MODEL RAMALAN PENUAAN SIHAT BERASASKAN PEMBELAJARAN MESIN

LIAW CHIN FEI

UNIVERSITI KEBANGSAAN MALAYSIA

MODEL RAMALAN PENUAAN SIHAT BERASASKAN PEMBELAJARAN MESIN

LIAW CHIN FEI

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

10 JUN 2024

LIAW CHIN FEI

P117805

PENGHARGAAN

Sepanjang masa yang saya membuat kajian ini, saya amat menghargai pendapat dan tunjuk ajar yang telah diberikan oleh penyelia latihan ilmiah saya, Prof. Madya Dr. Suhaila Binti Zainudin dan ingin memberikan setinggi-tinggi penghargaan kepada beliau sehingga latihan ilmiah ini berjaya disiapkan dengan sebaiknya.

Akhirnya, saya juga menghargai keluarga saya yang telah memberikan sokongan di sebalik pelbagai cabaran semasa membuat kajian ini.

ABSTRAK

Golongan penuaan di Malaysia telah mencapai sebanyak 7.2% daripada semua populasi tetapi masih mengalami kekurangan kajian yang spesifik kepada faktor kapasiti intrinsik yang dapat mempengaruhi penuaan sihat. Penyelidikan ini fokus untuk menganalisis faktor penuaan sihat menggunakan pembelajaran mesin ke atas set data yang mengandungi 43 pembolehubah utama berkaitan dengan kajian penuaan di Malaysia. Objektif utama penyelidikan ini adalah untuk mengenal pasti faktor utama dan seterusnya membangunkan model ramalan terbaik mengenai penuaan yang sihat dengan ketepatan tinggi serta model yang paling sesuai untuk menggambarkan penuaan sihat di Malaysia secara am dan untuk dunia secara umum. Bagi mencapai objektif tersebut, ujian *Chi-squared* dan nilai p telah dikira untuk menyiasat sama ada pembolehubah tersebut adalah signifikan secara statistik. Pembolehubah yang mengandungi nilai p kurang daripada 0.05 dipilih sebagai nilai p yang signifikan dan pembolehubah ini digunakan untuk membangunkan model ramalan. Tujuh model ramalan, Pohon Keputusan (DT), Mesin Vektor Sokongan (SVM), Naive Bayes (NB), Rangkaian Neural Tiruan (ANN), K-Nearest Neighbours (KNN), Hutan Rawak (RF), XGBoost (XGB) telah dibina dan dinilai menggunakan metrik ketepatan, kepersisan, ingatan, spesifisiti, skor F1 dan kawasan di bawah lengkung (AUC) antara kesemua model. Hasil akhir telah menunjukkan bahawa model pengelas Hutan Rawak (RF) memberikan prestasi yang terbaik antara semua model ramalan dan model tersebut telah mencapai ketepatan (*accuracy*) sebanyak 88.81%, kepersisan (*precision*) sebanyak 45.35%, ingatan (*recall*) sebanyak 54.15%, spesifisiti (*specificity*) sebanyak 92.47%, skor F1 (*F1-score*) sebanyak 48.09% dan AUC sebanyak 89.96%. Hasil kajian tersebut dapat mendedahkan faktor utama yang mempengaruhi penuaan sihat melalui dataset, dan dapat digunakan oleh kerajaan sebagai rujukan dalam merangka dasar terhadap isu penuaan sihat di Malaysia.

A Predictive Model of Healthy Ageing Based on Machine Learing

ABSTRACT

The ageing in Malaysia occupied as much as 7.2% of the entire population but there is still lack of specific studies on intrinsic capacity factors that can influence healthy aging. Hence, this research focused on analising the factors of healthy ageing with the using of machine learning on dataset which containing a total of 43 key variables related to the ageing study in Malaysia. The main objective of this research is to discover the main factors which are statistically significant to the target variable and discover the best predictive model on the healthy ageing with high accuracy and the most appropriate model to describe the healthy ageing. To achieve the objectives, chi-squared test and p-value had been calculated to investigate whether the variables are statistically significance. The variables that having a p-value lower than 0.05 is chosen as statistically significant and these variables were used to develop the predictive model. Seven predictive model were developed, which are Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Random Forest (RF) and XGBoost (XGB) and the comparison of metrics accuracy, precision, recall, specificity, F1-score and AUC were made among the models. The final result showing that Random Forest (RF) classifier outperform all the predictive models, with a score of 88.81% accuracy, 45.35% precision, 54.15% recall, 92.47% specificity, 48.09% F1-score and 89.96% AUC. The result able to reveal the main factors that affect the healthy ageing through the dataset, where the result can be used by the government as a reference on making policy for the healthy ageing issue.

ISI KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

SENARAI TATANAMA

ADL	*Activities of Daily Living*
ANN	*Artificial Neural Network*
AUC	*Area Under Curve*
DT	*Decision Tree*
FN	*False Negative*
FP	*False Positive*
FPR	*False Positive Rate*
IADL	*Instrumental Activities of Daily Living*
KDNK	Keluaran Dalam Negeri Kasar
KNN	*K-Nearest Neighbors*
NB	*Naive Bayes*
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
TN	*True Negative*
TP	*True Positive*
TPR	*True Positive Rate*
XGB	*Extreme Gradient Boosting*
WHO	*World Health Organization*

BAB I

PENDAHULUAN

1.1 PENGENALAN

Menurut Pertubuhan Kesihatan Sedunia (WHO), definisi populasi penuaan adalah merujuk kepada golongan penuaan yang berumur 65 tahun dan ke atas. Menjelang tahun 2030, satu daripada enam individu adalah berumur 60 tahun ke atas dan populasi penuaan global ini telah diramal akan meningkat pada masa depan (WHO 2022). Menurut rekod terkini daripada portal rasmi Kementerian Ekonomi Jabatan Perangkaan Malaysia (DOSM), golongan penuaan telah mencapai 7.2% daripada semua populasi di Malaysia (DOSM 2023). Perkembangan yang cepat dalam bidang perubatan pada era ini telah memanjangkan jangka hayat manusia. Tidak dapat dinafikan bahawa penemuan pelbagai vaksin dan antibiotik telah memainkan peranan penting dalam menyediakan jaminan asas terhadap kesihatan manusia dan pemanjangan hayat manusia (Michel & Francos 2022). Fenomena ini telah menyumbang kepada peningkatan populasi warga tua di peringkat global.

Oleh itu, kesihatan orang yang telah menua telah menjadi kebimbangan utama kerajaan kerana sokongan perubatan kepada kumpulan yang telah menua merupakan salah satu tekanan berat terhadap kewangan kerajaan. Penyelidikan sebelumnya telah menunjukkan bahawa peningkatan 1% daripada golongan penuaan akan menyebabkan penurunan 6.6% Keluaran Dalam Negara Kasar (KDNK) di Malaysia (Siti et al. 2021). Walau bagaimanapun, rawatan awal terhadap orang yang telah menua yang menghidap penyakit dapat mengurangkan perbelanjaan rawatan lanjutan warga tua. Pendekatan ini mengurangkan beban kewangan yang besar bagi sokongan perubatan dari kerajaan dan

meningkatkan produktiviti dalam golongan penuaan yang dijadikan sebagai kesan positif dalam ekonomi di Malaysia (Mazlynda et al. 2020).

1.2 PERNYATAAN MASALAH

Penuaan sihat boleh dipengaruhi oleh banyak faktor seperti demografi, sosioekonomi, kesejahteraan hidup dan sebagainya. Terdapatnya banyak kajian tentang isu penuaan sihat telah dijalankan di Malaysia, namun masih mengalami kekurangan kajian yang lebih fokus dan spesifik pada faktor kapasiti intrinsik yang mempengaruhi penuaan sihat di Malaysia. Oleh itu, kajian ini adalah bertujuan untuk mengenal pasti faktor-faktor utama dari segi kapasiti intrinsik yang mempengaruhi penuaan sihat di Malaysia. Hasil kajian ini berpotensi membantu kerajaan dalam merangka hala tuju yang lebih jelas dalam membuat dasar untuk menyelesaikan isu penuaan sihat di Malaysia.

Di samping itu, peramalan terhadap penuaan sihat juga penting dalam mengaplikasikan rawatan awal terhadap golongan penuaan yang mempunyai risiko tinggi menjadi penuaan yang kurang sihat. Oleh itu, model ramalan terhadap penuaan sihat yang dibangunkan bukan sahaja dapat meramal situasi penuaan sihat yang akan berlaku pada masa depan maka dapat membuat persediaan awal sebelum isu penuaan sihat boleh menjadi isu serius yang perlu ditangani oleh negara.

Dengan mengatasi isu penuaan ini, kesihatan dalam golongan penuaan perlu dikawal dengan baik dan kawalan ini boleh dicapai sekiranya penggunaan pembelajaran mesin dalam pembinaan model ramalan terhadap kesihatan golongan penuaan. Pembelajaran mesin adalah proses pembelajaran komputasi melalui analisis data. Biasanya, semakin banyak data yang dilatih dengan algoritma pembelajaran mesin, pola keluaran yang lebih umum dapat dihasilkan. Untuk mengelakkan konsep "sampah masuk, sampah keluar", data biasanya dinilai dahulu oleh penyelidik dan mengenal pasti sama ada data itu sah dan layak untuk diperoleh (Kilkenny & Robinson 2018). Pembelajaran mesin telah memainkan peranan penting dalam meramalkan status kesihatan orang yang telah menua untuk memahami situasi kesihatan dengan cepat dan menemui risiko potensi dalam badan seseorang.

Oleh itu, kajian ini berusaha untuk mengembangkan potensi pembelajaran mesin dalam membina model ramalan terhadap kesihatan golongan penuaan. Dengan memahami secara mendalam ke dalam data yang berkaitan dengan aspek-aspek seperti demografi, fizikal kapasiti dan faktor penyakit, model ramalan yang dibina dapat memberikan ramalan kesihatan yang lebih tepat dan boleh dipercayai. Hal ini disebabkan oleh penuaan sihat bukan sahaja dipengaruhi oleh genetik variasi tetapi juga boleh diakibatkan oleh persekitaran fizikal dan sosial (WHO 2022). Dengan pendekatan ini, kita dapat memajukan langkah-langkah proaktif dan penyelidikan awal, menyumbang kepada kesejahteraan golongan penuaan dan memastikan sistem kesihatan yang berkesan.

1.3 OBJEKTIF KAJIAN

Objektif kajian adalah seperti berikut :

1. Mengenalpasti faktor-faktor penuaan sihat dengan ujian *chi-squared* dan pengiraan nilai p.
2. Menilai model ramalan yang mempunyai ketepatan tinggi dan paling sesuai digunakan untuk menerangkan situasi penuaan sihat.

1.4 SKOP KAJIAN

Kajian ini melibatkan isu penuaan sihat yang dibahagikan kepada tiga aspek, iaitu aspek sosio-demografi yang membawa 13 pembolehubah, aspek kesihatan fisiologi dan metabolik yang membawa 25 pembolehubah serta aspek kapasiti fizikal yang membawa 5 pembolehubah.

Dalam kajian ini, ujian *chi-squared* dan nilai p akan dikira supaya dapat mengenal pasti keertian statistik hubungan antara kesemua pembolehubah terhadap pembolehubah sasaran. Beberapa algoritma seperti Pohon Keputusan (DT), Mesin Vektor Sokongan (SVM), Naive Bayes (NB), Rangkaian Neural Tiruan (ANN), *K*-

Nearest Neighbours (KNN), Hutan Rawak (RF), *XgBoost* (XGB) akan digunakan untuk membangunkan model ramalan untuk mengklasifikasikan penuaan sihat.

Metrik ketepatan, kepersisan, ingatan, spesifisiti, skor F1 dan kawasan di bawah lengkung (AUC) akan dibandingkan di antara kesemua model yang dibangkit untuk mengenal pasti algoritma yang mana satu paling sesuai dalam pembinaan model ramalan. Kelebihan dan kekurangan juga akan dibincangkan untuk memastikan tahap keserasian tertinggi dengan keperluan.

1.5 CABARAN KESIHATAN KALANGAN PENUAAN

Peningkatan jangka hayat manusia yang secara beransur dan penurunan kadar kesuburan telah membawa kepada fenomena penuaan global. Di bawah pengaruh kemajuan teknologi, penuaan secara global telah menjadi salah satu arah aliran masa depan di seluruh dunia. Kebanyakan negara maju mengalami fenomena ini. Hal ini kerana negara yang sangat maju menunjukkan kadar penuaan penduduk yang lebih tinggi disebabkan oleh kemudahan perubatan yang lebih lengkap dan kadar kesuburan yang rendah (Minjae et al. 2024).

Penuaan penduduk menjadi semakin biasa di setiap negara dan menghadapi isu kesihatan yang serius. Hal ini desebabkan oleh seseorang yang telah menua biasanya mempunyai risiko yang lebih tinggi untuk menghidap penyakit akibat daripada penuaan organ dan penurunan fungsi badan. Oleh itu, isu penuaan telah menjadi salah satu faktor penurunan kadar produktiviti dalam negara dan memberi impak negatif kepada pertumbuhan ekonomi sesuatu negara. Kajian terkini menunjukkan bahawa pertumbuhan KDNK di Malaysia mengalami penurunan dengan peningkatan pergantungan penuaan. Hubungan penyebab ini berlaku dalam jangka pendek dan jangka panjang pertumbuhan ekonomi di Malaysia. Walau bagaimanapun, modal insan dan penyertaan buruh yang melibatkan sejumlah besar orang yang telah menua menunjukkan impak positif yang signifikan dalam mendorong pertumbuhan ekonomi (Siti et al. 2021). Tidak dapat dinafikan bahawa sumbangan penuaan kepada pertumbuhan ekonomi suatu negara adalah sangat penting.

1.6 ORGANISASI TESIS

Tesis tersebut adalah terdiri daripada lima bab iaitu, pendahuluan, kajian kepustakaan, metodologi kajian, hasil kajian dan perbincangan, serta kesimpulan dan cadangan.

Dalam Bab 1, konsep asas tentang penuaan dan isu penuaan telah diterangkan secara keseluruhan. Bab ini juga telah menerangkan pernyataan masalah dan objektif utama untuk menjalankan penyelidikan terhadap penuaan sihat tersebut.

Bab 2 merupakan bab kajian kepustakaan yang terutama menerangkan penyelidikan semasa dan lepas yang telah dijalankan pada isu penuaan. Selain itu, kajian lepas tentang domain penuaan yang menggunakan pembelajaran mesin terhadap penuaan sihat juga dibincangkan dalam bab tersebut.

Bab 3 dalam kajian ini adalah bab metodologi kajian yang menerangkan kaedah dan teknik yang telah digunakan dalam kajian tersebut. Antaranya seperti penggunaan konsep CRISP-DM dalam menangani set data penuaan, dan penerangan tentang model ramalan yang menggunakan pembelajaran mesin. Selain itu, teknik lain seperti pemilihan faktor dengan menggunakan ujian Chi-squared dan nilai p, serta SMOTE dalam menangani masalah data tidak seimbang juga telah dibincang dalam bab tersebut.

Bab 4 adalah bab hasil kajian dan perbincangan. Bab tersebut telah menunjukkan kesemua hasil yang telah dibuat dari segi hasil proses pembersihan data, statistic perihalan dan prestasi bagi setiap model ramalan. Perbandingan model dan perbincangan yang lebih terperinci terhadap hasil kerja juga telah dibuat dalam bab tersebut.

Kajian ini telah diakhiri oleh Bab 5. Bab ini telah membuat rumusan terhadap semua bab secara keseluruhan dan memberikan cadangan yang dapat memperbaiki kajian tersebut pada masa depan.

BAB II

KAJIAN KEPUSTAKAAN

2.1 PENGENALAN

Bab ini memainkan peranan penting sebagai kajian menyeluruh ke atas domain pengetahuan yang berkaitan. Melalui kajian kepustakaan yang menyeluruh, asas kukuh dapat dibina untuk memahami konsep utama, metodologi, dan dapatan yang menyumbang kepada pemahaman penuaan sihat. Dengan merujuk kepada kajian yang terdahulu, rangkaian teori dan bukti empirikal akan memberikan pandangan terhadap keadaan pengetahuan semasa mengenai faktor-faktor yang mempengaruhi penuaan, corak berkaitan kesihatan dan model ramalan yang digunakan dalam konteks serupa. Bab ini membentuk asas untuk penyelidikan semasa dan membuka jalan bagi sumbangan baru yang ingin dilakukan dalam kajian ini.

2.2 PENJELAJAHAN FAKTOR KESIHATAN UTAMA DALAM PENUAAN

Terdapat banyak kajian mengenai penuaan penduduk yang sihat yang telah dilakukan sebelum ini yang menunjukkan bahawa status penuaan sihat telah dipengaruhi oleh demografi, sosio-demografi, gaya hidup, dan faktor penyakit. Dalam bahagian demografi, umur dan jantina biasanya digunakan sebagai petunjuk yang penting untuk kesihatan seseorang. Kajian sebelum ini telah menunjukkan bahawa peningkatan umur pada seseorang akan menyebabkan penurunan status kesihatan, antaranya adalah penurunan fungsi penglihatan dan pendengaran, masa reaksi menjadi lambat dan imbangan terjejas (Yuan 2024). Selain itu, jantina juga merupakan salah satu faktor

penting yang mempengaruhi status kesihatan seseorang (Allison et al. 2021). Menurut kajian tersebut, perbezaan jantina bukan sahaja mempunyai faktor biologi dari segi struktur badan yang mempunyai kebarangkalian yang berbeza antara lelaki dan perempuan dalam menghidapi penyakit tertentu, tetapi juga mempunyai faktor sosial dari segi gaya kehidupan antara lelaki dan perempuan. Oleh itu, jantina adalah satu indeks yang penting dalam menyiasat status penuaan sihat.

Dari bidang pendidikan pula, terdapatnya kajian telah menunjukkan bahawa seseorang yang mempunyai tahap pendidikan tinggi lebih cenderung berada dalam status kesihatan yang tinggi dan kadar kematian yang rendah (Viju & Wullianallur 2020). Manakala pendapatan merupakan faktor paling penting di antara semua pembolehubah sosio-ekonomi yang mempengaruhi status kesihatan di akhir hayat (Lili et al. 2024).

Kesihatan peribadi boleh dipengaruhi oleh penyakit, dan kebanyakan penyakit disebabkan oleh gaya hidup mereka. Sebagai contoh, merokok adalah tabiat buruk dan merupakan salah satu faktor utama kanser paru-paru. Bukan sahaja tabiat buruk dapat menyebabkan penyakit, gaya hidup yang mempengaruhi perasaan juga boleh menyebabkan penyakit. Kajian tentang kumpulan penuaan di Afrika sub-Sahara menunjukkan bahawa penuaan yang tinggal seorang diri mempunyai risiko lebih tinggi dalam mengalami tekanan psikologi yang boleh memburukkan kesihatan mereka (Razak et al. 2020).

2.3 KONSEP MODEL DAN TEKNIK DATA MANIPULASI

2.3.1 Konsep model

Beberapa algoritma telah digunakan dalam penyelidikan ini untuk membandingkan ketepatan dan kebolehlaksanaan bagi setiap algoritma. Model yang dicadangkan pada langkah ini adalah seperti berikut: Pohon Keputusan (Harsh et al. 2018), Mesin Vektor Sokongan (Pisner et al. 2020), Naive Bayes (Ismail et al. 2020), Rangkaian Neural Tiruan (Thomas et al. 2020), K-Nearest Neighbours (Venkateswarlu & Rekha 2024), Hutan Rawak (Matthias & Rosie 2020), XgBoost (Younus et al. 2024).

Pohon Keputusan (DT) merupakan salah satu algoritma pembelajaran mesin yang banyak digunakan dalam tugas klasifikasi. Algoritma ini beroperasi dengan cara memecah data ke dalam keputusan yang lebih kecil atau daun berdasarkan serangkaian pertanyaan dan keadaan (Harsh et al. 2018). Pohon ini membentuk struktur hierarki yang sangat jelas dan memberikan pemahaman yang baik tentang operasi model dalam membuat keputusan. Pohon keputusan akan menyelesaikan masalah dengan menganalisis ciri-ciri data dan memilih cabang keputusan terbaik.

Mesin Vektor Sokongan (SVM) adalah algoritma yang biasanya digunakan dalam masalah klasifikasi. Tekniknya adalah menggunakan hipersatah optimum dengan margin maksimum untuk memisahkan atau mengklasifikasikan titik data pencerapan. Kelebihan SVM adalah untuk mencapai prestasi yang agak seimbang dan mencegah daripada klasifikasi yang berlebihan atau kurang sesuai malah dalam dimensi set data yang tinggi (Pisner et al. 2020).

Naive Bayes (NB) adalah algoritma klasifikasi kebarangkalian yang bergantung pada teorem Bayes. Model tersebut dianggap "naive" kerana dianggap bahawa pembolehubah dalam set data adalah tidak bersandar antara satu sama lain (Ismail et al. 2020). Walaupun anggapan ini jarang memenuhi syarat atau peraturan dalam dunia sebenar, NB masih dapat memberikan hasil keputusan yang baik dan mudah dilaksana. Hal ini memyebabkan NB sering digunakan dalam klasifikasi teks, analisis sentimen, dan kegunan lain di mana kelas kebarangkalian diperlukan.

Algoritma Rangkaian Neural Tiruan (ANN) pada dasarnya adalah idea untuk meniru prosedur dalam otak manusia. Konsep asasnya adalah menghasilkan keputusan yang diingini dengan menggelung ralat keluaran sebagai rujukan dalam gelung kemasukan seterusnya. Kelebihan algoritma ini adalah untuk mengendalikan hubungan bukan linear yang kompleks di antara pembolehubah dengan kaedah "cuba dan ralat" (Thomas et al. 2020).

K-Nearest Neighbours (KNN) adalah algoritma pembelajaran mesin berselia yang digunakan untuk klasifikasi berdasarkan metrik jarak. Model ini beroperasi dengan cara mencari sejumlah k tetangga yang terdekat dari suatu titik data dan kemudian mengambil kelas (Venkateswarlu & Rekha 2024). Keputusan kelas adalah penuh bergantung pada tetangga majoriti yang terdekat. Algoritma tersebut sangat

efektif untuk masalah klasifikasi yang melibatkan data yang tidak berstruktur dan tidak linear .

Regresi Hutan Rawak (RF) merupakan algoritma pembelajaran mesin ensembel yang terdiri daripada siri ramalan pohon yang tumbuh pada kedalaman maksimum masing-masing. Kelebihan algoritma ini adalah mampu menilai julat nilai ramalan yang lebih luas yang dapat mencegah daripada pemodelan data yang berlebihan (Matthias & Rosie 2020).

Mankala XgBoost juga merupakan salah satu teknik canggih algoritma pohon keputusan ensembel yang berfungsi untuk meningkatkan kecerunan supaya mengurangkan kecerunan kerugian ketika model sedang dilaraskan. Kelebihan algoritma ini adalah kecekapan yang tinggi dalam menyesuaikan atau meningkatkan kecerunan tepat pada masanya (Younus et al. 2024).

2.3.2 Teknik pemilihan faktor

Pemilihan pembolehubah adalah proses mengenal pasti dan memilih subset pembolehubah daripada set data yang memenuhi kriteria untuk membina model ramalan dengan ketepatan yang tinggi (Chowdhury & Tanvir 2020). Objektif utama pemilihan pembolehubah adalah untuk meningkatkan prestasi model dengan mengurangkan dimensi data, dan mengurangkan masalah *overfitting*.

Dalam kajian ini, teknik yang telah digunakan adalah ujian *Chi-squared* dan pengiraan nilai p. Ujian *Chi-squared* merupakan ujian statistik yang digunakan untuk menentukan sama ada terdapatnya berkaitan yang signifikan antara dua pembolehubah kategorikal (Donald 2019). Statistik *Chi-squared* X^2 merupakan jumlah perbezaan kuasa antara frekuensi yang diperhatikan dan frekuensi yang dijangka, dinormalisasi dengan frekuensi yang dijangka,

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad ...(2.1)$$

di mana O_i ialah frekuensi yang diperhatikan dan E_i ialah frekuensi yang dijangka untuk setiap sel.

Manakala nilai p adalah nilai statistik yang mengukur kekuatan bukti untuk menolak hipotesis sifar atau hipotesis nol dalam ujian statistik. Nilai p digunakan untuk mengenalpasti bahawa terdapatnya bukti yang mencukupi untuk menolak hipotesis sifar yang menyatakan bahawa terdapat tiada kesan atau hubungan dalam data (Daniel 2021). Dalam konteks ujian statistik, nilai p menunjukkan kebarangkalian untuk mendapatkan hasil yang sama atau lebih spesifik daripada yang diperoleh, sekiranya hipotesis sifar adalah benar. Dalam kajian ini, interpretasi nilai p adalah merujuk kepada kajian Zahra et al. (2021) seperti berikut:

Nilai $p < 0.05$: Terdapat bukti yang mencukupi untuk menolak hipotesis sifar. Keputusan dianggap signifikan secara statistik;

Nilai $p \geq 0.05$: Tidak cukup bukti untuk menolak hipotesis sifar. Keputusan dianggap tidak signifikan secara statistik.

2.3.3 Teknik SMOTE

Teknik SMOTE ini merupakan teknik penyampelan semula yang kerap digunakan terutamanya apabila berurusan dengan dataset yang tidak seimbang di mana satu kelas adalah jauh lebih kurang diwakili. Matlamat utama SMOTE adalah untuk menghasilkan contoh sintetik bagi kelas minoriti supaya dapat mecapai keseimbangan pengagihan kelas pada pembolehubah sasaran (Elreedy & Atiya 2019).

2.4 KAJIAN TERKINI YANG BERKAITAN DENGAN RAMALAN TERHADAP PENUAAN SIHAT

Terdapat satu kajian terkini di China yang menggunakan faktor-faktor tertentu untuk membangunkan model ramalan status kesihatan populasi penuaan dengan menggunakan algoritma pembelajaran mesin. Kajian telah membuktikan hasil yang cemerlang dengan menggunakan rangkaian neural tiruan untuk menerangkan hubungan linear dan bukan linear di antara faktor kombinasi dalam meramalkan status penuaan sihat dengan ketepatan setinggi 69.9% (Qin et al. 2020).

Satu kajian yang serupa dalam meramalkan penuaan juga berjaya dijalankan di bandar Ahvaz, Iran. Kajian ini menggunakan tujuh jenis algoritma pembelajaran mesin untuk mengkelaskan penuaan yang berjaya dan mencadangkan pengkelas Hutan Rawak sebagai model terbaik dengan ketepatan tertinggi iaitu setinggi 95% ketepatan dalam mengkelaskan penuaan yang berjaya (Maryam et al. 2022).

Beard et al. (2019) telah mencadangkan jenis ukuran yang berbeza untuk membahagikan golongan penuaan di England dengan menggunakan kapasiti intrinsik. Kapasiti-kapasiti ini secara umumnya terbahagi kepada lima jenis, iaitu lokomotor, kognitif, psikologi, deria, dan daya hidup. Kajian telah membuktikan model ketepatan tinggi dengan ralat purata kuasa dua yang sangat rendah iaitu 0.02 sahaja dalam nilai ramalan dengan kaedah ini.

Selain daripada faktor-faktor sebelum ini, penggunaan bioinformatik dalam meramalkan umur biologi adalah teknik baru untuk mengenal pasti status penuaan sihat. Terdapat satu kajian terkini di China yang berkaitan dengan penggunaan kaedah berdasarkan pembelajaran mesin dengan menggunakan pengiraan umur biologi dengan beberapa biomarker. Kaedah ini telah meningkatkan ketepatan secara ketara dalam mengkelaskan penuaan yang berjaya kerana umur kronologi pastinya tidak mencukupi untuk menerangkan status kesihatan seseorang yang telah menua. Keputusan akhir di mana nilai R kuasa dua yang serendah 0.27 adalah cukup meyakinkan dengan membandingkan beberapa algoritma pembelajaran mesin pada data untuk mencari algoritma yang paling sesuai dan lebih baik menerangkan status penuaan sihat (Xingqi et al. 2021).

2.5 KAJIAN LEPAS DI DALAM DOMAIN PENUAAN YANG MENGGUNAKAN PEMBELAJARAN MESIN

Kajian pertama ini telah melibatkan pembinaan model ramalan dalam mengklasifikasikan penuaan berjaya terhadap golongan penuaan di negara Iran dari Januari 2016 hingga Ogos 2021 (Zahra et al. 2022). Konsep penuaan berjaya dalam kajian ini adalah merujuk kepada golongan penuaan yang mematuhi tiga teori

dicadangkan oleh Rowe dan Kahn (1997) iaitu aktif kegiatan kehidupan, tiada penyakit atau kurang upaya serta berfungsi dalam fizikal dan kognitif.

Demi mencapai objektif kajian ini (Zahra et al. 2022), tiga faktor utama telah dipilih sebagai pembolehuhah yang akan digunakan untuk mengklasifikasikan penuaan berjaya. Jadual 2.1 bawah telah merujuk kesemua pembolehubah yang digunakan dalam kajian tersebut.

Jadual 2.1 Pembolehubah yang digunakan di bawah faktor utama

Faktor	**Pembolehubah**
Sosiodemografik	Umur, jantina, tahap pendidikan, status perkahwinan, pekerjaan, tahap pendapatan dan situasi insurans.
Klinikal	Darah tinggi, kemalangan kardiovaskular, penyakit tulang, penyakit buah pinggang, penyakit hati, penyakit otot, tahap kemurungan, tahap pemulihan, penyakit mata, kencing manis, kanser dan penyakit lain.
Tingkah laku dan psikososial	Aktiviti kehidupan harian (ADL), kepuasan hidup, kualiti kehidupan, kesihatan gaya hidup, hubungan sosial dan interpersonal, gaya pemakanan, aktiviti fizikal, aktiviti pencegahan penyakit, dan pengurusan tekanan

Dalam kajian ini, teknik SMOTE telah digunakan dalam pemprosesan untuk menyelesaikan masalah data yang tidak seimbang. Proses pemilihan pembolehubah juga dijalankan dengan mengaplikasikan pengiraan *Chi-squared* dan nilai p. Proses ini adalah penting untuk menghapuskan data yang berlebihan dan tidak berkaitan. Kajian tersebut telah menetapkan pembolehubah yang mempunyai nilai p yang kurang daripada 0.05 sebagai statistik signifikan. Oleh itu, dalam kajian ini hanya 21 pembolehubah yang layak dapat dimasuki dalam proses pembinaan model ramalan.

Untuk mendapat ramalan yang lebih tepat, terdapatnya lima model telah dibina untuk membandingkan hasil yang terakhir. Antara model yang digunakan Dalam kajian ini adalah ANN, DT, NB, SVM dan KNN. Satu pembelajaran mesin ensembel juga dibina dengan menggabungkan beberapa model asas sebagai pelajar lemah. Dalam model tersebut, 30 KNN telah dipilih sebagai pelajar lemah yang menjadi model asas.

Penilaian bagi setiap model adalah diuji oleh beberapa metrik prestasi iaitu ketepatan, kepersisan, spesifisiti, ingatan dan skor F1. Demi mendapatkan hasil yang seimbang dan adil, teknik persilangan telah digunakan dalam mendapat purata bagi setiap metrik.

Hasil kajian telah menunjukkan bahawa model ANN mempunyai prestasi yang paling rendah manakala model ensembel yang berdasarkan KNN telah mencapai prestasi yang paling tinggi, iaitu sebanyak 93% ketepatan.

Kajian seterusnya menggunakan 975 rekod penuaan dari Januari 2019 ke Januari 2021 yang dikumpul dari pusat Hamdeli, Mehrjooyan dan Hasti di bandar Ahvaz (Ahmadi et al. 2023). Antaranya 751 rekod dan 224 rekod adalah diklasifikasikan sebagai penuaan tidak berjaya dan penuaan berjaya masing-masing. Dalam data tersebut, lelaki telah direkod seramai 515 orang dan perempuan seramai 460 orang.

Ahmadi et al. (2023) lebih fokus untuk mengklasifikasikan penuaan berjaya berdasarkan kehidupan sosial dan kesihatan mental penuaan. Antara pembolehubah yang digunakan dalam kajian ini adalah berdasarkan beberapa bahagian iaitu perasaan terhadap keadaan kesihatan sendiri, huhungan sosial secara rasmi, huhungan sosial secara tidak rasmi, kualiti kehidupan, kebergantungan individu, skala kepuasan terhadap kehidupan dan gaya kehidupan.

Kehilangan data yang melebihi 5% bagi setiap barisan data telah dikeluarkan dan bagi kehilangan nilai yang kurang daripada 5% telah dijangka oleh KNN algorithma dengan spesifik K=1, 3, 5. Proses pemilihan pembolehubah dalam kajian tersebut adalah memilih pembolehubah yang mempunyai nilai p kurang daripada 0.05. Pembolehubah yang telah dipilih adalah statistik signifikan terhadap pembolehubah keluaran.

Seterusnya, tujuh model ramalan telah dibina untuk mengklasifikasikan penuaan berjaya, antaranya ialah RF, *Ada-boost*, J-48, ANN, *XG-Boost*, SVM dan juga NB. Demi memastikan keputusan yang adil dalam kajian tersebut, persilangan 10-lipat telah digunakan dalam setiap algorithma.

Metrik yang telah digunakan menguji prestasi model ramalan dalam kajian ini adalah nilai ramalan positif, nilai ramalan negatif, ketepatan, kepersisan, spesifisiti, skor

F1 serta AUC. Keputusan akhir bagi kajian tersebut telah menunjukkan bahawa model RF mempunyai prestasi yang terbaik secara keseluruhan, di mana ketepatannya setinggi 97.05%.

Qin et al. (2020) telah melibatkan data yang agak besar dan merangkumi bidang yang berkenaan dengan sosiodemografik dan sosioekonomik untuk meramal status kesihatan dalam golongan penuaan. Data yang telah digunakan dalam kajian ini adalah data yang diekstrak 2013 dan 2015 China Health and *Retirement Longitudinal Surveys* (CHARLS). Jumlah akhir saiz sampel dalam kajian tersebut adalah sebanyak 29377 selepas data yang mempunyai kehilangan data yang melebihi 80% dikeluarkan.

Aspek yang terlibat dalam data ini telah merangkumi latar belakang demografi, keluarga, status dan fungsi kesihatan, penjagaan kesihatan dan insurans, status pekerjaan, pendapatan, pembelanjaan dan aset serta ciri-ciri perumahan. Disebabkan oleh data tersebut mempunyai pembolehubah yang lebih daripada 3500 dan data ini akan sukar diproses dalam pembinaan model ramalan. Teknik *Maximal Information Coefficient* (MIS) dan *Pearson Correlation Coefficient* (PCC) telah digunakan untuk memilih pembolehubah yang paling relevan (Aldaz et al. 2015). MIS telah memainkan peranan penting untuk mengukur kekuatan korelasi antara pembolehubah linear dan tidak linear. Manakala PCC adalah penting dalam mengukur korelasi antara dua pembolehubah linear. Akhirnya, 15 pembolehubah telah dipilih dalam kajian ini sebagai masukan dalam model latihan.

Dalam kajian ini, teknik persilangan 5-lipat telah digunakan untuk mendapat purata keputusan. Untuk penilaian prestasi model, dua metrik telah digunakan iaitu ketepatan bagi model klasifikasi dan MSE bagi model regresi. MSE merupakan teknik untuk mencari perbezaan nilai antara nilai sebenar dan nilai ramalan dan MSE ini adalah penting dalam menguji prestasi model ramalan regresi.

Kajian ini telah melibatkan dua jenis model ramalan iaitu, model klasifikasi dan model regresi. Bagi model klasifikasi, *Artificial Neural Network* (ANN), *Logistics Regression* (LR), *Support Vector Machine* (SVM), *XGBoost Classifier* (XGB) dan *Random Forest Classifier* (RF) telah digunakan. Manakala *Linear Support Vector Regression*, *Linear Regression*, *Random Forest Regressor* (RF), *XGBoost Regressor* (XGB) dan *Artificial Neural Network* (ANN) telah digunakan untuk model regresi.

Hasil kajian ini telah menunjukkan bahawa model ANN mempunyai prestasi yang paling tinggi antara kesemua model klasifikasi, di mana ketepatan telah mencapai 69.9%. Untuk model regresi dengan penggunaan pembolehubah skala ADL dan IADL, *Linear Support Vector Regression* mempunyai nilai MSE yang paling kecil manakala model regresi dengan penggunaan pembolehubah status kesihatan, Linear Regression mempunyai nilai MSE yang paling kecil.

Kajian seterusnya telah membuat ramalan terhadap golongan penuaan yang menghadapi keadaan kelemahan (Tarekegn et al. 2020). Set data yang digunakan dalam kajian ini merangkumi 1,095,612 subjek dan 64 pembolehubah. Set data tersebut seterusnya dipecahkan kepada enam bahagian, iaitu kematian, kecacatan, kemasukan hospital segera, patah tulang, kemasukan hospital yang boleh dicegah dan kemasukan jabatan kecemasan dalam kod merah.

Pembolehubah yang telah digunakan dalam kajian ini adalah berkaitan dengan faktor klinikal dan sosioekonomik yang mempengaruhi enam pembolehubah sasaran dalam keenam-enam bahagian data. Ujian *Chi-squared* dan nilai p telah dikira untuk setiap pembolehuhah dan hanya nilai p yang kurang daripada 0.05 telah dipilih untuk proses latihan. Teknik pensampelan semula telah digunakan dalam kajian tersebut untuk menyelesaikan masalah ketidakseimbangan data.

Empat metrik telah digunakan dalam kajian ini bagi menganalisis prestasi model. Antaranya ialah ketepatan, kepersisan, spesifisiti dan ingatan. Manakala model yang digunakan dalam kajian ini ialah ANN, SVM, RF, LR, DT dan genetic programming (GP). GP merupakan satu jenis algorithma evolusi dengan mengembangkan penyelesaian terhadap masalah yang kompleks. Teorinya adalah menghasilkan satu populasi program computer melalui banyak generasi sampai kriteria yang telah ditetapkan dan membuat keputusan yang terbaik melalui pilihan semulajadi. Teknik persilangan 10-lipat juga digunakan untuk menghasilkan keputusan yang adil.

Tarekegn et al. (2020) telah menunjukkan bahawa tiada satu model yang terbaik untuk meramal semua pembolehubah sasaran. Kuasa ramalan bagi setiap model adalah berbeza antara masalah. Dalam kajian tersebut, didapati bahawa set data yang digunakan ini lebih sesuai untuk meramal pembolehuhah sasaran kematian. Hal ini kerana, ketepatan secara keseluruhan dalam setiap model ramalan dibina terhadap

kematian adalah tertinggi antara keenam-enam pembolehubah sasaran, di mana ANN dan SVM mencapai ketepatan setinggi 78% dan 79% masing-masing.

2.5.1 Rumusan kajian lepas

Jadual 2.2 di bawah telah menunjukkan model yang telah digunakan dan hasil telah didapat oleh pengarang setiap kajian kes.

Jadual 2.2 Rumusan Kajian Kes

Jurnal	Model	Hasil
Developing a prediction model for successful aging among the elderly using machine learning algorithms (Ahmadi et al. 2023)	AB, XGB J-48, RF, ANN, SVM, NB	Model RF mencapai prestasi yang terbaik, di mana ketepatan setinggi 97.05%. Kajian ini telah berjaya membina model dalam meramal penuaan berjaya supaya dapat meningkatkan kualiti hidup penuaan.
Prediction of successful aging using ensemble machine learning algorithms (Zahra et al. 2022)	ANN, DT, SVM, NB, KNN, Ensembel-KNN	Model Ensembel-KNN mencapai prestasi yang terbaik, di mana ketepatan setinggi 89.62%. Kajian ini dapat menjadi sebagai panduan kepada kerajaan dalam mencadangkan dasar pada golongan penuaan.
Health status prediction for the elderly based on machine learning (Qin et al. 2020)	ANN, LR, SVM, XGB, RF	Model ANN mecapai prestasi yang terbaik, di mana ketepatan setinggi 69.9%. Peruntukan sumber penjagaan sosial yang terhad dapat dijalankan dengan lebih adil melalui hasil kajian tersebut.
Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches (Tarekegn et al. 2020)	ANN, GP, SVM, RF, LR, DT	Model ANN dan SVM mencapai prestasi yang lebih baik berbanding dengan model lain, di mana ketepatan setinggi 78% dan 79% masing-masing. Kajian ini telah membuktikan bahawa ketepatan ramalan prestasi model adalah berbeza secara signifikan dari segi masalah dan juga konteks metrik penilaian yang berbeza

2.6 KESIMPULAN

Secara kesimpulan, penjagaan kesihatan kepada orang yang telah menua adalah penting untuk memastikan tahap kesihatan masyarakat dan mengekalkan produktiviti penuaan. Sistem penjagaan kesihatan yang berkesan dipercayai dapat meningkatkan jumlah penduduk penuaan yang sihat.

Penggunaan ujian *chi-squared* dan nilai p dalam kajian ini adalah sangat penting untuk mengenalpasti kaitan yang signifikan secara statistik antara pembolehubah kategorikal. Hal ini kerana terdapat banyak faktor yang boleh mempengaruhi kesihatan orang yang telah menua secara langsung atau tidak langsung. Dengan memberikan perhatian terhadap faktor tersebut adalah penting untuk menentukan punca utama yang memanipulasi kesihatan orang yang telah menua. Dengan membuat perubahan pada punca utama mungkin dapat memastikan kesihatan masyarakat penuaan sentiasa berada pada tahap yang tinggi.

Selain itu, model ramalan mengenai penuaan sihat juga sangat penting untuk dibangunkan bagi meramalkan status kesihatan seseorang yang telah menua. Dengan penggunaan 7 model ramalan dalam kajian ini, perbandingan prestasi antara model ramalan dapat dinilai dan model ramalan yang mempunyai prestasi yang tertinggi dapat dicadangkan demi menerangkan situasi penuaan sihat di Malaysia. Bukan itu sahaja ketepatan tinggi model ramalan dapat membantu profesional perubatan untuk membuat penilaian terhadap orang yang telah menua yang berpotensi masuk dalam kategori penuaan yang tidak sihat. Oleh itu, seseorang yang telah menua boleh mendapat rawatan awal sebelum status kesihatan mereka merosot dengan kos perubatan yang lebih rendah. Langkah ini secara tidak langsung membantu kerajaan melepaskan sebahagian daripada tekanan kewangan negara.

Satu model ramalan yang berkesan adalah penting untuk mengenal pasti status kesihatan orang yang telah menua dan hasil ini akan memberikan bantuan besar kepada profesion untuk menilai kesihatan seseorang dengan lebih berkesan dan dapat memberikan rawatan lebih awal jika terdapat skor risiko kesihatan yang tinggi.

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

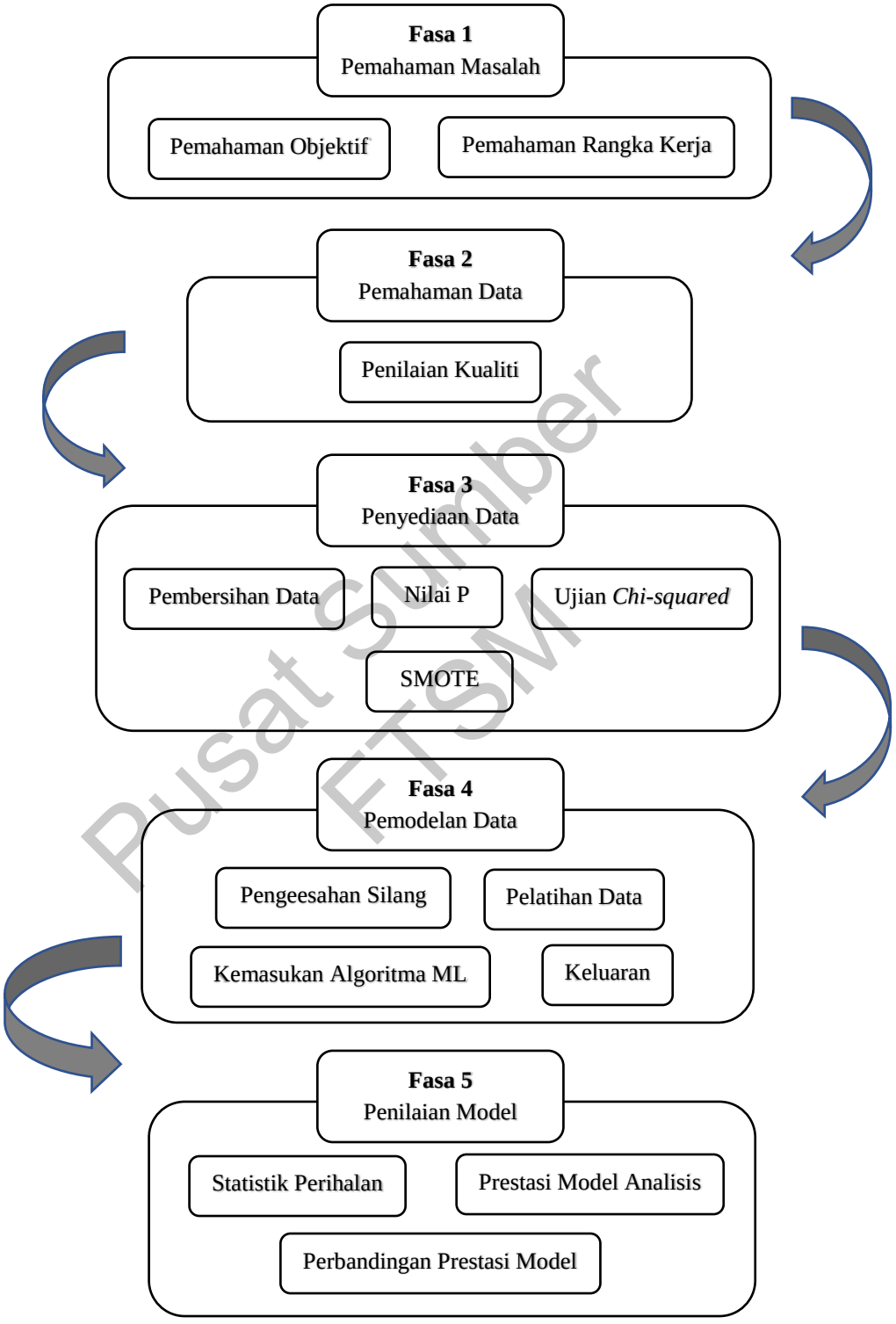
Bab ini telah menunjukkan langkah-langkah asas metodologi sains data dan memberikan panduan untuk melaksanakan data analisis yang melibatkan data pelbagai dimensi dan rumit. Dengan menerapkan pendekatan yang disiplin dan sistematik, metodologi ini bertujuan untuk mendedahkan naratif tersembunyi dalam data, membolehkan pemahaman yang lebih mendalam tentang fenomena yang sedang diselidik.

3.2 PROSES PEMAHAMAN PERSOALAN KAJIAN

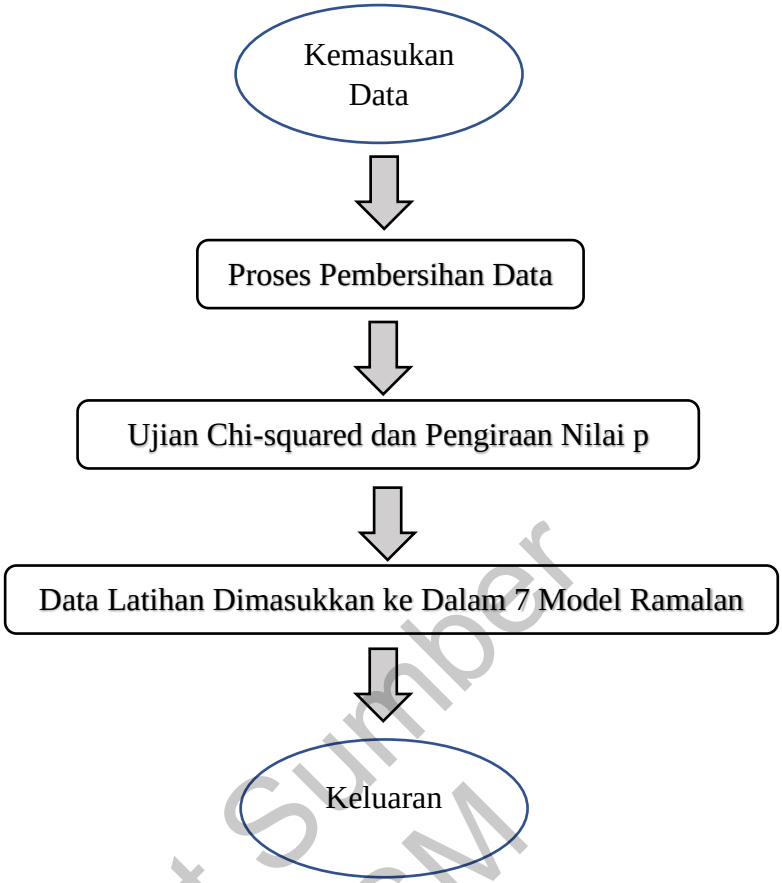
Metodologi yang digunakan dalam kajian ini adalah langkah asasi dalam prinsip sains data yang mengikuti pendekatan sistematik iaitu termasuknya pemahaman masalah, pemahaman data, penyediaan data, pemodelan, penilaian, dan pengegunaan. Setiap peringkat dalam rangka kerja komprehensif ini memberikan sumbangan unik kepada objektif keseluruhan untuk mendapat maklumat yang bermakna dari set data yang kompleks dan menyumbang dalam membuat keputusan yang berinformasi dan menganalisis sifat kebenaran data serta model ramalan.

Dalam kajian ini, proses penyelidikan telah dibahagikan kepada 5 fasa, iaitu pemahaman masalah, pemahaman data, penyediaan data, pemodelan data dan penilaian

model. Idea utama aliran kerja ini ditunjukkan dalam Rajah 3.1. Setiap peringkat telah diterangkan secara teliti dan jelas.



Rajah 3.1 Idea utama aliran kerja



Rajah 3.2 Rangka kerja umum dalam data manipulasi

Rajah 3.2 telah menunjukkan rangka kerja umum dalam data manipulasi, di mana data mentah akan menjalankan proses pembersihan data supaya mendapat data yang jelas dan meninkatkan keberkesanan semasa dimasukkan ke dalam proses data latihan. Seterusnya, ujian chi-squared dan pengiraan nilai p akan dijalankan untuk memilih pembolehubah yang signifikan secara statistik. Data ini kemudian akan dipecahkan kepada data latihan dan data ujian untuk menguji prestasi bagi setiap model ramalan.

3.2.1 Pemahaman Masalah

Sebelum projek ini dimulakan, objektif yang jelas perlu dikenal pasti dahulu bagi menyiapkan projek ini. Dalam projek ini, tujuan pertama adalah mengenal pasti ciri utama yang mendominasi kesihatan seseorang yang telah menua. Kaedah yang digunakan untuk mengatasi masalah ini adalah melibatkan ujian *Chi-squared* dan pengiraan nilai p. Ujian *Chi-squared* digunakan untuk menilai hubungan di antara

pembolehubah-pembolehubah kategorikal, sementara pengiraan nilai p membantu mengenal pasti keertian statistik hubungan tersebut. Selain itu, objektif utama projek ini adalah untuk mengklasifikasikan penuaan yang sihat dan mencapai tahap ketepatan yang tinggi. Masalah ini telah dilselesaikan dengan mengaplikasikan 7 model dan mencadangkan ramalan model yang paling sesuai dan model ini dapat menggambarkan corak sebenarnya pada kumpulan penuaan yang sihat dalam kajian ini.

3.2.2 Pemahaman Data

Dalam peringkat pemahaman data, satu siri set data mentah yang mengandungi maklumat tentang penuaan di Malaysia telah difahami melalui pelbagai aspek kapasiti intrinsik. Selain daripada pembolehubah sosiodemografi para peserta, pembolehubah termasuk penyakit, kapasiti fizikal, keupayaan kognitif, dan kesejahteraan psikologi untuk setiap individu penua juga dipilih dan dianalisis. Dalam peringkat ini, set data mentah telah dipratinjau supaya data yang dipilih adalah relevan dan konsisten serta kualiti set data mentah telah dinilai secara keseluruhan.

Jadual 3.1 telah menunjukkan pembolehubah asal dari data mentah dan telah menggambarkan nama bagi setiap pembolehubah yang tidak kemas serta tidak konsisten dalam bahasa yang digunakan dan keadaan ini akan susah semasa menjalankan proses pembersihan data. Oleh itu, pembolehubah telah dinamakan semula dalam bahasa Inggeris supaya pembolehubah yang akan digunakan pengekodan lebih mudah dijalankan nanti. Penerangan bagi setiap pembolehubah yang telah digunakan dalam kajian ini juga telah diterangkan dalam Jadual 3.1.

Jadual 3.1 Penerangan terhadap setiap pembolehubah

Pembolehubah Asal	Pembolehubah Baru	Penerangan
a2_4: (Negeri)	State	Negeri yang responden berasal dari
a5: (Jantina)	Gender	Jantina responden
age categories (5 years)	Age_category	Kategori umur di mana responden berada
a6: (Bangsa)	Ethnicity	Bangsa responden

bersambung...

...sambungan

a7: (Agama)	Religion	Agama responden
b1_1_1: (Status perkahwinan sekarang)	Marital_status	Status perkahwinan sekarang responden
b1_2_1_2: (Tahap pendidikan tertinggi)	Highest_education	Tahap pendidikan tertinggi yang diterima oleh responden
b1_5: (Adakah anda)	Smoker	Status merokok responden
a1: (Employment status)	Employment_status	Status pekerjaan responden
a2_3: (Job sector)	Job_sector	Sektor pekerjaan responden
a3_2: (Job category previously)	Job_category	Kategori pekerjaan responden
a5_1: (Total monthly income (Main))/	Total_monthly_main_income	Jumlah pendapatan utama bulanan responden
a5_2: (Total monthly income (Side))	Total_monthly_side_income	Jumlah pendapatan sampingan responden
Agegrpcat(SA UA MCI)	Ageing_group	Kategori penuaan sihat
cf_status	Cf_status	Status kognitif responden
b1_8_1: (Adakah anda sedang mengalami Tekanan darah tinggi)	Hypertension	Status tekanan darah tinggi responden
b1_8_2: (Adakah anda sedang mengalami Tinggi kolesterol)	High_cholesterol	Status tinggi kolesterol responden
b1_8_3: (Adakah anda sedang mengalami Kencing manis)	Diabetes	Status kencing manis responden
b1_8_4: (Adakah anda sedang mengalami Angin ahmar)	Stroke	Status angin ahmar responden
b1_8_5: (Adakah anda sedang mengalami Sakit sendi)	Arthritis	Status sakit sendi responden
b1_8_6: (Adakah anda sedang mengalami Penyakit jantung)	Heart_disease	Status penyakit jantung responden
b1_8_7: (Adakah anda sedang mengalami Katarak/ Glaucoma)	Cataract/Glaucoma	Status katarak/glaucoma responden

bersambung...

...sambungan

b1_8_8: (Adakah anda sedang mengalami Kegagalan buah pinggang)	Kidney_disease	Status kegagalan buah pinggang responden
b1_8_9: (Adakah anda sedang mengalami Lelah)	Asthma	Status lelah responden
b1810: (Adakah anda sedang mengalami Penyakit paru-paru kronik)	Respiratory_ailments	Status penyakit paru-paru kronik responden
b1811: (Adakah anda sedang mengalami Batuk kering)	Tuberculosis	Status batuk kering responden
b1812: (Adakah anda sedang mengalami Gout)	Gout	Status gout responden
b1813: (Adakah anda sedang mengalami Keretakan tulang pinggul)	Hip_fracture	Status keretakan tulang pinggul responden
b1814: (Adakah anda sedang mengalami masalah Sembelit)	Constipation	Status sembelit responden
b1815: (Adakah anda sedang mengalami Buasir)	Haemorrhoids	Status buasir responden
b1816: (Adakah anda sedang mengalami Gastrik atau ulser)	Gastric/Ulcer	Status gastrik/ulser responden
b1817: (Adakah anda sedang mengalami Masalah kelenjar tiroid)	Thyroid_gland_disease	Status kelenjar tiroid responden
b1820: (Adakah anda sedang mengalami Masalah kencing)	Urinary_incontinence	Status masalah kencing reponden
b1821: (Adakah anda sedang mengalami Masalah penglihatan dan pendengaran yang serius)	Vision/Hearing_problem	Status penglihatan dan pendengaran yang serius responden
b1822: (Adakah anda sedang mengalami Kesukaran mengunyah)	Chewing_problem	Status penyakit kesukaran mengunyah responden
b1823: (Adakah anda sedang mengalami Kurang selera makan)	Low_appetite	Status kurang selera makan responden

bersambung...

...sambungan

b1818_1o: (Sebutkan Kanser 1)	Cancer_1	Kanser yang dihidapi oleh responden
b1818_2o: (Sebutkan Kanser 2)	Cancer_2	Kanser kedua selain daripada pertama yang dihidapi oleh responden
b1819_1o: (Penyakit-penyakit lain 1)	Other_disease_1	Penyakit lain yang dihidapi oleh responden
b1819_2o: (Penyakit-penyakit lain 2)	Other_disease_2	Penyakit lain yang dihidapi oleh responden
ukmd8_ts	IADL_score	Jumlah markah aktiviti peralatan kehidupan harian diperolehi oleh responden
ADL_M: (ADL total marks)	ADL_score	Jumlah markah aktiviti kehidupan harian diperolehi oleh responden
d7_ts: (Jumlah Markah : SKALA KEMURUNGAN GERIATRIK)	Geriatric_depression_score	Jumlah markah bagi skala kemurungan geriatric diperolehi oleh responden
c4_1: (Eysenck Personality Questionnaire (EPQ) : Does your mood often go up and down?)	EPQ_Q1	Soal selidik EPQ bahawa sama ada perasaan sering naik dan turun pada resonden
c4_2: (Eysenck Personality Questionnaire (EPQ) : Do you ever feel ?just miserable? for no reason?)	EPQ_Q2	Soal selidik EPQ bahawa sama ada mempunyai perasaan menderita tanpa sebarang sebab pada responden
c4_3: (Eysenck Personality Questionnaire (EPQ) : Are you an irritable person?)	EPQ_Q3	Soal selidik EPQ bahawa sama ada responden merupakan seorang yang mudah marah
c4_4: (Eysenck Personality Questionnaire (EPQ) : Are your feelings easily hurt?)	EPQ_Q4	Soal selidik EPQ bahawa sama ada responden mudah disakiti hat
c4_5: (Eysenck Personality Questionnaire (EPQ) : Do you often feel ?fed-up??)	EPQ_Q5	Soal selidik EPQ bahawa sama ada responden selalu rasa penat

bersambung...

...sambungan

c4_6: (Eysenck Personality Questionnaire (EPQ) : Would you call yourself a nervous person?)	EPQ_Q6	Soal selidik EPQ bahawa sama ada responden rasa diri merupakan orang yang gementar
c4_7: (Eysenck Personality Questionnaire (EPQ) : Are you a worrier?)	EPQ_Q7	Soal selidik EPQ bahawa sama ada responden merupakan orang yang gelisah
c4_8: (Eysenck Personality Questionnaire (EPQ) : Would you call yourself tense or ?highly strung??)	EPQ_Q8	Soal selidik EPQ bahawa sama ada responden menyebut sendiri sebagai tegang
c4_9: (Eysenck Personality Questionnaire (EPQ) : Do you worry too long after an embarrassing experience?)	EPQ_Q9	Soal selidik EPQ bahawa sama ada responden rasa bimbang sepanjang masa selepas menemui pengalaman yang memalukan
c4_10: (Eysenck Personality Questionnaire (EPQ) : Do you suffer from ?nerves??)	EPQ_Q10	Soal selidik EPQ bahawa sama ada responden mempunyai masalah saraf
c4_11: (Eysenck Personality Questionnaire (EPQ) : Do you often feel lonely?)	EPQ_Q11	Soal selidik EPQ bahawa sama ada responden selalu rasa kesendirian
c4_12: (Eysenck Personality Questionnaire (EPQ) : Are you often troubled about feelings of guilt?)	EPQ_Q12	Soal selidik EPQ bahawa sama ada responden selalu berasa terganggu pada perasaan bersalah
c5_1: (Loneliness : How often do you feel that you lack companionship?)	Loneliness_Q1	Soal selidik tentang kesendirian bahawa kekerapan responden rasa sendiri jarang ditemani
c5_2: (Loneliness : How often do you feel left out?)	Loneliness_Q2	Soal selidik tentang kesendirian bahawa kekerapan responden rasa sendiri ditinggalkan
c5_3: (Loneliness : How often do you feel isolated from others?)	Loneliness_Q3	Soal selidik tentang kesendirian bahawa kekerapan responden rasa terasing daripada orang lain

3.2.3 Penyediaan Data

Setelah semua maklumat dapat difahami pada peringkat sebelumnya, beberapa langkah prapemprosesan telah dijalankan untuk memastikan set data bersih dan bersedia untuk diproses oleh pembelajaran mesin. Pembersihan data telah menjadi langkah pertama untuk memastikan kebersihan data dan data adalah bermaklumat. Nilai yang hilang dalam set data akan digantikan dengan min, mod atau nilai yang paling berkemungkinan mengikut keadaan. Seterusnya, kesahihan nilai telah diperiksa supaya nilai dalam data adalah berkonsisten. Oleh itu, nilai yang tidak sah akan digantikan dengan nilai yang berkemungkinan mengikut keadaan atau maklumat yang tidak sah akan dipadam.

Sebelum pembersihan data dijalankan, situasi dan kualiti bagi data mentah telah diperiksa dahulu supaya dapat mengetahui masalah secara keseluruhan. Kod yang digunakan untuk memeriksa data secara keseluruhan adalah seperti berikut:

```
df.info()
```

Rajah 3.3 Kod digunakan untuk memeriksa status data


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2329 entries, 0 to 2328
Data columns (total 58 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   State                     2322 non-null   object
 1   Gender                    2322 non-null   object
 2   Age_category              2322 non-null   object
 3   Ethnicity                 2322 non-null   object
 4   Religion                  2322 non-null   object
 5   Marital_status            2322 non-null   object
 6   Highest_education         2322 non-null   object
 7   Smoker                    2322 non-null   object
 8   Employment_status         2275 non-null   object
 9   Job_sector                532 non-null    object
 10  Job_category              1697 non-null   object
 11  Total_monthly_main_income 2275 non-null   float64
 12  Total_monthly_side_income 2268 non-null   float64
 13  Ageing_group              2322 non-null   object
 14  Cf_status                 815 non-null    object
 15  Hypertension              2322 non-null   object
 16  High_cholesterol          2322 non-null   object
 17  Diabetes                  2322 non-null   object
 18  Stroke                    2322 non-null   object
 19  Arthritis                 2322 non-null   object
 20  Heart_disease             2322 non-null   object
 21  Cataract/Glaucoma         2322 non-null   object
```

Rajah 3.4 Situasi data mentah secara keseluruhan

Rajah 3.4 telah menunjukkan bahawa data mentah tersebut mempunyai 58 pembolehubah dan 2329 baris data. Namun kebanyakan pembolehubah seperti “State”, “Gender”, “Age_category dan sebagainya hanya muncul 2322 baris data. Oleh itu, 7 baris data yang hilang telah dianggap membawa nilai nol secara keseluruhan. Demi memeriksa nilai nol secara lebih spesifik, kod yang berikut telah dimasukkan:

```
rows_with_null = df[df['State'].isnull()]
print(rows_with_null)
```

Rajah 3.5 Kod digunakan untuk mencari baris yang mengandungi nilai nol bagi setiap pembolehubah

	State	Gender	Age_category	Ethnicity	Religion	Marital_status	\
285	NaN	NaN	NaN	NaN	NaN	NaN	
483	NaN	NaN	NaN	NaN	NaN	NaN	
484	NaN	NaN	NaN	NaN	NaN	NaN	
485	NaN	NaN	NaN	NaN	NaN	NaN	
1217	NaN	NaN	NaN	NaN	NaN	NaN	
1623	NaN	NaN	NaN	NaN	NaN	NaN	
1812	NaN	NaN	NaN	NaN	NaN	NaN	

	Highest_education	Smoker	Employment_status	Job_sector	...	EPQ_Q6	\
285	NaN	NaN	NaN	NaN	...	NaN	
483	NaN	NaN	NaN	NaN	...	NaN	
484	NaN	NaN	NaN	NaN	...	NaN	
485	NaN	NaN	NaN	NaN	...	NaN	
1217	NaN	NaN	NaN	NaN	...	NaN	
1623	NaN	NaN	NaN	NaN	...	NaN	
1812	NaN	NaN	NaN	NaN	...	NaN	

	EPQ_Q7	EPQ_Q8	EPQ_Q9	EPQ_Q10	EPQ_Q11	EPQ_Q12	Loneliness_Q1	\
285	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
483	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
484	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
485	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1217	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1623	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1812	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Rajah 3.6 Nilai “NaN” dalam setiap pembolehubah

Hasil yang didapat adalah seperti Rajah 3.6 telah membuktikan bahawa 7 baris tersebut adalah tidak sah disebabkan kesemua nilai adalah nol dan telah dihapuskan dengan menggunakan kod seperti di bawah:

```
new_df = df.dropna(subset=['State'])
```

Rajah 3.7 Kod digunakan untuk menghapuskan baris mengandungi kesemua nilai nol

Seterusnya, setiap pembolehubah telah disemak dan dibersihkan satu persatu supaya dapat sedia untuk dimasukkan ke dalam model ramalan nanti.

```
Raw dataset value counts for
'Age_category':
Age_category
60-69    1329
70-79     853
80-89     136
90-99       3
11          1
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Age_category':
Age_category
60-69    1330
70-79     853
80-89     136
90-99       3
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.8 Pembersihan terhadap pembolehubah "Age_category"

Rajah 3.8 telah menunjukkan bahawa satu nilai "11" pada pembolehubah "Age_category" dalam data mentah dan nilai adalah dikira sebagai outlier disebabkan nilai ini tidak selaras dengan nilai yang sepatutnya bersifat julat seperti "60-69" atau sebagainya. Oleh itu, nilai "11" tersebut telah diganti dengan mod bagi pembolehubah ini dengan menggunakan kod di bawah:

```
age_mode_value = new_df['Age_category'].mode().iloc[0]
new_df['Age_category'] = new_df['Age_category'].replace(11,
age_mode_value)
```

Rajah 3.9 Kod digunakan untuk membersihkan pembolehubah 'Age_category'

```
Raw dataset value counts for
'Employment_status':
Employment_status
Retired
1309
Self employed
281
Housewife
257
Not working
177
Employed full time
128
Employed part time
94
NaN
47
999
18
Help family members
8
others
2
Students, trainiees or vocational
school       1
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Employment_status':
Employment_status
Retired
1309
Self employed
281
Housewife
257
Not working
177
Employed full time
128
Employed part time
94
Not stated
65
Help family members
8
others
2
Students, trainiees or vocational
school       1
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.10 Pembersihan terhadap pembolehubah "Employment_status"

Rajah 3.10 dan Rajah 3.12 telah menunjukkan tiga pembolehubah yang menerangkan pekerjaan bagi setiap individu iaitu 'Employment_status', 'Job_sector' dan 'Job_category'. Ketiga-tiga pembolehubah ini telah mengalami masalah yang sama iaitu mengandungi nilai yang tidak relevan iaitu 'NaN' dan '999'. Oleh itu, nilai tersebut telah diganti sebagai 'Not stated' yang menunjukkan bahawa peserta yang tidak menerangkan pekerjaannya pada bahagian ini. Kod yang digunakan until menjalankan penggantian ini adalah seperti di bawah:

```
columns_selected =
['Employment_status','Job_sector','Job_category']
for column_name in columns_selected:
    new_df[column_name] = new_df[column_name].replace([999, pd.NA,
None], 'Not stated')
```

Rajah 3.11 Kod digunakan untuk membersihkan pembolehubah 'Job_sector' dan 'Job_category'

```
Raw dataset value counts for
'Job_sector':
Job_sector
NaN                                  1790
Self                                  377
Private sector                         86
999                                    47
Public sector                          16
Non-governmental organization (NGO)     6
Name: count, dtype: int64
```

```
Clean dataset value counts for
'Job_sector':
Job_sector
Not stated                           1837
Self                                  377
Private sector                         86
Public sector                          16
Non-governmental organization (NGO)     6
Name: count, dtype: int6
```

(a) Data mentah

```
Raw dataset value counts for
'Job_category':
Job_category
Jobs skilled agricultural, forestry
and fishery    638
NaN                                  625
Basis jobs                           275
Sales and service jobs
263
Professional                         135
Jobs craft and related trades
112
Plant and machine operators and
assemblers          93
Military                              72
Technicians and associate
professionals                         55
Jobs clerical                         40
Manager                               11
999                                    3
Name: count, dtype: int64
```

```
Clean dataset value counts for
'Job_category':
Job_category
Jobs skilled agricultural, forestry
and fishery    638
Not stated                           628
Basis jobs                           275
Sales and service jobs
263
Professional                         135
Jobs craft and related trades
112
Plant and machine operators and
assemblers          93
Military                              72
Technicians and associate
professionals                         55
Jobs clerical                         40
Manager                               11
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.12 Pembersihan terhadap pembolehubah "Job_sector" dan "Job_category"

Rajah 3.14 telah menunjukkan bahawa dua pembolehubah yang menerangkan situasi pendapatan bagi peserta iaitu 'Total_monthly_main_income' dan 'Total_monthly_side_income'. Kedua-dua pembolehubah ini telah mengalami masalah terdapatnya terlalu banyak nilai unik dan data ini akan menjejaskan ketepatan sekiranya dimasukkan ke dalam model ramalan. Oleh itu, kedua-dua pembolehubah ini telah digabungkan dahulu melalui pengiraan tambahan menjadi 'Total_monthly_income' dan seterusnya semua nilai telah dibahagikan kepada tiga kategori iaitu kurang daripada RM1000, antara RM1001 dan RM2000 serta lebih daripada RM2000. Kod yang digunakan adalah seperti di bawah:

```
columns_selected =
['Total_monthly_main_income','Total_monthly_side_income']
for column_name in columns_selected:
    new_df[column_name] = new_df[column_name].fillna(0)
new_df.rename(columns={'Total_monthly_main_income':
'Total_monthly_income'}, inplace=True)
new_df['Total_monthly_income'] = new_df['Total_monthly_income'] +
new_df['Total_monthly_side_income']
new_df.drop(['Total_monthly_side_income'], axis=1, inplace=True)

conditions = [
    (new_df['Total_monthly_income'] <= 1000),
    (new_df['Total_monthly_income'] > 1000) &
(new_df['Total_monthly_income'] <= 2000),
    (new_df['Total_monthly_income'] > 2000)
]

labels = ["Below RM1000", "RM1001 - RM2000", "Above RM2000"]

new_df['Total_monthly_income'] = np.select(conditions, labels,
default='Unknown')
new_df.rename(columns={'Total_monthly_income': 'Income_category'},
inplace=True)
```

Rajah 3.13 Kod digunakan untuk membersihkan pembolehubah ‘Income_category’

```
Raw dataset value counts for
'Total_monthly_main_income':
Total_monthly_main_income
300.0     325
200.0     264
500.0     222
1000.0    205
100.0     178
          ...
840.0       1
790.0       1
740.0       1
210.0       1
770.0       1
Name:  count,  Length:  117,  dtype:
int64
```

```
Raw dataset value counts for
'Total_monthly_side_income':
Total_monthly_side_income
0.0       1567
100.0      144
200.0      144
300.0       95
500.0       64
          ...
240.0        1
230.0        1
8500.0       1
1100.0       1
450.0        1
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Income_category':
Income_category
Below RM1000       1752
RM1001 - RM2000     399
Above RM2000        171
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.14 Pembersihan terhadap pembolehubah "Total_monthly_main_income" dan "Total_monthly_side_income"


```
Raw   dataset   value   counts   for
'Ageing_group':
Ageing_group
Usual Aging                  1734
Mild Cognitive Impairment     334
Successful Aging              223
No Group                       31
Name: count, dtype: int64
```

(a) Data mentah

```
Clean   dataset   value   counts   for
'Ageing_group':
Ageing_group
Non-Successful Aging    2099
Successful Aging         223
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.15 Pembersihan terhadap pembolehubah "Ageing_group"

Rajah 3.15 telah menunjukkan bahawa golongan penuaan telah dibahagikan kepada empat kategori di pembolehubah 'Ageing_group' dalam data mentah. Demi memudahkan dan meringankan beban semasa diproses dalam model ramalan, empat kategori ini telah dipermudahkan kepada dua kategori sahaja untuk membezakan golongan penuaan sama ada berada dalam kategori berjaya atau tidak berjaya. Kod yang digunakan adalah seperti di bawah:

```
values_to_replace = ['Usual Aging', 'Mild Cognitive Impairment',
'No Group']
replacement_value = 'Non-Successful Aging'
new_df['Ageing_group'].replace(values_to_replace,
replacement_value, inplace=True)
```

Rajah 3.16 Kod digunakan untuk membersihkan pembolehubah ‘Ageing_group’

```
Raw dataset value counts for
'Hypertension':
Hypertension
Tidak                              1114
Ya (dianogsis oleh doktor)         1060
Ya                                  108
Tidak, tetapi pernah mengalami       40
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Hypertension':
Hypertension
1    1168
0    1154
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.17 Pembersihan terhadap pembolehubah “Hypertension”

Rajah 3.17 telah menunjukkan pembolehubah yang mewakili salah satu penyakit yang dihidapi oleh golongan penuaan. Penyakit tersebut telah menunjukkan 4 status. Status ini telah merangkumi peserta pernah menghidapi, masih menghidapi atau tidak menghidapi. Namun status ini tidak memberi makna banyak dalam pembelajaran mesin, oleh itu status tersebut telah dibahagikan kepada 2 bahagian, iaitu “0” yang mewakili golongan penuaan yang pernah menghidapi penyakit tersebut tetapi sudah pulih dan tidak pernah menghidapi penyakit tersebut, manakala “1” yang mewakili golongan penuaan yang masih menghidapi penyakit ini sama ada tengah menerima ramatan atau tidak menerima ramatan pada masa sekarang. Dalam kes ini, rajah tersebut hanya menunjukkan salah satu pembolehubah yang mewakili ‘Hypertension’ dan masih mempunyai 24 penyakit yang menggunakan kaedah yang sama untuk menjalankan data pembersihan ini. Kod yang digunakan adalah seperti di bawah:

```
value_mapping = {'Ya': 1, 'Ya (dianogsis oleh doktor)': 1,
'Tidak': 0, 'Tidak, tetapi pernah mengalami': 0}
columns_to_map = new_df.columns[14:35]
new_df[columns_to_map] =
new_df[columns_to_map].replace(value_mapping)

value_mapping = {'999': 0, '999.0': 0}
columns_to_map = new_df.columns[35:39]
new_df[columns_to_map] =
new_df[columns_to_map].replace(value_mapping, regex=True)
new_df[columns_to_map] = new_df[columns_to_map].replace('[^0]', 1,
regex=True)
```

Rajah 3.18 Kod digunakan untuk membersihkan pembolehubah tentang penyakit


```
Raw dataset value counts for 'IADL_score':
IADL_score
14.0     1029
13.0      427
12.0      289
11.0      172
10.0       96
9.0        73
NaN        55
8.0        53
7.0        28
6.0        27
5.0        24
4.0        16
3.0        15
2.0         8
1.0         5
888.0       3
0.0         2
Name: count, dtype: int64
```

```
Clean dataset value counts for 'IADL_score':
IADL_score
Independent    1917
Dependent       405
Name: count, dtype: int64
```

```
Raw dataset value counts for 'ADL_score':
ADL_score
6.0    2291
NaN      28
5.0       2
0.0       1
Name: count, dtype: int64
```

```
Clean dataset value counts for 'ADL_score':
ADL_score
Independent    2293
Dependent        29
Name: count, dtype: int64
```

(a) Data mentah (b) Data bersih

Rajah 3.19 Pembersihan terhadap pembolehubah “IADL_score” dan “ADL_score”

Rajah 3.19 telah menunjukkan dua pembolehubah ‘IADL_score’ dan ‘ADL_score’ yang menggambarkan tahap keupayaan golongan penuaan dalam aktiviti kehidupan harian. ‘IADL_score’ mempunyai julat markah 0-14 manakala ‘ADL_score’ mempunyai julat markah 0-6. Nilai ‘NaN’ dan ‘888.0’ telah dikira sebagai markah

kosong supaya dapat memastikan data adalah relevan. Seterusnya, disebabkan oleh kelemahan keberkesanan data berterusan dalam model kategori, kedua-dua pembolehubah ini telah dibahagikan semula kepada dua kategori sahaja, di mana 'IADL_score' yang mempunyai markah 10 atau kurang telah dikategori sebagai 'Dependent' dan markah lebih daripada 10 telah dikategori sebagai 'Independent', manakala 'ADL_score' yang mempunyai markah 4 atau kurang telah dikategori sebagai 'Dependent' dan markah lebih daripada 4 telah dikategori sebagai 'Independent'. Kod yang telah digunakan adalah seperti di bawah:

```
columns_to_replace = new_df.columns[39:41]
new_df[columns_to_replace] = new_df[columns_to_replace].fillna(0)
new_df[columns_to_replace] =
new_df[columns_to_replace].replace(888.0, 0)

conditions = [
    (new_df['IADL_score'] <= 10),
    (new_df['IADL_score'] > 10)
]

labels = ["Dependent", "Independent"]

new_df['IADL_score'] = np.select(conditions, labels,
default='Unknown')

conditions = [
    (new_df['ADL_score'] <= 4),
    (new_df['ADL_score'] > 4)
]

labels = ["Dependent", "Independent"]

new_df['ADL_score'] = np.select(conditions, labels,
default='Unknown')
```


Rajah 3.20 Kod digunakan untuk membersihkan pembolehubah 'IADL_score' dan 'ADL_score'

```
Raw dataset value counts for
'Geriatric_depression_score':
Geriatric_depression_score
2.0      534
1.0      446
3.0      393
0.0      289
4.0      229
5.0      149
6.0       76
888.0     58
7.0       44
8.0       36
9.0       22
10.0      19
11.0      13
12.0       7
13.0       4
14.0       3
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Geriatric_depression_score':
Geriatric_depression_score
No depression        1949
Mild depression       327
Severe depression      46
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.21 Pembersihan terhadap pembolehubah "Geriatric_depression_score"

Rajah 3.21 di atas telah menunjukkan pembolehubah 'Geriatric_depression_score' yang mewakili status kemurungan dalam golongan penuaan. Pembolehubah ini mempunyai markah 0 hingga 14 dan nilai '888.0' yang tidak relevan. Nilai ini telah diganti kepada '0' yang menggambarkan status kemurungan yang berada dalam keadaan sihat. Selepas itu, markah tersebut telah dikategori semula supaya dapat memudahkan proses dalam pembelajaran mesin nanti. Dalam keadaan tersebut, golongan penuaan yang mendapat markah 4 atau kurang telah dikategori sebagai 'No depression', markah lebih daripada 4 dan tidak melebihi 9 telah dikategori sebagai 'Mild depression', serta markah yang lebih daripada 9 telah dikategori sebagai 'Severe depression'. Kod yang dimasukkan untuk mendapatkan hasil ini adalah seperti berikut:

```
columns_to_replace = new_df.columns[41]
new_df[columns_to_replace] = new_df[columns_to_replace].fillna(0)
new_df[columns_to_replace] =
new_df[columns_to_replace].replace(888.0, 0)

conditions = [
    (new_df['Geriatric_depression_score'] <= 4),
    (new_df['Geriatric_depression_score'] > 4) &
(new_df['Geriatric_depression_score'] <= 9),
    (new_df['Geriatric_depression_score'] > 9)
]
labels = ["No depression", "Mild depression","Severe depression"]

new_df['Geriatric_depression_score'] = np.select(conditions,
labels, default='Unknown')
```

Rajah 3.22 Kod digunakan untuk membersihkan pembolehubah 'Geriatric_depression_score'

```
Raw dataset value counts for
'EPQ_Q1':
EPQ_Q1
No     1604
Yes     667
NaN      51
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'EPQ_score':
EPQ_score
Low level         1820
Moderate level     320
High level         182
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.23 Pembersihan terhadap pembolehubah "EPQ_score"

Rajah 3.23 di atas telah menunjukkan pembolehubah 'EPQ_Q1' yang merupakan soalan pertama dalam ujian EPQ (*Eysenck Personality Questionnaire*) dan bahagian mempunyai jumlah 12 soalan yang diakhiri oleh pembolehubah 'EPQ_Q12'. Dalam siri soalan ini, 'Yes' dan 'No' telah ditukar kepada '1' dan '0' masing-masing supaya lebih mudah ditafsir. Seterusnya, 12 soalan ini telah digabungkan untuk mendapat jumlah markah EPQ bagi setiap penuaan dan markah tersebut telah dibahagikan kepada tiga kategori, di mana markah 3 atau kurang telah dikategori sebagai 'Low level', markah yang lebih daripada 3 dan tidak melebihi 7 telah dikategori sebagai 'Moderate Level' serta markah yang lebih daripada 7 telah dikategori sebagai 'High Level'. Kod yang telah digunakan adalah seperti di bawah:

```
columns_to_replace = new_df.columns[42:54]
value_mapping = {'Yes': 1, 'No': 0}
new_df[columns_to_replace] = new_df[columns_to_replace].fillna(0)
new_df[columns_to_replace] =
new_df[columns_to_replace].replace(value_mapping)

new_df['EPQ_score'] = new_df.loc[:,
'EPQ_Q1':'EPQ_Q12'].sum(axis=1)
new_df.drop(new_df.loc[:, 'EPQ_Q1':'EPQ_Q12'], axis=1,
inplace=True)
```

Rajah 3.24 Kod digunakan untuk membersihkan pembolehubah 'EPQ_score'

```
Raw dataset value counts for
'Loneliness_Q1':
Loneliness_Q1
Hardly ever          2045
Some of the time      190
NaN                    53
Often                  34
Name: count, dtype: int64
```

(a) Data mentah

```
Clean dataset value counts for
'Loneliness_score':
Loneliness_score
Low level          2084
Moderate level      206
High level           32
Name: count, dtype: int64
```

(b) Data bersih

Rajah 3.25 Pembersihan terhadap pembolehubah "Loneliness_score"

Rajah 3.25 telah menunjukkan pembolehubah 'Loneliness_Q1' yang menggambarkan salah satu soalan terhadap tahap kesendirian dalam golongan penuaan. Soal selidik tersebut adalah terdiri daripada 3 soalan yang bermula dari 'Loneliness_Q1' ke 'Loneliness_Q3'. Penyelarasan nilai telah dibuat untuk ketiga-tiga pembolehubah ini, di mana 'Hardly ever' kepada '1', 'Some of the time' kepada '2', 'Often' kepada '3' dan 'NaN' kepada '0'. Nilai yang besar menunjukkan tahap kesendirian yang tinggi dalam golongan penuaan. Seterusnya, pengiraan tambahan telah dijalankan pada ketiga-tiga pembolehubah tersebut dan menghasilkan satu pembolehubah baru iaitu 'Loneliness_score', dan penyelarasan nilai juga dijalankan dan dibahagikan kepada tiga kategori, di mana markah 3 atau kurang berada dalam kategori 'Low level', markah lebih daripada 3 atau tidak melebihi 6 berada dalam kategori 'Moderate level, serta markah lebih daripada 6 berada dalam kategori 'High level'. Kod yang digunakan adalah seperti berikut:

```
columns_to_replace = new_df.columns[54:57]
value_mapping = {'Hardly ever': 1, 'Some of the time': 2, 'Often':
3}
new_df[columns_to_replace] = new_df[columns_to_replace].fillna(0)
new_df[columns_to_replace] =
new_df[columns_to_replace].replace(value_mapping)

new_df['Loneliness_score'] = new_df.loc[:,
'Loneliness_Q1':'Loneliness_Q3'].sum(axis=1)
new_df.drop(new_df.loc[:, 'Loneliness_Q1':'Loneliness_Q3'],
axis=1, inplace=True)
```

Rajah 3.26 Kod digunaka untuk membersihkan pembolehubah 'Loneliness_score'

Seterusnya, pemilihan pembolehubah dianggap sebagai langkah penting dalam pembersihan data. Pembolehubah yang kurang penting akan dikeluarkan dari set data mentah untuk mengurangkan beban kerja dalam pemprosesan data. Tindakan ini dapat meningkatkan kecekapan dan mengurangkan masa pemprosesan apabila diproses oleh pembelajaran mesin. Demi mencapai objektif ini, ujian *Chi-squared* dan nilai p telah dikira untuk setiap pembolehubah terhadap pembolehubah sasaran, kod yang dimasukkan adalah seperti di bawah:


```
chi2_results = pd.DataFrame(columns=['Variable', 'Chi-square', 'P-
value'])

for column in df.columns:
    if column != 'Ageing_group':

        contingency_table = pd.crosstab(df[column],
df['Ageing_group'])
        chi2, p, _, _ = chi2_contingency(contingency_table)
```

Rajah 3.27 Kod digunakan untuk membuat ujian *chi-squared* dan pengiraan nilai p

Selepas itu, hanya pembolehubah yang memiliki nilai p yang kurang daripada 0.05 akan dipilih ke prosedur yang seterusnya dan hasilnya adalah seperti di dalam Rajah 3.28, kod yang digunakan adalah seperti di bawah:

```
relevant_features = chi2_results[chi2_results['P-
value'].astype(float) < 0.05]['Variable']
```

Rajah 3.28 Kod digunakan untuk memilih pembolehubah yang mempunyai nilai p < 0.05

	Variable	Chi-square	P-value
0	State	12.714070	5.297572e-03
1	Gender	6.096451	1.354537e-02
2	Age_category	20.246539	1.508982e-04
5	Marital_status	12.512292	5.819288e-03
6	Highest_education	123.284146	6.949769e-23
10	Job_category	55.424878	2.630332e-08
11	Income_category	53.510787	2.400381e-12
12	Cf_status	15.830207	3.651861e-04
13	Hypertension	247.463610	9.276978e-56
14	High_cholesterol	25.485905	4.456267e-07
15	Diabetes	85.434405	2.395166e-20
17	Arthritis	8.853247	2.925679e-03
18	Heart_disease	27.085269	1.946757e-07
19	Cataract/Glaucoma	4.150183	4.163003e-02
26	Constipation	6.029834	1.406606e-02
30	Urinary_incontinence	7.182286	7.362684e-03
38	IADL_score	309.982938	9.763492e-58
40	Geriatric_depression_score	65.298640	1.353066e-08
41	EPQ_score	21.402449	4.478881e-02

Rajah 3.29 Pembolehubah yang memiliki nilai p yang kurang daripada 0.05

3.2.4 Pemodelan Data

Dalam peringkat ini, satu set data yang bersih telah bersedia untuk diproses dan dilatih oleh model pembelajaran mesin. Demi memastikan set data dilatih dengan adil, teknik persilangan telah digunakan sebagai pendekatan pengesahan. Teknik persilangan dapat memastikan data yang dilatih tidak ditinggalkan dan mencegah penilaian tidak berat sebelah dengan berkesan dalam proses menganalisis data (Lei 2019). Pengekodan adalah seperti di bawah, dengan menggunakan persilangan 10-lipat:

```
skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
```

3.30 Kod digunakan untuk menjalankan persilangan 10-lipat

Kajian ini telah mengalami masalah ketidakseimbangan kelas dalam data di mana kelas ‘Non-Successful Aging’ mempunyai 1890 data manakala kelas ‘Successful Aging’ hanya mempunyai 200 data. Oleh itu, teknik SMOTE juga telah diaplikasikan untuk menangani masalah ketidakseimbangan kelas pada pembolehubah sasaran. Pengekodan bagi teknik SMOTE adalah seperti di bawah:

```
smote = SMOTE(sampling_strategy='auto', random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train,
y_train)
```

Rajah 3.31 Kod digunakan untuk mengaplikasikan teknik SMOTE

Kod tersebut telah menunjukkan cara untuk menyimbangkan kelas dalam data latihan dengan aplikasi teknik SMOTE. Dalam proses ini, kedua-dua kelas telah mecapai keseimbangan dan mempunyai 1890 data masing-masing.

Seterusnya, data yang bersedia untuk dilatih telah dimasukkan ke dalam algoritma. Sebagai contoh yang menunjukkan kod untuk memasukkan ke dalam model RF:

```
model = RandomForestClassifier(random_state=42)
model.fit(X_train_resampled, y_train_resampled)
```

Rajah 3.32 Kod digunakan untuk memasukkan data latihan ke dalam model

Ramalan terhadap pembolehubah sasaran di y_test telah dibuat selepas latihan telah disiapkan dalam model, kod adalah seperti di bawah:

```
y_prob = model.predict_proba(X_test)[:, 1]
```

Rajah 3.33 Kod digunakan untuk membuat ramalan pada data ujian

3.2.5 Penilaian Model

Korelasi bertugas untuk menggambarkan kekuatan hubungan antara pembolehubah. Oleh itu, korelasi semua pembolehubah akan dinilai untuk menyiasat pembolehubah yang paling berkorelasi dengan penuaan sihat.

Petunjuk paling biasa yang digunakan oleh penyelidik untuk menilai prestasi model klasifikasi adalah metrik seperti ketepatan, kepersisan, ingatan serta spesifisiti, dan metrik ini juga akan digunakan dalam kajian ini. Selain daripada itu, metrik Skor F1 dan kawasan di bawah lengkung (AUC) juga akan dikira bagi setiap model ramalan. Berdasarkan matriks kekeliruan dalam Rajah 3.34, kita dapat mendapatkan semua nilai TP, FP, TN, dan FN. Nilai-nilai ini kemudian digunakan dalam formula bagi setiap metrik untuk menilai prestasi model.

		Nilai Sebenar	
		Positif	**Negatif**
Nilai Ramalan	**Positif**	**TP (Positif Benar)**	**FP (Positif Palsu)**
	Negatif	**FN (Negatif Palsu)**	**TN (Negatif Benar)**

Rajah 3.34 Matriks kekeliruan

Formula bagi setiap metrik adalah seperti di bawah:

$$\text{Ketepatan} = \frac{TP+TN}{TP+TN+FP+FN} \quad ...(3.1)$$

$$\text{Kepersisan} = \frac{TP}{TP+FP} \quad ...(3.2)$$

$$\text{Ingatan} = \frac{TP}{TP+FN} \quad ...(3.3)$$

$$\text{Spesifisiti} = \frac{TN}{TN+FP} \quad ...(3.4)$$

$$\text{Skor F1} = \frac{TP}{TP+1/2(FP+FN)} \quad ...(3.5)$$

$$\text{AUC} = \int TPR\, d(FPR) \quad ...(3.6)$$

Di sini, TP, FP, TN, FN merujuk kepada positif benar, positif palsu, negatif benar dan negatif palsu masing-masing. Manakala TPR dan FPR merujuk kepada kadar positif benar dan kadar positif palsu masing-masing.

Seterusnya, langkah yang terakhir dalam bab metodologi ini adalah melakukan penilaian yang teliti terhadap beberapa model pembelajaran mesin untuk menganalisis prestasi kesemua model dalam meramalkan keputusan penuaan sihat. Proses perbandingan model yang teliti ini bertujuan untuk mengenal pasti algoritma yang paling sesuai dan bukan sahaja dapat mengklasifikasikan penuaan sihat dengan tepat tetapi juga menunjukkan kemampuan secara keseluruhan yang kuat terhadap data yang belum pernah dilihat sebelumnya. Model yang dipilih akan memainkan peranan penting dalam menentukan strategi implementasi untuk aplikasi praktikal seterusnya.

Sebarang proses yang dinyatakan dalam bab metodologi ini telah dilakukan melalui laman web Google Colaboratory (https://research.google.com/colaboratory).

Laman web ini adalah menggunakan kod Python untuk pengaturcaraan. Manakala stesen kerja yang digunakan adalah Komputer Riba keluaran Asus melalui pemprosesan Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz dengan NVIDIA GeForce 940M (2GB DDR3). Memori capaian rawak sebanyak 8.0 GB dengan sistem operasi 64bit.

3.3 KESIMPULAN

Secara kesimpulan, metodologi yang diterangkan dalam bab ini menyediakan kerangka sistematik untuk menangani objektif penyelidikan. Peringkat pemahaman masalah, pemahaman data, penyediaan data, pemodelan, penilaian, dan pengegunaan telah dirangka dengan teliti untuk memastikan pendekatan menyeluruh dan teliti dalam mengkaji pembolehubah yang mempengaruhi penuaan sihat. Penggunaan teknik seperti SMOTE untuk mengendalikan ketidakseimbangan kelas dan aplikasi pelbagai algoritma pembelajaran mesin menyumbang kepada metodologi yang kukuh untuk menghasilkan keputusan yang tepat dan boleh dipercayai dalam mengklasifikasikan penuaan sihat. Bab seterusnya akan membincangkan hasil dan perbincangan yang menerangkan wawasan selepas diperoleh daripada pelaksanaan metodologi tersebut.

BAB IV

HASIL KAJIAN DAN PERBINCANGAN

4.1 PENGENALAN

Pembentangan hasil merupakan fasa penting dalam mendedahkan hasil kajian yang dijalankan dalam penyelidikan ini secara menyeluruh. Dalam bab ini, kita merenung hasil utama yang diperoleh daripada pemeriksaan teliti terhadap dataset penuaan. Hasil ini akan berfungsi sebagai asas bagi perbincangan dan penafsiran yang mengikutinya, menerangkan faktor-faktor yang mempengaruhi penuaan yang sihat dan memberikan pandangan terhadap dinamika kompleks dalam dataset. Selain itu, bab ini akan mendedahkan kesemua hasil dapatan daripada ketujuh-tujuh model ramalan untuk perbandingan keputusan antara satu sama lain.

4.2 PROSES PEMBERSIHAN DATA

Proses pembersihan data telah dijalankan dalam data mentah penuaan untuk memastikan data adalah bersih, rapi dan bersedia dalam pembinaan model ramalan penuaan sihat. Pemilihan pembolehubah secara manual telah dijalankan untuk memastikan hanya pembolehubah yang berkaitan dengan domain keupayaan dipilih dalam kajian ini, antaranya adalah pembolehubah dari aspek socio-demografi, penunjuk kesihatan dan kesejahteraan psikologi. Akhirnya, terdapat 58 pembolehubah telah dipilih termasuk pembolehubah sasaran dan sebanyak 2329 data yang tersedia.

Dalam proses pembersihan data, didapati bahawa terdapat 7 barisan data mempunyai nilai hilang dan telah dipadamkan supaya dapat memastikan semua data adalah bermaklumat dan memberi makna. Data yang bermasalah seperti nilai yang tidak bermakna atau nilai kosong telah digantikan dengan nilai mod masing-masing. Nilai-nilai yang muncul sebagai nol juga diatasi berdasarkan situasi. Seterusnya, terdapat beberapa pembolehubah yang menunjukkan nilai berterusan telah diringkaskan dengan membentuk beberapa kategori supaya senang dianalisis, antaranya ialah:

1. Kategori pendapatan
2. Markah ujian IADL
3. Markah ujian ADL
4. Skala kemurungan geriatrik
5. Markah soal selidik personaliti Eysenck
6. Markah kesendirian

Pembolehubah yang berkaitan dengan penyakit juga diringkaskan dari 4 kategori kepada 2 kategori supaya senang difahamkan. Pembolehubah sasaran juga diringkaskan kepada 2 kategori sahaja, iaitu penuaan yang tidak berjaya dan penuaan yang berjaya, hal ini disebabkan oleh kajian ini adalah untuk pembinaan model ramalan penuaan sihat.

Jadual 4.1 di bawah telah menunjukkan masalah yang didapati pada pembolehubah dan hasil selepas pembersihan juga dijelaskan secara terperinci. Sebelum setiap pembolehubah diperiksa, set data telah dipratinjau secara keseluruhan dan didapati bahawa terdapat tujuh baris data yang menunjuk nilai ‘NaN’ pada setiap pembolehubah. Oleh itu, tujuh baris data ini telah dikenal pasti bahawa tidak mempunyai sebarang maklumat dan ketujuh-tujuh baris ini telah dkeluarkan. Seterusnya, pemeriksaan terhadap setiap pembolehubah telah dijalankan.

Jadual 4.1 Masalah dan penyelesaian bagi setiap pembolehubah

Pembolehubah	Masalah	Penyelesaian
Age_category	Pengesanan satu *outlier* yang bernilai '11'	Dibetulkan dengan mod yang bernilai '60-69'
Employment_status	Pengesanan 47 nilai '*NaN*' dan 18 nilai '999'	Nilai tersebut ditukar menjadi '*Not stated*'
Job_sector	Pengesanan 1790 nilai '*NaN*' dan 47 nilai '999'	Nilai tersebut ditukar menjadi '*Not stated*'
Job_category	Pengesanan 625 nilai '*NaN*' dan 3 nilai '999'	Nilai tersebut ditukar menjadi '*Not stated*'
Total_monthly_main_income dan *Total_monthly_side_income*	Pengesanan 61 nilai '*NaN*'	Nilai tersebut ditukar kepada '0'
	Kedua-dua pembolehubah ini tidak memberi makna banyak sekiranya diasingkan	Kedua-dua pembolehubah ini digabungkan dengan menggunakan pengiraan tambahan dan dinamakan sebagai '*Total_monthly_income*'
	Nilai terlalu spesifik	Semula nilai dikategori semula menjadi '*Below* RM1000', 'RM1001 – RM2000' dan '*Above* RM2000'
Cf_status	Pengesanan 1507 nilai '*NaN*'	Nilai tersebut ditukar menjadi '*No dianogsis*'
Hypertension, High_cholesterol, Diabetes, Stroke, Arthritis, Heart_disease, Cataract/Glaucoma, Kidney_disease, Asthma, Respiratory_ailments, Tuberculosis, Gout, Hip_fracture, Constipation, Haemorrhoids, Gastric/Ulcer, Thyroid_gland_disease, Urinary_incontinence, Vision/Hearing_problem, Chewing_problem, Low_appetite,	Terdapat 4 jenis nilai berbeza	Jenis nilai dipermudahkan menjadi 2 jenis, iaitu 'Ya' dan 'Ya (*diagnosis* oleh doktor)' ditukar menjadi '1', manakala 'Tidak' dan 'Tidak, tetapi pernah mengalami' ditukar menjadi '0'

bersambung...

...sambungan

Cancer_1, Cancer_2, Other_disease_1, Other_disease_2		
IADL_score	Pengesanan 55 nilai '*NaN*' dan 3 nilai '888.0' tetapi nilai yang berkemungkinan adalah dalam julat 0-14.	Nilai tersebut dikira sebagai '0'
	Nilai berterusan kurang sesuai dalam membina model klasifikasi	Semua nilai dikategorikan semula di mana nilai sama atau kurang daripada 10 menjadi '*Dependent*' dan nilai lebih daripada 10 menjadi '*Independent*'
ADL_score	Pengesanan 28 nilai '*NaN*' tetapi nilai yang berkemungkinan adalah dalam julat 0-6.	Nilai tersebut dikira sebagai '0'
	Nilai berterusan kurang sesuai dalam membina model klasifikasi	Semua nilai dikategorikan semula di mana nilai sama atau kurang daripada 4 menjadi '*Dependent*' dan nilai lebih daripada 4 menjadi '*Independent*'
Geriatric_depression_score	Pengesanan 58 nilai '888.0' tetapi nilai yang berkemungkinan adalah dalam julat 0-14.	Nilai tersebut dikira sebagai '0'
	Nilai berterusan kurang sesuai dalam membina model klasifikasi	Semua nilai dikategorikan semula di mana nilai sama atau kurang daripada 4 menjadi '*No depression*', nilai di antara 4 dan 9 atau sama dengan 9 menjadi '*Mild depression*' dan nilai lebih daripada 9 menjadi '*Severe depression*'
EPQ_Q1, EPQ_Q2, EPQ_Q3, EPQ_Q4, EPQ_Q5, EPQ_Q6,	Pengesanan 50 hingga 59 nilai '*NaN*' masing-masing	Nilai tersebut dikira '0'

bersambung...

...sambungan

EPQ_Q7, EPQ_Q8, EPQ_Q9, EPQ_Q10, EPQ_Q11, EPQ_Q12	Nilai perlu ditukar menjadi nombor supaya pengiraan boleh dijalankan	Nilai 'Yes' ditukar kepada '1' manakala nilai 'No' ditukar kepada '0'
	Data yang berdimensi tinggi akan meningkatkan beban semasa dimasuki ke dalam latihan pembelajaran mesin	12 pembolehubah tersebut telah digabungkan dengan pengiraan tambahan dan dinama sebagai '*EPQ_score*'. Seterusnya semua nilai dikategorikan semula di mana nilai sama atau kurang daripada 3 menjadi '*Low level*', nilai di antara 3 dan 7 atau sama dengan 7 menjadi '*Moderate level*' dan nilai lebih daripada 7 menjadi '*High level*'
Loneliness_Q1, Loneliness_Q2 dan *Loneliness_Q3*	Pengesanan 53 hingga 54 nilai '*NaN*' masing-masing	Nilai tersebut dikira '0'
	Nilai perlu ditukar menjadi nombor supaya pengiraan boleh dijalankan	Nilai '*Hardly ever*' ditukar kepada '1', nilai '*Some of the time*' ditukar kepada '2' dan nilai '*often*' ditukar kepada '3'
	Data yang berdimensi tinggi akan meningkatkan beban semasa dimasuki ke dalam latihan pembelajaran mesin	3 pembolehubah tersebut telah digabungkan dengan pengiraan tambahan dan dinama sebagai '*Loneliness_score*'. Seterusnya, 3 pembolehubah tersebut telah digabungkan dengan pengiraan tambahan dan dinama sebagai '*Loneliness_score*'. Seterusnya semua nilai dikategorikan semula di mana nilai sama atau kurang

bersambung...

...sambungan

	daripada 3 menjadi '*Low level*', nilai di antara 3 dan 6 atau sama dengan 6 menjadi '*Moderate level*' dan nilai lebih daripada 6 menjadi '*High level*'

4.3 PERBINCANGAN PROSES PEMBERSIHAN DATA

Proses pembersihan data yang dijalankan telah memainkan peranan yang penting dalam meningkatkan kualiti data dan memastikan integriti maklumat bagi pembinaan model ramalan penuaan sihat. Pembersihan data bukan sahaja untuk memastikan data rapi dan senang dianalitik tetapi juga langkah kritikal dalam meningkatkan kuasa ramalan dalam model (Ridzuan 2019). Pemilihan pembolehubah secara manual merupakan langkah yang sangat wajib untuk memilih faktor-faktor yang berkaitan dengan domain kapasiti intrinsik sahaja dapat menjadi fokus dalam analisis ini. Dimensi data yang lebih rendah membolehkan pembinaan model pengelas yang ringkas dengan keupayaan interpretasi yang lebih baik dan prestasi yang baik (Akhiat et al. 2020). Oleh itu, keputusan untuk memilih 58 pembolehubah, termasuk sasaran utama, dan kesemua 2329 data menunjukkan tindakan yang sistematik dan berlandaskan pertimbangan yang matang. Pembuangan 7 barisan data yang mengandungi nilai hilang supaya dapat memastikan dataset yang digunakan adalah sepenuhnya bermaklumat, menghasilkan satu set data yang mantap dan boleh dipercayai.

Kajian ini juga menunjukkan kepakaran dalam menangani masalah nilai yang hilang, di mana strategi penggantian dengan nilai mod dan penyelesaian bagi nilai nol diterapkan secara sistematik. Pengasingan pembolehubah yang bersifat nilai berterusan kepada kategori-kategori yang lebih ringkas adalah langkah yang kritikal untuk memudahkan analisis. Misalnya, pengasingan kategori pendapatan, markah ujian IADL dan ADL, serta skala kemurungan geriatrik dan skala kesendirian memberi kejelasan dalam memahami hubungan antara faktor-faktor tersebut dengan penuaan. Kesimpulannya, keputusan ini bukan sahaja dapat meningkatkan data boleh percaya,

tetapi juga membantu mengurangkan kompleksiti dataset bagi memudahkan interpretasi dan pemodelan dalam fasa seterusnya.

4.4 STATISTIK PERIHALAN

Selepas melaksanakan satu siri proses pembersihan data pada dataset penuaan, terdapat sejumlah 2322 peserta yang layak untuk menyertai analisis data yang selanjutnya dan pembinaan model ramalan penuaan sihat. Terdapat jumlah keseluruhan 44 pembolehubah yang akan digunakan yang berkaitan dengan objektif penyelidikan, merangkumi domain keupayaan intrinsik seperti socio-demografi, penunjuk kesihatan, dan kesejahteraan psikologi.

Pembolehubah sasaran dalam dataset ini ialah kumpulan penuaan yang boleh diklasifikasikan kepada 2 kategori, iaitu penuaan yang tidak berjaya dengan 2099 peserta, yang membentuk kira-kira 90% daripada jumlah peserta keseluruhan dan baki 10%, iaitu 223 peserta, dikategorikan sebagai penuaan yang berjaya. Dalam dataset ini, sebanyak 52% daripada peserta adalah wanita yang berjumlah 1208 orang, manakala 48% yang selebihnya adalah lelaki dengan jumlah 1114 peserta. Dengan menguraikan dataset mengikut kumpulan umur, didapati bahawa 60.37% daripada peserta berada dalam julat umur 60-69 tahun dengan jumlah keseluruhan 1330 peserta. Kategori usia 70-79 mengambil bahagian sekitar 30.36%, dengan jumlah keseluruhan 853 peserta. Kategori umur 80-89 mewakili 6.86%, termasuk 136 peserta. Manakala julat umur 90-99 merupakan kategori umur yang paling kecil, hanya menyumbang kira-kira 0.13% dengan 3 peserta.

Jadual 4.2 Nilai dan frekuensi bagi setiap pembolehubah dalam aspek sosio-demografi

No.	Nama Pembolehubah	Jenis Pembolehubah	Nilai	Frekuensi
1	Negeri	Polinomial	Kelantan	724
			Selangor	573
			Perak	543
			Johor	482
2	Jantina	Binomial	Lelaki	1114
			Perempuan	1208

bersambung...

...sambungan

3	Kumpulan Umur	Polinomial	60-69	1330
			70-79	853
			80-89	136
			90-99	3
4	Bangsa	Polinomial	Melayu	1447
			Cina	750
			India	120
			Lain-lain	5
5	Agama	Polinomial	Islam	1452
			Buddha	627
			Hindu	99
			Kristian	96
			Lain-lain	48
6	Status Perkahwinan	Polinomial	Berkahwin	1585
			Balu/Duda	655
			Bujang	41
			Bercerai/Berpisah	41
7	Tahap Pendidikan Tertinggi	Polinomial	Sekolah Rendah	1338
			Tidak Bersekolah	492
			LCE/SRP/PMR	192
			SPM	188
			HSC/STPM/ SIJIL	64
			Ijazah	37
			Diploma	5
			Sarjana	5
			PHD	1
8	Perokok	Polinomial	Tidak merokok	1630
			Merokok	399
			Bekas perokok	293
9	Status Pekerjaan	Polinomial	Bersara	1309
			Bekerja sendiri	281
			Suri rumah	257
			Tidak bekerja	177
			Bekerja sepenuh masa	128
			Bekerja sambilan	94
			Tidak dinyatakan	65

bersambung...

...sambungan

			Membantu ahli keluarga	8
			Lain-lain	2
			Pelajar, pelatih atau sekolah vokasional	1
10	Sektor Pekerjaan	Polinomial	Tidak dinyatakan	1837
			Sendiri	377
			Sektor swasta	86
			Sektor awam	16
			Pertubuhan bukan kerajaan	6
11	Kategori Pekerjaan	Polinomial	Pertanian, perhutanan dan perikanan	638
			Tidak dinyatakan	628
			Pekerjaan asas	275
			Jualan dan perkhidmatan	263
			Profesional	135
			Kraf tangan dan perdagangan yang berkaitan	112
			Pengendali dan pemasang loji dan mesin	93
			Tentera	72
			Juruteknik dan profesional bersekutu	55
			Perkeranian	40
			Pengurus	11
12	Kategori Pendapatan	Polinomial	RM1000 ke bawah	1752
			RM1001-RM2000	399
			RM2000 ke atas	171
13	Status Kognitif	Polinomial	Tiada diagnosis	1507
			Kognitif biasa	490
			Kognitif lemah	325

Dari Jadual 4.2, peserta-peserta dalam kajian ini adalah terdiri daripada pelbagai lapisan sosio-demografi. Kelantan telah menyumbang jumlah peserta yang terbanyak,

iaitu seramai 724 orang, diikuti oleh Selangor, Perak dan Johor. Peserta yang berasal dari keturunan Melayu merupakan bangsa majoriti dalam kajian ini, iaitu seramai 1447 orang, telah mecapai lebih kurang 62% dari kesemua peserta, dan diikuti dengan Cina, India dan lain-lain. Oleh itu, sebahagian besar peserta adalah beragama Islam. Kebanyakan peserta dalam kajian ini sudah berkahwin dan sebahagian daripada mereka ialah balu, serta sekelompok kecil peserta adalah bujang atau sudah bercerai. Dalam aspek tahap pendidikan, kebanyakan peserta hanya memiliki pendidikan sekolah rendah, iaitu seramai 1338 orang, menduduki lebih separuh daripada kesemua peserta dan 492 orang tidak pernah bersekolah. 70% daripada peserta tidak pernah merokok, iaitu seramai 1630 orang. Secara keseluruhan, profil sosio-demografi peserta memberikan gambaran yang komprehensif tentang pelbagai ciri-ciri di kalangan mereka.

Jadual 4.3 Nilai dan frekuensi bagi setiap pembolehubah dalam aspek kesihatan fisiologi dan metabolik

No.	Nama Pembolehubah	Jenis Pembolehubah	Nilai	Frekuensi
1	Tekanan Darah Tinggi	Binomial	0	1154
			1	1168
2	Kolesterol Tinggi	Binomial	0	1620
			1	702
3	Kencing Manis	Binomial	0	1717
			1	605
4	Angin Ahmar	Binomial	0	2277
			1	45
5	Sakit Sendi	Binomial	0	1741
			1	581
6	Penyakit Jantung	Binomial	0	2083
			1	239
7	Katarak/Glaucoma	Binomial	0	2104
			1	218
8	Kegagalan Buah Pinggang	Binomial	0	2282
			1	40
9	Lelah	Binomial	0	2132
			1	190
10	Penyakit Paru-paru Kronik	Binomial	0	2303
			1	19

bersambung...

...sambungan

11	Batuk Kering	Binomial	0	2300
			1	22
12	Gout	Binomial	0	2218
			1	104
13	Keretakan Tulang Pinggul	Binomial	0	2300
			1	22
14	Sembelit	Binomial	0	2121
			1	201
15	Buasir	Binomial	0	2271
			1	51
16	Gastrik/Ulser	Binomial	0	2008
			1	314
17	Masalah Kelenjar Tiroid	Binomial	0	2282
			1	40
18	Masalah Kencing	Binomial	0	2095
			1	227
19	Masalah Penglihatan/ Pendengaran yang Serius	Binomial	0	2029
			1	293
20	Kesukaran Mengunyah	Binomial	0	2187
			1	135
21	Kurang Selera Makan	Binomial	0	2158
			1	164
22	Kanser 1	Binomial	0	2287
			1	35
23	Kanser 2	Binomial	0	2320
			1	2
24	Penyakit Lain 1	Binomial	0	2148
			1	174
25	Penyakit Lain 2	Binomial	0	2265
			1	57

Berdasarkan Jadual 4.3, didapati bahawa lebih daripada separuh peserta dalam kajian ini mempunyai masalah tekanan darah tinggi, iaitu seramai 1168 orang dan telah mecapai kira-kira 50.3% daripada keseluruhan. Selain daripada itu, terdapat beberapa penyakit juga dikira aktif dalam kalangan penuaan, antaranya ialah kolesterol tinggi seramai 702 (30.2%) orang, kencing manis seramai 605 (26.1%) orang dan sakit sendi

seramai 581 (25%) orang. Manakala penyakit lain adalah jarang dihidapi dalam kalangan penuaan dalam kajian ini.

Jadual 4.4 Nilai dan frekuensi bagi setiap pembolehubah dalam aspek kapasiti fizikal

No.	Nama Pembolehubah	Jenis Pembolehubah	Nilai	Frekuensi
1	Markah Ujian IADL	Polinomial	Bebas	1917
			Bergantung	405
2	Markah Ujian ADL	Polinomial	Bebas	2293
			Bergantung	29
3	Skala Kemurungan Geriatrik	Polinomial	Tiada kemurungan	1949
			Kemurungan ringan	327
			Kemurungan teruk	46
4	Markah Soal Selidik Personaliti Eysenck	Polinomial	Tahap rendah	1820
			Tahap sederhana	320
			Tahap tinggi	182
5	Markah Kesendirian	Polinomial	Tahap rendah	2084
			Tahap sederhana	206
			Tahap tinggi	32

Jadual 4.4 menunjukkan data dari aspek kapasiti fizikal bagi semua peserta. Skala IADL dalam kajian ini adalah berjulat 0-14. Peserta yang mempunyai 10 markah atau ke atas akan dilabelkan sebagai bebas dan yang ke bawah adalah sebagai bergantung. Dalam kajian ini, hanya terdapat 405 peserta adalah sama ada sedikit atau keseluruhan bergantung kepada orang lain untuk membantu mereka dalam aktiviti peralatan kehidupan harian. Skala ADL adalah berjulat 0-6, 5 markah atau ke atas akan dikategorikan sebagai bebas dan yang lain adalah bergantung. Berdasarkan Jadual 4.4, hanya 29 orang peserta perlu bergantung kepada orang lain dalam aktiviti kehidupan harian. Selain itu, tahap kemurungan bagi setiap peserta juga dilabelkan berdasarkan skala kemurungan geriatrik dan data telah menunjukkan bahawa kebanyakan peserta tidak mengalami sebarang kemurungan, iaitu seramai 1949 orang dan hanya seramai 46 orang mengalami kemurungan yang teruk. Soal selidik personaliti Eysenck dan kesendirian juga dijalankan demi penilaian psikologi dan kesihatan mental. Kedua-dua ujian telah menunjukkan majoriti peserta dalam kajian ini ada pada tahap rendah.

4.5 PERBINCANGAN STATISTIK PERIHALAN

Analisis statistik perihalan yang dilaksanakan dapat memberikan gambaran komprehensif mengenai ciri-ciri peserta dalam kajian ini dan langkah ini dapat membentuk asas yang kukuh untuk pemahaman mendalam mengenai golongan penuaan yang dikaji. Jumlah peserta yang layak selepas proses pembersihan data sebanyak 2322 orang menunjukkan kerangka kajian yang kukuh dan bermakna. Keputusan menunjukkan bahawa kajian ini melibatkan peserta dari pelbagai lapisan sosio-demografi, dengan Kelantan menyumbang jumlah peserta yang terbanyak.

Perbandingan antara kumpulan umur menunjukkan bahawa peserta telah mendominasi dalam kategori 60-69 tahun yang merangkumi lebih daripada 60% daripada keseluruhan peserta. Hasil ini telah mencerminkan corak penuaan dalam masyarakat yang dikaji di Malaysia dan menggambarkan perbezaan dalam struktur demografi kumpulan yang dipilih. Sementara itu, aspek demografi seperti agama, status perkahwinan, dan tahap pendidikan memberikan wawasan yang lebih mendalam ke dalam konteks sosio-budaya peserta. Menurut Annamika et al., socio-demografi memainkan peranan penting dalam tingkah laku pencarian kesihatan. Tahap pendapatan yang tinggi atau tahap pendidikan yang tinggi lebih cenderung dalam kesedaran kesihatan dan memiliki kemahuan tinggi dalam mementingkan penjagaan kesihatan (Annamika et. al. 2015). Oleh itu, kepelbagaian aspek demografi antara peserta merupakan salah satu kunci dan index kritikal dalam pembinaan model ramalan terhadap penuaan sihat.

Analisis berdasarkan status kesihatan menunjukkan bahawa beberapa penyakit kronik, seperti tekanan darah tinggi, kolesterol tinggi dan kencing manis, memiliki prevalensi dalam golongan penuaan. Menurut Singh et al., tekanan darah sistolik, jumlah kolesterol dan glukosa plasma puasa menunjukkan peningkatan semasa umur individu meningkat (Singh et al. 2012). Indeks-indeks ini merupakan indeks utama dalam diagnosis tekanan darah tinggi, kolesterol tinggi dan kencing manis. Fenomena ini telah menekankan bahawa keperluan pada pemantauan kesihatan yang lebih efektif dalam golongan penuaan. Walaupun demikian, peratusan tinggi dalam golongan penuaan yang tidak mengalami kemurungan, menunjukkan tahap personaliti dan

kesendirian yang rendah adalah faktor positif yang telah menyumbang kepada kualiti kesihatan mental golongan penuaan dalam kajian ini.

Selanjutnya, penelitian mendalam ke atas kebergantungan peserta dalam aktiviti harian, seperti yang dicerminkan dalam skala IADL dan ADL, memberikan gambaran tentang tahap kemandirian dalam kehidupan seharian. Keputusan menunjukkan majoriti peserta mengekalkan tahap kemandirian yang tinggi, walaupun terdapat segelintir yang memerlukan bantuan dalam aktiviti tertentu. Penekanan terhadap kebolehpercayaan dan interpretasi data ini membantu menyusun landasan untuk pembinaan model ramalan penuaan sihat yang lebih terarah dan relevan dengan keperluan populasi ini.

4.6 PEMILIHAN FAKTOR

Dalam Jadual 4.5, ujian Chi-squared telah dijalankan pada semua pembolehubah untuk menyiasat faktor-faktor yang paling penting yang berkaitan dengan pembolehubah sasaran, iaitu kumpulan penuaan. Nilai p juga telah dihitung bagi setiap pembolehubah untuk mengenal pasti sama ada pembolehubah tersebut adalah secara statistik bererti.

Jadual 4.5 Ujian *Chi-squared* dan nilai p bagi setiap pembolehubah

No.	Nama Pembolehubah	Ujian *Chi-squared*	Nilai p
1	Negeri	12.71	<0.01
2	Jantina	6.09	0.013
3	Kumpulan Umur	20.25	<0.01
4	Bangsa	2.49	0.476
5	Agama	4.69	0.320
6	Status Perkahwinan	12.51	<0.01
7	Tahap Pendidikan Tertinggi	123.28	<0.01
8	Perokok	0.39	0.823
9	Status Pekerjaan	13.09	0.158
10	Sektor Pekerjaan	8.58	0.072
11	Kategori Pekerjaan	55.42	<0.01
12	Kategori Pendapatan	53.51	<0.01
13	Status Kognitif	15.83	<0.01
14	Tekanan Darah Tinggi	247.46	<0.01

bersambung...

...sambungan

15	Kolesterol Tinggi	25.49	<0.01
16	Kencing Manis	85.43	<0.01
17	Angin Ahmar	3.81	0.051
18	Sakit Sendi	8.85	<0.01
19	Penyakit Jantung	27.09	<0.01
20	Katarak/Glaucoma	4.15	0.042
21	Kegagalan Buah Pinggang	0.53	0.468
22	Lelah	0.49	0.480
23	Penyakit Paru-paru Kronik	1.07	0.300
24	Batuk Kering	1.37	0.241
25	Gout	0.72	0.397
26	Keretakan Tulang Pinggul	0.00	1.000
27	Sembelit	6.03	0.014
28	Buasir	0.04	0.848
29	Gastrik/Ulser	3.18	0.075
30	Masalah Kelenjar Tiroid	1.61	0.205
31	Masalah Kencing	7.18	<0.01
32	Masalah Penglihatan/ Pendengaran yang Serius	3.36	0.067
33	Kesukaran Mengunyah	3.79	0.052
34	Kurang Selera Makan	0.79	0.372
35	Kanser 1	2.74	0.098
36	Kanser 2	0.00	1.000
37	Penyakit Lain 1	0.74	0.390
38	Penyakit Lain 2	3.36	0.067
39	Markah Ujian IADL	50.79	<0.01
40	Markah Ujian ADL	2.10	0.15
41	Skala Kemurungan Geriatrik	47.21	<0.01
42	Markah Soal Selidik Personaliti Eysenck	7.01	0.03
43	Markah Kesendirian	1.82	0.40

Dalam kajian ini, nilai $p < 0.05$ akan dianggap sebagai signifikan secara statistik dan akan diambil untuk langkah seterusnya dalam pembinaan model ramalan penuaan sihat. Terdapatnya 19 pembolehubah telah dipilih sebagai pembolehubah yang layak untuk masuk ke dalam proses pembinaan model ramalan penuaan sihat. Antaranya ialah:

1. Negeri
2. Jantina

3. Kumpulan umur
4. Status perkahwinan
5. Tahap pendidikan tertinggi
6. Kategori pekerjaan
7. Kategori pendapatan
8. Status kognitif
9. Tekanan darah tinggi
10. Kolesterol tinggi
11. Kencing manis
12. Sakit sendi
13. Penyakit jantung
14. Katarak/glaucoma
15. Sembelit
16. Masalah kencing
17. Markah ujian IADL
18. Skala kemurungan Geriatrik
19. Markah soal selidik personaliti Eysenck.

4.7 PERBINCANGAN PEMILIHAN FAKTOR

Jadual 4.5 menunjukkan bahawa ujian *Chi-squared* telah dijalankan secara menyeluruh pada semua pembolehubah untuk mengkaji hubungan antara faktor-faktor tertentu dan pembolehubah sasaran utama, iaitu kumpulan penuaan. Proses ini membantu mengenal pasti faktor-faktor yang paling penting dan signifikan dalam konteks penuaan dalam kajian ini. Ujian *Chi-squared* dan pengiraan nilai p merupakan alat strategi dalam

menjelaskan hubungan antara pembolehubah. Ujian *Chi-squared* dua dapat menghasilkan kuasa statistik yang kuat dengan memeriksa hubungan keertian antara dua pembolehubah kategorikal terutamanya saiz sampel yang besar dan ujian ini dapat membantu pengkaji menguji hipotesis tentang pembolehubah yang diukur pada tahap nominal (Chen 2021).

Mankala nilai p digunakan sebagai penanda untuk menguji keertian statistik pada setiap pembolehubah. Nilai p yang kurang daripada 0.05 diambil sebagai ambang keputusan untuk menentukan keertian statistik. Dalam konteks ini, nilai p yang rendah merujuk kepada hubungan yang lebih kuat dan bererti antara setiap pembolehubah dengan kumpulan penuaan.

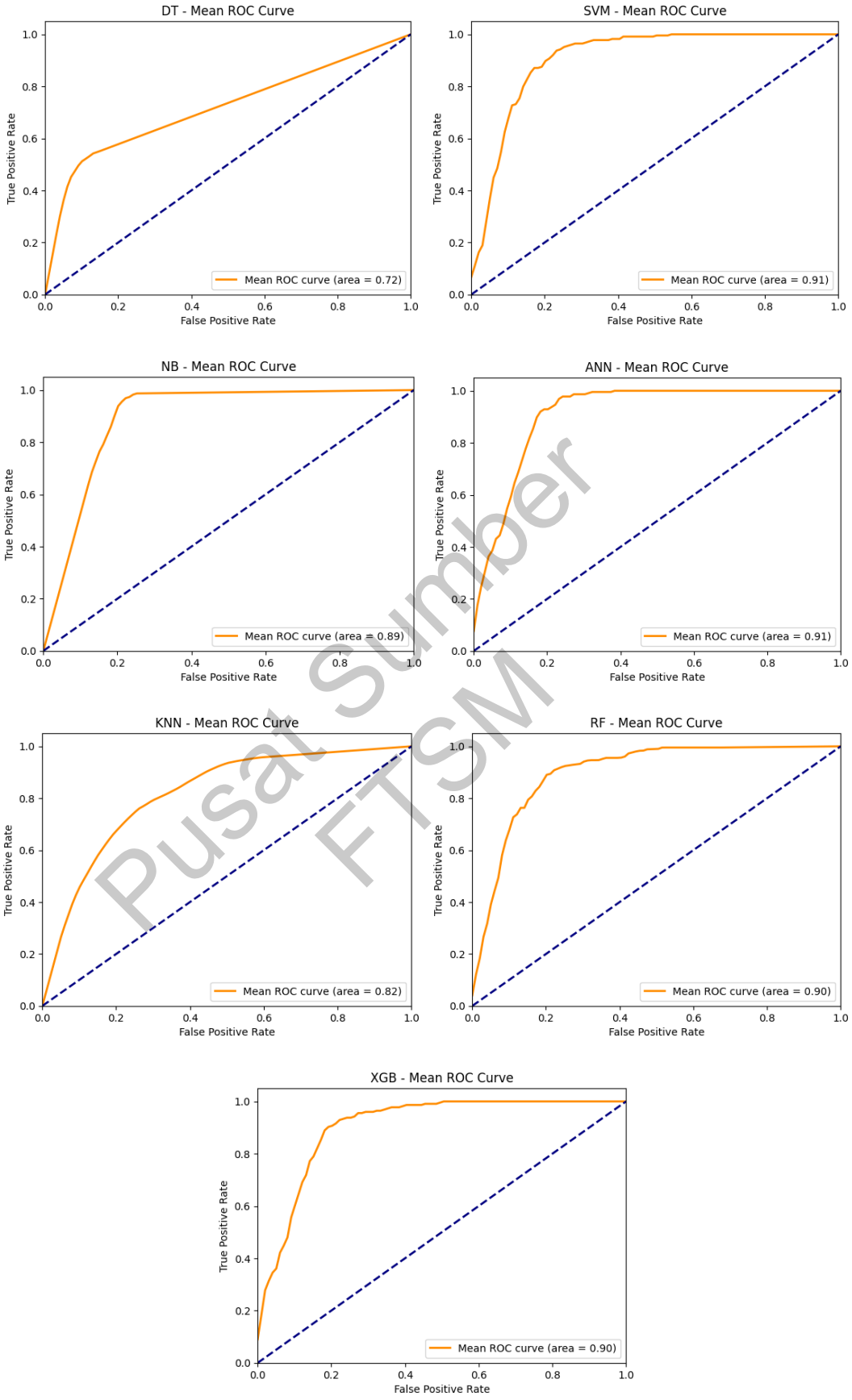
Dengan demikian, aspek-aspek yang dikenal pasti memainkan peranan penting dalam membentuk kualiti penuaan telah dipilih sebagai pembolehubah input kepada model ramalan penuaan sihat. Analisis yang lebih mendalam dan model ramalan yang dibina berdasarkan pembolehubah ini akan memberikan wawasan yang lebih dalam dan membantu dalam penyelidikan penuaan sihat pada masa depan.

4.8 PRESTASI MODEL RAMALAN

Jadual 4.6 telah menunjukkan beberapa metrik yang digunakan dalam pengujian model ramalan dalam kalangan penuaan berserta dengan Rajah 4.1 yang merujuk kepada graf kawasan di bawah lengkung bagi setiap model ramalan.

Jadual 4.6 Metrik ketepatan, kepersisan, ingatan, spesifisiti, skor F1 dan kawasan di bawah lengkung (AUC) bagi setiap model ramalan

Model	Ketepatan	Kepersisan	Ingatan	Spesifisiti	Skor F1	AUC
DT	87.73	40.27	51.07	91.62	44.55	71.80
SVM	87.51	41.09	62.75	90.14	49.11	90.67
NB	82.30	33.88	87.00	81.80	48.70	89.42
ANN	88.16	42.12	50.10	92.19	44.89	91.30
KNN	73.34	23.46	77.59	72.89	35.98	81.58
RF	88.81	45.35	54.15	92.47	48.09	89.96
XGB	88.24	42.16	50.55	92.24	45.08	90.43



Rajah 4.1 Graf kawasan di bawah lengkung (AUC) bagi setiap model ramalan

Antara model ramalan tersebut, model Naive Bayes (NB) dan *K-Nearest Neighbours* (KNN) tidak mempunyai prestasi yang cukup baik dari segi ketepatan model berbanding dengan model lain, iaitu 82.30% dan 73.34% masing-masing. Namun begitu, kedua-dua model ini mencapai metrik ingatan yang paling tinggi dalam kesemua model, iaitu sebanyak 87.00% dan 77.59% masing-masing. Manakala model Pohon Keputusan (DT) juga tidak dikira sebagai model ramalan yang bagus dalam kajian ini, hal ini kerana model ini mempunyai kawasan di bawah lengkung yang terlalu rendah, hanya mencapai 71.80% walaupun telah mencapai ketepatan yang agak tinggi sebanyak 87.73%.

Dari segi ketepatan model dan kawasan di bawah lengkung, keempat-empat model ramalan ini telah menunjukkan prestasi yang memuaskan dan tidak banyak berbeza antara satu sama lain, iaitu Mesin Vektor Sokongan (SVM), Rangkaian Neural Tiruan (ANN), XgBoost (XGB) dan Hutan Rawak (RF), masing-masing mecapai ketepatan antara 87% - 89% dan kawasan di bawah lengkung antara 89% - 92%. Walau bagaimanapun, Hutan Rawak dipilih sebagai model ramalan yang terbaik dalam kajian ini. Hal ini disebabkan model ini bukan sahaja menunjukkan kepersisan dan spesifisiti yang tertinggi antara semua model, model ini juga menunjukkan keseimbangan yang baik antara ketepatan, kawasan di bawah lengkung dan skor F1 yang agak tinggi berbanding dengan model lain.

4.9 PERBINCANGAN PRESTASI MODEL RAMALAN

Analisis mendalam terhadap Jadual 4.6 menunjukkan perbandingan prestasi yang teliti antara model ramalan penuaan yang digunakan dalam kajian ini. Terdapat variasi yang ketara dalam prestasi setiap model, dan hal ini memberikan pandangan yang kaya tentang keberkesanan model masing-masing.

Model Naive Bayes (NB) dan *K-Nearest Neighbours* (KNN) menunjukkan ketepatan yang agak rendah berbanding dengan model lain. Hal ini kerana KNN terlalu bergantung kepada kedekatan dalam ruang fitur tanpa memahami struktur yang lebih mendalam. Sementara itu, NB membuat anggapan ringkas yang mungkin kurang sesuai dengan keadaan sebenar dataset dan cenderung memberikan prestasi yang baik dalam