptimumkan pengeluaran, penjadualan dan inventori dalam rantaian bekalan terhadap sesuatu keperluan.

2.4.5 Teknik Kejuruteraan & Pengekstrakan Data

Di dalam kajian ini, kerangka kerja kajian adalah menerangkan sub-proses bagi fasa analisis utama yang dirancang. Menerusi penetapan jenis analisis yang ingin dijalankan, pelbagai peringkat proses dilakukan dalam memenuhi skop kerangka kajian ini. Kerangka kajian ini telah diterangkan mengikut peringkat sub-proses yang dijalankan oleh kajian terdahulu.. Keseluruhan kerangka kerja kajian yang melibatkan elemen asas proses bagi projek data sains yang dicadangkan oleh Lo (2021) menerangkan proses kajian secara hierarki mengikut analisa yang telah dijalankan dengan perincian yang dijalankan dalam kerangka kajian ini.

Penelitian teknik dilakukan di dalam kajian oleh Melo et al. (2022) digunakan dalam kajian ini yang mana mencadangkan penggunaan teknik kejuruteraan ciri dan teknik pengekstrakan untuk memilih ciri yang paling penting daripada pangkalan data. Pemodelan dan ramalan beban dijalankan menggunakan mesin utama teknik pembelajaran. Metodologi yang digunakan terdiri daripada urutan langkah yang bertujuan untuk menjalankan pemprosesan data, mengekstrak data dan memilih ciri peramal yang paling tepat untuk meramalkan penggunaan elektrik.

a. Peringkat 1: Proses Penemuan Pengetahuan dalam *Knowledge Discovery In Databases* (KDD)

Pada peringkat pertama ini dikenali sebagai kajian awal projek yang akan mengendalikan *domain* permasalahan isu yang ingin dikaji dengan penemuaan

hipotesis atau andaian oleh pakar atau dikenali sebagai *domain-expert*. Kajian telah dilakukan oleh penyelidik Benítez & Díez (2022) yang menyentuh berkenaan teknik perlombongan data dengan menjelaskan ianya dianggap sebagai peringkat perantaraan dalam yang lebih dengan mengemukakan istilah Proses Penemuan Pengetahuan dalam Pangkalan Data, juga dikenali sebagai KDD. Proses KDD ini ditakrifkan sebagai yang sesuatu penting dengan menjalankan proses mengenal pasti corak data yang sah, baru, berpotensi berguna, dan akhirnya boleh difahami. Proses ini melibatkan pengekstrakan pengetahuan daripada set data yang besar. Pada permulaan proses, pakar atau penganalisis data melihat set data dan membuat jangkaan atau mengemukakan hipotesis tentang matlamat proses penemuan pengetahuan. Untuk melakukan hipotesis ini, tindakan kajian dilakukan terhadap model atau hubungan yang mungkin wujud antara pelbagai atribut atau ciri objek data yang diperlukan. Hipotesis awal ini adalah langkah asas dalam menjayakan analisis data seterusnya iaitu satu keputusan diperoleh atau pun tidak. Jika tiada keputusan dapat dibuat, hipotesis baru akan dikemukakan dan proses KDD akan dimulakan semula.

b. Peringkat 2: Pengumpulan Data

Selepas keputusan di dalam proses KDD diperoleh, tindakan selanjutnya adalah memulakan penyediakan set data sedia ada untuk dianalisis yang mana dipanggil sebagai data mentah yang akan diproses dan dipastikan dapat dibawa kepada peringkat analisis seterusnya. Kajian yang dibuat oleh Kotronoulas et al. (2023) menunjukkan makluman proses pengumpulan data ini adalah dengan pembinaan atribut dan penjelasannya dari kajian ini menakrifkan atribut sebagai apa-apa yang boleh diukur dengan cara yang berbeza. Secara teorinya, atribut akan termasuk data kuantitatif yang memerlukan analisis statistik yang melaporkan kepada rumusan nilai dan unit ukuran adalah asas bagi pengelasan awal yang diperlukan. Ditambah juga dari kajikan Melo et al. (2022) yang mengemukakan penghasilan satu set atribut peramal dan atribut objektif membentuk pangkalan data yang diperlukan. Apabila pangkalan data ini dibentuk, proses berkenaan teknik ciri kejuruteraan dan pengekstrakan teknik akan dilakukan.

c. Peringkat 3: Pra-pemprosesan Data

Daripada analisa yang dilakukan oleh Kaur & Kaur (2017), pra-pemprosesan dan *data mining* menggunakan RStudio telah dijalankan untuk melakukan analisis korelasi bagi kemalangan jalan raya melalui teknik visualisasi eksploratif. Metodologi ini digunakan menggunakan IDE *Integrated Development Environment* (Rstudio) R yang merupakan alat pengkomputeran grafik dan statistik pada set data yang mana proses pra-pemprosesan dilakukan pada pelbagai parameter data jalan raya bagi analisis korelasi dan teknik visualisasi eksploratif untuk menganalisis dan meramalkan hasil yang berguna yang membantu mengurangkan kemalangan dan menentukan keadaan jalan raya. Penelitian dalam menyediakan pemprosesan kuantitatif, telah dilakukan dalam kajian oleh Kotronoulas et al. (2023) dengan data penyelidikan disemak dengan teliti untuk kesilapan dan nilai yang hilang, dan kemudian analisis data kuantitatif melibatkan aplikasi statistik. Data untuk sesetengah variabel mungkin seolah-olah dibahagikan secara normal. Walau bagaimanapun, untuk variabel lain data mungkin kelihatan lebih meluncur ke kanan atau ke kiri, pengukuran corak pusat hendaklah dikira, ia sentiasa merupakan idea yang baik untuk menilai perkongsian data. Di mana data kelihatan dibahagikan secara normal, sama ada purata atau median boleh dianggarkan untuk memberikan maklumat yang sama dan sekiranya data meluncur ke kanan atau kiri, maka median adalah pilihan yang lebih baik dalam menggantikan data yang hilang. Ini kerana purata mudah dipengaruhi oleh nilai yang terlalu kecil atau terlalu besar dalam koleksi data dan boleh memberikan ringkasan palsu daripada variabel yang sedang dipelajari. Sebaliknya, median tidak dipengaruhi oleh nombor ekstrem. Untuk penggantian nilai menggunakan mod ini boleh digunakan bagi meringkaskan variabel kategori dengan membandingkan dua atau lebih kumpulan.

Langkah seterusnya diambil dalam kajian ini dengan melihat model rangka kerja pra-pemprosesan yang dilakukan oleh Parizad (2020) ini dibuat setelah *gateway* data mengimport data ke *cloud*. Aktiviti seperti Integrasi data, pembersihan, transformasi, dan pengesanan data hilang dilakukan. Penyelidikan ini bermula dengan pengenalan entiti data yang mengalami pertindihan di mana kajian melakukan penghapusan data yang sama dan pembersihan data dengan mengenal pasti data yang hilang sebagai contoh kesilapan meter pintar dengan hasil yang tidak normal iaitu

ukuran bacaan yang salah. Kajian ini juga menggunakan konsep standardisasi data dalam fasa ketiga sekaligus berlaku transformasi pada data. Akhir sekali, teknologi dan algoritma perlu digunakan dalam mengenal pasti data palsu. Selain itu, pendekatan dalam pengenalan corak beban dan ramalan dilakukan dengan menggunakan sejumlah besar data dari tahun-tahun yang lalu, termasuk data beban, data cuaca ,hari, minggu, hujung minggu, musim, dan lain-lain. Oleh itu, pra-pemprosesan diperlukan dalam kajian bagi membersihkan dan menyediakan data untuk meramalkan algoritma yang akan digunakan.

Penambahan beberapa sub-proses kerja dibuat di dalam kajian Mohammedqasim et al. (2023) seperti *feature encoding* dan data normalisasi dengan pembelajaran mesin beroperasi dalam bentuk nombor yang memerlukan kajian untuk menggunakan pengekodan kategori label variabel untuk semua rekod. Hasil pengekodan ini akan menetapkan nombor unik untuk semua kategori data yang ada. Kajian oleh Bora & Baruah (2023) yang juga mengunakan pendekatan *label encoding* untuk mengubah ciri kategori kepada nilai numerik yang menjadi nombor unik megikut kategori bagi penggunaan analisa seterusnya. Selain itu, di dalam kajian Mohammedqasim et al. (2023) juga menjelaskan pengisian nilai ciri yang hilang adalah dengan purata nilai ciri dan ditambah dengan pendekatan normalisasi disebabkan sampel ciri dalam set data itu mempunyai julat yang berlainan yang menghasilkan ketidakstabilan dalam hasil klasifikasi. Fungsi bagi penggunaan normalisasi ini dilakukan pada data sampel adalah untuk mengekalkan data dalam pengedaran yang seragam. Hal ini bertepatan dengan kajian yang dilakukan oleh Chanal & Steiner (2021) yang menjelaskan bahawa pendekatan skala ini terdiri daripada skala data dalam julat 0 sehingga 1, di mana ia juga dikenali sebagai *Min-Max scaler* iaitu menggunakan data minimum dan maksimum sebagai sempadan dan data penukaran semula. Salah satu kelebihannya ialah ia membolehkan untuk memasukkan ciri-ciri *interval* yang sama yang boleh menjadi sangat pelbagai sambil mengekalkan semua maklumat kepada pembahagian jarak antara titik yang membolehkan mengekalkan ciri-ciri dengan nilai kecil berbanding dengan yang dengan nilai besar. Ianya bersesuaian digunakan bagi normalisasi bagi pendekatan ramalan klasifikasi yang mempunyai tanda sasaran mengikut kelas 0 dan 1 atau lebih yang mana diseragamakan dengan julat input data menggunakan kaedah normalisasi menerusi *Min-Max scaler.*

Namun begitu bagi analisa ramalan menggunakan pendekatan regresi, penggunakan normalisasi ini tidak digunakan. Ini diterangkan di dalam penyelidikan oleh Rauch et al. (2022) bahawa praprosesan data, normalisasi yang paling penting melalui *biomarkers* dan jarak masa yang sama data yang tersebar, adalah penting untuk regresi dan ramalan. Dalam tradisi, normalisasi hampir tidak pernah mudah dalam menggunakan purata parameter yang tersedia. Oleh itu, purata dan beban *standard* digunakan untuk menghasilkan perkiraan beban yang lebih baik dengan menukar parameter dalam proses normalisasi yang mana faktor-faktor sedemikian perlu dipertimbangkan. Sebagaimana yang diterangkan oleh Chanal & Steiner (2021) , skala *standard* adalah bertujuan untuk mengubah ciri-ciri supaya mereka mempunyai purata kosong dan penyimpangan *standard*. Skala *standard* membolehkan pembentukan data berpusat dan menjadikannya mudah untuk digunakan dengan teknik pembelajaran mesin statistik.

d. Peringkat 4: Teknik pemilihan ciri

Penyelidikan oleh Hayes et al. (2015) menerangkan teknik permilihan ciri yang lebih sesuai bagi set data yang memerlukan penglibatan hubung kait antara variabel di mana kajian ini mempunyai dua set data besar yang terdiri daripada rekod meter pintar digunakan untuk menilai korelasi antara permintaan dan variabel yang mempengaruhinya, dan untuk menjalankan ramalan beban. Elemen yang mempengaruhi permintaan dalam jangka pendek sering jatuh ke dalam tiga kategori berkaitan masa, sejarah dan berkaitan cuaca. Bagi menganalisis hubungan antara variabel dalam model untuk ramalan beban, satu metrik yang biasa digunakan untuk kekuatan hubungan *linear* antara dua variabel ialah koefisien korelasi *Pearson* di mana parameter x dan y menunjukkan nilai purata variabel input dan sasaran yang dinilai dalam analisis. Kadar korelasi boleh bervariasi antara 1 iaitu korelasi positif sempurna dan -1 iaitu korelasi negatif sempurna manakala nilai yang mendekati 0 adalah hasil daripada hubungan yang lemah antara variabel yang disiasat. Kajian pendekatan yang sama ditambah oleh Panwar et al. (2017) dengan tujuan penyelidikan adalah untuk membina mekanisme pengesanan penyusupan yang menggunakan pelbagai metodologi pembelajaran mesin untuk mengenal pasti anomali atau penyusutan pada set data CICIDS-2017. Set data in menggabungkan pelbagai serangan kontemporari dan telah

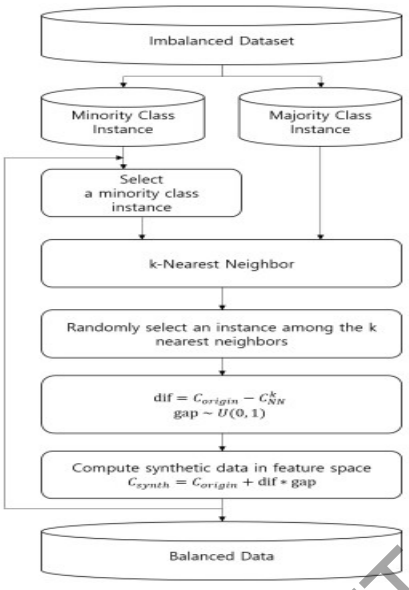
ditubuhkan dalam pengklasifikasian binari. Proses pemilihan ciri melibatkan mengenal pasti ciri-ciri yang relevan dan mengurangkan dimensi set data dengan menghapuskan atribut luar, pertindihan, atau tidak berkaitan.

e. Peringkat 3: Teknik Set Pengesahan

Di bawah model yang disyorkan, menggunakan sifat-sifat yang diperlukan daripada set data CICIDS2017 dan selepas melalui prosedur pembersihan, pengekodan, label, skala ciri dan normalisasi, kajian dilakukan dengan memastikan bahawa set data bersedia untuk dilatih. Set pengesahan yang digunakan dalam set data CICIDS-2017 ialah teknik yang dibahagikan kepada dua bahagian 80% daripada data latihan dan 20% daripada data ujian. Begitu juga kajian oleh Disaggregation et al. (2021) yang mencadangkan penglibatan dalam rangka kerja pembelajaran mesin, ia menyampaikan penilaian penuh menggunakan kaedah latihan atau ujian-perpecahan senario yang menggunakan 90–10% daripada setiap subset. Dalam keadaan ini, data latihan dipilih secara rawak dan dibahagikan kepada 10 kali di mana pada setiap iterasi pengesahan, satu digunakan untuk ujian, manakala sisanya digunakan untuk latihan. Angka-angka ini adalah purata untuk 10 prosedur pengesahan silang.

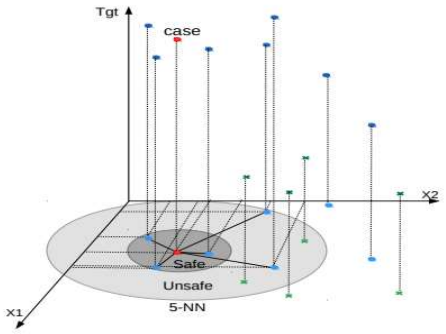
f. Peringkat 3: Teknik Pensampelan

Melalui teknik pensampelan yang pernah dilakukan, satu teknik yang sering digunakan bagi situasi kelas tidak seimbang adalah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) iaitu teknik yang biasa digunakan di dalam mana-mana kajian yang dilakukan ke atas set data ramalan mempunyai pengelasan iaitu klasifikasi. Antara pendekatan kaedah SMOTE ini dilakukan oleh Lee et al. (2017) adalah pendekatan SMOTE dengan mencipta data sintetik menggunakan KNN dan pengedaran kebarangkalian yang seragam dengan membahagikan data yang diberikan kepada kelas majoriti dan minoriti. Di setiap kelas minoriti mempunyai jiran terdekatnya yang dipilih secara rawak di antara jiran k yang hampir bagi mencapai keseimbangan data dengan nilai majoriti. Rajah di bawah menunjukkan proses kerja SMOTE yang dilakukan dalam kajian tersebut.



Rajah 2.2 SMOTE algoritma proses Lee et al (2017)

Berhubung dengan pendekatan SMOTE ini, kajian daripada Hassanzadeh et al. (2023) yang menggunakan data daripada rekod maklumat perubatan pesakit trauma di hospital Besar di Wilayah Hamadan, Barat Iran. Teknik SMOTE iaitu teknik *oversampling* yang mencipta data sintetik untuk kelas minoriti berdasarkan jiran k terdekatnya sehingga nisbah data sintetis baru sifat data adalah lebih seimbang antara majoriti dan minoriti kelas. Dalam kajian ini turut menggabungkannya dengan kaedah-kaedah lain seperti *Synthetic Minority Oversampling Technique Iterative-Partitioning Filter* (SMOTE-IPF) dan *Synthetic Minority Oversampling Technique Local Outlier Factor* (SMOTE-LOF) yang digunakan untuk menangani isu *noisy*. Selain itu, ditambah juga modifikasi SMOTE dengan penciptaan serupa dengan *borderline* SMOTE dan *Support Vector Machine Synthetic Minority Oversampling Technique* (SVM-SMOTE).



Rajah 2.3 Contoh aplikasi SMOGN algoritma Branco (2017)

Masalah regresi yang tidak seimbang telah dipelajari dalam penyelidikan oleh Branco (2017) yang mendahului tahap pembelajaran. *Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise* (SMOGN) ialah teknik yang menggabungkan teknik *undersampling* dengan dua teknik *oversampling* iaitu *Synthetic Minority Oversampling Technique For Regression* (SmoteR) dan pengenalan *Gaussian Noise*. Apabila contoh *seed* dan jiran k terdekat yang dipilih adalah sangat hampir, SMOGN akan menghasilkan contoh sintetik baru dengan SmoteR dan akan mengemukakan penggunaan *Gaussian Noise* apabila kedua-dua jiran k terdekat adalah lebih jauh. Lanjutan daripada teknik yang dicadangkan dalam kajian tersebut, kajian daripada Steininger et al. (2021) juga melakukan perbandingan teknik SMOGN dengan yang lain iaitu *DenseLoss* dan *DenseWeight* bagi melihat sejauh mana hasil yang lebih baik di mana dapat diadaptasi dari model ini. Menurut kajian ini, penerangan SMOGN dinyatakan bahawa SMOGN mengulangi melalui semua sampel yang jarang atau kumpulan minoriti dan memilih antara pengambilan sampel berasaskan interpolasi SmoteR dan pengambilan sampel berasaskan bunyi Gaussian yang bersandarkan kepada jarak jiran k terdekat. Interpolasi SmoteR digunakan untuk sampel berhampiran manakala titik data jauh lain diisi dengan bunyi *noisy* Gaussian.

2.4.6 Penyelidikan Konsep Model Ramalan Klasifikasi

Pembangunan model adalah suatu pembinaan asas bagi pembelajaran mesin dengan mengenal pasti ciri model dalam memahami kesesuaian model terhadap analisa data yang ingin dilakukan. Pemilihan beberapa model kajian telah dilakukan bagi menganalisis kepelbagaian model yang ada dan melakukan perbandingan setiap daripadanya. Antara kajian yang menerangkan model *Naive Bayes(NB)* daripada Budiman et al.(2020) dengan memberikan pendedahan bahawa model NB adalah berkebolehan dalam menyelesaikan isu klasifikasi menggunakan kebarangkalian dan kaedah statistikal dengan mengira frekuensi dan gabungan nilai data. Ia didasarkan pada *Teorem Bayes* di mana teori ini menganggap bahawa semua label kelas set data yang mempunyai nilai-nilai atribut secara tidak bergantung antara satu sama lain. Oleh itu, model *Naive Bayes* dipanggil *Naive* kerana ia menganggap ciri-ciri tanpa had antara satu sama lain walaupun hakikat bahawa ramalan ini jarang berlaku. Walau bagaimanapun, model ini masih mampu mencapai ketepatan klasifikasi yang kuat

walaupun dalam keadaan di mana ramalan adalah tidak tepat. Selain itu, NB sesuai untuk menyelesaikan masalah ramalan pelbagai kelas di mana ia boleh melakukan lebih baik daripada model lain dan ia sesuai untuk variabel input nominal atau kategori daripada variabel numerik. Namun, algoritma ini juga mempunyai masalah apabila masalah frekuensi kosong berlaku dalam set data latihan dan perlu memohon penambahan satu nilai kepada frekuensinya kelas kosong. Ia boleh menjadi perkiraan yang salah, oleh itu ramalan dengan meletakkan hasil kebarangkalian disampaikan. Berikut adalah formula bagi *Naive Bayes* Budiman et al.(2020) ini :

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \ldots(2.1)$$

Di mana $P(C_i|X)$ mewakili kebarangkalian ramalan C_i berlaku apabila X ialah benar, manakala $P(X|C_i)$ mewakili kebarangkalian ramalan X berlaku apabila C_i ialah benar dengan $P(C_i)$ dan$P(X)$ adalah masing- masing merupakan kebarangkalian untuk X dan $C_i$secara berasingan antara satu sama lain.

Kajian model klasifikasi *Random Forest* (RF) oleh Polan et al. (2021) menerangkan Teknik *Random Forest* secara konvensional menggunakan *Decision Tree* dengan *Classification and Regression Trees* (CART) tanpa memotong sebagai klasifikasi asas dan menggabungkan *Bagging* dan pemilihan ciri acak untuk meningkatkan kepelbagaian model *Decision Tree*. Prinsip ini daripada set sampel asal menggunakan kaedah *Bootstrap* untuk mengekstrak latihan yang ditetapkan ke dalam model *Decision Tree*, kemudian disebarkan kepada dalam *Random Forest* untuk sampel input baru. Akhirnya, suara majoriti mutlak pada mekanisme pemilihan *Decision Tree* akan menentukan hasil penilaian akhir. Pembelajaran bersepadu daripada *Decision Tree*, N pada set data ini menghasilkan klasifikator *Random forest*. Algoritma CART menggunakan kaedah pemisahan binari secara iteratif pada setiap ciri dengan membahagikan ruang ciri kepada unit-unit yang terhad dan menentukan distribusi kebarangkalian yang dijangka. CART memilih ciri-ciri menggunakan koefisien *Gini* yang mengandaikan set data D mempunyai kategori k dan kategori C_k, koefisiennya ialah ditunjukkan seperti formula Polan et al. (2021) di bawah:

$$Gini(D) = 1 - \sum_{k=1}^{k} \left(\frac{C_k}{D}\right)^2 \quad \ldots(2.2)$$

$$f(x) = argmax\left\{\sum_{i=1}^{n} T(x) = y\right\}, y = 1,2 \ldots c \quad \ldots(2.3)$$

Model *Random forest* diperolehi dengan mengintegrasikan *Decision Tree*, yang masing-masing dilatih pada set data latihan n, untuk tujuan klasifikasi. Fungsi *predict* $f(x)$ digunakan untuk tujuan pengklasifikasian, di mana ia mengenal pasti klasifikasi yang paling sesuai untuk data terkini seperti yang ditunjukkan pada formula 2.3 di atas.

Bagi tambahan kajian berkenaan *K Nearest Neighbor*(KNN), penyelidikan daripada Kasaraneni et al. yang menerangkan pembinaan klasifikator binari KNN, bilangan jiran terdekat k untuk corak p dan teknik pengiraan untuk jiran paling dekat k pada keahlian kelas diberikan. Parameter ini akan mengurangkan kesilapan kategori. Kebanyakan jiran terdekat pola k mendefinisikan kelas keahlian q. Kaedah KNN mengira jarak corak menggunakan metrik seperti *Euclidean, Cosine, Manhattan*, dan lain-lain. Formula bagi KNN menggunakan jarak *Euclidean* Kasaraneni et al. (2022) :

$$dist(p,q) = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2} \quad \ldots(2.4)$$

di mana $dist$ adalah Euclidean distance, m ialah bilangan titik, dan p_i mewakili titik corak p dan q_i mewakili titik corak q.

Model *Support Vector Classification (SVC)* diterangkan dalam kajian yang dibuat oleh Huang et al. (2022). Pembelajaran statistik SVM teori dimensi *Vector Classification* (VC) dan prinsip pengurangan struktur risiko menjadikannya sangat berkesan dalam menyelesaikan masalah pengiktirafan corak dengan sedikit sampel, *nonlinearity*, dimensi tinggi, dan minimum tempatan. SVM pada mulanya memaparkan vektor input ke ruang ciri dimensi tinggi menggunakan fungsi kernel dan kemudian mencari jarak yang terdekat dengan sempadan antara sampel positif dan negatif. Inilah cara set sampel latihan yang dioptimalkan mencari *hyperplane* yang terbaik untuk generalisasi. Vektor sokongan ialah titik sempadan positif dan negatif. Akhirnya,

sampel yang akan dinilai dijalankan ke ruang dimensi tinggi dan diklasifikasikan. SVM menggunakan fungsi kernel untuk membezakan input yang tidak dapat dibezakan secara linear dalam ruang dimensi yang lebih tinggi. Fungsi kernel menentukan kerumitan ruang model SVM, yang memberi kesan kepada prestasi algoritma. SVM perlu menentukan titik *hyperplane* di mana pendakatan SVM sama ada *linear* dan *non-linear* digunakan. SVC secara *linear* membahagikan data dan memilih vektor ekstrem untuk menjana *hyperplane* yang optimum. Selebihnya SVC *Polynomial* (Poly), *Radial Basis Function* (RBF) dan *Sigmoid* tidak boleh dipisahkan secara *linear*. SVC poly memfokuskan pada data global, tetapi model ini adalah kompleks dalam mengira pada pesanan yang tinggi. SVC RBF pula adalah lebih fleksibel, berseragam dan berlokalisasi namun begitu, model RBF ini mudah untuk meginterprestasi dengan kurang baik iaitu mudah terjadi *overfitting*. Akhir bagi SVC bersama fungsi *kernel* adalah SVC Sigmoid yang mana, model ini mencari nilai optimum global dengan baik namun penetapan di dalam pemilihan parameter di dalam model ini sangat mempengaruhi kesan sama ada menjadi lebih atau sebaliknya. Berikut, di bawah adalah formula bagi model jenis SVC ini.

Formula bagi SVC Linear oleh Huang et al. (2022) adalah seperti berikut:

$$y(x) = w^T x_i + b \quad \ldots(2.5)$$

Formula bagi SVC Poly oleh Huang et al. (2022) adalah seperti berikut:

$$K(x, x_i) = [x * x_i + 1]^d \quad \ldots(2.6)$$

Formula bagi SVC RBF oleh Huang et al. (2022) adalah seperti berikut:

$$K(x_i, x_J) = exp\left\{-\frac{|x_i - x_j|^2}{2\sigma^2}\right\} \quad \ldots(2.7)$$

Formula bagi SVC Sigmoid oleh Huang et al. (2022) adalah seperti berikut:

$$K(x, x_i) = \tanh\left[v(x. x_i + c\right] \quad \ldots(2.8)$$

2.4.7 Penyelidikan Konsep Model Ramalan Regresi

Penyelidikan ke atas model-model regresi dibincangkan di dalam kebanyakan kajian ramalan pada hari ini. Antara model asas di dalam regresi adalah linear regresi yang mana diterangkan konsep melalui kajian oleh Ahmad (2017) iaitu regresi linear ialah kaedah statistik yang digunakan untuk menubuhkan hubungan antara dua variabel berdasarkan data yang tersedia dalam persamaan linear. Variabel yang dipengaruhi oleh variabel bebas biasanya dipanggil variabel bergantung. Pengenalan nilai purata variabel X dan Y akan dicapai melalui penggunaan teknik analisis regresi. Persamaan berikut menyediakan korelasi antara dua variabel yang dinyatakan sebagai x dan y. Formula *Linear Regression* (LR) oleh Groß & Möller (2023) seperti berikut:

$$y = \beta_0 + \beta_1 \mathrm{z} + \beta_2 x_1 + \cdots + \beta_{w+1} x_w + \varepsilon \qquad \ldots(2.9)$$

di mana z mewakili nilai z_i = 0 jika berkaitan dengan y_i, y mewakili kumpulan 1 dengan z_i = 1, y_i mewakili kumpulan 2, i = 1,..., $n_1 + n_2$, w mewakili parameter bebas x_1,..., x_w and β ialah *intercep*t ordinal yang paling sesuai. Kesilapan variabel ε adalah menganggap untuk mengikuti pengedaran normal dengan ramalan 0 dan varian σ2.

Ridge regression (RR) yang diterangkan oleh Kenney et al. (2023) adalah teknik regresi yang serupa dengan algoritma *Multiple Linear Regression* (MLR). Walau bagaimanapun, dalam RR, faktor penyesuaian β_j dimasukkan sebagai hukuman dalam fungsi meminimunkan kesalahan. Jumlah persegi yang diubahsuai disenaraikan seperti berikut, di mana λ adalah faktor pemberat yang dioptimalkan oleh Kenney et al. (2023).

$$RSS_{RR} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 + \lambda \sum_{m=1}^{M} \beta_m^2 \qquad \ldots(2.10)$$

Di dalam kajian yang sama, turut menerangkan berkenaan model *k-Nearest Neighbors Regression* (KNNR). Bagi KNNR, penulis menyatakan bahawa model ini adalah teknik yang mengukur jarak *Euclidian* antara titik hampir dan semua titik di dalam set latihan. KNNR tidak membuat apa-apa jangkaan mengenai pengedaran data dalam ruang dimensi n. Sebaik sahaja semua jarak diperolehi, algoritma akan melakukan purata nilai jiran k yang terdekat yang ingin diperolehi.

Formula bagi KNNR oleh Kenney et al. (2023) ini adalah seperti berikut:

$$\hat{y}_n = \frac{1}{K}\left(\sum_{k=1}^{K} \min\left(ED_n, k\right)\right) \quad \ldots(2.11)$$

Variabel K menandakan jumlah jiran terdekat dan digunakan untuk mengira hasil yang dijangka. Nilai minimum antara ED_n dan k dinyatakan sebagai $\min(ED_n, k)$. Matlamatnya ialah untuk menentukan jarak *Euclidean* terkecil antara sebatian hampir dan titik latihan.

Untuk model *Random Forest Regression* (RFR), kajian yang sama juga menerangkan konsep model algoritma menggunakan regresi *decision tree*, yang menggunakan ciri-ciri dalam struktur *if-then-else* dan koleksi regresor *decision tree*. Ini melibatkan pelbagai *decision tree* dan nilai ramalan purata. Formula bagi RFR oleh Kenney et al. (2023) ini adalah seperti berikut:

$$\hat{y}_n = \frac{1}{B}\sum_{b=1}^{B} T_b(X_n) \quad \ldots(2.12)$$

Mesin Vektor Sokongan (SVM) ialah model pembelajaran mesin yang beroperasi di atas prinsip pembelajaran terkawal dan mempunyai atribut yang berbeza berbanding dengan model lain dalam bidang pembelajaran mesin. Pelbagai kajian telah dilakukan dengan menggunakan model SVM ini. Satu kajian penyelidikan oleh Molla et al.(2023) menjelaskan konsep SVM ini di mana ianya juga digunakan dalam penggunaan Klasifikasi Vektor Sokongan (SVC) dan regresi adalah boleh dilakukan di mana Regresi Vektor Sokongan (SVR). mempunyai keupayaan untuk secara berkesan menangkap hubungan *non-linear* yang rumit dalam ruang *input-feature*,dengan kerumitan pengiraan daripada dimensi ruang input. Model ini menunjukkan tahap ketepatan yang tinggi dalam ramalan dan menunjukkan tahap yang memuaskan secara keseluruhannya. Mengikut banyak manfaatnya, ia berpotensi boleh digunakan untuk tujuan sasaran hasil. Konsep SVR menggunakan proses pencarian untuk vektor input *Received Signal Strength Indication* (RSSI) yang sebanding dari set latihan untuk mengenal pasti vektor RSSI paling sepadan daripada set latihan. Ini membolehkan

sistem untuk membuat penentuan yang bermakna mengenai estimasi lokasi sasaran yang sesuai. Fungsi *kernel* yang berbeza boleh digunakan untuk menyelesaikan masalah yang dimaksudkan menggunakan Regresi Vektor Sokongan (SVR). Kajian ini mengkaji keberkesanan rangka kerja SVR dalam tugas lokalisasi sasaran dengan menggunakan fungsi kernel yang biasa digunakan. Keterangan bagi formula bagi SVR bersama kernel ini ditunjukkan di bawah. Di mana, $k(z, z_i)$ adalah *Kernel function*, d adalah *degree polynomial*, γ dan c adalah nilai tetap kernal dan z mewakili nilai input vektor.

Formula SVR linear oleh Molla et al. (2023) adalah seperti berikut:

$$k(z, z_i) = z_i^T .\mathrm{z} \qquad \ldots(2.13)$$

Formula SVR RBF oleh Molla et al. (2023) adalah seperti berikut:

$$k(z, z_i) = \exp\left(-\gamma \parallel z - z_i \parallel^2\right) \qquad \ldots(2.14)$$

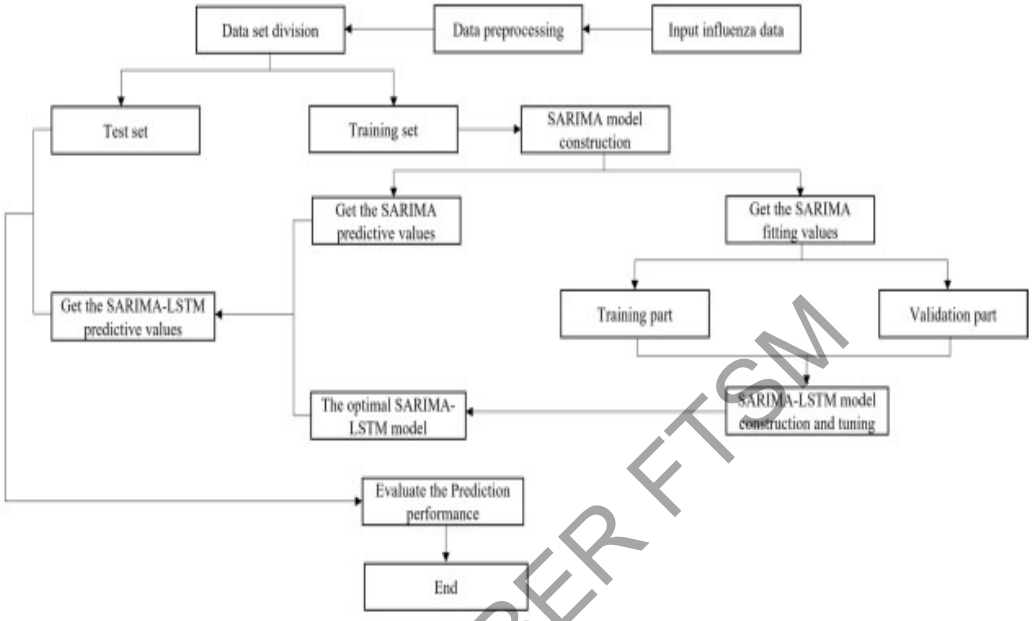
Formula SVR Poly oleh Molla et al. (2023) adalah seperti berikut:

$$k(z, z_i) = (\gamma(z_i^T .\mathrm{z}) + \mathrm{c})^d) \qquad \ldots(2.15)$$

2.4.8 Penyelidikan Konsep Siri Masa

Model *Seasonal Autoregressive Integrated Moving Average* (SARIMA) diterangkan dalam kajian oleh Zhao et al. (2023), yang merupakan model yang digunakan secara meluas dalam analisis siri masa, biasanya diformulasikan sebagai SARIMA (p, d, q) (P, D, Q, s). Bagi set parameter (p, d, q) adalah mewakili parameter data tidak bermusim manakala bagi (P, D, Q, s) mewakili parameter data bermusim. Untuk setiap keterangan parameter, dijelaskan dengan parameter P menandakan urutan p, manakala parameter Q menunjukkan urutan q yang masing – masing digunakan untuk menandakan urutan auto-regresif dan purata bergerak. Di samping itu, pada parameter d dan D menunjukkan tahap perbezaan dalam corak dan musim, masing-masing. Parameter P, Q, dan s menandakan urutan masing-masing *auto-regressive* secara musim, purata bergerak musim, dan tahap tempoh musim. Penyelidikan ini menggunakan peratusan

mingguan penyakit serupa *influenza* (ILI%) yang berkisar dari minggu pertama tahun 2010 hingga minggu akhir tahun 2018 untuk membina model SARIMA.



Rajah 2.4 Proses ramalan model SARIMA-LSTM oleh Zhao et al. (2023)

Prosedur ini merangkumi tahap-tahap proses di dalamnya. Data asal telah menjalani prosedur pra-pemprosesan data, selepas itu ia dibahagikan kepada satu set latihan dan satu set ujian. Model SARIMA yang optimal dibina menggunakan set latihan yang merangkumi minggu pertama 2010 hingga minggu ke-52 2018. Penggunaan set ujian digunakan untuk mengesahkan keberkesanan model. Untuk menangani keupayaan prediktif yang terhad, model SARIMA musim digunakan dalam meramalkan komponen *non-linear* dan ketepatan suboptimal prediktif siri asal. Penyelidikan ini telah membangunkan tiga model iaitu SARIMA, rangkaian saraf SARIMA hibrid dan *Long-Short Term Memory neural network* (SARIMA-LSTM), dan gabungan SARIMA-LSTM berdasarkan *Singular Spectrum Analysis* (SSA-SARIMA-LSTM). Model-model ini digunakan untuk menghasilkan ramalan dan menentukan mod optimal. Rajah di atas menunjukkan rangka kerja keseluruhan bagi model ini.

2.5 KESIMPULAN

Di dalam penemuan metodologi yang bersesuaian dengan kajian, satu pendekatan perolehan fikiran perlu ada dalam mengkaji setiap permasalahan kajian sehingga penyelesaian dan kaedah yang ingin dilaksanakan berjaya mencapai objektif sesuatu kajian. Konsep atau bidang kajian merupakan asas kepada pembinaan model metodologi ini bagi sesuatu kerangka kerja menyeluruh sebelum sesuatu kajian dijalankan. Bagi melihat kepada konsep kajian yang ingin dijalankan, adalah melihat kepada permasalahan dan *domain* isu yang akan dikaji. Di dalam kajian ini, permasalahan yang ada boleh membimbangkan beberapa kelompok yang berperanan dan pakar di dalam sesuatu bidang dalam melihatkan kepada kelemahan atau masalah pada sistem sedia ada sekiranya tidak dikaji punca dan asas sesuatu isu yang timbul. Selain itu, kajian dari penyelidikan sebelum ini menerangkan ketersediaan data penggunaan ialah metodologi yang digunakan untuk menstrukturkan semula utama, sekunder dan *tertier* corak , peristiwa, isu-isu baru muncul, dan kemungkinan keputusan masa depan. Oleh yang sedemikian, kajian ini dijalankan adalah berdasarkan penemuan dalam kajian lepas tentang perincian setiap fasa analisis secara lebih luas bagi memberikan suatu pendekatan baru di dalam menyediakan pengetahuan berhubung dengan kepelbagaian analisa yang telah dijalankan.

BAB III

KAEDAH DAN METODOLOGI KAJIAN

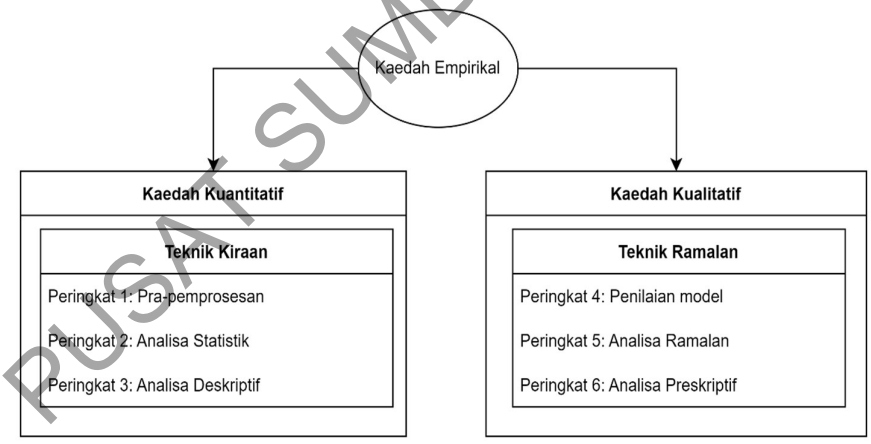
3.1 PENGENALAN

Pada peringkat metodologi, kajian menerangkan kaedah atau teknik yang digunakan sepanjang kajian, berdasarkan tujuan dan skop kajian yang dinyatakan. Metodologi adalah elemen penting dalam penyelidikan. Oleh itu, bahagian ini menerangkan kaedah dan teknik penyelidikan untuk mengatur langkah-langkah dalam proses penyelidikan dengan lebih baik. Metodologi menerangkan bagaimana sesuatu masalah dikenal pasti dan mengapa kaedah itu digunakan. Metodologi ini membantu kajian untuk memahami aplikasi konsep dan kaedah penyelidikan dengan lebih mendalam dan untuk menerangkan proses penyelidikan dengan lebih sistematik.

Konsep asas utama di dalam penghuraian kaedah metodologi kajian ini adalah menerangkan secara menyeluruh dan diberikan pemakluman spesifikasi yang lebih mendalam berkenaan pendekatan yang digunakan di dalam kajian ini. Tujuan bagi pendedahan aspek di dalam metodologi kajian ini adalah untuk memastikan objektif kajian dapat dicapai. Bagi pembentangan kaedah dan metodologi kajian, penerangan asas kajian diterangkan pada bahagian 3.2. Seterusnya, bab ini diikuti dengan penerangan berkenaan fasa-fasa kajian dan kerangka utama kajian berserta peringkat sub-proses kajian pada bahagian 3.3 dan 3.4. Bagi menerangkan konsep struktur data diterangkan pada bahagian 3.5 dan kajian instrumen pula diterangkan pada bahagian 3.6. Bahagian akhir di bahagian 3.7, merumuskan kaedah dan metodologi kajian ini secara menyeluruh.

3.2 KONSEP AWAL KAJIAN

Bagi mencari penyelesaian dan menangani permasalahan ini, kaedah yang diusulkan di dalam kajian ini menjalankan satu kajian empirikal yang mana kajian adalah berdasarkan suatu pemerhatian dan eksperimen dan bukan bergantung hanya kepada teori. Selain itu, kajian tambahan asas analisis yang utama disusun bagi dihasilkan di dalam kajian ini. Bagi asas utama kajian empirikal ini boleh dibahagikan kepada beberapa bahagian iaitu kajian kuantitatif, kualitatif dan seterusnya gabungan kedua kajian tersebut. Kajian berkenaan teknik kajian empirikal yang menggabungkan teknik kajian membawa hasil eksperimen dibahagikan kepada dua kajian iaitu dari kaedah ciri kuantitatif dan ciri kualitatif. Bagi ciri kuantitatif, kaedah analisis sewaktu pra-pemprosesan adalah dijalankan manakala kaedah kualitatif ditunjukkan semasa melakukan kekeliruan matriks selepas proses pengelasan dalam melihat kebenaran ramalan *True Positive* (TP) dan *True Negative* (TN) . Rajah 3.1 di bawah menunjukkan aliran proses bagi kaedah ini.



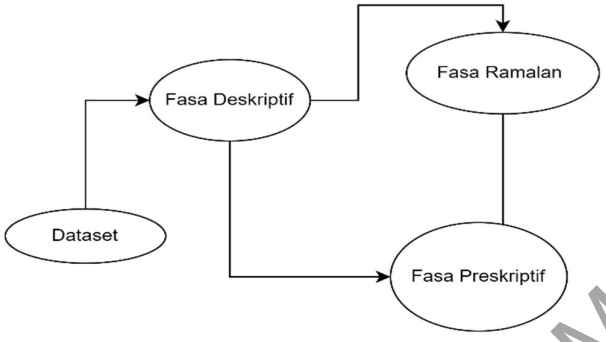
Rajah 3.1 Kaedah Empirikal (Gabungan Teknik)

3.3 PEMBAHAGIAN ANALISA PENYELIDIKAN

3.3.1 Pembentukan Jenis Analisis Fasa

Pendekatan kajian ini adalah dengan membentuk skop analisis kepada beberapa bahagian bagi memboleh fungsi utama analisis dapat ditetapkan. Di dalam kajian ini, tiga pembahagian fasa utama dibentuk iaitu bermula dari fasa deskriptif, fasa ramalan

dan berakhir pada fasa preskriptif. Rajah 3.2 berikut menunjukkan carta alir dan hubungan di antara fasa analisis yang dijalankan dalam kajian ini.



Rajah 3.2 Fasa Analisis dan hubungannya

3.3.2 Definisi Fasa Deskriptif

Fasa ini merupakan tempoh awal permulaan kajian yang dikenali sebagai prasyarat untuk analisis selanjutnya. Secara lebih menyeluruh, analisis kuantitatif berlaku di dalam fasa ini selaras dengan fungsinya dalam mengukur bilangan atau masa yang diperoleh. Perlaksanaan ciri pengekstrakan data berlaku melalui aktiviti pra-pemprosesan data, pembahagian keseluruhan data kepada beberapa bahagian dan mencipta *cluster* data yang sama coraknya atau sebaliknya. Analisis di dalam fasa ini bertujuan untuk menyediakan set data yang telah dibersihkan, menerangkan taburan data mentah yang disediakan, laporan perhubungan antara atribut dan jenis pangkalan data. Pengelompokan biasanya merupakan langkah akhir di dalam analisa ini.

3.3.3 Definisi Fasa Ramalan

Pada peringkat ini, proses pembangunan model ramalan akan dilakukan. Di fasa ini juga disebut sebagai pembelajaran mesin dengan melihat kepada konsep masalah kajian ini, dapat mengkaji sejauh mana masalah ketersediaan data profil beban pengguna berjaya diperoleh oleh meter RMR ini. Menerusi kajian ini, dua pendekatan model dibentuk iaitu pendekatan model ramalan iaitu regresi dan model pengelasan atau klasifikasi. Bagi pembangunan model regresi, nilai sasaran adalah nilai jumlah peratusan ketersediaan data pada setiap meter dengan mempunyai jenis nilai angka berterusan. Manakala bagi nilai sasaran yang mempunyai label nilai binari, pendekatan model klasifikasi dijalankan. Di dalam model ini, sasaran nilai adalah status lengkap

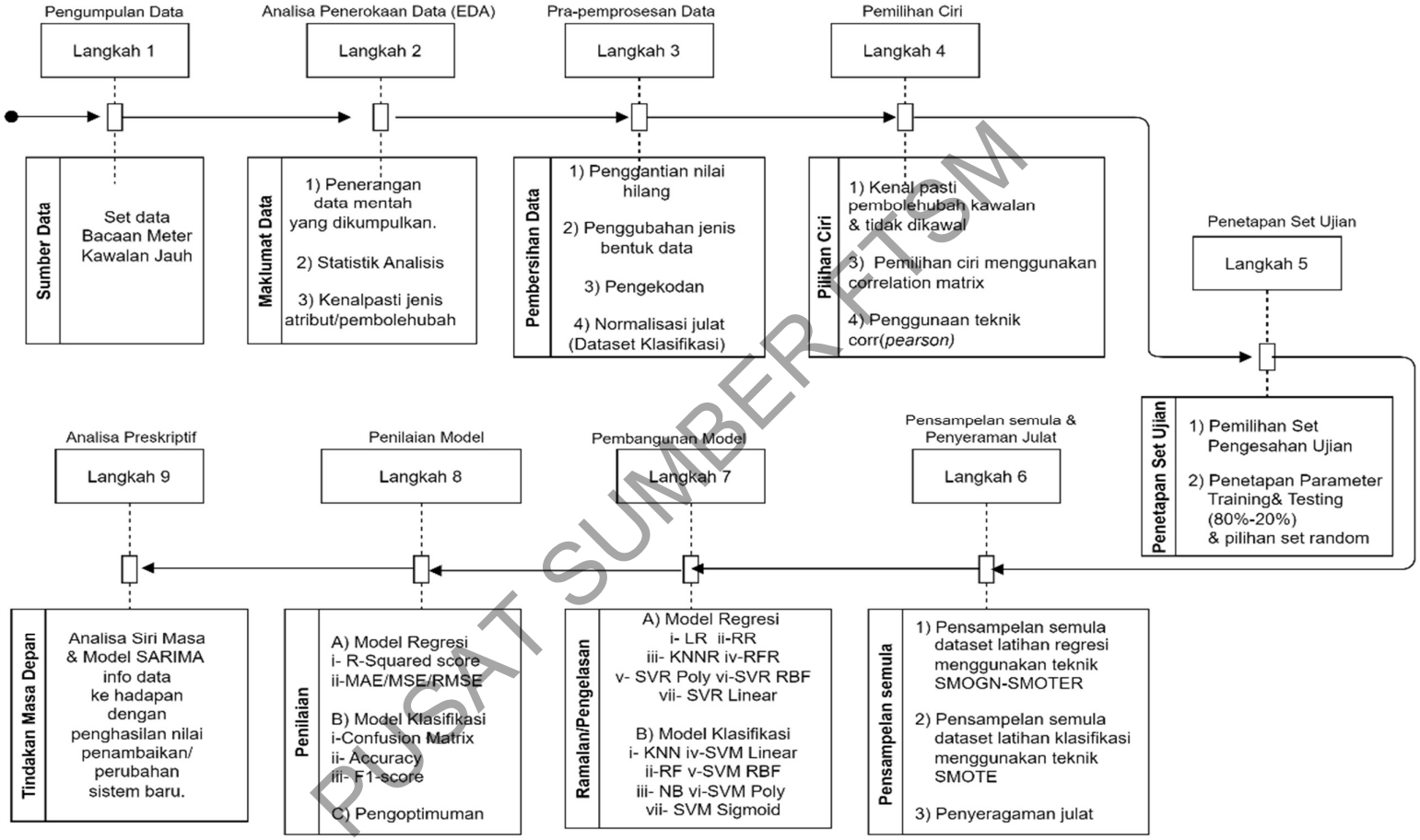
ketersediaan data sama ada 1 atau 0. Sekiranya status ketersediaan data adalah 1, ianya dengan merujuk kepada jumlah peratusan ketersediaan data bersamaan dengan 100 iaitu ketersediaan data adalah penuh. Sebaliknya, status ketersediaan data dengan nilai 0 adalah bersandar kepada nilai ketersediaan data kurang daripada 100 yang mana menunjukkan data tidak berjaya dibawa balik sepenuhnya dari meter RMR kepada sistem pusat dengan baik. Di sini, analisis kualitatif akan dilakukan dalam melihat kepada hasil jangkaan yang terjadi terhadap keputusan ramalan.

3.3.4 Definisi Fasa Preskriptif

Bagi analisis prekriptif ini membolehkan pelbagai tindakan cadangan yang boleh membantu menghasilkan perbagai hasil keputusan mengikut kepada isu jangkaan hasil dari analisis sebelum ini. Ringkasnya, analisis ini bertujuan untuk mengukur kesan keputusan masa depan untuk memberi nasihat tentang hasil yang mungkin sebelum keputusan benar-benar dibuat. Pada tahap terbaiknya, analitik preskriptif meramalkan bukan sahaja apa yang akan berlaku tetapi juga punca permasalahan yang dapat dikaitkan dari analisa di fasa deskriptif sebelum ini. Oleh itu kajian berkenaan siri masa juga diterang di dalam bahagian ini sebagai mengambil analisa yang lebih meluas dengan mengabungkan kajian di peringkat fasa deskriptif dan ramalan.

3.4 KERANGKA KERJA UTAMA

Secara amnya, proses analisis secara analitik ini adalah dihasilkan dengan merujuk kepada saluran proses secara terperinci dan berturutan. Kajian ini telah membentangkan perincian langkah kerja bagi menerangkan fungsi dan peranan setiap peringkat yang dijalankan. Bagi kajian proses ini, beberapa teknik analisis disusunkan dan gabungan di antara beberapa proses ini telah terhasil satu kerangka kerja berdasarkan perlombongan data pembelajaran mesin. Rajah 3.3 di bawah menjelaskan proses keseluruhan bagi kerangka kerja utama yang dilakukan.



Rajah 3.3 Kerangka Kerja Keseluruhan

3.4.1 Langkah 1.0 : Proses Pengumpulan Data

Di bahagian proses pengumpulan data dan dikenali sebagai pangkalan data merupakan langkah pertama dalam melakukan sesuatu kajian. Konsep dan permasalahan isu diproseskan pada peringkat kajian ini. Perincian setiap sub-proses yang berlaku dinyatakan di dalam bahagian berikutnya.

3.4.2 Langkah 1.1 : Permasalahan Isu & Hipotesis Keputusan

Untuk memahami domain kajian yang dijalankan, perbincangan bersama pakar bidang yang berkaitan telah dijalankan bagi mengetahui keperluan kajian dilakukan. Permasalahan isu menerangkan domain kajian adalah berasaskan kesediaan profil beban pengguna dalam penggunaan sistem meter kawalan jauh (RMR) di dalam sektor grid kuasa ini. Bagi kajian utama ini, melibatkan ramalan peratusan ketersediaan data profil pengguna yang berjaya dibawa balik oleh meter RMR. Isu permasalahan dikaji dan andaian hipotesis dilakukan yang mana melibatkan jumlah peratusan data lengkap atau tidak lengkap. Begitu juga dengan kajian selanjutnya melalui pembahagian status kepada data yang baik atau data yang kurang baik. Andaian kepada keputusan dilakukan di samping mencari atribut ubah yang berkaitan dengan permasalahan.

3.4.3 Langkah 1.2 : Integrasi dan Pengumpulan Data

Proses pengumpulan dan mengintegrasikan data dilakukan pada langkah ini. Bagi kajian ini, pemilihan data berstruktur diambil dengan melihat kepada sumber data yang diperoleh. Bagi kajian ini, sasaran ramalan telah ditetapkan di mana proses analitik adalah *Supervise Learning* yang mempunyai label atau ramalan hasil masalah. Integrasi data berlaku dengan mengumpulkan sejarah data pada tempoh masa beberapa tahun kebelakangan dan menyatukan dengan maklumat pemboleh ubah lain yang diperlukan. Perincian berkenaan struktur dan jenis data dijelaskan dalam bahagian selanjutnya.

3.4.4 Langkah 2.0: Analisa Penerokaan Data (EDA)

Pada langkah ini merupakan analisa awal yang akan dijalankan dengan memerlukan kepada gabungan proses aktiviti-aktiviti yang diterangkan di dalam topik ini. Penilaian dan analisa kuantitatif amat penting pada peringkat kajian ini. Pelbagai analisa

dijalankan secara visualisasi dijalankan di dalam peringkat kajian ini. Di dalam rajah-rajah di bawah ditunjukan penggunaan *syntax* di dalam *R-Studio* dalam melakukan *Exploratory data analysis* (EDA) ini.

```
#Visualization using Mosaic Map
png(file="table(data2$state, data2$data_availability_bucket).png",
    width=1200, height=350)
mosaicplot(table(data2$state, data2$data_availability_bucket),
           color = TRUE,
           xlab = "State", # label for x-axis
           ylab = "data_availability_bucket" # label for y-axis
)
dev.off()
```

Rajah 3.4 *Syntax* visualisasi menggunakan *Mosaic Map*

```
#Visualization using Histogram
png(file="hist(data8$day).png",width=600, height=350)
hist(data8$day, col="blue")
dev.off()

png(file="hist_data9$data_availability.png",width=600, height=350)
hist(data9$data_availability)
hist(data9$data_availability, xlab = 'data_availability', ylab = 'Number of
abline(v = mean(data9$data_availability), col='red', lwd = 3)
lines(density(data9$data_availability), col = 'green', lwd = 3)
hist( x = data9$data_availability,
      main = "2010 AFL margins", # title of the plot
      xlab = "Margin",           # set the x-axis label
      density = 10,              # draw shading lines: 10 per inch
      angle = 40,                # set the angle of the shading lines is 40
      border = "gray20",         # set the colour of the borders of the bars
      col = "gray80",            # set the colour of the shading lines
      labels = TRUE,             # add frequency labels to each bar
      ylim = c(0,100)            # change the scale of the y-axis
)
hist(data9$data_availability)
dev.off()
```

Rajah 3.5 *Syntax* visualisasi menggunakan *Histogram plot*

```
#Visualization using Density plots
png(file="hist_data9$data_availability_2.png",width=600, height=350)
density_values <- density(data9$data_availability)
# Create a plot with appropriate limits
plot(density_values,
     xlim = c(min(data9$data_availability), max(data9$data_availability)),
     ylim = c(0, max(density_values$y)),
     main = "Distribution of Data Availability")
# Add the density line
abline(v = mean(data9$data_availability), col='red', lwd = 3)
lines(density_values, col = 'green', lwd = 3)
dev.off()
```

Rajah 3.6 *Syntax* visualisasi menggunakan *Density plot*

Analisis data penerokaan, dalam statistik, ialah kaedah menganalisis set data untuk menggambarkan ciri pentingnya, biasanya menggunakan grafik statistik dan kaedah visualisasi data lain. Analisis statik bertujuan untuk menganalisis data dalam tempoh masa tertentu dengan melakukan pengkajian data yang direkodkan.

```
#Statistical Analysis
library(summarytools)
library(pastecs)

sink(file = "summary(data2)_output.txt")
summary(data2)
sink(file = NULL)

sink(file = "data2$data_availability_bucket_output.txt")
by(data2, data2$data_availability_bucket, summary)
sink(file = NULL)

sink(file = "data2$state_output.txt")
freq(data2$state)
sink(file = NULL)

sink(file = "data2$voltage_level_output.txt")
freq(data2$voltage_level)
sink(file = NULL)

sink(file = "data2$meter_brand_output.txt")
freq(data2$meter_brand)
sink(file = NULL)

sink(file = "dfSummary(data2).txt")
dfSummary(data2)
sink(file = NULL)

sink(file = "stat.desc(data2).txt")
stat.desc(data2)
sink(file = NULL)
```

Rajah 3.7 *Syntax* bagi Analisa Statistik data2

3.4.5 Langkah 3.0 : Pra-pemprosesan data

Di dalam pra-pemprosesan data ini, terhadap pelbagai aktiviti yang berperanan setiap satunya adalah berbeza di antaranya. Setiap kategori aktiviti ini boleh dikelaskan sebagai sub-proses di dalam langkah pra-pemprosesan data sebagai objektif untuk melakukan pembersihan data untuk penyediaan data kepada aktiviti pembelajaran mesin selepas langkah ini. Penggunaan instrumen kajian di setiap aktiviti ini adalah berbeza di mana langkah 3.1 sehingga langkah 3.3 adalah menggunakan perisian *R-studio* manakala langkah 3.4 dan selanjutnya adalah menggunakan perisian *Pyhton.*

3.4.6 Langkah 3.1 : Penggantian nilai hilang pra-pemprosesan data

Aspek pembersihan data dilakukan di dalam langkah ini setelah kajian analisis statistik dijalankan. Dengan proses statistik sebelum ini, maklumat kehilangan nilai akan ditunjukkan. Pendekatan kaedah gantian nilai yang hilang ini juga berbeza mengikut kepada jenis pemboleh ubah tertentu. Kaedah penggantian nilai dilaksanakan mengikut jenis data yang ditetapkan. Penggantian nilai hilang di dalam kajian ini adalah dengan menggunakan dua kaedah iaitu menggantikan secara nilai median dan juga nilai mod. Konsep gantian secara *median* ini diterangkan dengan menunjukkan nilai tengah di antara set data disusun daripada terkecil kepada terbesar. Untuk set data kajian yang digunakan, melalui analisa penerokaan data yang dibuat sebelum ini, hasil pemerhatian melihat kebanyakan nilai di atribut dalam set data tidak memiliki nilai *symetry* atau

mempunyai *skewed* nilai. Oleh itu, penggunaan median dipilih berbanding *mean* bagi teknik gantikan data angka yang hilang.

```
#To find all the rows in a data frame with at least one NA
unique (unlist (lapply (data4, function (x) which (is.na (x)))))
sum(is.na(data4))

# Replace the missing value with median in all dataframe
data5 <- data4 %>%
  mutate_if(is.numeric, function(x) ifelse(is.na(x), median(x, na.rm = T),

#To re-confirm all the rows in a data frame with at least one NA
unique (unlist (lapply (data5, function (x) which (is.na (x)))))
sum(is.na(data5))

#To re-confirm all the rows in a data frame with at least one Null
unique (unlist (lapply (data5, function (x) which (is.null (x)))))
sum(is.null(data5))
```

Rajah 3.8 Teknik penggantian menggunakan *median*

Bagi penggantian nilai hilang dengan penggunaan *mode*, teknik ini dipraktikkan bagi jenis data yang mempunyai nilai karakter. Teknik penggantian dengan *mode* ini akan melakukan gantian dengan nilai karakter yang paling kerap ditemui dalam set data tersebut.

```
Mode <- function(x) {
  ux <- na.omit(unique(x))#excludes NA values
  tab <- tabulate(match(x,ux)); ux[tab == max(tab) ]
}

#test on columns
Mode(data5$comm_type)
Mode(data5$aging)

#replace the NA by the mode(factor data)
data5$comm_type[is.na(data5$comm_type)]<-Mode(data5$comm_type)
data5$aging[is.na(data5$aging)]<-Mode(data5$aging)

data5v2 <- data5
sapply(data5v2, function(x) sum(is.na(x)))
data5v2 %>% summarise_all(~ sum(is.na(.)))
```

Rajah 3.9 Teknik penggantian menggunakan *mode*

3.4.7 Langkah 3.2 :Penggubahan jenis data pra-pemprosesan data

Aktiviti penggubahan jenis data dilakukan dalam kajian bagi memastikan penyeragaman jenis data untuk memudahkan analisa lanjut yang diperlukan. Penggubahan jenis data ini diberikan contoh sekiranya data jenis *string* ditukarkan kepada jenis *integer* atau lain-lain jenis data.

```
#Change data types
data8$total_missing_interval <- as.numeric(
  as.character(data8$total_missing_interval))
data8$voltage_level <- as.numeric(
  as.character(data8$voltage_level))
```

Rajah 3.10 *Syntax* pengubahan data *types*

3.4.8 Langkah 3.3 : Perubahan dan Pengekstrakan data pra-pemprosesan data

Bagi kajian di bahagian ini, pengekstrakan data dilakukan pada atribut tertentu bagi menambahkan atribut yang menjelaskan maklumat kepada model ramalan ini. Untuk aktiviti di peringkat ini, data maklumat tarikh meter RMR dipasang di tapak premis dicambahkan lagi kepada hari, masa dan tahun.

```
#remove duplicate
library(dplyr)
data3 <- distinct(data2)
data3

#Transform date into year,month & day
library(tidyverse)
library(lubridate)
data4 = data3 %>%
  mutate(date = dmy(meter_installation_date)) %>%
  mutate_at(vars(date), funs(year, month, day))
```

Rajah 3.11 *Syntax* bagi pengekstrakan data pada tarikh meter di pasang

Nilai data ini adalah penting dalam menjalankan analisis siri masa dalam melakukan penyelidikan yang lebih terperinci mengikut kepada perbandingan hari, bulan dan tahun bagi permasalahan isu yang berlaku di dalam ketersediaan data profil beban pengguna ini. Selain itu, dengan penggunaan teknik pengekstrakan ciri ini, ianya akan memberi kesan ke dalam perubahan dimensi data menjadi lebih kecil di mana ciri baharu yang terhasil daripada gabungan ciri asal ini mempunyai dimensi yang lebih kecil.

3.4.9 Langkah 3.4 : Pengekodan pra-pemprosesan data

Di dalam langkah ini, proses pengekodan dilakukan bagi tujuan menukarkan penyediaan set data asal kepada set data yang boleh dijalani kepada kesemua jenis model ramalan di peringkat pembelajaran mesin. Ciri dengan nilai kategori boleh ditukar kepada nilai berangka dengan menggunakan satu salah satu kaedah pengekodan yang yang boleh memainkan peranan penukaran kesemua nilai kategori yang diperlukan. Untuk memulakan proses pengekodan di dalam kajian ini, pendekatan yang berbeza untuk menangani data kategori di ambil. Sebagai alternatif, tindakan untuk memilih data tertentu untuk dialih keluar daripada set data supaya peranan pengekodan dapat dijalankan dalam mengubah data kepada nilai numerikal yang berfungsi untuk analisa seterusnya. Di dalam hal ini, semakan terhadap set data dilakukan terlebih

dahulu bagi melihat secara keseluruhan nilai unik di setiap atribut. Apabila nilai data hampir sepenuhnya unik dengan bilangan yang tinggi, hanya menurunkan potensi data dan membuang data ini adalah pendekatan berguna untuk proses pengekodan ini.

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder

labelencoder = LabelEncoder()

df_le['State'] = labelencoder.fit_transform(df_le['State'])
df_le['RateCategory'] = labelencoder.fit_transform(df_le['RateCategory'])
df_le['DeviceCat'] = labelencoder.fit_transform(df_le['DeviceCat'])
df_le['CommType'] = labelencoder.fit_transform(df_le['CommType'])
df_le['SOStatus'] = labelencoder.fit_transform(df_le['SOStatus'])
df_le['ServiceMode'] = labelencoder.fit_transform(df_le['ServiceMode'])
df_le['MeterBrand'] = labelencoder.fit_transform(df_le['MeterBrand'])
df_le['DailyCallStatus'] = labelencoder.fit_transform(df_le['DailyCallStatus'])
df_le['ServicModeSuccessStatus'] = labelencoder.fit_transform(df_le['ServicModeSuccessStatus'])
df_le['FailureReason'] = labelencoder.fit_transform(df_le['FailureReason'])
df_le['DataAvailabilityBucket'] = labelencoder.fit_transform(df_le['DataAvailabilityBucket'])
```

Rajah 3.12 Pengekodan label menggunakan kaedah *LabelEncoder*

Menerusi Rajah 3.12 di atas, kaedah pengekodan yang digunakan adalah kaedah pengekodan label iaitu *LabelEncoder*. Untuk memindahkan pemboleh ubah kategori ke label *integer* disebabkan oleh algoritma pembelajaran mesin tidak dapat berinteraksi langsung dengan nilai kategori. Justeru, pengekodan mesti dilakukan menjadi nilai numerikal. *LabelEncoder* mempunyai fungsi mengurutkan nilai kategorikal mengikut urutan yang diberikan atau mengikut alfabet, dan kemudian memberikan angka unik untuk setiap nilai. Sebagai contoh, apabila kajian ini mengambil daftar negeri Selangor, Melaka, dan Kedah, *LabelEncoder* akan mengubahnya menjadi 0, 1, dan 2 berturut-turut. Penggunaan angka ini boleh dijadikan sebagai input untuk algoritma pembelajaran mesin. Kaedah pengekodan label menggunakan *LabelEncoder* memberikan nilai kepada setiap kategori data mengikut hierarkinya dan hanya boleh digunakan untuk data kategorikal tanpa hubungan ordinal yang jelas. Oleh itu, kaedah ini adalah menjadi pilihan memandangkan ketiadaan jenis ordinal di dalam kategorikal data di dalam set kajian ini.

3.4.10 Langkah 3.5 : Normalisasi Julat pra-pemprosesan data

Di bahagian akhir proses pra-pemprosesan ini adalah dengan menjalankan aktiviti normalisasi julat bagi tujuan untuk membolehkan perbandingan antara ketersediaan data dengan magnitud yang berbeza. Dalam kajian ini, penggunaan nilai puncak maksimum dan nilai minimum setiap data digunakan dalam kaedah skala untuk

menormalkan nilai set data ini. Kaedah ini juga dipraktikkan menggunakan skala normalisasi dengan penggunaan *MinMaxScaler*. Kepentingan penggunaan kaedah normalisasi ini adalah penting pada peringkat kajian ini memandangkan pengedaran data tidak diketahui dengan jelas sama ada telah sepadan dengan pengedaran ciri *Gaussian* atau pun tidak. Ciri taburan *Gaussian* ini adalah penting dalam proses pembangunan model pembelajaran mesin selanjutnya. Selain daripada itu, peranan normalisasi ini digunakan untuk mengalih keluar ciri yang tidak diingini daripada set data yang dikenali sebagai *outlier*. Penormalan ini akan menjejaskan *outlier* disebabkan proses penjulatan berskala ini akan menghasilkan julat sempadan. Melihat kepada beberapa model ramalan yang akan dijalankan, pendekatan teknik normalisasi ini juga akan dilakukan dengan melihat kepada keperluan dan ciri kaedah ramalan yang digunakan seperti ramalan klasifikasi. Oleh itu, bagi kajian penyelidikan ramalan regresi, pendekatan normalisasi tidak dilakukan. Dalam situasi ini, di mana sasaran output mewakili frekuensi atau peratusan , teknik normalisasi keseluruhan data adalah tidak bersesuaian. Begitu juga jika sasaran menunjukkan pengedaran atau julat yang unik, ianya adalah tidak signifikan dalam konteks isu yang sedang berlaku. Oleh hal yang sedemikian, kajian ini melakukan pengekalan pada set data bagi model ramalan regresi dan normalisasi hanya kepada set data untuk model klasifikasi sahaja.

3.4.11 Langkah 4.0 : Pemilihan ciri

Fungsi bahagian ini adalah sangat penting dalam mengemukakan suatu kaedah pemilihan ciri data bagi melihat hubungan di antara ciri yang menyumbang kepada ramalan ketersediaan data atau tidak. Untuk memastikan ciri-ciri atribut ini tidak menjejaskan hasil model ramalan, pemilihan ciri tertentu perlu dikenal pasti. Jika tiada langkah pemilihan ciri ini, kerumitan kepada model ini dalam memacu data semasa proses pembelajaran mesin yang mana secara tidak langsung membawa kepada tinggi masa pengkomputeran, dan menghasilkan juga variasi bias yang tinggi di dalam hasil keputusan. Bagi menjalankan proses pemilihan ciri tersebut, kajian ini telah menggunakan teknik pemilihan ciri dengan menggunakan kaedah *correlation matrix* yang merupakan satu kaedah korelasi matrik iaitu mengukur hubungan *linear* melibatkan dua atribut. Teknik yang digunakan ini dikenali sebagai *Pearson's correlation coefficient* yang mengkaji cara atribut dalam model ramalan beban berinteraksi antara satu sama lain.

Formula Pearson's correlation coefficient Hayes et al. (2015) adalah seperti berikut:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2 \sum_{i=1}^{n}(y_i - y)^2}} \quad \ldots(0.1)$$

Merujuk kepada formua di atas, di mana i = 1, .., dan n ialah nilai index pada x_i dan y_i dan atribut bebas masing-masing. Parameter x dan y mewakili nilai purata atribut bersandar yang dikira di dalam kajian ini. Kekuatan perkaitan hubungan linear dua atribut diukur melalui teknik ini. Bagi makluman berkenaan *correlation matrik* ini digunakan dengan penggunaan *corr()* di dalam Pandas di mana konsep perkaitan ciri berlaku berasaskan pada setiap subset atribut terhadap sasaran atribut y. Di peringkat langkah ini, penyelidikan untuk melihat sejauh mana setiap pemboleh ubah ini berhubung dapat dilihat apabila hasil pekali hampir kepada 1 dan -1, ianya menunjukkan korelasi yang sempurna dalam mengetahui ciri atribut terhadap ketersediaan data dapat dilakukan. Untuk membentuk subset ciri, ciri yang mempunyai sedikit korelasi iaitu nilai menghampiri 0 dikategorikan sebagai ciri berlebihan yang boleh dihapuskan.

3.4.12 Langkah 5.0 : Penetapan set ujian

Langkah di dalam bahagian ini menyediakan set data bagi klasifikasi Algoritma Pembelajaran Mesin yang akan digunakan. Semasa melatih model pembelajaran mesin, set data yang tersedia biasanya dibahagikan menjadi dua bahagian utama. Yang pertama ialah subset pelatihan, dan yang kedua ialah subset pengujian atau validasi. Model dilatih dengan subset pelatihan, manakala subset pengujian digunakan untuk menguji keberhasilan model yang dilatih. Analisa kajian bagi kedua-dua model ramalan regresi dan klasifikasi dilakukan pemisahan set data terlebih dahulu. Sekiranya model regresi dibentuk, data dengan sasaran ramalan iaitu peratusan ketersediaan data diasingkan mengikut pemisahan set data kepada dua bahagian. Pendekatan yang sama juga diambil bagi membentuk model klasifikasi dengan tujuan pengelasan status wujud ketersediaan data ataupun tidak. Melalui kajian ini, perimbangan pembahagi antara dua subset ditentukan mengunakan perimbangan secara amnya iaitu 80% untuk pelatihan dan 20% untuk pengujian. Dengan data yang tersedia dan kompleksiti isu yang dibincangkan, 80% dipilih menjadi subset latihan bagi membolehkan data set dilatih

lebih daripada separuh data untuk memboleh model ujian telah dilatih dan diajar oleh majoriti keadaan data terlebih dahulu. Oleh itu, 80% data pertama data akan menjadi subset pelatihan dan 20% data terakhir akan menjadi subset pengujian. Di dalam peringkat ini juga dilakukan penyeragaman julat selepas pembahagian set data dilakukan. Penggunaan *StandardScaler* yang mana ciri-ciri kumpulan data diubah menghasilkan *mean* yang kebanyakkan 0 dan *standard deviation* iaitu 1. Proses penyeragaman juga ini bermanfaat ke atas algoritma pembelajaran mesin yang apabila data ditetapkan pada skala yang sama. Teknik *fit()* pada objek skala dengan menyediakan *X_train1* sebagai argumen. Metodologi ini juga digunakan untuk menentukan *standard deviation* dan *mean* dalam *X_train1*, yang akan digunakan untuk penskalaan. Kemudian, teknik *transformasi()* mengubah *X_train1* menggunakan *standard deviation* dan *mean* yang telah dikira sebelum ini. Hasilnya disimpan oleh variabel *X_train*. Kemudian, sekali lagi dilakukan *transformasi()* pada *X_test1* menggunakan *standard deviation* dan *mean* yang telah dikira daripada *X_train1*. Dengan menggunakan *StandardScaler*, kedua-dua *X_train* dan *X_test* akan mempunyai *mean* 0 dan *standard deviation* 1 yang membolehkan perbandingan, pemprosesan, dan pengiraan jarak yang lebih konsisten di antara cirinya.

3.4.13 Langkah 6.0: Pensampelan semula

Pada peringkat ini, pensampelan semula berlaku hanya kepada set data latihan dan tidak digunakan di dalam set data ujian. Keadaan di dalam set latihan yang telah dibahagi oleh pembahagi set sebelum ini, juga memberikan keadaan kandungan kelas yang sama dari set data asal. Hal ini kerana ia boleh memberi kesan kepada prestasi kepada algoritma latihan set pembelajaran mesin. Untuk menangani isu ketidakseimbangan kelas dalam subset latihan ini, pendekatan pensampelan semula atau pemerolehan semula digunakan dengan menggunakan teknik SMOTE dan SMOGN- SMOTER yang digunakan di dalam penyelidikan. Kaedah pensampelan semula yang paling biasa digunakan ialah SMOTE bagi mengimbangi set data kepada data yang seimbang. Untuk mencapai matlamat ini, penggunaan kaedah SMOTE digunakan di dalam sampel model klasifikasi dalam pengelasan kelas sasaran dengan melakukan sampel minoriti dipilih secara rawak di mana set sampel baru untuk kelas ini dicipta dengan menggabungkan ciri-ciri sampel minoriti yang sedia ada sehingga jumlah sampel minoriti meningkat dan menjadi seimbang dengan set sampel majoriti.

```
#Importing SMOTE
from imblearn.over_sampling import SMOTE
#Create an oversampled training data
sm =  SMOTE(random_state=101)
#X_res_sm,y_res_sm = smote.fit_resample(X_train,y_train)
X_train_res, y_train_res = sm.fit_resample(X_train1,y_train1)
ax = y_train_res.value_counts().plot.pie(autopct='%.2f')
_ = ax.set_title("SMOTE-sampling")
```

Rajah 3.13 Penggunaan Teknik SMOTE di dalam Model Klasifikasi

Penggunaan SMOTE adalah sebagai sebuah teknik pensampelan yang dikenali sebagai *imbalanced-learn* dan teknik ini mengimport kelas SMOTE daripada modul *over_sampling* di *imblearn* dalam kod ini. SMOTE yang digunakan pada ketetapan *random_state* bersamaan 101 adalah sebuah argumen *random_state* yang di set kepada nilai 101. Penggunaan *random_state* ini akan menghasilkan hasil yang sama untuk setiap latihan set. Dengan kata lain, algoritma SMOTE akan menggunakan urutan yang sama untuk menghasilkan sampel set baru bagi memastikan bahawa keputusan SMOTE adalah konsisten dan boleh diulang semula apabila *random_state* yang sama digunakan. Bagi teknik SMOGN.SMOTER yang digunakan dalam kajian ini adalah merujuk kepada pembangunan model ramalan regresi yang berbeza dengan model klasifikasi sebelum ini. Namun, ketidakseimbangan kelas dalam set data juga berlaku semasa pembelajaran mesin ramalan regresi. Keperluan bagi menggunakan *Imbalance* yang berfokus kepada regresi dijalankan dalam kajian ini. Pendekatan kaedah pensampelan semula ini menggabungkan idea SMOTE dengan kaedah regulasi yang ditingkatkan, menghasilkan sampel sintetik untuk kelas minoriti. Rajah di bawah menunjukkan teknik SMOGN-SMOTER digunakan dengan mengaplikasikan pendekatan ini di dalam set data latihan. Untuk mencapai matlamat ini, teknik SMOGN.SMOTER ini membantu mengenal ciri-ciri sampel sintetis dinilai dalam meningkatkan tahap ketidakpastiannya dan menghasilkan sampel sintetis yang lebih realistik dan mengurangkan kemungkinan *over-fitting.*


```
## conduct smogn
train_smogn = smogn.smoter(

    data=train_df,
    y='DataAvailability'
)
```

Rajah 3.14 Teknik SMOGN-SMOTER di dalam Model Regresi

3.4.14 Langkah 7.0 : Pembangunan Model Regresi dan Klasifikasi

Pada peringkat langkah ini, pembangunan model akan dilakukan mengikut kepada kaedah yang ditetapkan. Kajian model regresi dan klasifikasi yang dijalankan pada set data meter RMR ini memerlukan pengesahan latihan terlebih dahulu daripada beberapa jenis algoritma yang berbeza karakter dan fungsinya dalam melakukan ramalan. Oleh itu, kepelbagaian jenis model ramalan diperlukan bagi kajian ini dalam melihat sejauh mana model ramalan ini bertindak dan perbandingan ciri setiap model dapat dilakukan.

3.4.15 Langkah 7.1 : Pembangunan model-model regresi

Pembangunan model-model regresi yang dilakukan dalam kajian ini adalah dipilih melalui kebiasaan model yang sering digunakan dan juga percambahan model SVM dilakukan di dalam bahagian ini. Oleh itu, tujuh model yang dipilih dalam kajian ini iaitu diantaranya ialah *Linear Regression, Ridge Regression*, *Random Forest Regressor, K-Nearest Neighbors Regressor*, SVR *Linear*, SVR *polynomia*l dan SVR *Radial Basis Function*.

```
from sklearn.linear_model import LinearRegression
# Instantiate LinearRegression object
linear_reg = LinearRegression()
# Train model on training data
linear_reg.fit(X_train_res, y_train_res)
# Predict trained model on test data
prediction_1 = linear_reg.predict(X_train_res)##oof preds on my trainset
prediction_2 = linear_reg.predict(X_train)##oof preds on my trainset
prediction = linear_reg.predict(X_test)##oof preds on my whole test set
```

Rajah 3.15 Model *Linear Regression* (LR)

Bagi model *linear regression* menggunakan kod perolehan daripada *LinearRegression*. Variabel tidak bergantung *X* dan bergantung *y* dihubungkan melalui penggunaan *linear_reg.fit* dengan persamaan garis lurus yang menunjukkan model asas regresi linear.

```
from sklearn.linear_model import Ridge
# Instantiate Ridge Regression object
#ridge = Ridge(alpha=0.05, normalize=True)
ridge = Ridge(alpha=0.05)
# Train model on training data
ridge.fit(X_train_res, y_train_res)
# Predict trained model on test data
ridge_prediction = ridge.predict(X_test)
```

Rajah 3.16 Model *Ridge Regression* (RR)

Pemilihan model *Ridge Regression* dilakukan bagi melihat perbezaan prestasi dengan *Linear Regression* sebelum ini. Kaedah *Ridge Regression* menggunakan regularisasi untuk mengurangkan *overfitting* dalam regresi linear regresi. Di dalam kajian ini, *Ridge Regression* menggunakan juga penggunaan *alpha* bersamaan 0.05 yang berfungsi sebagai parameter yang menentukan tahap regularisasi yang akan digunakan pada model. Dengan membatasi nilai koefisien regresi, *alpha* meningkatkan kestabilan model dan mengurangkan *overfitting*.

```
from sklearn.neighbors import KNeighborsRegressor
knn_model = KNeighborsRegressor(n_neighbors=3)
# Train model on training data
knn_model.fit(X_train_res, y_train_res)
# Predict trained model on test data
prediction_1 = knn_model.predict(X_train_res)##oof preds on my trainset
prediction_2 = knn_model.predict(X_train)##oof preds on my trainset
prediction = knn_model.predict(X_test)##oof preds on my whole test set
# Calculate the performance
y_pred_knn= knn_model.predict(X_test)
rsquared = knn_model.score(X_test, y_test)
```

Rajah 3.17 Model *K Nearest Neighours Regressor* (KNN)

Model *K Nearest Neighbors Regressor* (KNN) digunakan dalam model ini dalam melakukan analisis regresi dalam kajian ini. Pemodelan menggunakan KNN ini ialah ramalan regresi berdasarkan suara majoriti jiran terdekat dengan meletakkan jiran terdekatnya dalam ruang ciri sebagai ramalan. Penetapan objek *K Neighbors Regressor* pada bahagian knn = *K Neighbours Regressor* r(*n_neighbors*=3) sebagai langkah untuk menetapkan parameter *n_neightbors* kepada 3. Ini menunjukkan bahawa model ini ingin menggunakan tiga jiran terdekat dalam algoritma KNN.


```
 #Import the model we are using
from sklearn.ensemble import RandomForestRegressor
# Instantiate model with 1000 decision trees
rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
# Train the model on training data
rf.fit(X_train_res, y_train_res)
# Predict trained model on test data
prediction_1 = rf.predict(X_train_res)##oof preds on my trainset
prediction_2 = rf.predict(X_train)##oof preds on my trainset
prediction = rf.predict(X_test)##oof preds on my whole test set
# Calculate the performance
y_pred_rf= rf.predict(X_test)
rsquared = rf.score(X_test, y_test)
```

Rajah 3.18 Model *Random Forest* Regressor (RF)

Random Forest Regressor adalah model yang digunakan bagi analisa kajian regresi ini. Pemilihan model ini menggunakan *library* daripada *scikit-learn* di mana ini

termasuk kelas *RandomForestRegressor* yang terdiri daripada satu siri pokok keputusan yang dihasilkan secara rawak. Setiap pokok keputusan menghasilkan ramalan berdasarkan maklumat yang diberikan. Penetapan *n_estimators* = 1000 dan *random_state* = 42 adalah bertujuan menentukan jumlah pohon sebanyak 1000 dalam *Random Forest* dan menganalisis secara rawak pada nilai tertentu iaitu 42. Oleh itu, penggunaan parameter tersebut, membolehkan model analisa *Random Forest Regressor* dengan 1000 pohon dan menggunakan bilangan rawak yang sama setiap kali menjalankan latihan model pada nilai 42 yang ditetapkan dalam kajian ini.

```
import numpy as np
from sklearn import svm
from sklearn.svm import SVR
import matplotlib.pyplot as plt
svr_lin = svm.SVR(kernel="linear",C=1, gamma="scale")
svr_lin.fit(X_train_res,y_train_res)
y_pred_lin = svr_lin.predict(X_test)
```

Rajah 3.19 Model SVR Linear

Dalam konteks regresi, pendekatan *Support Vector Machine Regression* (SVR) digunakan. SVR ialah istilah yang merujuk kepada algoritma pembelajaran mesin yang digunakan dalam tugas regresi yang bertujuan untuk membangunkan model yang dapat meramalkan nilai output berterusan berdasarkan input. Algoritma ini juga mencari hiperpaksi optimum yang linear di mana hiperpaksi berfungsi sebagai garis regresi yang terbaik menghampiri titik data iaitu *support vector.* Dengan kata lain, SVR linear ialah SVR tanpa kernel yang mana ianya adalah salah satu variasi SVR di mana fungsi kernel tidak digunakan dan model regresi yang dibangunkan. Objek SVR dengan *kernel linear* dibuat seketika melalui penggunaan sintaks SVR*(kernel='linear')* hanya parameter C memainkan peranan dalam mengawal model. Parameter gamma tidak menjejaskan kernel linear kerana kernel linear tidak memperkenalkan perubahan dalam jarak dengan parameter gamma. Bagi penggunaan parameter C ianya berfungsi sebagai mengawal kompromi antara toleransi kesilapan dan kerumitan model. Nilai C yang lebih besar memberi hukuman yang lebih tinggi dan signifikan untuk kesilapan, yang membawa kepada model yang lebih ketat dan kecenderungan untuk menjadi *overfitting*. Sebaliknya, nilai C yang lebih kecil memberi toleransi kesilapan yang lebih besar, yang membawa kepada model yang lebih longgar dan cenderung untuk *underfitting.* Oleh itu, di dalam kajian ini , pemilihan parameter C dengan nilai 1 adalah sebagai *default*

yang secara amnya dianggap sebagai nilai moderat dan seimbang yang digunakan dalam kajian.Model ini kemudiannya dilatih dengan menggunakan kaedah *svr.fit(X, y)*. Selepas model dilatih, model digunakan untuk membuat ramalan pada data baharu menggunakan *svr.predict()*. Dalam contoh di atas, kajian ini menggunakan model terlatih untuk meramal nilai output.

```
import numpy as np
from sklearn import svm
from sklearn.svm import SVR
import matplotlib.pyplot as plt
svr_rbf = svm.SVC(kernel='rbf', C=1,gamma='scale')
svr_rbf.fit(X_train_res,y_train_res)
y_pred_rbf = svr_rbf.predict(X_test)
```

Rajah 3.20 Model SVR RBF

SVR RBF *(Radial Base Function)* ialah algoritma *Support Vector Regression* yang menggunakan fungsi *kernel rbf* sebagai asasnya di mana dibuat melalui penggunaan sintaks SVR*(kernel='rbf')*. Dalam konteks yang lebih mudah difahami, RBF SVR digunakan untuk melakukan regresi, iaitu meramalkan nilai output berdasarkan data input. Fungsi *kernel rbf* memodelkan hubungan antara input dan output menggunakan fungsi dasar *radial* yang mencerminkan jarak relatif antara titik data. Dengan menggunakan SVR RBF, kajian ini boleh mewakili hubungan *non-linear* antara input dan output yang berguna dalam kes-kes apabila hubungan antara variabel input dan bukan dari hubungan yang *linear*. Penggunaan parameter C bersamaan 1 dan *gamma* adalah *scale* yang mana berperanan sebagai nilai yang signifikan terhadap model ini. Ini bermakna bahawa nilai C menjelaskan sejauh mana model akan mengikuti atau memperbaiki kesilapan latihan. Parameter C mendefinisikan kompromi antara hukuman untuk kesilapan dan jumlah titik data yang boleh diterima sebagai vektor sokongan iaitu *support vectors*. Jadi, dalam konteks SVR dengan kernel RBF, parameter C memainkan peranan penting dalam mengawal kerumitan model dan toleransi ralat. Begitu juga dengan *gamma* menggunakan *scale* yang merujuk kepada penyesuaian automatik skala parameter *gamma* berdasarkan data dengan mengawal sejauh mana setiap titik data mempengaruhi pembentukan model. Nilai *gamma* yang lebih besar mengakibatkan julat pengaruh yang lebih kecil, dengan titik data yang lebih dekat. Sebaliknya, nilai *gamma* yang lebih kecil memberikan pelbagai pengaruh yang lebih besar, termasuk titik data yang lebih jauh. Apabila nilai *gamma* ditetapkan kepada

scale, skala *gamma* dikira secara automatik berdasarkan pengiraan *default* yang dijalankan dalam pustaka Scikit-learn di *library* Python.

```
import numpy as np
from sklearn import svm
from sklearn.svm import SVR
import matplotlib.pyplot as plt
svr_poly = svm.SVC(kernel='poly', C=1,gamma='scale')
svr_poly.fit(X_train_res,y_train_res)
y_pred_poly = svr_poly.predict(X_test)
```

Rajah 3.21 Model SVR Poly

Rajah di atas menunjukkan penggunaan cambahan SVR lain yang digunakan dalam kajian ini yang menggunakan SVR Polinomial merujuk kepada istilah *Support Vector Regression Polynomial.* SVR polinomial digunakan untuk melakukan regresi menggunakan fungsi dasar polinomial yang berperanan menukarkan ruang ciri asal kepada ruang ciri dimensi yang lebih tinggi. Dalam kajian ini, SVR polinomial berfungsi sebagai model regresi *non-linear* yang mampu menangani hubungan kompleks antara variabel input dan output. *Kernel* polinomial membolehkan SVR untuk mewakili hubungan *non-linear* dengan mencari corak dan struktur dalam data yang tidak boleh disusun secara *linear.* Model ini menggunakan parameter utama yang akan ditentukan ialah darjah polinomial dan konstan *kernel* yang menentukan berat relatif sumbangan antara elemen linear dan *non-linear* dalam fungsi dasar polinomial. Walau bagaimanapun, dalam senario kajian ini, tiada parameter darjah polinomial digunakan dan hanya parameter C=1 dan *gamma=scale* digunakan, yang menyarankan bahawa nilai darjah polinomial menggunakan nilai *default.* Menurut *library* Scikit-learn, nilai *default* untuk darjah polinomial ialah 3. Oleh itu, dengan menggunakan C=1 dan *gamma=scale* tanpa mendefinisikan nilai darjah polinomial secara signifikannya, menawarkan penetapan *default* pada darjah polinomial iaitu 3 dalam model SVR polinomial ini.

3.4.16 Langkah 7.2 : Pembangunan model-model klasifikasi

Pembangunan model-model klasifikasi yang dilakukan dalam kajian ini adalah dipilih melalui kebiasaan model yang sering digunakan dan juga percambahan model SVM dilakukan di dalam bahagian ini. Tujuh model yang terlibat antaranya adalah *K-Nearest Neighbors*(KNN), *Naive Bayes*(NB), *Random Forest* (RF), *Support Vector Classifier* dengan *linear kernel* (SVC Linear), *Support Vector Classifier* dengan *radial basis*

kernel (SVC RBF), *Support Vector Classifier* dengan *polynomial kernel* (SVC Poly) dan *Support Vector Classifier* dengan *sigmoid kernel* (SVC Sigmoid).

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train_res, y_train_res.values.ravel())
pred= knn.predict(X_test)
pred_prob = knn.predict_proba(X_test)
```

Rajah 3.22 Model Klasifikasi *K-Nearest Neighbors*(KNN)

Pemodelan menggunakan KNN ini ialah dengan konsep kaedah bukan parametrik yang mengklasifikasikan data berdasarkan suara majoriti jiran terdekat. Dalam kajian ini, klasifikasi model KNN dibina bertujuan untuk meramalkan label kelas atau sampel dengan meletakkan jiran terdekatnya dalam ruang ciri dan menggunakan majoriti daripada kelas-kelas berdekatan sebagai ramalan. Berhubung kepada penetapan objek *KNeighborsClassifier* pada bahagian *n_neighbors=3* sebagai langkah untuk menetapkan parameter *n_neightbors* kepada 3. Ini menunjukkan bahawa model ini ingin menggunakan tiga jiran terdekat dalam algoritma KNN.

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(X_train_res, y_train_res.values.ravel())
prediction = rf.predict(X_test)
prediction_prob = rf.predict_proba(X_test)
```

Rajah 3.23 Model Klasifikasi Random Forest (RF)

Random Forest telah diimplementasikan dalam model yang digunakan dalam kajian ini di mana ini termasuk kelas *RandomForestClassifier* untuk membantu dengan masalah klasifikasi. *Random Forest* ialah pendekatan pembelajaran mesin yang biasa digunakan untuk menangani klasifikasi. Menggunakan konsep kumpulan pembelajaran, teknik ini menggabungkan banyak model yang lebih mudah menjadi satu model yang kuat. Tugas utama *Random Forest* ialah untuk meramalkan atau mengklasifikasikan data berdasarkan input. Algoritma ini terdiri daripada satu siri pokok keputusan yang dihasilkan secara rawak. Setiap pokok keputusan menghasilkan ramalan berdasarkan maklumat yang diberikan. Pengelasan untuk setiap pokok keputusan kemudian dibuat sebagai klasifikasi akhir di dalam model ini.

```
#fitting Naïve Bayes Classifier to the training set using Gaussian
from sklearn.naive_bayes import GaussianNB
NB= GaussianNB()
NB.fit(X_train_res,y_train_res)
```

Rajah 3.24 Model Klasifikasi *Naive Bayes* (NB)

Pelbagai *library* Python boleh digunakan untuk melaksanakan algoritma *Naive Bayes*. Algoritma *GaussianNB* digunakan untuk melatih model dalam penyelidikan ini, menggunakan model *Naive Bayes* untuk set data latihan. Penggunaan GaussianNB berdasarkan sifat numerik data kajian, yang selaras dengan pengedaran *Gaussian* sebagai hasil dalam tindakan teknik standardisasi penyeragaman sebelum ini. Penggunaan *teorema Bayes* adalah asas algoritma ini kerana korelasi yang signifikan yang wujud antara atribut yang digunakan menunjukkan pengiraan probabiliti kelas untuk konteks klasifikasi ini.

```
from sklearn import svm
svr_linear=svm.SVC(kernel='linear' , gamma='scale', C=1,probability=True)
svr_linear.fit(X_train_res,y_train_res)
svr_linear.predict(X_test)
svr_prob_linear = svr_linear.predict_proba(X_test)
y_pred_linear=svr_linear.predict(X_test)
svr_linear_score = svr_linear.score(X_test, y_test)
```

Rajah 3.25 *Support Vector Classifier* dengan *linear kernel* (SVC Linear)

Implementasikan *Support Vector Classifier* (SVC) dengan *linear kernel,* menggunakan parameter C=1, *gamma=scale*, dan *probability=true.* Penggunaan parameter ini diambil menggunakan *probabilities* adalah kebarangkalian prediksi untuk setiap kelas bagi model klasifikasi. Parameter dengan ketetapan C=1 dan gamma=scale telah diterangkan di dalam model regresi yang mana fungsi dan peranannya sama bagi model klasifikasi ini. Model klasifikasi SVM di dalam setiap kernal kajian ini menggunakan parameter C untuk mengawal dasar toleransi ralat. Nilai C yang tinggi boleh membawa kepada sempadan keputusan yang lebih ketat dalam penilaian yang salah dan begitu juga sebaliknya. Pada masa yang sama, parameter *gamma* model klasifikasi mesin vektor sokongan (SVM) mengawal sejauh mana kesan latihan tunggal mempengaruhi mode yang mana sampel pelatihan yang dekat dengan titik pembelajaran memiliki pengaruh yang lebih besar dan begitu juga sebaliknya.

```
from sklearn import svm
svr_rbf=svm.SVC(kernel='rbf' , gamma='scale', C=1, probability=True)
svr_rbf.fit(X_train_res,y_train_res)
svr_rbf.predict(X_test)
y_pred_rbf=svr_rbf.predict(X_test)
svr_prob_rbf = svr_rbf.predict_proba(X_test)
```

Rajah 3.26 *Support Vector Classifier* dengan *radial basis kernel* (SVC RBF)

Untuk membina model SVC RBF, pendekatan SVC objek menggunakan kernel RBF melalui *SVC (kernel='rbf')* dengan *hyperparameter* berikut iaitu C=1, *gamma=skale*, dan *probability=True.* Penggunaan *Support Vector Classification* (SVC) dengan *Radial Basis Function* (RBF) kernel digunakan untuk menukar data ke ruang dimensi yang lebih tinggi melalui penggunaan fungsi basis *radial* dengan mengenal pasti sempadan keputusan *non-linear* yang kompleks.

```
from sklearn import svm
svr_poly=svm.SVC(kernel='poly' , gamma='scale', C=1, probability=True)
svr_poly.fit(X_train_res,y_train_res)
svr_poly.predict(X_test)
y_pred_poly=svr_poly.predict(X_test)
svr_prob_poly = svr_poly.predict_proba(X_test)
```

Rajah 3.27 *Support Vector Classifier* dengan *polynomial kernel* (SVC Poly)

Metodologi object SVC menggunakan kernel *polynomial* melalui penggunaan *SVC (kernel='poly')* dan menggabungkan *hyperparameter* C=1, *gamma=skale*, dan *probability=True* untuk membina model SVC *polynomal.* Fungsi polinomial digunakan dalam gabungan dengan SVC yang mempunyai pendekatan polinomial untuk memudahkan pemaparan data ke dimensi ruang yang lebih tinggi.

```
from sklearn import svm
svr_sigmoid=svm.SVC(kernel='sigmoid' , gamma='scale', C=1 , probability=True)
svr_sigmoid.fit(X_train_res,y_train_res)
svr_sigmoid.predict(X_test)
y_pred_sigmoid=svr_sigmoid.predict(X_test)
svr_prob_sigmoid = svr_sigmoid.predict_proba(X_test)
```

Rajah 3.28 *Support Vector Classifier* dengan *sigmoid kernel* (SVC Sigmoid)

Model SVC Sigmoid dibina dengan menggunakan kod *sigmoid (kernel = 'sigmoid')*, yang dicapai dengan menggabungkan *hyperparameter* berikut: C = 1, *gamma = skala*, dan *probability = True*. Di samping itu, model Sigmoid mempunyai peranan dalam proses menukar data kepada tempat-tempat dengan dimensi yang lebih tinggi. Selain data linear, model ini bertindak dalam mengendalikan maklumat *non-linear*.

3.4.17 Langkah 8.0: Penilaian Keputusan

Langkah penilaian keputusan bagi kajian ini, dibahagikan kepada dua kaedah penilaian iaitu dalam ramalan regresi dan klasifikasi di mana masing-masing mempunyai kaedah penilaian mengikut keperluan model yang dibina. Oleh itu, analisa pendekatan ramalan setiap model di dalam kajian ini, diterangkan di dalam peringkat ini bagi melihat keberkesanan penilaian terhadap model algoritma yang dibina.

3.4.18 Langkah 8.1: Kaedah penilaian ramalan regresi

Kaedah bagi penilaian model ramalan regresi di dalam kajian ini akan melalui penilaian menerusi dua kaedah iaitu *Regression Evalution Metrics* dan *R Squared*, R^2. *Regression Evalution Metrics* mempunyai beberapa matrik yang berkaitan iaitu MAE, MSE, dan RMSE atau dalam istilah lain sebagai *Mean Absolute Error, Mean Squared Error* dan *Root Mean Squared Error* manakala *R-Squared* sebagai R^2 adalah dua kaedah ramalan yang digunakan kebanyakan masa. Berhubungan penjelasan kedua-dua pendekatan, ianya mempunyai fungsi dan peranan dalam menilai hasil pembelajaran mesin yang dibuat.

```
# Calculate the performance
rsquared = linear_reg.score(X_test, y_test)
print('The R-Squared score for linear regression model is {:.2f}%'.format(rsquared*100))
```

Rajah 3.29 Penilaian Regresi menggunakan *R-Squared* (R^2)

R-Squared (R^2) adalah kaedah sebagai pengukur sejauh mana ramalan yang dihasilkan berkorelasi dengan data sebenar yang diukur oleh *R-squared*. *R-squared* ialah pengukuran korelasi antara data sebenar dan ramalan. Nilai *R-Squared* dengan skor tertinggi yang ialah 1, yang menunjukkan perpaduan sempurna antara ramalan dan data sebenar, manakala skor *R-squared* yang mempunyai skor terendah adalah 0. Mengira perbandingan antara jumlah variasi dalam data yang boleh dijelaskan oleh model ramalan dengan jumlah keseluruhan perubahan dalam data diperlukan untuk menganggarkan nilai *R-Squared.* Semakin besar nilai statistik *R-Squared*, semakin baik ramalan sesuai dengan data sebenar.

```
#evaluating regression
from sklearn import metrics
print('MAE, Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred_lr))
print('MSE, Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred_lr))
print('RMSE, Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred_lr)))
```

Rajah 3.30 Penilaian regresi menggunakan *Regression evaluation metrics*

Apabila mengira kesilapan dalam ramalan yang dinyatakan dalam nilai berterusan, penilaian regresi dapat dilakukan menggunakan kaedah *regression evalution metrics* yang mempunyai penilaian berkenaan MAE, MSE, dan RMSE. Matrik ini menerangkan tentang magnitud perbezaan di antara ramalan dan nilai sebenar dan matrik ini adalah statistik pilihan yang sering digunakan dalam pengiraan prestasi model regresi. Keberkesanan ramalan apabila keputusan korelasi dengan nilai MAE, MSE dan RMSE yang lebih rendah diperolehi. *Error performances metrik* ini memaparkan ketidaktepatan perbandingan ramalan, manakala *R-Squared* matrik menggambarkan tahap korelasi yang wujud antara ramalan dan data sebenar keseluruhan. Kedua-dua pendekatan ini sering digunakan bersama-sama bagi mencapai pemahaman yang lebih komprehensif mengenai keberkesanan ramalan.

3.4.19 Langkah 8.2: Kaedah penilaian klasifikasi

Dalam konteks model klasifikasi berasaskan pembelajaran mesin, pelbagai matrik penilaian biasanya digunakan untuk menilai keberkesanan model dalam hal prestasi. Terdapat beberapa matrik penting untuk menilai model klasifikasi iaitu ketepatan, kejituan, dapatan semula, Skor F1, dan Lengkungan di bawah *ROC*.


```
# Evaluate the accuracy of the model
knn_score = knn.score(X_test, y_test.values.ravel())
print('The accuracy of kNN is {:.2f}%'.format(knn_score*100))
```

Rajah 3.31 Penilaikan ketepatan model klasifikasi

Metodologi penilaian ketepatan merangkumi matrik penilai yang mengukur sejauh mana model ini boleh meramalkan dengan tepat kategori data yang dilabel. Tahap ketepatan dinyatakan oleh persamaan berikut. Ketepatan merujuk kepada keupayaan model untuk secara tepat meramalkan klasifikasi data. Kaedah untuk menunjukkan ketepatan ramalan adalah dengan mengira perbandingan ramalan yang betul dengan saiz sampel keseluruhan.

```
from scikitplot.metrics import plot_confusion_matrix as plt_confusion
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_recall_fscore_support
#Plot confusion matrix using the plt_confusion method imported earlier
plt_confusion(y_test, pred)
```

Rajah 3.32 Penilaikan model klasifikasi (*Confusion Matrik*)

Matrik Kekeliruan ialah kaedah yang digunakan untuk menilai keberkesanan model klasifikasi. Ia melibatkan perbandingan ramalan model dengan nilai sebenar data yang sedang diperiksa, dan disenaraikan dalam bentuk jadual. Matrik yang disebutkan di atas dinyatakan sebagai kuantiti jumlah ramalan positif yang tepat dan bilangan ramalan negatif yang tidak tepat. Matrik Kekeliruan terdiri dari empat kategori utama iaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Bagi penerangan TP ianya merujuk kepada bilangan hasil yang tepat dengan ramalan positif yang dijangkakan oleh model manakala TN adalah bilangan hasil yang betul yang diramalkan sebagai negatif oleh model. Berbeza bagi FP, ianya merujuk kepada bilangan hasil positif yang diramal dengan tidak tepat oleh model di mana model meramal suatu keadaan yang termasuk dalam kelas positif, tetapi sebenarnya ianya termasuk di dalam kelas negatif. FP ini juga dikenali sebagai Kesilapan Jenis I. Begitu juga konsep yang sama berlaku di dalam FN yang merujuk kepada hasil ramalan yang telah diklasifikasikan sebagai negatif oleh model tetapi tidak memberikan hasil sebenar negatif dan ia juga dikenali sebagai Kesilapan Jenis II.

```
from sklearn.metrics import classification_report
target_names = ['Yes', 'No']
print(classification_report(y_test, pred , target_names=target_names))
```

Rajah 3.33 Penilaikan model klasifikasi (*Classification report)*

Penilaian dalam klasifikasi juga boleh ditambah menerusi *Classification report* dalam menyediakan ringkasan yang komprehensif mengenai keberkesanan model. Laporan klasifikasi biasanya terdiri daripada pelbagai matrik pengukuran, termasuk ketepatan, amaran, skor F1, dan sokongan untuk setiap kelas sasaran. Penilaian keupayaan model untuk membezakan contoh positif dari kelas yang diberikan diukur dengan mengingat, juga dikenali sebagai kepekaan atau kadar positif sebenar. Perbandingan ramalan positif benar dikira dengan membahagikan bilangan sampel yang positif dengan jumlah sampel keseluruhan yang dijangka positif. Makna pengingat diperkuat apabila matlamat utama adalah untuk mengurangkan kejadian negatif palsu. Skor F1 ialah ukuran statistik yang mewakili purata harmoni matrik ketepatan dan

mengingat. Pengukuran yang diberikan menunjukkan keseimbangan antara keupayaan untuk mendapatkan maklumat yang bernas (memanggil semula) dan ketepatan maklumat yang diperolehi (*precision*). Skor F1 merupakan matrik yang berharga untuk mencapai keseimbangan antara jumlah positif palsu dan negatif palsu.

```
from scikitplot.metrics import plot_roc_curve as plt_roc
# Plot the ROC curve using plt_roc method imported earlier
plt_roc(y_test, pred_prob, title='kNN ROC Curve', figsize=(8,8))
plt.show()
```

Rajah 3.34 Penilaikan model klasifikasi (*ROC Curve*)

Hasil penilaian juga dapat dibuat melalui pemerhatian graf yang dihasilkan di dalam *ROC Curve* yang mana berfungsi untuk menilai keupayaan model bagi membezakan antara kelas positif dan negatif. Ini menunjukkan keupayaan model untuk mengkategorikan sampel melalui skor kebarangkalian yang terjadi di dalam set ramalan dan juga hasil sebenar ujian.

3.4.20 Langkah 9.0: Analisa Preskriptif dengan Siri Masa

Untuk menjalankan analisis preskriptif, kajian ini menggunakan penemuan deskriptif dan prediktif, bersama-sama dengan data yang diperolehi daripada beberapa musim. Menggabungkan analisis musim untuk menggambarkan corak musim adalah pertimbangan penting dalam analisis siri masa. Teknik-teknik seperti pemecahan musim dan *corak* boleh digunakan untuk mencapai resolusi siri musim bersama-sama dengan analisis yang dinyatakan di atas. Analisa seterusnya adalah analisa ke depan dengan menggunakan model SARIMA di mana model ini adalah daripada istilah *Seasonal Autoregressive Integrated Moving Average* yang menggunakan pendekatan *Seasonal* atau musim dalam melakukan situasi corak pada masa hadapan. Bagi analisa menggunakan pendekatan SARIMA, pembinaan model ini memerlukan kepada penglibatan fungsi *rolling mean* yang bertujuan untuk mendapatkan kiraan gerakan corak dengan *window* bersamaan 12 yang mana peranan ini dapat membantu model SARIMA dalam membentuk satu jangkaan nilai pada masa hadapan.

```
import statsmodels.api as sm

model=sm.tsa.statespace.SARIMAX(data1['DataAvailability'],order=(1, 0, 1),seasonal_order=(1,0,1,12))
results=model.fit()
```

Rajah 3.35 Komponen p,d,q dan P,D,Q,S di dalam model SARIMA

Kajian penelitian terhadap ramalan *corak* dan corak dengan menggunakan model SARIMA dilakukan dalam melihat hasil jangkaan pola isu pada tahun-tahun ke hadapan. Melalui kajian ini, ini memerlukan kepada penglibatan parameter p,d,q dan P,D,Q,S di mana parameter ini amat penting dalam pembinaan model SARIMA. Bagi komponen p,d,q adalah merujuk kepada bukan susunan mengikut musim yang mana p mewakili *auto regresif,* d mewakili perbezaan, dan q mewakili pergerakan rawak. Bagi komponen P,D,Q, S pula adalah susunan yang dilakukan mengikut musim di mana P adalah *auto regresif* musim, D adalah perbezaan mengikut musim, Q adalah gerakan rawak mengikut musim, dan S adalah mewakili pola masa dalam setahun iaitu 12 bulan. Justeru, dengan model SARIMA (1,0,1) dan (1,0,1,12) ini adalah menunjukkan model ini, mempunyai *auto regresif* gabungan bukan bermusim dan bermusim di samping tiada perbezaan d and D di dalam sampel kajian.

3.5 STRUKTUR DATA

Maklumat kajian ini adalah mengenai kajian ketersediaan data pembacaan meter jauh yang mengandungi beberapa prosedur kegagalan dan petunjuk berdasarkan data yang diperolehi. Pengukuran keadaan ketersediaan meter ini berasal daripada kajian eksplorasi pada sampel data 122,053 meter, yang merangkumi 16 tahun rekod mengikut status sampel data dari 2006 hingga 2022. Dalam set data ini terdiri daripada maklumat tentang jenis peranti meter, tarikh pemasangan meter, jenama meter, isu kegagalan, jumlah data *interval* yang hilang dan maklumat - maklumat lain yang relevan. Pendekatan kajian ini adalah bertujuan mencari keserasian antara sampel data dengan sampel corak taburan operasi yang mana kajian analitik ini tidak menggunakan data sebenar bagi sesebuah operasi di dalam sesebuah organisasi. Penggunaan sampel yang merupakan bukan data sebenar ini, secara tidak langsung berdasarkan kepada senario yang mungkin terjadi dengan situasi sampel kegagalan sistem meter peranti. Sampel statistik set data ini diperolehi berdasarkan corak daripada pengumpulan data tahunan daripada data yang dimodifikasi iaitu data sebenar kepada sampel data baru yang tidak berkaitan identiti sebenar di mana rujukan asal diperolehi daripada salah satu utiliti elektrik di Malaysia. Sampel ini terdiri daripada makluman tempoh 12 bulan setiap tahun dan memaparkan prestasi meter yang berlaku. Melalui data ini, penggunaan sistem komunikasi dua arah antara stesen pusat dan pengguna telah dicatat dalam

mengira maklumat penggunaan beban masa nyata. Peranan membaca meter jauh RMR ini mengandungi pilihan voltan daripada voltan rendah, voltan tinggi, dan voltan menengah. Set data ini mempunyai 27 atribut dan 122,053 sampel data dengan beberapa nilai yang hilang. Terdapat 5 jenis aktiviti meter yang dicatatkan iaitu pemasangan meter baru, penggantian meter, pengukur penggunaan semula, meter tidak aktif dan penghapusan meter. Setiap jenis meter terdiri daripada jenama meter yang berbeza dan menggunakan jenis sambungan komunikasi sama ada 3.5G, 4G, GSM atau GPRS. Berdasarkan maklumat ini, pemeriksaan set data ini berguna untuk mengetahui aspek kecenderungan yang mempengaruhi prestasi pemodelan ramalan ini.

3.5.1 Penyediaan Data

Data penggunaan meter bacaan jauh RMR ini menggambarkan ketersediaan data pada peranti yang boleh membawa semula pembacaan profil beban pengguna untuk semua pengguna pada tempoh masa yang ditetapkan. Untuk tujuan menjalankan kajian kes ini, set data bacaan meter jauh (RMR), mengandungi rekod bacaan ketersediaan data pengguna dengan dapatan sampel daripada utiliti elektrik. Ianya berdasarkan sampel data yang direkodkan pada ketersediaan data RMR yang boleh diklasifikasikan sebagai tidak mencapai ketersediaannya 100% daripada data profil beban apabila wujud kejanggalan di dalam pembaca profil data yang tidak normal. Antara kejanggalan yang ditemui adalah seperti peningkatan nilai kehilangan *interval*, kesilapan pada akses rangkaian komunikasi, jurang profil data, masa bacaan terakhir yang tidak dikemas kini dan lain-lain. Pengumpulan data bacaan meter RMR, dijalankan untuk menyediakan maklumat yang diperlukan untuk membuat pemantauan jumlah ketersediaan data profil yang berjaya dicatat dan untuk membahagikan kelas ketersediaan data di Semenanjung Malaysia berdasarkan tiga kategori peringkat data, secara khususnya iaitu 0-90%, 90%, dan 100%.

3.5.2 Komponen Data

Data yang disediakan adalah dianjurkan dengan cara yang tidak terperinci atau tidak menggunakan data asal mengenai pengguna, ID kad SIM, dan ID IP. Ini dilakukan adalah bagi melindungi maklumat peribadi yang dikaitkan dengan pelanggan. Data yang digunakan dengan saiz sampel ditentukan oleh jumlah meter, iaitu 122,053, dan

pengukuran dibuat daripada pelbagai jenis meter dan tahap voltan. Hasil pengumpulan data mentah ini menyediakan beberapa ciri dan jenis data yang terdiri daripada jenis nominal, ratio, ordinal, dan interval. Maklumat mengenai setiap jenis data kajian ini disenaraikan dengan lebih terperinci dalam Jadual 3.1 berserta dengan penerangannya.

Jadual 3.1 Senarai Atribut dan Penerangannya

Atribut	Jenis Data	Penerangan
State	Nominal	Pendekatan untuk Negara
Voltage Level	Ratio	Jenis Tahap voltan rendah
Rate Category	Nominal	Jenis Tarif kategori Bill
Installation Type	Ratio	Jenis pemasangan meter jenis berdasarkan voltan
Logical device no.	Ratio	Titik logik penghantaran perkhidmatan pelanggan
Device No	Ratio	Nombor ID Meter
Device Cat.	Nominal	Jenis kategori peranti berdasarkan CT ratio
Register Group	Nominal	Kumpulan Meter Program
Meter Installation Date	Interval	Tarikh meter dipasang di tapak
SIMCardNumber	Ratio	ID Simcard dalam peranti modem
IP No	Ratio	Serial ID untuk sambungan IP dalam sistem
Comm Type	Nominal	Rangkaian komunikasi meter
Status of Meter	Nominal	Status konfigurasi
SO Status	Nominal	Jenis aktiviti meter
Reason for incomplete configuration	Nominal	Status kegagalan konfigurasi
Aging	Nominal	Status penuaan tindakan data tidak selesai
Service Mode	Ordinal	Kedudukan mod perkhidmatan lokasi meter
MeterBrand	Nominal	Jenis jenama meter
Daily Call Status	Nominal	Status kejayaan panggilan harian
Service Mode Success Status	Nominal	Status kejayaan mod perkhidmatan
Daily Call Error Code	Ratio	Kod kesilapan panggilan harian
Daily Call Error Code Activity	Ratio	Aktiviti Kod kesilapan panggilan harian
Failure Reason	Nominal	Status isu kegagalan
Data Availability Bucket	Ordinal	Kedudukan tahap ketersediaan data
Total Missing Interval	Ratio	Bilangan kehilangan selang masa data
Data Availability	Ratio	Peratusan ketersediaan data
Status Data Availability	Ratio	Kelas status ketersediaan data lengkap

3.6 INSTRUMEN PENYELIDIKAN

Untuk menjelaskan instrumen yang digunakan dalam penyelidikan, terdapat tiga penggunaan alatan penting bagi menjalankan kajian ini. Fungsi setiap alatan analisa ini telah menyediakan analisa data yang mudah difahami dengan matlamat meningkatkan ketepatan dan kecekapan untuk proses pembelajaran. Peranan setiap instrumen ini adalah melibatkan perisian Microsoft Excel, *R-Studio* dan *Python* yang mempunyai keupayaan tugas tersendiri. Apabila melihat pelbagai peringkat analisis diperlukan, pendekatan penggunaan alatan yang berbeza pada tugasan di peringkat tersebut adalah ditetapkan. Pada peringkat awal kajian, data pemprosesan dikumpulkan menggunakan *Microsoft Excel* dalam mengumpulkan jenis maklumat data diperoleh daripada pakar domain. Ia telah dianalisis dengan menggunakan bahasa perisian *R-Studio* dan *Python*. Aktiviti tugasan menggunakan *R-Studio* adalah termasuk pra-pemprosesan, analisis statistik dan analisis deskriptif. Selain itu, penggunaan perisian *Python* adalah bagi menjalankan tugasan berkenaan analisis prediktif dan preskriptif iaitu analisa utama dalam kajian ini. Jadual 3.2 seperti yang ditunjukkan di bawah, adalah senarai instrumen berserta fungsinya bagi penerangan tugasan di dalam kajian ini.

Jadual 3.2 Fungsi alatan dan instrumen yang digunakan dalam kajian mengikut kepada keperluan setiap peringkat analisis

Alatan & Instrumen	Fungsi	Penerangan
Microsoft Excel	Data *Pre-review*	Pengumpulan makluman data daripada *domain expert* yang diperlukan dan menstruktur semula maklumat baru untuk keselamatan data
R-Studio	Data Integrasi & Visualisasi	Kompilasi data dengan menjana set data mentah, melakukan proses pra-pemprosesan, Statistik Data & Visualisasi Data
Python 3.10.12	Data Pembangunan Model Ramalan	Penyediaan data bagi proses Pembelajaran Mesin dengan Model Algoritma dan pembentukan Analisis Siri Masa

3.7 RUMUSAN

Menerusi kajian ini, metodologi kajian telah diterangkan di dalam bahagian ini di mana proses awal sehingga akhir kajian dibentangkan bagi memberikan penekanan aspek penting di dalam pembangunan kajian secara lebih komprehensif. Penekanan yang

diterangkan adalah merangkumi aspek permasalahan isu dengan pendedahan pelbagai kaedah analisis yang digunakan. Analisa data yang direkodkan ini, membolehkan melakukan pencarian maklumat bersih daripada data mentah dalam merumuskan ramalan dan pandangan yang boleh dilaksanakan untuk penambahbaikan sistem. Konsep kajian ini adalah selari dengan pendekatan yang digunakan di dalam konteks analisa sains data yang umumnya mempunyai kitaran hayat lima peringkat yang terdiri daripada pemerolehan data iaitu kemasukan dan pengambilan data, kemudian kepada penyimpanan dan pemprosesan data melalui pergudangan data, pembersihan data, dan penyusunan data. Seterusnya, data proses dicipta melalui perlombongan data iaitu pengelompokan atau pengelasan dan pemodelan data. Ia dibawa ke dalam bentuk komunikasi melalui pelaporan dan visualisasi data. Akhirnya, proses menganalisis data dibuat dengan pengesahan penerokaan atau ramalan. Pendekatan ini diperkuat lagi oleh penggunaan teknik pembelajaran mesin, dengan itu membawa kepada rangka kerja analitik yang lebih berkesan. Sehubungan dengan itu, penyelidikan yang diusulkan di dalam kajian ini adalah untuk menyiasat komponen analitik penting dalam prosedur perlombongan data, dengan penekanan utama kepada penggunaan metodologi pembelajaran mesin untuk tujuan ramalan dan pengelasan.

BAB IV

DAPATAN KAJIAN

4.1 ANALISA DAN DAPATAN DESKRIPTIF

Proses pembersihan data dilakukan untuk membetulkan sebarang ketidaksesuaian yang terdapat dalam data. Proses mengenal pasti nilai yang tidak tepat, tidak mencukupi, atau tidak bersesuaian dalam set data yang diberikan biasanya dipanggil pembersihan data. Proses pembersihan ini akan menangani nilai yang telah didiagnosis sebelumnya melalui pelaksanaan teknik seperti penggantian nilai, perubahan corak data, dan penghapusan atribut. Teknik pra-prosesan data mempunyai keupayaan untuk mengintegrasikan pelbagai prosedur seperti pembersihan dan proses lain yang berkaitan. Di bahagian ini, analisa dan dapatan kajian berkenaan proses awal dalam kajian dianalisa dan diperolehi prestasi nilai yang membantu keperluan analisa seterusnya.

4.1.1 Analisa Hasil 1.0: Analisa Penerokaan Data(EDA)

Pada peringkat penerokaan *Exploratory Data Analysis* (EDA), berlaku proses di mana data diselidiki dengan menggunakan berbagai kaedah statistik dan visualisasi dalam memahami pola, hubungan, dan sifat data yang ada. Tujuan dari EDA ini adalah untuk menemukan kriteria yang bermanfaat dan membantu dalam merungkaikan persoalan yang ada berkenaan data mentah dalam kajian ini. Melalui penerokaan EDA ini, data dianalisasi menggunakan grafik yang bersesuaian untuk memvisualisasikan data yang mana dengan visualisasi dapat membantu dalam memahami taburan data, korelasi antara atribut, dan pola-pola lain yang ada. Dalam bahagian ini, set data yang digunakan

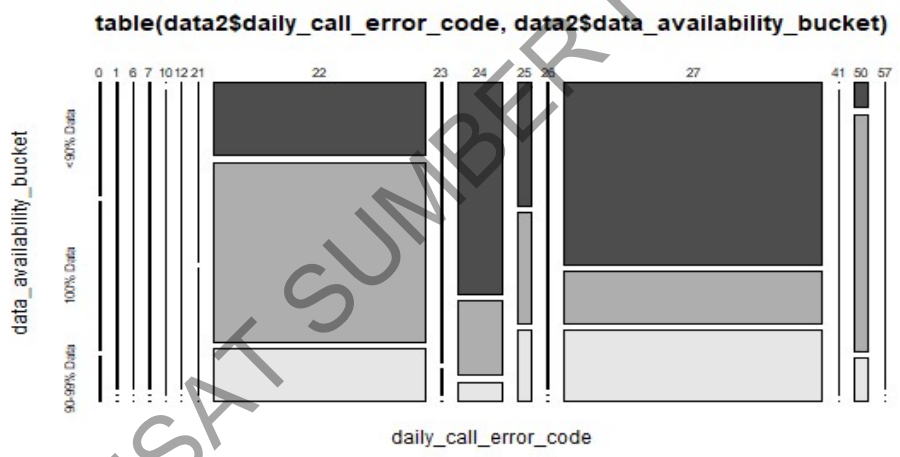
dalam kajian ini diperkenalkan terhasil dati ciri yang diekstrak daripadanya dan diterangkan secara lebih bermaklumat. Begitu juga, proses pemeriksaan keseluruhan data dapat diperolehi dengan bantuan pendekatan analisa yang dilakukan dalam kajian ini peringkat ini.

Pada permulaan analisa, dengan menggunakan *Mosaic Map* menerangkan beberapa komposisi awal berkenaan data mentah berkaitan meter RMR dalam memahami data yang kompleks dengan menggabungkan informasi yang tersebar menjadi satu gambaran data yang komprehensif. Hal ini menunjukkan analisa awal dalam menginterprestasi data yang digunakan di dalam kajian ini. Antaranya, di bawah menerangkan hubungan kolerasi antara dua atribut berkenaan jumlah meter RMR yang dianalisa dalam kajian dengan mengikut beberapa jenama meter dan peratus kesediaan data. Berdasarkan Rajah 4.1 ini menunjukkan jenama L meter RMR ini yang mempunyai bilangan terbesar dan diikuti dengan jenama P Premier, jenama P Sprint, jenama A, jenama E, jenama I dan jenama M dengan menunjukkan juga jumlah terbesar di dalam *bucket* ketersediaan data adalah peratusan 100% diikuti dengan peratusan <90% dan yang paling kurang adalah peratusan 90%-99%.

Rajah 4.1 *Mosaic Map* berkaitan jenama meter mengikut *data availability bucket*

Di dalam rajah 4.2 di bawah, menjelaskan *mosaic map* berkaitan kuantiti isu *daily call error code* yang mempengaruhi bilangan ketersediaan data dalam operasi meter RMR ini. Setiap *error code* di dalam *daily call* ini mempunyai punca kesalahan

tersendiri dalam operasi kerja meter RMR. Contoh yang ditunjukkan adalah *daily call error code* 27 adalah paling tinggi terjadi di dalam ketersediaan data yang menyumbang peratusan ketersediaan data <90 % di mana kesilapan ini kebanyakkan berpunca daripada masalah *signal* di modem meter RMR sama ada berpunca atas sebab tertentu iaitu lokaliti meter, jenis komunikasi dan sebagainya. Selain daripada itu, *daily error code* 22 diikuti dengan 24, 25 dan 50 juga menunjukkan bilangan yang besar terjadi di dalam kesilapan dalam sambungan *call* berbanding *error code* yang lain yang turut menyumbang kepada penurunan prestasi peratusan ketersediaan data. Kesilapan yang sering berlaku dalam *error code* ini termasuk berkenaan isu *network* data kabel meter, kesalahan *baurate,* kesalahan *password* meter, kesalahan IP no dan lain-lain yang menyumbang kepada gangguan semasa *daily call* data pengguna ke pusat data.



Rajah 4.2 *Mosaic Map* berkaitan *daily call error code* mengikut *data availability bucket*

Bagi pendekatan analisa *remote meter reading* (RMR) ini, beberapa teknologi komunikasi yang digunakan untuk menghantar data dari peranti meteran ke pusat kawalan adalah ditunjukkan di dalam Rajah 4.3. Salahnya adalah GSM yang merupakan komunikasi selular memungkinkan penghantaran data melalui jaringan selular dengan menggunakan pendekatan kad SIM berdasarkan nombor telefon pada perantian meter. Namun teknologi ini adalah teknologi lama yang berjumlah paling rendah penggunaannya di dalam set data ini disebabkan masih digunakan di sesetengah perantian meter. Teknologi kedua yang dievolusikan dari GSM ini adalah teknologi GPRS dengan kecepatan transfer data GPRS biasanya lebih tinggi daripada GSM yang menyebabkan pengiriman data yang lebih besar dan lebih kompleks. Teknologi GPRS

ini masih banyak dan pilihan kedua digunakan dalam set data meter ini. Melihat kepada teknologi yang semakin berkembang, penggunaan teknologi jaringan selular yang lebih canggih digunakan iaitu penggunaan 3.5G dan 4G di mana jaringan 4G adalah yang paling besar jaringanya digunakan di dalam kajian set ini. Dalam konteks *remote meter reading* (RMR), GSM, GPRS, 3.5G, dan 4G digunakan sebagai teknologi komunikasi perantian meter adalah bergantung kepada infrastruktur jaringan sedia ada di lokasi dalam memenuhi keperluan spesifik operasi sistem RMR ini

Rajah 4.3 *Mosaic Map* berkaitan jenis jaringan komunikasi mengikut *data availability bucket*

4.1.2 Analisa Hasil 1.1: Analisa Statistik

Analisis statistik ialah cabang analisis data daripada penerokaan data (EDA) yang melibatkan pengumpulan, interprestasi dan persembahan data untuk mendedahkan korelasi, pola, dan matlamat kajian. Matlamat analisis statistik ini ialah untuk mengkaji corak, dan hubungan menggunakan analisis kuantitatif dengan mengeluarkan temuan yang bermakna, dan membuat kesimpulan mengenai populasi atau fenomena berdasarkan data sampel. Bentuk-bentuk penyelidikan deskriptif statistik ini merupakan subset kajian terhadap pemerhatian yang digunakan untuk menentukan gabungan kumpulan, atau pengurangan dimensi kepada nilai maklumat baru. Dalam kajian ini, menggunakan statistik deskriptif ini adalah langkah utama di dalam mencari ukuran kecenderungan pusat yang merupakan *mean, median* dan kepelbagaian *standard deviation* yang akan menyediakan ringkasan ciri-ciri utama set data . Pendekatan ini membolehkan penyelidikan ini untuk memahami pengedaran, penyebaran, dan corak

dalam data dengan lebih komprehensif. Dalam konteks ciri yang berterusan, nilai digunakan untuk menggantikan nilai dengan menggunakan *mean* atau *median* set. Selain daripada itu, mode item digunakan dalam proses ini dengan memilih nilai yang paling kerap muncul dalam set data. Oleh itu, dalam laporan analisis ciri-ciri ditunjukkan dalam jadual untuk menonjolkan nilai-nilai yang paling penting untuk digunakan dalam proses analisa selanjutnya.

```
R 4.2.2 · C:/Users/User/Desktop/MODS/Project/
> summary(data2)
    state           voltage_level   rate_category      installation_type logical_device_no
 Length:122053      Min.   :1.000   Length:122053      Min.   : 1.000    Min.   :       0
 Class :character   1st Qu.:3.000   Class :character   1st Qu.: 1.000    1st Qu.:   60006
 Mode  :character   Median :3.000   Mode  :character   Median : 1.000    Median :  672283
                    Mean   :2.911                      Mean   : 2.543    Mean   : 2472611
                    3rd Qu.:3.000                      3rd Qu.: 1.000    3rd Qu.: 4018223
                    Max.   :3.000                      Max.   :29.000    Max.   :13798372

   device_no          device_cat        register_group     meter_installation_date
 Min.   :2.808e+05   Length:122053      Length:122053      Length:122053
 1st Qu.:3.610e+08   Class :character   Class :character   Class :character
 Median :6.190e+08   Mode  :character   Mode  :character   Mode  :character
 Mean   :6.107e+08
 3rd Qu.:8.160e+08
 Max.   :6.132e+09

 sim_card_number        ip_no            comm_type          status_of_meter
 Min.   :4.77e+17   Min.   :1.010e+06   Length:122053      Length:122053
 1st Qu.:4.77e+17   1st Qu.:1.010e+09   Class :character   Class :character
 Median :4.77e+17   Median :1.010e+09   Mode  :character   Mode  :character
 Mean   :4.77e+17   Mean   :4.364e+09
 3rd Qu.:4.77e+17   3rd Qu.:1.010e+10
 Max.   :4.77e+17   Max.   :1.011e+10

  so_status         reason_for_incomplete_configuration    aging           service_mode
 Length:122053      Length:122053                      Length:122053      Length:122053
 Class :character   Class :character                   Class :character   Class :character
 Mode  :character   Mode  :character                   Mode  :character   Mode  :character
```

Rajah 4.4 Analisa statistik keseluruhan bagi set data kajian

Analisa statistik akhir ini menunjukkan penelitian terhadap atribut yang ada dengan menjelaskan nilai, frekuensi, graf, data yang wujud atau *valid* dan *missing value* yang ada di dalam set data kajian ini. Analisa keseluruhan set data ini, memberi maklumat berkenaan kekerapan atau frekuensi data yang kerap digunakan sebagai pendekatan *mode* dan juga maklumat berkenaan nilai purata *mean* dan juga *median* yang bole membantu kajian dalam mendapatkan ketepatan nilai yang diperlukan bagi proses pengkajian lanjut isu yang ada.

4.1.3 Analisa Hasil 2.0: Pra-pemprosesan

Pra-pemprosesan yang dilakukan di dalam data kajian ini adalah merangkumi beberapa proses tambahan analisa tindakan bagi membaiki set data dengan pendekatan kesediaan data untuk meningkatkan keberhasilan hasil ramalan model dan keupayaan melakukan ramalan kepada data yang tidak dilihat atau data set yang baru.

4.1.4 Analisa Hasil 2.1: Penggantian nilai hilang pra-pemprosesan

Konsep di dalam tugasan penggantian nilai hilang ini adalah bertujuan melakukan pembersihan data dengan membetulkan ketidaksamaan data atau kejanggalan yang ada di dalam data. Pemprosesan data juga boleh dijelaskan lagi dengan bentuk sesuatu proses dalam mencari dan mengenal pasti nilai yang salah, tidak lengkap dan tidak konsisten daripada koleksi data yang dikumpulkan. Melalui proses pembersihan ini, nilai yang telah didapati sebelumnya akan ditangani dengan menggantikan nilai baru, mengemas kini corak data, atau menghilangkan atribut yang tidak perlu dalam menjalankan aktiviti secara interaktif dalam setiap pembersihan dan melakukan tugas untuk membantu dan mengenal pasti set data dan kemudian membuat koreksi kepada data. Analisis statistik dalam Rajah 4.5 dan 4.6 menunjukkan set data bagi setiap atribut pada keadaan data sebelum dilakukan penggantian nilai dengan selepas dilakukan penggantian nilai. Kaedah penggantian nilai bagi setiap numerikal adalah dengan diganti dengan nilai median disebabkan kebanyakkan nilai data menunjukkan keadaan nilai yang tidak disebarkan secara simetri, di mana *median* adalah pilihan yang lebih baik kerana ia mewakili nilai pusat atau tengah dari nilai tertinggi hingga terendah. Begitu juga dengan pendekatan bagi nilai hilang bagi kategorikal data ini dengan dilakukan gantian data hilang bersama nilai *mode* data iaitu nilai kategori kerap di dalam set data ini. Tahap nilai hilang dalam set data ini ditunjukan seperti di dalam rajah di bawah dengan nilai hilang oleh atribut *daily call error code, comm type* dan *aging.*


```
R 4.2.2 · C:/Users/User/Desktop/MODS/Project/
> #count total missing values in each column of data frame
> sapply(data2, function(x) sum(is.na(x)))
                              state                        voltage_level
                                  0                                    0
                      rate_category                    installation_type
                                  0                                    0
                   logical_device_no                            device_no
                                  0                                    0
                         device_cat                       register_group
                                  0                                    0
             meter_installation_date                     sim_card_number
                                  0                                    0
                               ip_no                            comm_type
                                  0                                  217
                     status_of_meter                           so_status
                                  0                                    0
reason_for_incomplete_configuration                               aging
                                  0                                117187
                        service_mode                         meter_brand
                                  0                                    0
                   daily_call_status          servic_mode_success_status
                                  0                                    0
              daily_call_error_code       daily_call_error_code_activity
                             114160                               114160
                      failure_reason           data_availability_bucket
                                  0                                    0
               total_missing_interval                    data_availability
                                  0                                    0
            status_data_availability
                                  0
```

Rajah 4.5 Bilangan nilai hilang sebelum penggantian nilai

Rajah 4.6 Bilangan nilai hilang selepas penggantian nilai

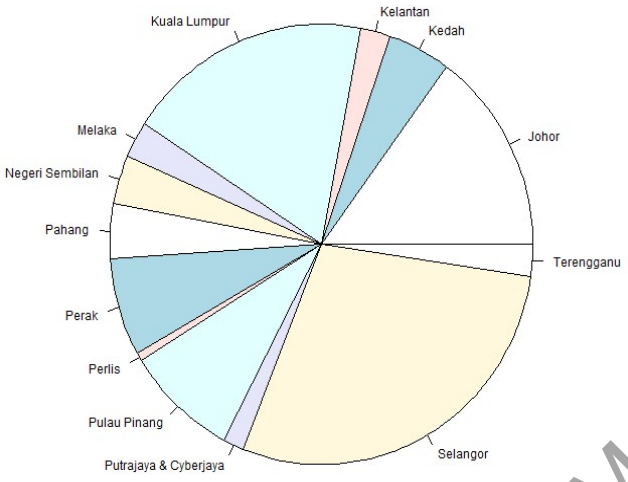
Perbezaan nilai data hilang ditunjukkan di dalam rajah di atas dengan penggantian nilai hilang numerikal *daily call error code* dengan *median* data manakala nilai hilang kategorikal data *comm type* dan *aging* diganti dengan nilai *mode* data masing-masing.

4.1.5 Analisa Hasil 2.2: Penggubahan dan pengekstrakan data baru pra-pemprosesan

Hasil pemprosesan penggubahan jenis karakter data asal kepada karakter dengan ciri yang baru turut dilakukan di dalam peringkat ini. Perubahan yang dibuat adalah bertujuan bagi memenuhi keperluan karakter baru ini bersesuaian dengan adaptasi terhadap analisa – analisa yang diperlukan di dalam kajian ini. Sebagai contoh, perubahan jenis karakter *string* bagi atribut *total missing value* dan *voltage level* telah ditukarkan kepada karakter *integer* bagi keperluan keseragaman nilai untuk analisa. Selain itu, pengekstrakan data baru melalui tambahan atribut bagi nilai tarikh meter dipasang, status wujud atau *val*id meter, *IPNo*, *Register Group* dan *Si*mCard.

4.1.6 Analisa Hasil 3.0: Deskripsi analisa menggunakan *Pie Chart*

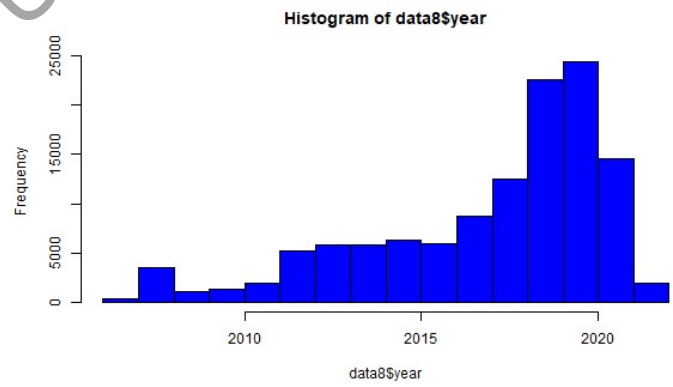
Bagi langkah deskriptif analisis selepas pra-pemprosesan data dilakukan, beberapa pendekatan visualisasi nilai lain ditunjukkan dalam model kajian ini bagi melihat secara lebih terperinci aspek data yang diperlukan.



Rajah 4.7 *Pie Chart* yang menerangkan jumlah set data berdasarkan negeri

Info grafik yang ditunjukkan di dalam *Pie Chart* ini adalah menunjukkan, nilaian penglibatan negeri yang terdapat dalam set data ini yang memberikan jumlah terbesar bilangan adalah direkodkan bagi negeri Selangor, diikuti dengan Kuala Lumpur, Johor dan Pulau Pinang. Jumlah meter RMR yang ditunjukkan menjelaskan lagi bahawa jumlah meter RMR di dalam kajian set ini, telah banyak dilakukan pemasangan di empat negeri tersebut. Kepadatan penduduk dan kawasan domestik utama bagi sesebuah negeri menyebabkan keperluan meter RMR dipasang selaras dengan peningkatan pembinaan infrastruktur yang berlaku.

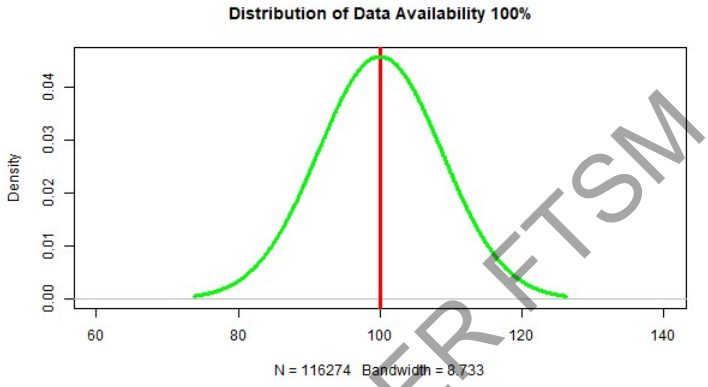
4.1.7 Analisa Hasil 3.1: Deskripsi analisa menggunakan *Histogram*



Rajah 4.8 *Histogram* berdasarkan tahun dan frekuensi ketersediaan data

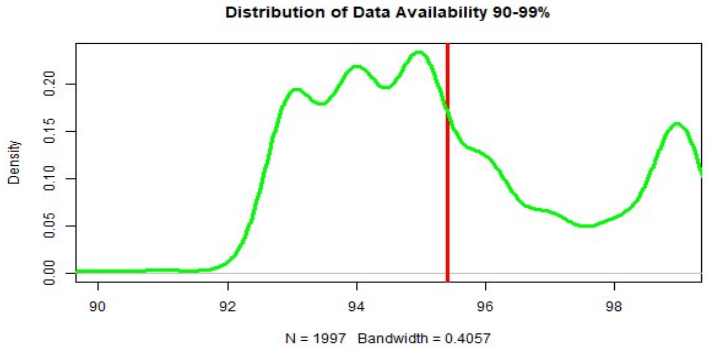
Analisa berkaitan siri masa juga dijalankan dengan melihat secara ringkas melalui *histogram* yang memperlihatkan bilangan frekuensi atau bilangan meter RMR pada tahun yang direkodkan. Rajah 4.8 di atas menerangkan bahawa nilai tertinggi set data berlaku pada di antara tahun 2019 dan 2020 dengan frekuensi hampir kepada 25000.

4.1.8 Analisa Hasil 3.2: Deskripsi analisa menggunakan *Density Distribution*



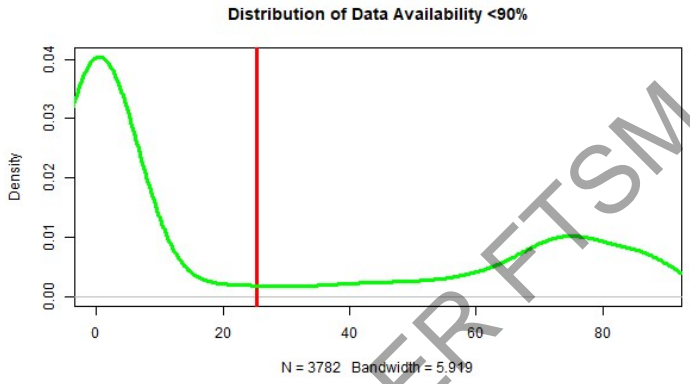
Rajah 4.9 Graf *Density Distribution* 100% Ketersediaan Data

Melalui taburan data secara pembahagian status peratusan, bagi pembahagian pertama iaitu 100% ketersediaan data, graf menunjukkan purata nilai di dalam pembahagian data ini, adalah 100% dengan jumlah set data sebanyak 116274 bilangan dan julat *bandwidth* iaitu 8.733. Bilangan yang agak tinggi di dalam bahagian 100% data menunjukkan, data mentah dalam kajian ini diperolehi dengan status data yang tidak mempunyai permasalahan isu. Namun, jumlah set data ini membantu dalam memahami keadaan dan kriteria yang perlu ada bagi kajian ini.



Rajah 4.10 Graf *Density Distribution* 90-99% Ketersediaan Data

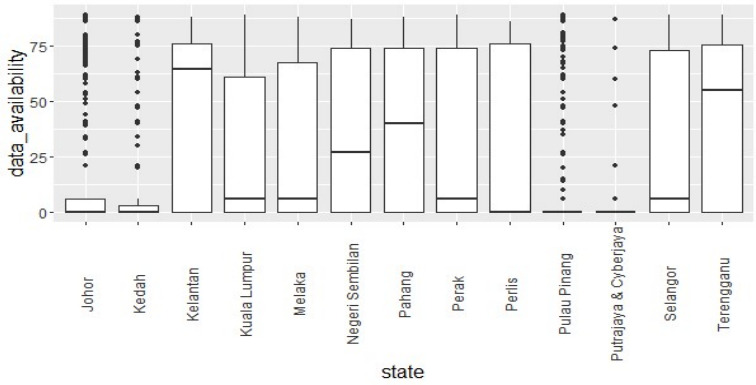
Bagi pembahagian kedua iaitu 90-99% ketersediaan data, graf menunjukkan purata nilai di dalam pembahagian data ini, adalah di antara nilai 94% sehingga 96% di mana jumlah set data yang berapa di dalam set bahagian ini adalah sebanyak 1997 bilangan dan julat bandwidth iaitu 0.4057. Bilangan yang agak tinggi di dalam bahagian ini ditunjukkan di antara peratusan 92% sehingga 96%.



Rajah 4.11 Graf *Density Distribution* <90% Ketersediaan Data

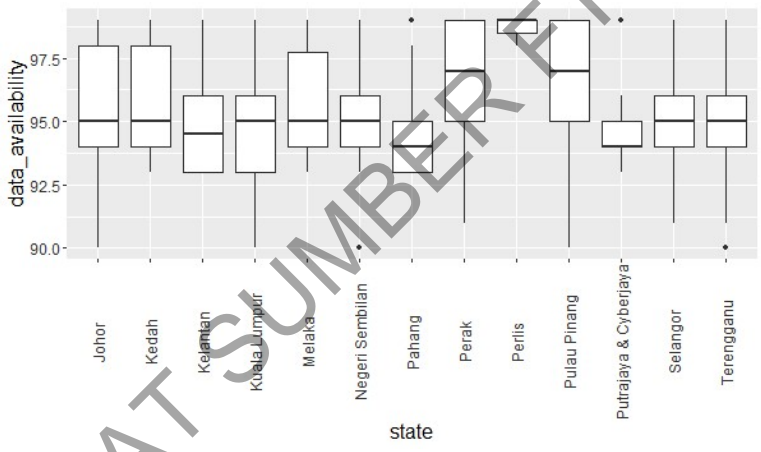
Pembahagian ketiga iaitu pembahagian akhir bagi set data kajian ini adalah peratusan ketersediaan data yang kurang daripada 90%. Rajah 4.11 di atas menunjukkan kepadatan set data adalah sebanyak 3782 bilangan dengan analisa *bandwidth* iaitu 5.919. Pemerhatian hasil graf, menunjukkan nilai purata data berada di antara nilai 20% sehingga nilai 40% dengan kepadatan yang tinggi berlaku pada nilai di antara 0% sehingga 20% dan kepadatan berkurang pada peratusan seterusnya.

4.1.9 Analisa Hasil 3.3: Deskripsi analisa menggunakan Boxplot



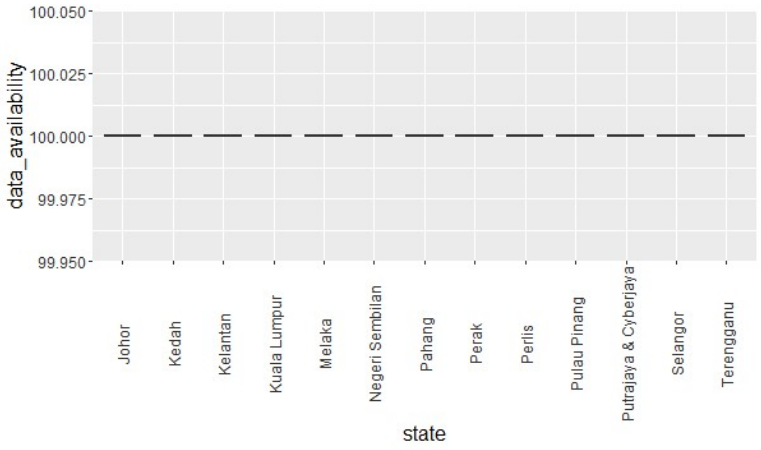
Rajah 4.12 *Boxplot* bagi ketersedian data negeri < 90%

Analisa menggunakan *boxplot* dilakukan dalam kajian ini, bagi melihat perbandingan antara kategori negeri, dalam memahami perbezaan antara median, *quartile*, serta adanya nilai ekstrem atau *outlier*. Melihat kepada hasil *boxplot* dalam rajah di atas, menunjukkan hasil taburan data di dalam bahagian ketersediaan data kurang daripada 90% di mana nilai median setiap negeri adalah berbeza-beza. Setiap negeri menunjukkan nilai data berada di dalam nilai maksimum dan minimum kecuali kepada data yang berlaku di negeri Johor, Kedah, Pulau Pinang dan Putrajaya Cyberjaya. Nilai pada negeri tersebut mempunyai nilai yang tinggi pada skala minimum nilai ketersediaan data yang menyebabkan nilai ekstrem lain berada di luar maksimum nilai data.



Rajah 4.13 *Boxplot* bagi ketersedian data negeri 90% hingga 99%

Menerusi hasil *boxplot* pada bahagian di antara 90% sehingga 99% ketersediaan data pada julat peratusan di dalam bahagian ini, adalah pada nilai median yang melebihi 93% kebanyakan di setiap negeri yang direkodkan. Untuk hasil bagi maksimum median pada set data ini, menunjukkan nilai median tertinggi adalah di negeri Perlis, dan diikuti oleh Perak dan Pulau Pinang. Walau bagaimanapun, terdapat juga julat nilai ekstrem yang berlaku pada maksimum skala data pada negeri tertentu iaitu Pahang dan Putrajaya Cyberjaya. Ini ditambah juga dengan beberapa nilai ekstrem kecil berlaku pada skala minimum data iaitu di Negeri Sembilan dan Terengganu.



Rajah 4.14 *Boxplot* bagi ketersediaan data negeri 100%

Hasil *boxplot* bagi bahagian ketersediaan data 100%, menunjukkan nilai median pada 100% dan tiada nilai *quartile*, maksimum dan minimum disebabkan set data ini hanya mempunyai nilai 100% data berbanding dengan hasil *boxplot* di dua pembahagian nilai sebelum ini.

4.1.10 Analisa Hasil 4.0: Pengekodan

Hasil analisa sebelum ini yang memperlihatkan isu kategori data tidak dikodkan yang mana sekiranya tidak dilakukan penambahbaikan dari segi pengekodan, data yang dihasilkan untuk pembangunan model adalah tidak lengkap pra-pemprosesannya dan akan mempengaruhi kesan yang tidak membantu dalam langkah ramalan model. Sebelum pengekodan dilakukan, maklumat unik setiap atribut diperlukan bagi mengenal pasti bilangan kuantiti nilai yang perlu dikodkan semula.

Menerusi hasil analisa, mendapati nilaian atribut yang tidak memberikan nilai unik yang stabil pada setiap atribut akan merumitkan proses pengekodkan terhadap semua dataset. Justeru, tindakan bagi penetapan membuang set data kategori yang mempunyai nilai unik hampir keseluruhan data dan kardinalitinya adalah sangat tinggi. Bagi tindakan ini, beberapa atribut yang terlibat untuk dikeluarkan daripada set data adalah *LogicalDeviceNo* dengan nilai unik sebanyak 79461, *DeviceNo* dengan nilai unik sebanyak 46436, *IPNo* sebanyak 4059, *MeterInstallationDate* dan *Date* iaitu sebanyak 4958 setiapnya. Seterusnya proses pengekodan dilakukan pada set data yang tinggal dengan menggunakan pendekatan *LabelEncoder()* bagi ciri jenis data kategori.

4.1.11 Analisa Hasil 5.0: Normalisasi Julat

Proses normalisasi julat adalah proses akhir yang perlu dilakukan di dalam pra-pemprosesan data. Namun proses ini tidak melibatkan semua bentuk analisa ramalan yang diperlukan memandangkan normalisasi julat ini hanya perlu dilakukan bagi pada set data model ramalan secara klasifikasi dan tidak dilakukan ke atas model ramalan regresi. Hal ini berlaku disebabkan oleh normalisasi sangat berguna apabila ciri-ciri data mempunyai skala atau unit yang berbeza, dan perlu untuk menormalkan data ke skala yang sama bagi mengelakkan ciri-ciri tertentu daripada mendominasi proses pembelajaran kerana skala yang lebih besar. Walau bagaimanapun, bagi model regresi dengan sesetengah masalah regresi, pendekatan normalisasi tidak diperlukan disebabkan oleh normalisasi yang dilakukan akan menyebabkan nilai data mempunyai julat hanya antara 0 dan 1, dan juga pendekatan ini tidak sesuai ke atas model regresi disebabkan oleh atribut sasaran mewakili bilangan atau peratusan, yang mana menormalkan ia mungkin tidak sesuai. Oleh itu, di dalam kajian ini, kaedah normalisasi model klasifikasi dilakukan menggunakan kaedah seperti dalam Rajah 4.15 di bawah menggunakan skala min-max, di mana nilai ditukar kepada julat antara 0 dan 1.

```
from sklearn.preprocessing import MinMaxScaler
from pylab import *
sc_x = MinMaxScaler()
df_scaled= sc_x.fit_transform(df_le)
df_scaled
```

```
array([[0.91666667, 1.        , 0.14285714, ..., 0.        , 0.09090909,
        0.9       ],
       [0.91666667, 1.        , 0.14285714, ..., 0.        , 0.18181818,
        0.76666667],
       [0.58333333, 1.        , 0.14285714, ..., 0.        , 0.27272727,
        0.        ],
       ...,
       [0.41666667, 1.        , 0.14285714, ..., 1.        , 0.09090909,
        0.46666667],
       [0.        , 1.        , 0.14285714, ..., 1.        , 0.09090909,
        0.46666667],
       [0.91666667, 1.        , 0.14285714, ..., 1.        , 0.09090909,
        0.46666667]])
```

Rajah 4.15 Teknik Normalisasi ke atas set data Model Klasifikasi

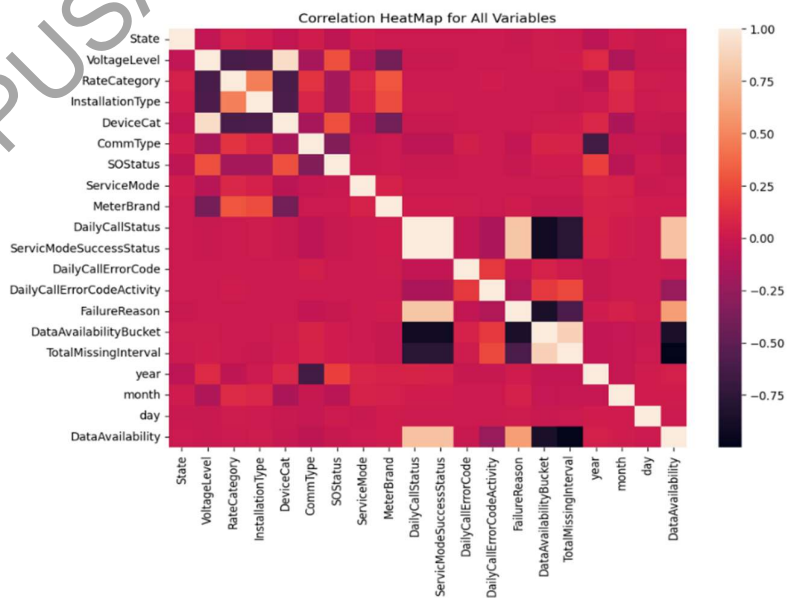
Analisa keseluruhan data set data bagi kedua-dua pendekatan ramalan adalah ditunjukkan. Penerangan berkaitan pendekatan ini adalah yang menunjukkan nilai normalisasi berlaku di dalam set pembangunan secara klasifikasi dengan nilai julat di antara 0 dan 1 manakala normalisasi tidak berlaku ke atas set ramalan regresi iaitu dengan mengekalkan data mentah yang diperolehi sebelum ini.

4.1.12 Analisa Hasil 6.0: Pemilihan Ciri

Pemilihan ciri ialah proses memilih subset atribut yang berkaitan berdasarkan dalam set data sedia ada yang akan digunakan dalam model pembelajaran mesin. Tujuan pemilihan atribut ini adalah untuk meningkatkan prestasi model dalam hal mengurangkan saiz data yang tidak menyediakan hubungan kolerasi yang baik dengan ramalan. Dengan melibatkan hanya atribut yang berkaitan boleh membantu mempercepatkan proses data latihan dengan dimensi data yang telah berkurang serta mengelakkan *overfitting* atau penyesuaian yang berlebihan kepada hasil pembelajaran mesin ini.

4.1.13 Analisa Hasil 6.1: Ciri Model Regresi

Bagi pemilihan ciri yang dibuat, kaedah korelasi menggunakan pendekatan *Pearson* adalah metod yang digunakan dalam kajian ini bagi kedua-dua model ramalan kajian dengan cara mengukur hubungan linear antara dua atribut input dan sasaran. Korelasi *Pearson,* adalah berkaitan hubungan linear antara dua variabel numerik. Oleh itu, pendekatan ini adalah sesuai dengan kedua-dua kajian regresi dan klasifikasi ini setelah proses pengekodan dilakukan ke atas set data tersebut. Untuk pemilihan ciri bagi model regresi ditunjukkan yang menerangkan bahawa sejauh mana hubungan linear sesuatu atribut terhadap sasaran iaitu atribut *DataAvailability*.



Rajah 4.16 Korelasi *Heatmap* bagi setiap atribut Model Regresi

Korelasi yang efisien melalui *Pearson* adalah dikira sebagai nilai yang signifikan apabila korelasi menunjukkan antara -1 hingga 1, dengan nilai 1 menunjukkan hubungan kolerasi positif *linear* yang tepat sebagaimana nilai -1 menunjukkan hubungan kolerasi *linear* negatif yang tepat. Bagi 0 menunjukkan tidak adanya hubungan linear satu atribut tersebut dengan sasaran nilai *DataAvailability*. Rajah di atas menunjukkan analisa kolerasi yang visualkan melalui *heatmap* secara matrik bagi setiap atribut di dalam model regresi. *Heatmap* ini dapat menjelaskan secara jelas pola korelasi di dalam set data ini, di mana hubungan di antara setiap atribut ditunjukkan dengan ciri setiap warna adalah mempunyai penanda aras nilai korelasinya.

```
corr_matrix_stdrd["DataAvailability"].sort_values(ascending=False)

#filter selection
```

```
DataAvailability              1.000000
ServicModeSuccessStatus       0.782313
DailyCallStatus               0.782071
FailureReason                 0.603553
year                          0.044723
month                         0.024542
MeterBrand                    0.018532
InstallationType              0.010762
RateCategory                  0.004986
day                           0.003373
ServiceMode                  -0.000124
State                        -0.003023
DailyCallErrorCode           -0.015101
VoltageLevel                 -0.015592
DeviceCat                    -0.016332
SOStatus                     -0.023199
CommType                     -0.059886
DailyCallErrorCodeActivity   -0.235229
DataAvailabilityBucket       -0.863807
TotalMissingInterval         -0.999976
Name: DataAvailability, dtype: float64
```

Rajah 4.17 Matrik *Corr(Pearson)* bagi *DataAvailability*

Pendekatan *Corr(Pearson)* dilakukan sebab teknik dalam mengira hubungan setiap atribut dengan menghitung matriks korelasi antara mereka. Hasilnya adalah sebuah *corr_matrik* yang berisi korelasi antara setiap pasangan atribut input dan hasil sasaran. Rajah 4.17 menunjukkan bahawa nilai yang paling tinggi korelasi positif dalam set data ini adalah atribut *ServiceModeSuccessStatus* iaitu nilainya 0.782313. Bagi nilai korelasi negatif paling sempurna adalah atribut *TotalMissingInterval* dengan korelasinya sebanyak -0.999976 iaitu menghampiri nilai korelasi -1. Hasil analisis korelasi ini juga menunjukkan bahawa atribut yang tidak mempunyai hubungan dengan sasaran adalah ditunjukkan kepada atribut *ServiceMode* dengan korelasi -0.000124 iaitu nilai terendah yang didapati di dalam set data ini. Setelah perkiraan korelasi antara variabel ini diperolehi, tindakan yang perlu bagi tujuan pemilihan ciri dibuat adalah dengan

membuang ciri yang tidak mempunyai korelasi yang baik iaitu memiliki korelasi rendah atau tidak relevan dengan atribut sasaran. Oleh itu, kajian bagi set data regresi ini mengambil langkah dengan menghapuskan nilaian atribut *ServiceMod*e di atas kolerasi terendah yang dimilikinya.

4.1.14 Analisa Hasil 6.2: Ciri Model Klasifikasi

Untuk pemilihan ciri bagi model klasifikasi, proses yang sama dilakukan seperti pemilihan ciri model regresi. Perubahan hanya dilakukan kepada perubahan atribut sasaran kepada sasaran pengelasan dari atribut *StatusDataAvailability*.

Rajah 4.18 Korelasi *Heatmap* bagi setiap atribut Model Klasifikasi

Heatmap di dalam Rajah 4.18 bagi pembentukan ciri model klasifikasi ditunjukkan di dalam analisis korelasi ini dengan memberikan ciri warna berserta penanda korelasi pada setiap atribut dan hasil sasaran iaitu *StatusDataAvailability.* Matrik penilaian korelasi juga ditunjukkan dalam Rajah 4.19 di bawah, menunjukkan nilai korelasi positif tertinggi adalah atribut *FailureReason* iaitu 0.945351, nilai negatif terendah adalah daripada atribut *DataAvailabilityBucket* iaitu -0.959267 dan nilai kolerasi yang menghampiri 0 iaitu atribut *ServiceMode* iaitu 0.001287. Atribut *ServiceMode* menjadi ciri yang dihapuskan bagi penyediaan data analisa seterusnya.

```
corr_matrix['StatusDataAvailability'].sort_values(ascending=False)

#filter selection
```

```
StatusDataAvailability        1.000000
FailureReason                 0.945351
DailyCallStatus               0.913416
ServicModeSuccessStatus       0.912743
DataAvailability              0.750858
year                          0.040687
month                         0.039824
MeterBrand                    0.023938
day                           0.010696
InstallationType              0.004617
ServiceMode                   0.001287
RateCategory                 -0.005672
VoltageLevel                 -0.006008
DeviceCat                    -0.008176
State                        -0.010070
SOStatus                     -0.016901
DailyCallErrorCode           -0.059756
CommType                     -0.061292
DailyCallErrorCodeActivity   -0.157893
TotalMissingInterval         -0.747165
DataAvailabilityBucket       -0.959267
Name: StatusDataAvailability, dtype: float64
```

Rajah 4.19 Matrik *Corr(Pearson)* bagi *StatusDataAvailability*

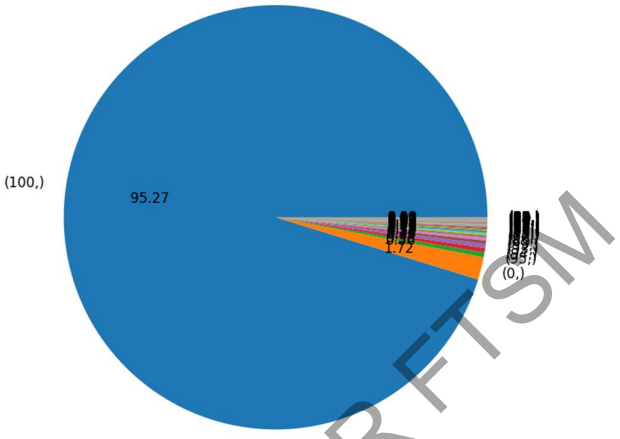
4.1.15 Analisa Hasil 7.0: Pensampelan Semula

Kaedah pensampelan semula adalah salah satu teknik di dalam pembelajaran mesin sebelum dilakukan latihan ke atas model algoritma yang dipilih. Penggunaan kaedah ini amat penting sekiranya di dapati set data kajian adalah tidak seimbang sasaran hasilnya kerana pembelajaran yang dibuat tidak berjaya melihat keseimbangan ramalan yang lebih tepat disebabkan tidak dapat menguasai secara lebih menyeluruh keadaan set data dalam semua keadaan data. Namun, pensampelan semula ini hanya dilakukan ke atas set latihan kajian dan tidak dilakukan ke atas keseluruhan data atau set data ujian. Kajian ini menggunakan beberapa teknik untuk menangani ketidakseimbangan kelas atau hasil sasaran. Oleh itu, penggunaan pelaksanaan kajian ini membantu analisis pensampelan data sasaran yang tidak seimbang dan secara tidak langsung meningkatkan keberkesanan model.

4.1.16 Analisa Hasil 7.1: Teknik SMOGN-SMOTER bagi Model Regresi

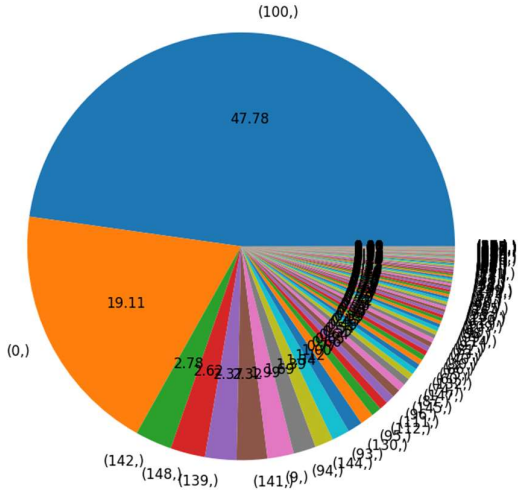
Kajian ini memperkenalkan pendekatan baharu untuk menangani set data regresi tidak seimbang melalui pendekatan yang dikenali sebagai SMOGN-SMOTER. Teknik ini mempunyai metodologi menggabungkan satu pendekatan untuk mengurangkan saiz kumpulan majoriti dengan mengimbangkan dan meningkatkan saiz kumpulan minoriti. Penggunaan kaedah ini adalah lebih kompleks berbanding kaedah pensampelan yang lain apabila pendekatan gabungan SMOGN-SMOTER ini mendatangkan hasil yang

lebih baik namun terlebih dahulu perlu dilihat kepada karakter atribut lain di dalam data. Bagi kajian ini, sampel atribut yang digunakan adalah nilai numerikal yang telah melalui proses pengekodan. Oleh itu, teknik ini adalah sesuai dijalankan apabila setiap atribut data berbentuk nilai numerikal digunakan.

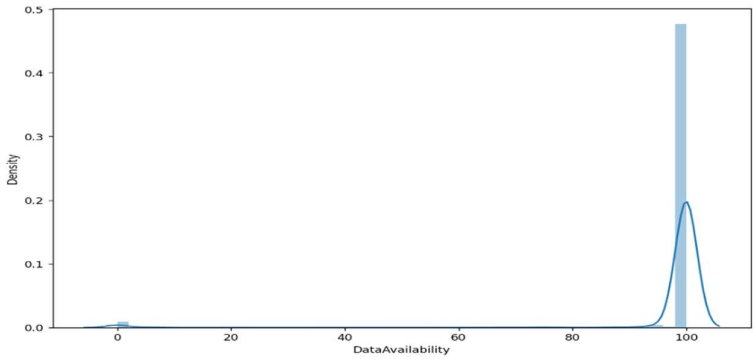


Rajah 4.20 *Pie Chart* Sasaran *DataAvailability* sebelum teknik pensampelan SMOGN-SMOTER

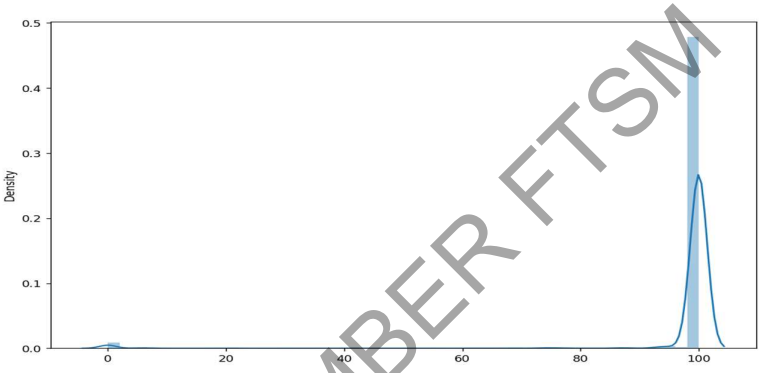
Hasil keseluruhan set data regresi ini yang diperoleh daripada teknik ini ditunjukkan dalam Rajah 4.20 dan 4.21. *Pie chart* ini menjelaskan bagaimana pensampelan dilakukan ke atas set data dan peratusan nilai bagi sasaran telah diseimbangkan dengan sempurna dan membantu bagi penggunaan proses seterusnya.



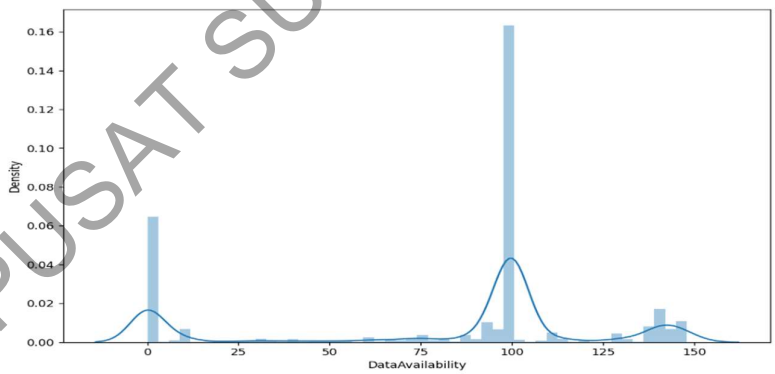
Rajah 4.21 *Pie Chart* Sasaran *DataAvailability* selepas pensampelan teknik SMOGN-SMOTER



Rajah 4.22 *Density* graf bagi *y_test* tanpa pensampelan



Rajah 4.23 *Density* graf bagi *y_train* sebelum pensampelan

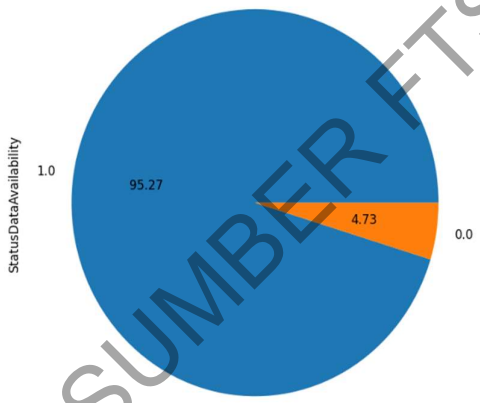


Rajah 4.24 *Density* graf bagi *y_train* selepas pensampelan

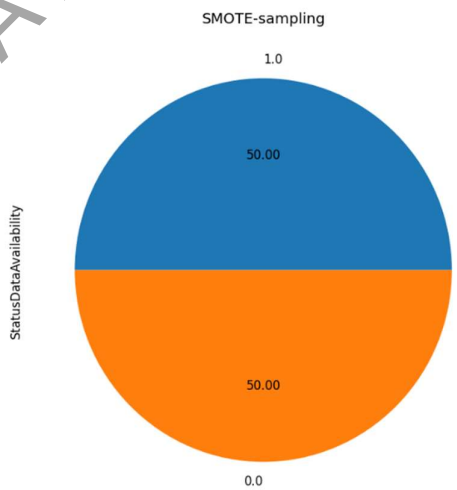
Selain itu, pensampelan semula dilakukan kepada set data latihan dan tidak dijalankan ke atas set data *testing* pada dalam Rajah 4.22 di atas, di mana *y_test* dengan penggunaan data asal tanpa pensampelan. Untuk teknik pensampelan ini, hanya dijalankan dalam data set latihan seperti yang ditunjukkan di dalam Rajah 4.23 dan 4.24 seperti kesan pensampelan data ini membawa kepada pemerhatian berkenaan perbandingan graf density taburan data untuk strategi sampel semula terhadap sasaran *DataAvailability*

4.1.17 Analisa Hasil 7.2: Teknik SMOTE bagi Model Klasifikasi

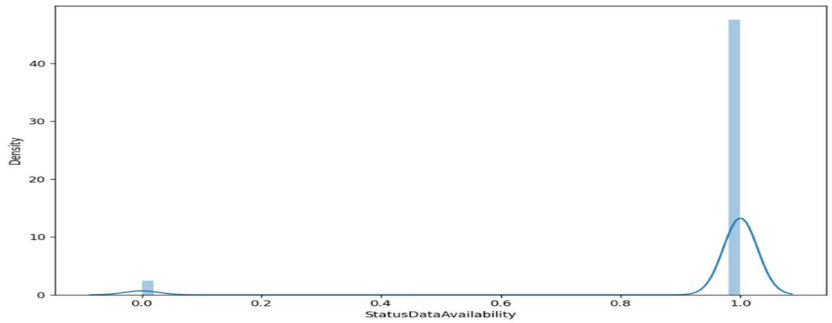
Kajian bagi pensampelan klasifikasi turut dijalankan ke atas *Status DataAvailability* dengan menggunakan kaedah SMOTE bagi membantu kepada penyediaan data seimbang. Rajah 4.25 menjelaskan bahawa 95.27% adalah dari rekod data yang pada kelas 1 di mana 100% *DataAvailability* berjaya dibawa balik oleh meter RMR ini. Namun, 4.73% adalah data kelas 0 iaitu kurang 99% dan ke bawah iaitu peratusan data yang tidak berjaya diperoleh sepenuhnya. Perubahan dapat dilihat pada Rajah 4.26 selepas proses pensampelan dilakukan, pembahagian data dapat dibahagikan kepada dua kelas pembahagian dengan keseimbangan diperolehi ke atas ke dua-dua kelas ini.



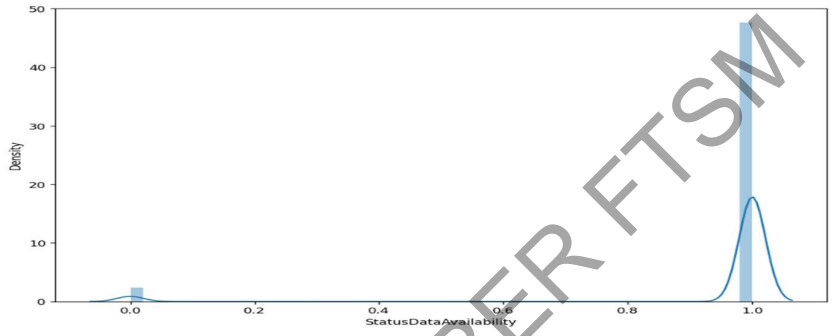
Rajah 4.25 *Pie Chart* Sasaran *DataAvailability* sebelum teknik SMOTE



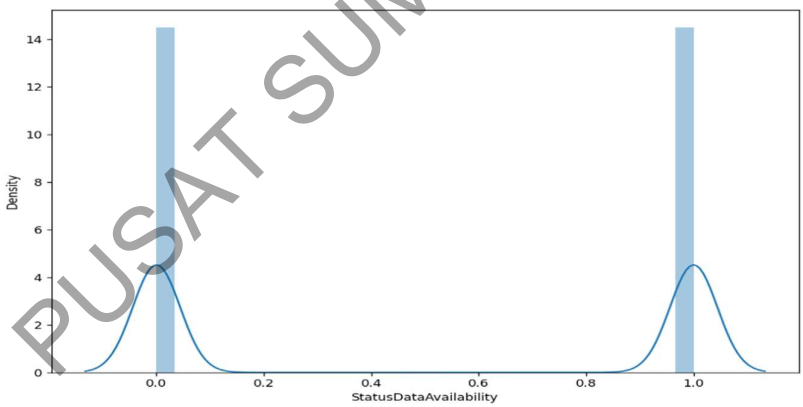
Rajah 4.26 *Pie Chart* Sasaran *DataAvailability* selepas teknik SMOTE



Rajah 4.27 *Density* graf bagi *y_test* tanpa pensampelan



Rajah 4.28 *Density* graf bagi *y_train* sebelum pensampelan



Rajah 4.29 *Density* graf bagi *y_train* selepas pensampelan

Penerangan berkenaan dengan pensampelan juga boleh dilihat menerusi hasil graf density yang ditunjukkan di dalam rajah yang mana data set latihan, *y_test* bagi *StatusDataAvailability* ditunjukkan dalam Rajah 4.27, *y_train* sebelum dilakukan pensampelan iaitu Rajah 4.28 dan *y_train* selepas dilakukan pensampelan pada Rajah 4.29. Bagi data *y_test* tiada pensampelan dilakukan ke atas set ujian dan rajah di atas menunjukkan keadaan asal *y_test* tanpa aktiviti pensamplean dilakukan. Sebagaimana keadaan asal berlaku pada *y_test,* keadaan yang sama juga berlaku kepada data set

latihan *y_train* pada awal sebelum pensampelan dilakukan. Perbezaan dapat dilihat pada pensampelan yang berlaku ke atas data set latihan iaitu *y_train* dengan mengurangkan data majoriti dan meningkatkan data minoriti bagi hasil keseimbangan yang diperlukan.

4.2 ANALISA DAN DAPATAN PREDIKTIF

Peringkat analisis seterusnya di dalam penyelidikan bagi kajian ini adalah peringkat utama yang penting dalam menggunakan pendekatan analisa analitik berhubung dengan isu kajian yang dijalankan ini. Di dalam konteks analisa prediktif, ramalan hasil menggunakan kaedah pembelajaran mesin dengan pendekatan beberapa model algoritma diperlukan bagi menjadikan analisa kajian yang lebih komprehensif. Oleh itu, pembinaan model bagi dua jenis kajian iaitu kajian berkenaan set data yang perlu diramal peratusan ketersediaan dalam sesuatu senario meter RMR dan kajian dalam membina pengelasan antara set data yang lengkap ketersediaan data dan yang tidak lengkap ketersediaan data. Analisa ramalan ini adalah amat penting dalam mencari model algoritma yang terbaik bagi menjalankan analisa harian isu yang berkenaan.

4.2.1 Analisa Hasil 1.0: Model Regresi

Pembangunan beberapa model regresi dilakukan dalam kajian ini bagi melihat keseluruhan analisa prestasi model dalam menentukan keberhasilan model melakukan tugasan ramalan yang diperlukan. Oleh itu, di dalam analisa model regresi, hasil ramalan model dilakukan penilaian bagi melihat sejauh mana prestasi model-model ini dengan kaedah penilaian menggunakan analisa *Regression Performance Error* iaitu nilai bagi MAE, MSE, dan RMSE. Di samping itu, penilaian menggunakan R^2 bagi melihat seberapa baik model regresi tersebut adalah baik dengan set data yang disediakan. Nilai-nilai persegi R^2 yang lebih tinggi ini menjelaskan kesesuaian model dengan data adalah lebih baik. Justeru, dalam penilaian model regresi ini, gabungan antara analisa *performance error* dan kebolehupayaan model dapat menyesuaikan dengan ciri-ciri data adalah penting dalam mengkaji keberhasilan setiap model.

4.2.2 Analisa Hasil 1.1: Hasil keputusan R^2 dan *Regression Performance Error*

Jadual 4.1 Perbandingan Prestasi Model Ramalan Regresi

Model	R^2	MAE	MSE	RMSE
LR	89.48%	1.7454	22.0798	4.6989
RR	89.48%	1.7454	22.0797	4.6989
KNN	87.95%	0.7134	25.3012	5.0300
RF	93.42%	0.5234	13.8147	3.7168
SVR Linear	79.72%	1.1781	42.5892	6.5260
SVR RBF	97.03%	1.0111	44.8852	6.6996
SVR Poly	97.00%	1.0264	45.9026	6.7751

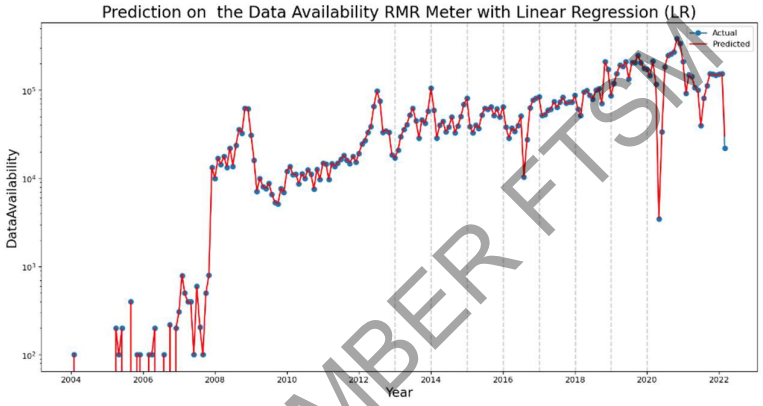
Penilaian model LR ditunjukan di dalam rajah di 4.1 dengan menjelaskan model LR dengan hasil R^2 iaitu 89.48% di mana peratusan model boleh dapat membina kolerasi bagi kedua-dua atribut input dan sasaran. Analisa kesilapan regresi bagi LR model menunjukkan MAE iaitu 1.745, MSE iaitu 22.079 dan RMSE iaitu 4.698. Model RR yang dibinakan menunjukkan prestasi bagi R^2 sebanyak 89.48% dengan nilai MAE iaitu 1.745, MSE iaitu 22.07, dan RMSE iaitu 4.698. Dengan nilai MAE yang mengira purata perbezaan mutlak hasil ramalan dan hasil sebenar, ditambah dengan MSE dan RMSE yang mengira purata perbezaan bersegi dan akar persegi ini menampakkan nilai yang rendah bagi analisa kesilapan ini, namun kesesuaian antara hubungan data masih ditahap yang perlu ditingkatkan. Untuk analisa seterusnya, penilaian terhadap model KNN dengan hasil R^2 adalah 87.95% dengan purata perbezaan mutlak MAE iaitu 0.713 manakala purata perbezaan persegi MSE 25.3011 dan akar perbezaan persegi iaitu 5.030. Model ini menunjukkan MAE yang rendah namun peningkatan di dalam perbezaan RMSE yang agak tinggi menyebabkan peratusan kesesuaian model dengan data menurun dari peratusan yang tinggi. Untuk analisa seterusnya, penilaian terhadap model KNN dengan hasil R^2 adalah 87.95% dengan purata perbezaan mutlak MAE iaitu 0.713 manakala purata perbezaan persegi MSE 25.3011 dan akar perbezaan persegi iaitu 5.030. Model ini menunjukkan MAE yang rendah namun peningkatan di dalam perbezaan RMSE yang agak tinggi menyebabkan peratusan kesesuaian model dengan data menurun dari peratusan yang lebih baik. Selain itu, model RF turut dilakukan analisa ramalan hasilnya dengan hasil yang menunjukkan nilai bagi R^2 adalah 93.42%. Seterusnya, penilaian dengan purata perbezaan mutlak MAE iaitu

0.5234 manakala purata perbezaan persegi MSE adalah 13.8146 dan akar purata perbezaan persegi iaitu 3.7168. Model ini menunjukkan kesilapan purata perbezaan yang rendah, sekaligus meningkatkan prestasi bagi R^2.

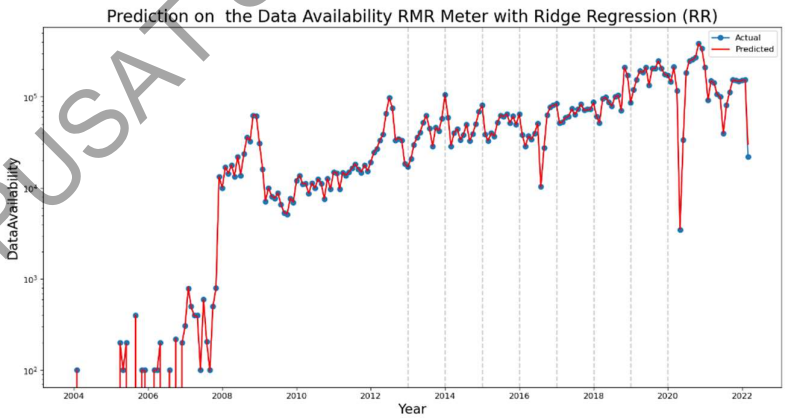
Bagi analisa model SVR, penilaian ke atas SVR linear dibuat dengan hasil R^2 adalah 79.72% dan purata perbezaan mutlak MAE sebanyak 1.178, di samping purata perbezaan persegi MSE dan akar purata perbezaan persegi masing – masing adalah 42.5892 dan 6.526. Model SVR Linear di dapati kurang menunjukkan persamaan dalam kesesuaian set data yang ditunjukkan dengan nilai perbezaan mutlak dan persegi yang agak besar sekaligus juga menyebakan nilai RMSE meningkat dan menyumbang kepada penurunan nilai R^2. Model SVR bersama kernel RBF iaitu *kernel radial basis function* adalah digunakan dalam mencari persamaan kesesuaian data di dalam ruang dimensi yang lebih besar. Penilaian ke atas SVR RBF mendapati bahawa skor bagi R^2 ialah 97.03% dengan kesalahan purata mutlak MAE sebanyak 1.0111 manakala purata persegi kesalah MSE iaitu 44.8852 dan akar persegi purata kesalahan RMSE adalah 6.6996. Walaupun nilai MSE dan RMSE adalah sedikit tinggi, MAE bagi model ini menunjukkan penurunan perbezaan nilai purata kesilapan hasil ramalan dan yang sebenar. Peningkatan nilai bagi R^2 yang agak tinggi menunjukkan model SVR model ini membantu dalam menyesuaikan model dengan set data dengan mengambil kira juga hubungan non-linear data sehinggakan pengiraan ramalan menjadi semakin tepat. Persamaan bagi SVR poly berlaku apabila SVR poly ini menggunakan *polynomial function* dalam membantu hubungan non linear di antara input dan sasaran ramalan. Untuk penilaian di dalam SVR poly ini, hasil kajian menunjukkan R^2 yang ditunjukkan adalah agak tinggi di mana memperolehi 97.00% bersama dengan nilai MAE iaitu 1.0264 dan MSE serta RMSE dengan nilai sebanyak 45.9025 dan 6.7751. Model SVR poly ini juga menjelaskan bahawa model ini menghubungkan hubungan setiap atribut input dan sasaran dengan menggunakan pendekatan fungsi *polynomial* yang mana kesesuaian data berlaku di dalam ruangan dimensi yang lebih besar dengan dalam mendapatkan *hyperplane* yang optimum bagi model ini.

4.2.3 Analisa Hasil 1.2: Hasil Graf Sebenar dan Ramalan Regresi

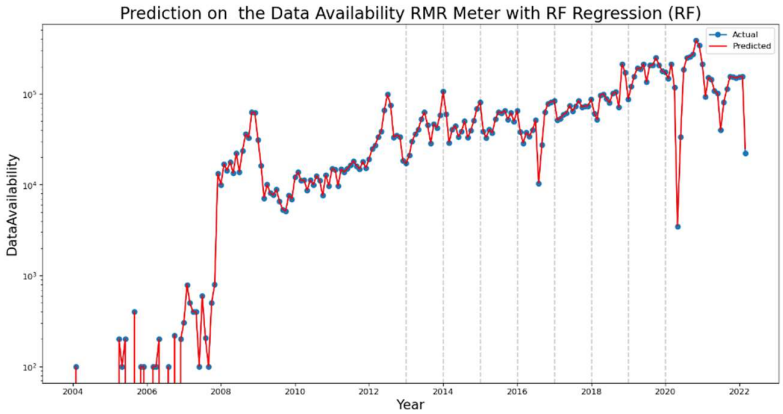
Pembinaan model regresi di dalam kajian ini memberikan beberapa hasil yang boleh dilakukan perbandingan prestasinya. Untuk melihat sejauh mana hasil setiap model, implementasi hasil set latihan dalam setiap model boleh dilihat melalui prestasi perbezaan antara hasil sebenar dan hasil ramalan yang diperolehi. Di dalam bahagian ini, pemerhatian graf bagi nilai ramalan dan nilai sebenar dilakukan seperti di dalam rajah di bawah.



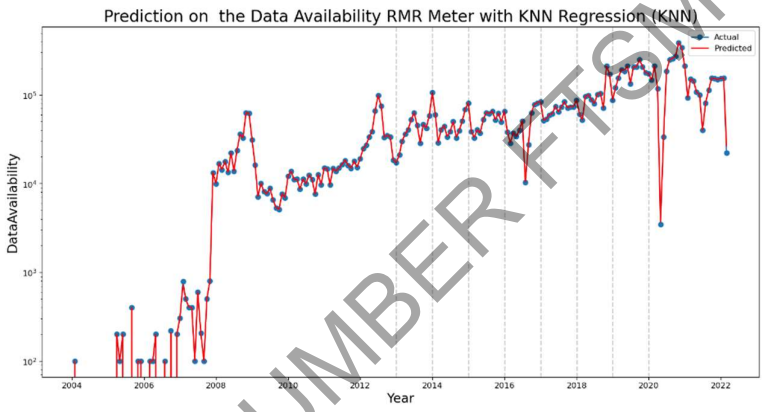
Rajah 4.30 Perbandingan di antara hasil ramalan dan hasil sebenar (LR)



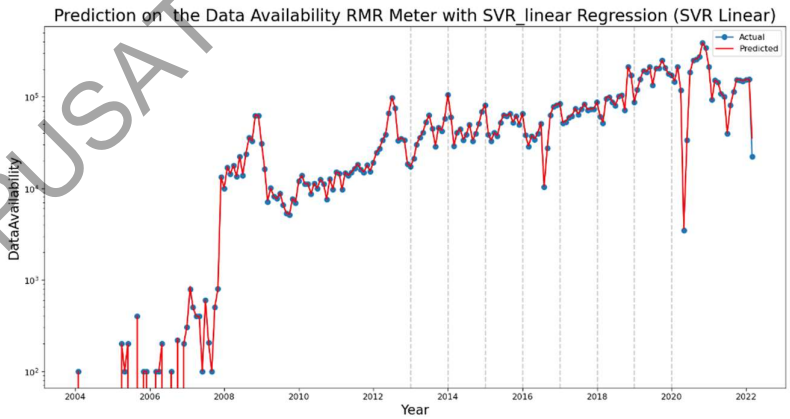
Rajah 4.31 Perbandingan di antara hasil ramalan dan hasil sebenar (RR)



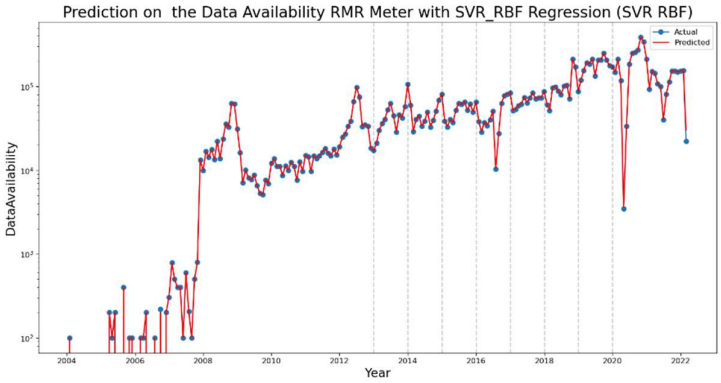
Rajah 4.32 Perbandingan di antara hasil ramalan dan hasil sebenar (RF)



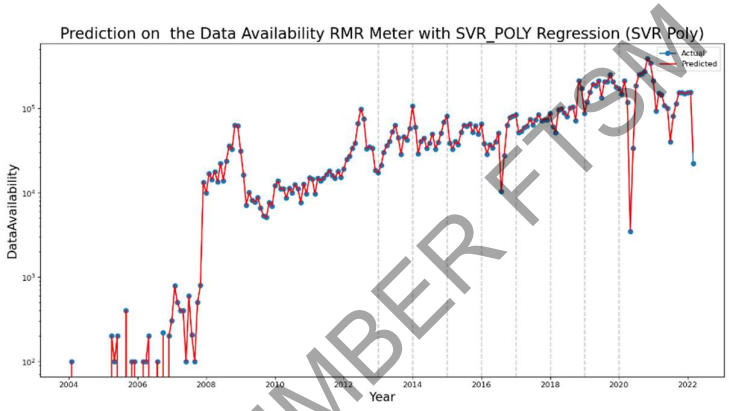
Rajah 4.33 Perbandingan di antara hasil ramalan dan hasil sebenar (KNN)



Rajah 4.34 Perbandingan di antara hasil ramalan dan hasil sebenar (SVR Linear)



Rajah 4.35 Perbandingan di antara hasil ramalan dan hasil sebenar (SVR RBF)



Rajah 4.36 Perbandingan di antara hasil ramalan dan hasil sebenar (SVR Poly)

Melalui pemerhatian ke atas analisa graf yang dijalankan, setiap model regresi ini menunjukkan nilai yang hampir sama atau tepat di antara hasil ramalan dan nilai sebenar. Perbezaan di dalam graf adalah tidak ketara memandangkan skala digit perbezaan adalah sangat kecil *decimal* pada setiap model. Oleh itu, penambahan analisa menerusi gabungan hasil setiap model dipaparkan di dalam bentuk *dataframe*.

date	RAW	LR	RR	RF	KNN	SVR_LIN	SVR_RBF	SVR_POLY
2004-01-31	100	102.187788	102.187802	100.000000	100.000000	100.100259	100	100
2004-02-29	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0
2004-03-31	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0
2004-04-30	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0
2004-05-31	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0
...	...	...	...	...	...	...	...	...
2021-10-31	152168	150652.331685	150652.338119	152280.683855	152474.000000	152768.390669	152488	152488
2021-11-30	148579	147456.878697	147456.882149	148856.363808	148965.333333	149771.104660	149264	149264
2021-12-31	153208	151813.513160	151813.518564	153353.226256	153662.000000	154174.964307	153832	153832
2022-01-31	154559	153065.962701	153065.970755	154901.716335	155229.000000	156198.745945	155591	155591
2022-02-28	22116	30163.820846	30163.802999	25007.169337	26662.000000	35305.601280	30246	30246

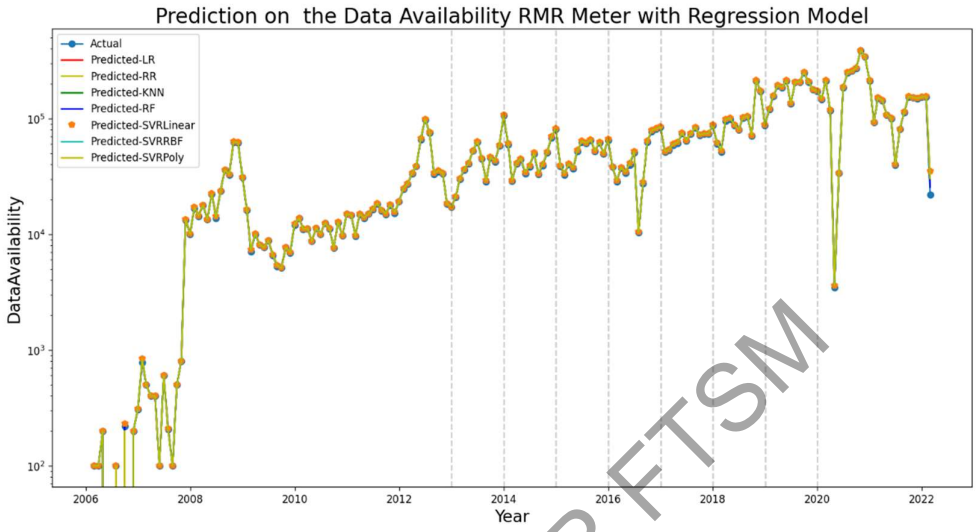
Rajah 4.37 *Dataframe* bagi ramalan hasil setiap model dan hasil sebenar

4.2.4 Analisa Hasil 1.3: Hasil keputusan Perbandingan Model Regresi

Bagi analisa akhir untuk hasil daripada model regresi ini, perbandingan hasil ditunjukkan di dalam Jadual 4.1 bagi melihat secara keseluruhan hasil prestasi setiap model. Menerusi penilaian nilai di dalam jadual, model SVR RBF menunjukkan nilai R^2 yang tertinggi iaitu pada 97.03% diikuti oleh dua model terbaik selepasnya iaitu SVR Poly dan RF yang masing-masing memperoleh 97.00% dan 93.42%. Dua model kajian yang menunjukkan nilai R^2 yang sama sebanyak 89.48% iaitu model LR dan RR. Pemilihan dua model yang terbaik antara dua model ini dibuat dengan melihat kepada hasil MAE,MSE dan RMSE. Justeru, model RR dilihat lebih baik jika dibandingkan dengan LR apabila terdapat perbezaan nilai kecil di dalam MSE yang lebih kecil di dalam RR berbanding LR. Selanjutnya model KNN yang menunjukkan nilai R^2 pada 87.95% dan mendapati model terakhir adalah model SVR Linear dengan nilai R^2 hanya sebanyak 79.72%. Perolehan hasil dari R^2 menunjukkan model yang terbaik dalam membentuk hubungan antara dua atribut input dan sasaran dapat memberikan gambaran seberapa baik model adalah sesuai dengan data yang disediakan dan boleh dicadangkan kepada model SVR RBF bagi kajian penyelidikan isu ini. Antara lain yang boleh disimpulkan, adalah jenis data bagi kajian ini adalah lebih berdimensi tinggi dan mempunyai banyak atribut tidak linear di dalam set data sehinggakan peningkatan kolerasi hubungan data dan hasil boleh ditingkatkan dengan penggunaan *kernal* fungsi RBF atau Polinomial ini.

Selain daripada analisa R^2, penilaian ke atas kesilapan dalam ramalan hasil dan nilai sebenar turut dijalankan dengan prestasi MAE terendah adalah daripada model RF iaitu 0.5234. Seterusnya diikuti dengan model KNN, SVR RBF, SVR Poly, SVR Linear, RR dan LR dengan nilai MAE setiapnya adalah, 0.7134, 1.0111, 1.0264, 1.1781, dan 1.7454. Namun hasil analisa ini juga menunjukkan nilai MSE dan RMSE yang agak tinggi bagi sesetengah model contohnya model SVR Linear, SVR RBF dan SVR Poly. Ini kerana, model daripada *Support Vector Machine* ini adalah algoritma yang sangat sensitif dengan *outlier* yang terdapat di dalam set data kerana konsep SVM ini adalah model yang mencari nilai maksimum margin. Oleh itu, kajian ini boleh dilakukan

analisa lanjut berkenaan perubahan *kernal* dengan parameter *gamma* atau C yang lebih sesuai.



Rajah 4.38 Graf keseluruhan model yang dilakukan dalam kajian ini.

Penilaian akhir bagi analisa model regresi ini juga dilakukan secara lebih menyeluruh hasil kajian daripada model dengan perbezaan hasil ramalan dan hasil sebenar bagi setiap model yang dijalankan kajian hasilnya. Perincian graf pada bahagian sebelum ini yang menerangkan hasil ramalan mempunyai perbezaan *decimal* nilai yang kecil. Ini menunjukkan, kesemua model ramalan di dalam kajian ini adalah baik dalam dan mendapati model yang terbaik adalah berdasarkan keseluruhan gabungan analisa yang dijalankan.

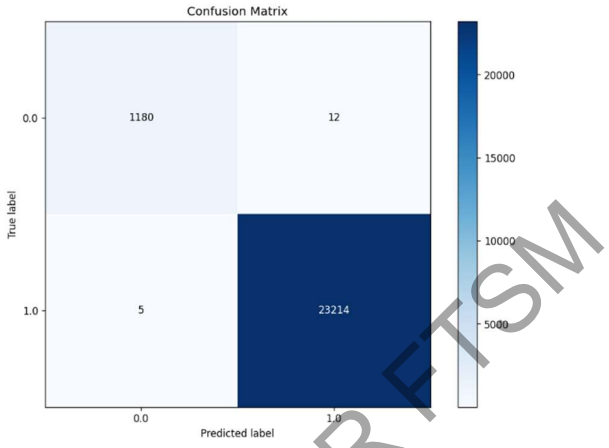
4.2.5 Analisa Hasil 2.0: Analisa Hasil Model Klasifikasi

Perolehan hasil daripada analisa klasifikasi dibentangkan di dalam bahagian ini dengan kaedah penilaian dan pemerhatian hasil dapatan dalam kajian. Antara penilaian yang dilakukan adalah analisa *Confusion Matrik, Classification Report, ROC Curve* dan perbandingan prestasi model ramalan.

4.2.6 Analisa Hasil 2.1: Analisa Hasil Model Kekeliruan Matrik (*Confusion Matrik*)

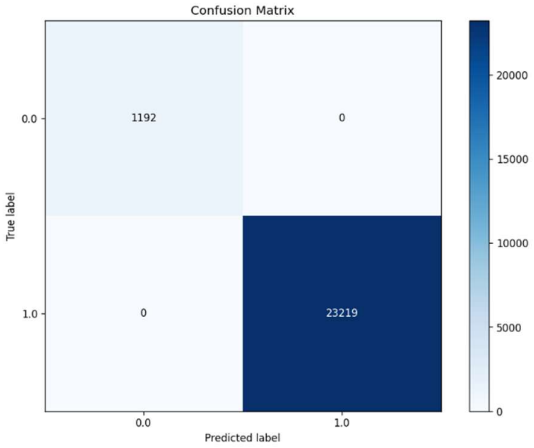
Matrik Kekeliruan adalah penilaian teknik yang digunakan dalam kajian analisa prestasi hasil ramalan klasifikasi dan analisa kesilapan matrik melalui penggunaan terma TP,

FP, TN, FN yang akan menunjukkan bilangan betul dan salah bagi hasil ramalan. Bagi kajian ini, tujuh model klasifikasi ramalan dijalankan dengan setiapnya mempunyai hasil berdasarkan sifat algoritma model. Melalui pembelajaran mesin dengan menggunakan algoritma KNN, Matrik Kekeliruan seperti di dalam Rajah 4.39 ini.



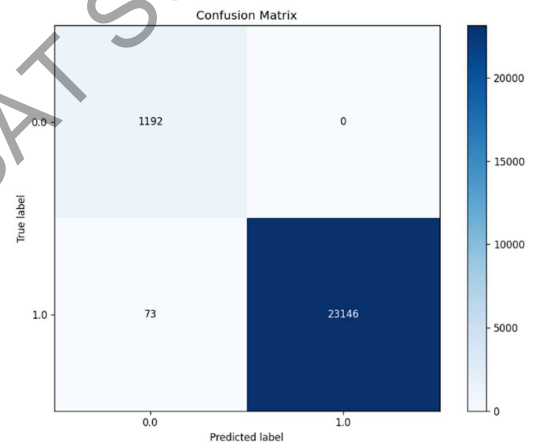
Rajah 4.39 Matrik Kekeliruan bagi KNN model

Hasil daripada nilai Matrik Kekeliruan melalui KNN model menunjukkan, dua pembahagian kelas matrik yang dinilai ketepatan hasil sebenar dan hasil ramalan melalui analisa model ini. Bagi bahagian kelas 0, bilangan bagi TN yang menerangkan bahawa nilaian sebenar adalah negatif dan hasil ramalan juga melakukan ramalan kelas yang betul sebagai nilaian negatif adalah sebanyak 1180, manakala bilangan FP mewakili nilaian sebenar yang sepatutnya adalah negatif namun hasil ramalan adalah salah dengan menjangka sebagai positif iaitu sebanyak 12 bilangan. Sebagai bahagian kelas 1, bilangan bagi TP seterusnya menunjukkan 23214 bilangan yang mewakili hasil sebenar sebagai positif dan meramal dengan betul sebagai positif juga, manakala bilangan bagi FN menunjukkan sebanyak 5 bilangan yang mana menerangkan berkenaan hasil sebenar adalah positif, tetapi hasil ramalan memberikan nilaian yang salah dengan jangkaan sebagai negatif.



Rajah 4.40 Matrik Kekeliruan bagi RF model

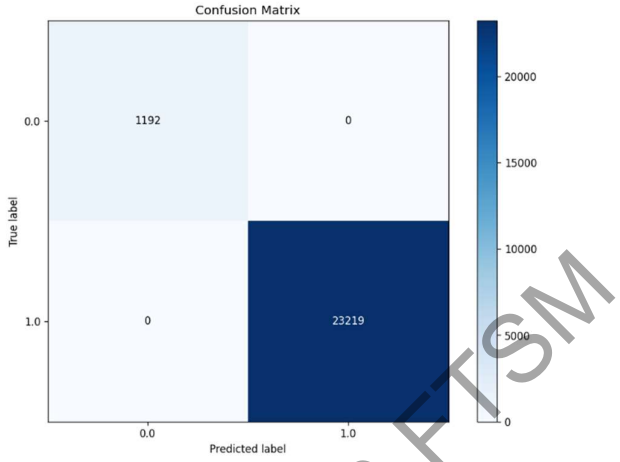
Bagi model klasifikasi RF, analisa kajian daripada pendekatan Matrik Kekeliruan dalam Rajah 4.40 di atas, menghasilkan bilangan TN sebanyak 1192 dan bilangan TP iaitu 23219 manakala tiada bilangan bagi FP dan FN. Ini menunjukkan hasil ramalan yang dilakukan adalah tepat ramalannya dengan hasil sebenar iaitu tiada kesilapan ramalan FP dan FN di dalam kedua-dua kelas negatif dan positif bagi kelas 0 yang mewakili ketersediaan data tidak lengkap dan kelas 1 yang mewakili ketersediaan data yang lengkap.



Rajah 4.41 Matrik Kekeliruan bagi NB model

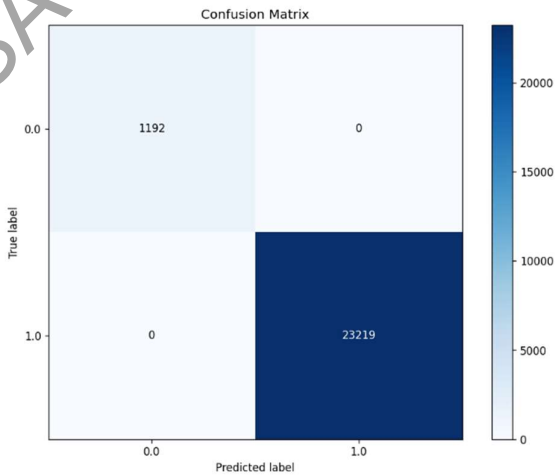
Penilaian Matrik Kekeliruan dalam Rajah 4.41 di atas, menunjukkan model NB mempunyai nilai TP sebanyak 23146, FN sebanyak 73 manakala TN sebanyak 1192 dan tiada bilangan bagi FP. Dengan hasil keputusan ini, memberikan prestasi model ini berjaya meramal kelas negatif 0 dengan tepat terhadap nilai sebenar yang mewakili status ketersediaan data tidak lengkap. Bagi ramalan kelas positif 1, wujud kesilapan

dalam analisa ramalan sebanyak 73 daripada data asal yang berada di dalam kelas positif 1 iaitu ketersediaan adalah data lengkap, namun terdapat kesilapan dalam meramal dengan ramalan ketersediaan data adalah tidak lengkap.



Rajah 4.42 Matrik Kekeliruan bagi SVC Linear model

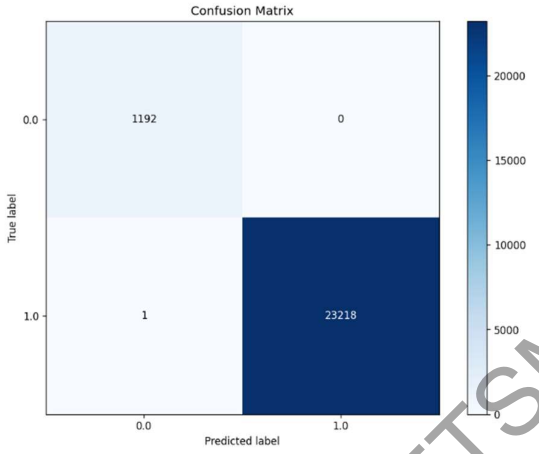
Untuk hasil penilaian daripada Matrik Kekeliruan dalam Rajah 4.42, yang dijalankan ke atas model SVC Linear menunjukkan model ini berjaya meramal kedua-dua kelas positif dan negatif dengan tepat iaitu bersamaan dengan hasil sebenar yang ditetapkan. Bilangan bagi TP adalah sebanyak 1192 manakala bilangan bagi TN sebanyak 23219 dan tiada kesilapan berlaku di dalam terma FN dan FP.



Rajah 4.43 Matrik Kekeliruan bagi SVC Poly model

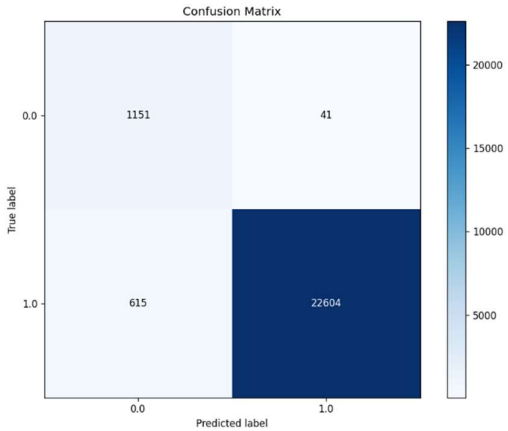
Sebagaimana prestasi hasil di dalam model SVC Linear, hasil yang sama berlaku di dalam model SVC Poly yang mana prestasi model adalah berjaya meramal

dengan tepat bagi kedua-dua kelas positif 1 dan negatif 0 dengan bilangan bagi TP adalah sebanyak 1192 manakala bilangan bagi TN sebanyak 23219.



Rajah 4.44 Matrik Kekeliruan bagi SVC RBF model

Untuk prestasi model SVC RBF, menunjukkan hasil ramalan berjaya dilakukan dengan tepat bagi kelas negatif 0 dengan bilangan TN sebanyak 1192 tanpa bilangan FP manakala ramalan bagi kelas positif 1 menunjukkan bilangan TP sebanyak 23218 dengan 1 bilangan FN wujud dalam hasil ramalan ini. Kesilapan dengan hanya 1 nilai FN ini mewakili nilai positif yang sepatutnya model diramalkan kepada kelas positif 1 ketersediaan data lengkap, namun model menghadapi kesilapan dalam melakukan ramalan kepada ketersediaan data tidak lengkap iaitu yang sepatutnya tidak diramal dalam kelas negatif 0.



Rajah 4.45 Matrik Kekeliruan bagi SVC Sigmoid model

Penilaian Matrik Kekeliruan bagi model SVC Sigmoid menunjukkan model meramal kelas positif 1 dengan hasil bilangan TP iaitu 22604 bersama bilangan FN iaitu 615. Begitu juga dengan hasil penilaian terhadap kelas negatif 0 dengan bilangan TN sebanyak 1151 bersama bilangan FP iaitu 41. Model SVC Sigmoid ini menunjukkan model dapat membuat ramalan namun berlaku juga ketidaktepatan dan perbezaan yang ketara antara hasil ramalan dan hasil sebenar.

4.2.7 Analisa Hasil 2.2: Analisa Hasil Model Laporan Klasifikasi (*Classification Report*)

Bagi perbincangan hasil ramalan model yang selanjutnya, penilaian melalui *Classification Report* turut dilakukan dalam kajian ini dengan membentangkan hasil prestasi model dari segi *Accuracy, Precision, Recall, F1-Score* dan *Support* matrik yang mana masing-masing membawa penilaian ramalan model kepada perincian analisa prestasi sejauh mana keberhasilan model tersebut. Definasi bagi setiap terma penilaian laporan klasifikasi ini dapat diterangkan dengan *Accuracy* sebagai ketepatan model meramal yang benar secara keseluruhan, *Precision* sebagai pendekatan hasil ramalan positif yang benar dihasilkan, *Recall* sebagai analisa hasil sebenar positif yang benar, *F1 score* sebagai penilaian keseimbangan hasil positif antara *Precision* dan *Recall*, dan akhirnya *Support* sebagai analisa jumlah kelas sebenar yang ada di dalam set data . Bagi hasil penilaian ke atas model, Rajah 4.46 di bawah menunjukkan prestasi model KNN dengan memberikan nilai hasil mengikut kelas di mana dengan jumlah bilangan 23219 sampel di dalam kelas 1, mempunyai nilai *precision, recall, f1-score* adalah 1 manakala jumlah bilangan 1192 sampel di dala kelas 0 mempunyai nilai 1 bagi *precision* dengan *recall* dan *f1-score* masing-masing adalah 0.99. Hasil keseluruhan penilaian model ini menunjukkan nilai ketepatan sebanyak 99.93% dan nilai *F1-Score* juga sebanyak 99.93%.


```
Accuracy: 99.93035926426612
F1 Score: 99.93045655838444
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99      1192
         1.0       1.00      1.00      1.00     23219

    accuracy                           1.00     24411
   macro avg       1.00      0.99      1.00     24411
weighted avg       1.00      1.00      1.00     24411
```

Rajah 4.46 *Classification Report* bagi KNN model

Penilaian seterusnya ke atas model RF yang menunjukkan nilaian 1 kepada semua terma di dalam analisa laporan klasifikasi ini dan ketepatan model sebanyak 100% dan *F1-Score* sebanyak 100%.

```
Accuracy: 100.0
F1 Score: 100.0
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1192
         1.0       1.00      1.00      1.00     23219

    accuracy                           1.00     24411
   macro avg       1.00      1.00      1.00     24411
weighted avg       1.00      1.00      1.00     24411
```

Rajah 4.47 *Classification Report* bagi RF model

```
Accuracy: 99.70095448773094
F1 Score: 99.69674743080233
              precision    recall  f1-score   support

         0.0       0.94      1.00      0.97      1192
         1.0       1.00      1.00      1.00     23219

    accuracy                           1.00     24411
   macro avg       0.97      1.00      0.98     24411
weighted avg       1.00      1.00      1.00     24411
```

Rajah 4.48 *Classification Report* bagi NB model

Analisa terhadap model NB turut dijalankan dengan penilaian terhadap kelas 1 menunjukkan hasil 1 kepada semua terma di dalam laporan klasifikasi manakala di dalam kelas 0 hasil 1 berlaku hanya kepada nilai di *recall*, dengan *precision* dan f*1-score* memperoleh nilai 0.94 dan 0.97. Bagi keseluruhan model, nilai *accuracy* sebanyak 99.70% dan nilai *F1 Score* sebanyak 99.70%.

```
Accuracy: 100.0
F1 Score: 100.0
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1192
         1.0       1.00      1.00      1.00     23219

    accuracy                           1.00     24411
   macro avg       1.00      1.00      1.00     24411
weighted avg       1.00      1.00      1.00     24411
```

Rajah 4.49 *Classification Report* bagi SVC Linear model

Model ramalan SVC linear turut dilakukan penilaian dengan makluman klasifikasi model yang sangat baik di mana nilai *accuracy* dan *F1-Score* adalah 100%