

PENGELASAN HASIL ANGIOPLASTI
MENGUNAKAN PEMBELAJARAN MESIN
BERGABUNG

NOOR SAKIRAH BINTI KHAMIS

UNIVERSITI KEBANGSAAN MALAYSIA

PENGELASAN HASIL ANGIOPLASTI MENGGUNAKAN PEMBELAJARAN
MESIN BERGABUNG

NOOR SAKIRAH BINTI KHAMIS

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

PENGAKUAN

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

09 Februari 2025

NOOR SAKIRAH BINTI KHAMIS
P127064

PENGHARGAAN

Alhamdulillah, setinggi-tinggi kesyukuran dipanjatkan ke hadrat Ilahi atas limpah rahmat dan keberkatan-Nya membolehkan saya menyiapkan kajian ini sebagai syarat akhir bagi Ijazah Sarjana Sains Data. Segala puji dan syukur juga diiringkan atas kekuatan dan ketabahan yang dikurniakan sepanjang proses penyelidikan ini.

Setinggi-tinggi penghargaan dan ucapan terima kasih saya tujukan kepada penyelia saya, Ts. Dr. Nor Samsiah Sani diatas tunjuk ajar, bimbingan serta sokongan yang tidak berbelah bahagi sepanjang tempoh penyeliaan kajian ini. Kepakaran dan nasihat beliau amat membantu dalam memastikan projek ini berjalan lancar dan memenuhi standard akademik yang ditetapkan.

Tidak lupa juga ucapan terima kasih ditujukan kepada semua pensyarah di Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia yang telah mencurahkan ilmu dan panduan sepanjang pengajian saya dalam bidang Sarjana Sains Data. Bimbingan dan sokongan mereka telah memberikan saya landasan yang kukuh untuk menghasilkan kajian ini.

Jutaan penghargaan juga saya tujukan kepada *Clinical Research Department* (CRD), Institut Jantung Negara (IJN) yang telah membekalkan sumber data perubatan sebagai sumber data asas yang amat membantu dalam memastikan kajian ini dapat dilaksanakan. Selain itu, saya ingin merakamkan terima kasih kepada pihak penaja biasiswa, Jabatan Perkhidmatan Awam (JPA) yang telah menyediakan pembiayaan penuh untuk pengajian saya.

Akhir sekali, setulus penghargaan ditujukan kepada suami tercinta, anak-anak dan mak ayah tersayang yang sentiasa memberikan sokongan moral, doa, serta dorongan sepanjang perjalanan pengajian ini. Tidak dilupakan juga rakan-rakan seperjuangan yang sentiasa sedia membantu dan berkongsi idea sepanjang tempoh pembelajaran.

Semoga setiap usaha yang dicurahkan ini memberikan manfaat dan menjadi sumbangan bermakna kepada bidang ilmu dan masyarakat. Sekian, terima kasih.

ABSTRAK

Angioplasti merupakan satu prosedur perubatan yang secara teknikalnya dikenali sebagai intervensi koronari perkutan (PCI). Ia kebiasaannya digunakan untuk merawat keadaan yang berkaitan dengan saluran darah yang sempit atau tersumbat, terutamanya di dalam jantung. Kajian ini bertujuan membangunkan model pembelajaran mesin bergabung bagi meramal hasil prosedur angioplasti dengan mengenal pasti faktor risiko utama yang mempengaruhi keputusan rawatan. Tiga pembelajaran mesin bergabung utama digunakan, iaitu hutan rawak, *gradient boosting* dan XGBoost, dengan penilaian prestasi melalui pendekatan *train-test split* (TTS) dan *cross-validation* (CV) 10-lipat. Proses kajian melibatkan empat fasa utama: pembersihan data, penalaan hiperparameter menggunakan kaedah *grid search*, penilaian prestasi model berdasarkan metrik seperti ketepatan, kepekaan, kejituan, Skor F1, dan AUC, serta analisis pemilihan fitur menggunakan SHAP (*Shapley Additive Explanations*). Hasil kajian menunjukkan bahawa model XGBoost (TTS Terbaik) adalah model terbaik berdasarkan metrik kritikal. Model ini mencatatkan kepekaan (*recall*) tertinggi sebanyak 57.92%, skor F1 tertinggi iaitu 0.6929, dan AUC tertinggi sebanyak 0.8791, menjadikannya paling sesuai untuk mengenal pasti pesakit berisiko tinggi. Prestasi cemerlang model ini dalam senario multi-kelas menunjukkan keupayaannya dalam mengimbangi kesemua kelas yang terlibat, bukan hanya tertumpu pada satu kelas dominan. Dalam aplikasi perubatan, kepekaan merupakan metrik yang sangat penting kerana ia menilai keupayaan model untuk mengenal pasti kes positif dengan tepat, sekaligus memainkan peranan besar dalam pengurusan risiko pesakit angioplasti. Dalam kajian ini, model XGBoost (TTS Terbaik) mencatatkan kepekaan tertinggi sebanyak 57.92%, skor F1 tertinggi iaitu 0.6929, dan AUC tertinggi sebanyak 0.8791, menjadikannya model paling sesuai untuk mengenal pasti pesakit berisiko tinggi. Analisis statistik menggunakan ujian-t berpasangan 5x2-lipat menunjukkan bahawa perbezaan ketepatan antara model XGBoost (TTS Terbaik) dan model-model lain adalah signifikan dalam kebanyakan perbandingan. Ini menegaskan bahawa model ini mempunyai kelebihan statistik yang kukuh berbanding model lain, sekaligus mengukuhkan lagi kebolehpercayaan dan kestabilannya dalam meramalkan hasil angioplasti. Sebaliknya, model Hutan Rawak (CV Terbaik) dikenalpasti sebagai model yang paling lemah. Model ini mencatatkan kepekaan terendah sebanyak 48.09% dan skor F1 yang rendah iaitu 0.5903 walaupun mencatatkan ketepatan keseluruhan sebanyak 91.83%. Analisis SHAP mengenal pasti BaselineCreatinine, MaxStentDiameter dan TotalStentLength sebagai fitur paling signifikan yang menyumbang kepada kebarangkalian hasil angioplasti berisiko tinggi, di mana "berisiko tinggi" didefinisikan sebagai pesakit yang diklasifikasikan dengan keputusan 0 untuk fitur Death. Sebaliknya, nilai sistolik yang rendah menunjukkan hubungan positif dengan kebarangkalian hasil yang lebih baik.

CLASSIFICATION OF ANGIOPLASTY OUTCOMES USING ENSEMBLE MACHINE LEARNING

ABSTRACT

Angioplasty is a medical procedure technically known as percutaneous coronary intervention (PCI). It is commonly used to treat conditions related to narrowed or blocked blood vessels, particularly in the heart. This study aims to develop an ensemble machine learning model to predict the outcomes of angioplasty procedures by identifying key risk factors that influence treatment decisions. Three primary ensemble machine learning methods were utilized: Random Forest, Gradient Boosting, and XGBoost, with performance evaluation conducted using the train-test split (TTS) and 10-fold cross-validation (CV) approaches. The study process involved four main phases: data cleaning, hyperparameter tuning using grid search, model performance evaluation based on metrics such as accuracy, recall, precision, F1-score, and AUC, as well as feature selection analysis using SHAP (Shapley Additive Explanations). The results indicate that the XGBoost (Best TTS) model is the best-performing model based on critical metrics. This model achieved the highest recall of 57.92%, the highest F1-score of 0.6929, and the highest AUC of 0.8791, making it the most suitable for identifying high-risk patients. The exceptional performance of this model in a multi-class scenario demonstrates its ability to balance all involved classes rather than focusing solely on a dominant class. In medical applications, recall is a crucial metric as it assesses a model's ability to accurately identify positive cases, thereby playing a significant role in managing angioplasty patients' risk. In this study, the XGBoost (Best TTS) model recorded the highest recall (57.92%), the highest F1-score (0.6929), and the highest AUC (0.8791), making it the most suitable model for identifying high-risk patients. Statistical analysis using the 5x2-fold paired t-test revealed that the accuracy differences between XGBoost (Best TTS) and other models were statistically significant in most comparisons. This reinforces the model's statistical superiority over other models, further strengthening its reliability and stability in predicting angioplasty outcomes. Conversely, the Random Forest (Best CV) model was identified as the weakest model. It recorded the lowest recall at 48.09% and the lowest F1-score of 0.5903, despite achieving an overall accuracy of 91.83%. SHAP analysis identified BaselineCreatinine, MaxStentDiameter, and TotalStentLength as the most significant features contributing to the probability of high-risk angioplasty outcomes, where "high-risk" is defined as patients classified with a score of 3 in the Death feature. Conversely, lower systolic values were found to have a positive relationship with better outcomes.

KANDUNGAN

	Halaman
PENGAKUAN	ii
PENGHARGAAN	iii
ABSTRAK	iv
ABSTRACT	v
KANDUNGAN	vi
SENARAI JADUAL	ix
SENARAI ILUSTRASI	xi
SENARAI SINGKATAN	xiii
BAB I PENDAHULUAN	
1.1 Pengenalan	1
1.2 Latar Belakang kajian	5
1.3 Permasalahan Kajian	8
1.4 Objektif Kajian	9
1.5 Skop Kajian	9
1.6 Kepentingan Kajian	9
1.7 Struktur Tesis	10
BAB II KAJIAN KESUSASTERAAN	
2.1 Pengenalan	11
2.2 Kajian Mengenai Pembelajaran Mesin Menggunakan Data Prosedur <i>Percutaneous Coronary Intervention</i> (PCI)	11
2.3 Model Pembelajaran Mesin Bergabung	26
2.3.1 Hutan Rawak	28
2.3.2 Gradient Boosting	29
2.3.3 eXtreme Gradient-Boosting (XGBoost)	30
2.4 Shapley Additive Explanation (SHAP)	32
2.5 Kesimpulan	35
BAB III METODOLOGI KAJIAN	
3.1 Pengenalan	37

3.2	Kerangka Penyelidikan	37
3.3	Pemahaman Data	39
3.4	Penyediaan Data	47
	3.4.1 Pengurusan Data Hilang	48
	3.4.2 Penyingkiran Outlier untuk Data Numerik	50
	3.4.3 Pemetaan Semula dan Pengekodan untuk Data Kategori	52
	3.4.4 Penerangan Set Data Akhir	62
3.5	Pemodelan	67
	3.5.1 Pemilihan Teknik Pemodelan	67
	3.5.2 Pembahagian dan Penyeimbangan Data	71
	3.5.3 Penalaan Hiperparameter	73
3.6	Penilaian	79
	3.6.1 Metrik Utama Untuk Penilaian Prestasi Model	79
	3.6.2 Ujian Statistik Pada Prestasi Model	82
	3.6.3 Pemilihan Fitur	83
	3.6.4 Shapley Additive Explanations (SHAP)	84
3.7	Kesimpulan	85
BAB IV	DAPATAN KAJIAN	
4.1	Pengenalan	86
4.2	Prestasi Model	86
4.3	Pengujian Statistik	97
4.4	Pemilihan Fitur	100
4.5	SHAP	104
4.6	Kesimpulan	109
BAB V	RUMUSAN	
5.1	Pendahuluan	110
5.2	Rumusan Kajian	110
5.3	Implikasi	112
5.4	Kekangan	113
5.5	Cadangan	114
5.6	Kesimpulan	115
RUJUKAN		116

LAMPIRAN

Lampiran A	Surat Kebenaran Institut Jantung Negara (IJN)	123
Lampiran B	Fitur-fitur yang Dikeluarkan Di Peringkat Penyediaan Data	125

LIBRARY FTSM

SENARAI JADUAL

No. Jadual		Halaman
Jadual 1.1	Penerangan bagi kategori data	9
Jadual 2.1	Ringkasan kajian literatur berkaitan aplikasi pembelajaran mesin dalam meramal hasil intervensi koronari perkutaneus (PCI)	22
Jadual 3.1	Jenis dan penerangan fitur bagi Set Data A (prosedur PCI)	40
Jadual 3.2	Jenis dan penerangan fitur bagi Set Data B (fitur lesi yang dirawat)	44
Jadual 3.3	Penggantian istilah bagi fitur dalam set data	48
Jadual 3.4	Peratusan nilai yang hilang bagi setiap fitur	49
Jadual 3.5	Kelas label Death	52
Jadual 3.6	Pengekoden binari untuk Gender	53
Jadual 3.7	Pemetaan semula dan pengekoden untuk Ethnicity	54
Jadual 3.8	Pengekoden binari untuk OHA	54
Jadual 3.9	Pengekoden binari untuk Insulin	55
Jadual 3.10	Pengekoden binari untuk DietTherapy	55
Jadual 3.11	Pengekoden binari untuk Hypertension	56
Jadual 3.12	Pengekoden binari untuk PrevPCI	56
Jadual 3.13	Pengekoden binari untuk CVAdisease	57
Jadual 3.14	Pengekoden binari untuk PrevCABG	57
Jadual 3.15	Pengekoden binari untuk PeripheralVASCdisease	57
Jadual 3.16	Pengekoden label untuk SmokingStatus	58
Jadual 3.17	Pengekoden binari untuk HeartFailureHist	59
Jadual 3.18	Pengekoden label untuk PCIstatus	59
Jadual 3.19	Pengekoden label untuk CCSscore	60
Jadual 3.20	Pengekoden label untuk NYHA	60

Jadual 3.21	Pengekoden label untuk KillipClass	61
Jadual 3.22	Pengekoden binari untuk SideStent	61
Jadual 3.23	Statistik deskriptif bagi data numerik setelah pra-pemprosesan	64
Jadual 3.24	Statistik deskriptif bagi data kategori setelah pra-pemprosesan	64
Jadual 3.25	Penalaan hiperparameter untuk hutan rawak	75
Jadual 3.26	Penalaan hiperparameter untuk <i>gradient boosting</i>	77
Jadual 3.27	Penalaan hiperparameter untuk XGBoost	78
Jadual 4.1	Kombinasi hiperparameter dengan nilai ketepatan ujian tertinggi bagi model hutan rawak dengan TTS dan CV	88
Jadual 4.2	Perbandingan antara nilai hiperparameter lalai dengan hiperparameter terbaik bagi model hutan rawak (TTS) dan (CV)	89
Jadual 4.3	Kombinasi hiperparameter dengan nilai ketepatan ujian tertinggi bagi model <i>gradient boosting</i> dengan TTS dan CV	91
Jadual 4.4	Perbandingan antara nilai hiperparameter lalai dengan hiperparameter terbaik bagi model <i>gradient boosting</i> (TTS) dan (CV)	91
Jadual 4.5	Kombinasi hiperparameter dengan nilai ketepatan ujian tertinggi bagi model XGBoost dengan TTS dan CV	93
Jadual 4.6	Perbandingan antara nilai hiperparameter lalai dengan hiperparameter terbaik bagi model XGBoost (TTS) dan XGBoost (CV)	94
Jadual 4.7	Perbandingan prestasi model hutan rawak, <i>gradient boosting</i> dan XGBoost dengan pendekatan <i>train-test split</i> dan <i>cross-validation</i> berdasarkan 26 fitur	96
Jadual 4.8	Hasil ujian-t berpasangan 5x2-lipat dengan validasi silang bagi konfigurasi lalai dengan konfigurasi hiperparameter terbaik	98
Jadual 4.9	Hasil ujian-t berpasangan 5x2-lipat dengan validasi silang untuk perbandingan prestasi antara model	98
Jadual 4.10	Perbandingan kedudukan ciri untuk semua model dengan <i>train-test split</i> dan <i>cross-validation</i> berdasarkan kepentingan ciri	102

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 1.1	Gambaran angioplasti	3
Rajah 1.2	Pemasangan stent	4
Rajah 1.3	Punca utama kematian di Malaysia bagi tahun 1991-2022	6
Rajah 1.4	Punca utama kematian mengikut kumpulan umur, jantina dan kumpulan etnik bagi tahun 2022	7
Rajah 2.1	Pembelajaran gabungan	27
Rajah 2.2	Struktur hutan rawak	28
Rajah 2.3	Struktur <i>gradient boosting</i>	30
Rajah 2.4	Struktur XGBoost	32
Rajah 2.5	Tafsiran model menggunakan SHAP	33
Rajah 3.1	Carta alir keseluruhan kajian ini.	39
Rajah 3.2	Boxplot bagi Systolic, Diastolic dan MaxStentDiameter dengan <i>outlier</i>	51
Rajah 3.3	Boxplot bagi Systolic, Diastolic dan MaxStentDiameter tanpa <i>outlier</i>	51
Rajah 3.4	Taburan fitur bagi set data pra-pemprosesan menggunakan histogram	66
Rajah 3.5	Kolerasi peta haba Spearman antara fitur dengan pembolehubah sasaran bagi set data IJN	67
Rajah 3.6	Seni bina model hutan rawak yang dicadangkan	69
Rajah 3.7	Seni bina model <i>gradient boosting</i> yang dicadangkan	70
Rajah 3.8	Seni bina model XGBoost yang dicadangkan	71
Rajah 3.9	Sebelum dan selepas penyeimbangan data	73
Rajah 4.1	Ketepatan latihan dan ketepatan ujian bagi model hutan rawak dengan <i>train-test split</i> berdasarkan gabungan pelbagai hiperparameter	87

Rajah 4.2	Ketepatan latihan dan ketepatan ujian bagi model hutan rawak dengan <i>cross-validation</i> berdasarkan gabungan pelbagai hiperparameter	87
Rajah 4.3	Ketepatan latihan dan ketepatan ujian bagi model <i>gradient boosting</i> dengan <i>train-test split</i> berdasarkan gabungan pelbagai hiperparameter	90
Rajah 4.4	Ketepatan latihan dan ketepatan ujian bagi model <i>gradient boosting</i> dengan <i>cross-validation</i> berdasarkan gabungan pelbagai hiperparameter	90
Rajah 4.5	Ketepatan latihan dan ketepatan ujian bagi model XGBoost dengan <i>train-test split</i> berdasarkan gabungan pelbagai hiperparameter	92
Rajah 4.6	Ketepatan latihan dan ketepatan ujian bagi model XGBoost dengan <i>cross-validation</i> berdasarkan gabungan pelbagai hiperparameter	93
Rajah 4.7	Kepentingan fitur untuk model XGBoost dengan TTS	100
Rajah 4.8	Nilai SHAP bagi model XGBoost dengan <i>train-test split</i> ditunjukkan dalam plot beeswarm untuk Kelas 0	105
Rajah 4.9	Nilai SHAP bagi model XGBoost dengan <i>train-test split</i> ditunjukkan dalam plot beeswarm untuk Kelas 1	106
Rajah 4.10	Nilai SHAP bagi model XGBoost dengan <i>train-test split</i> ditunjukkan dalam plot beeswarm untuk Kelas 2	107
Rajah 4.11	Nilai SHAP bagi model XGBoost dengan <i>train-test split</i> ditunjukkan dalam plot beeswarm untuk Kelas 3	108

SENARAI SINGKATAN

AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
AKI	Acute Kidney Injury
ANN	Artificial Neural Networks
AUC	Area Under the Curve
CABG	Coronary Artery Bypass Grafting
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
CVE	Cerebrovascular Events
DT	Decision Tree
ERT	Extremely Randomized Trees
ICU	Intensive Care Unit
IHD	Ischemic Heart Disease
IJN	Institut Jantung Negara
IJNREC	IJN Research Ethics Committee
kNN	k-Nearest Neighbors
LMCA	Left Main Coronary Artery
LR	Logistic Regression
LSTM	Long Short-Term Memory
MACE	Major Adverse Cardiovascular Events
MDI	Mean Decrease in Impurity
ML	Machine Learning
NB	Naive Bayes
NN	Neural Network
NR	No-Reflow

PCI	Percutaneous Coronary Intervention
RF	Random Forest
SHAP	Shapley Additive Explanations
STEMI	ST-Elevated Myocardial Infarction
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

LIBRARY FTSM

BAB I

PENDAHULUAN

1.1 PENGENALAN

Kecerdasan buatan (AI) dan pembelajaran mesin (ML) telah mencapai kemajuan ketara dalam bidang perubatan, terutamanya dalam meramal dan mengenal pasti keperluan kesihatan, kelaziman penyakit, serta tindak balas imun melalui aplikasi teknologi terkini. Walaupun terdapat keraguan terhadap amalan praktikal dan kebolehpercayaan kaedah ML dalam sektor kesihatan, penggunaan teknologi ini terus berkembang dengan pesat.

Pembelajaran mesin (Ghazal et al. 2021) merupakan satu cabang kecerdasan buatan (AI) (Janiesch et al. 2021) yang melibatkan kajian algoritma pintar yang mempunyai keupayaan untuk memperoleh pengetahuan melalui data yang diberikan dan memberikan hasil cadangan baharu berdasarkan pengalaman melalui data yang dibekalkan semasa latihan yang boleh digunakan untuk melaksanakan aktiviti seperti ramalan, keputusan dan klasifikasi (Mohammad et al. 2019). Definisi klasik oleh Tom Mitchell menyatakan bahawa pembelajaran berlaku apabila prestasi program komputer terhadap satu tugas meningkat seiring dengan pengalaman. Dalam konteks ini, ML menggunakan pelbagai model algoritma dan teknik statistik untuk menyelesaikan masalah tanpa memerlukan pengaturcaraan khusus, yang menjadikannya berguna dalam pelbagai bidang termasuk kesihatan. Impak AI dan ML telah merentas pelbagai bidang dan kehidupan seharian dan kesan transformatif seperti ini dijangka turut mempengaruhi sektor kesihatan dan perubatan. Pembelajaran mesin berkeupayaan menggunakan pelbagai model algoritma dan teknik statistik untuk menangani masalah tanpa memerlukan pengaturcaraan khusus (Ji-Peng et al. 2021). Kebanyakan model

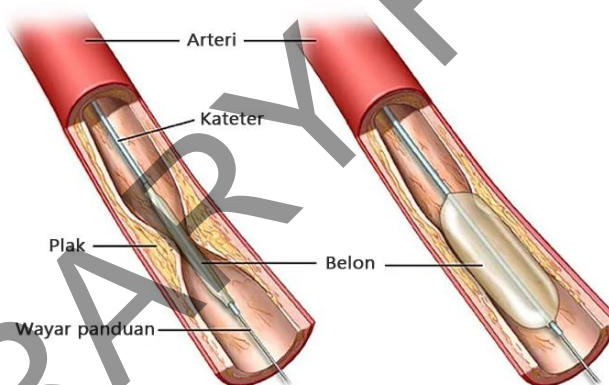
pembelajaran mesin adalah berdasarkan struktur satu lapisan dan memerlukan pra-pemrosesan data yang meluas dan mengekstrakan fitur sebelum data boleh digunakan pada model (Alpaydin 2020). Pra-pemrosesan yang meluas ini adalah penting untuk memastikan model mampu membuat ramalan dengan baik.

Penyakit arteri koronari (CAD) juga dikenali sebagai penyakit jantung koronari (CHD), penyakit jantung iskemia (IHD), iskemia miokardial atau *myocardial ischemia* ringkasnya dipanggil penyakit jantung melibatkan pengurangan aliran darah ke otot jantung akibat dari peningkatan plak aterosklerosis dalam arteri jantung. CAD terdapat dalam pelbagai jenis termasuklah angina stabil, angina tidak stabil, dan infark miokardium (Chang et al. 2019). Gejala yang sering dikaitkan dengan CAD adalah sakit dada atau ketidakselesaan yang mungkin merebak ke bahu, lengan, belakang badan, leher, atau rahang (Parashar et al. 2024). Adakalanya pesakit mungkin mengalami simptom seperti pedih ulu hati. Gejala ini kebiasaannya berlaku ketika sedang bersenam atau sedang mengalami tekanan emosi dan ianya berlangsung selama kurang daripada beberapa minit dan keadaan akan bertambah baik dengan cara berehat. Selain itu, sesak nafas juga mungkin petanda awal dialami pesakit dan dalam kes tertentu, mungkin tidak ada gejala yang ketara sama sekali. Dalam kebanyakan kes, penyakit ini mula dikesan setelah mengalami serangan jantung. Komplikasi lain yang berpotensi timbul dari keadaan ini termasuklah kegagalan jantung atau aritmia (denyutan jantung tidak normal) (Kim et al. 2021). Pada tahun 2015, CAD menjejaskan lebih 110 juta orang dan mengakibatkan 8.9 juta kematian. Ia membentuk 15.6% daripada semua kematian dan menjadikannya punca kematian paling tinggi di seluruh dunia (Rethemiotaki. 2024).

Angioplasti (*Angioplasty*) merupakan satu prosedur perubatan yang secara teknikalnya dikenali sebagai intervensi koronari perkutan (PCI), kebiasaannya digunakan untuk merawat keadaan yang berkaitan dengan saluran darah yang sempit atau tersumbat, terutamanya di dalam jantung. Angioplasti koronari mula diperkenalkan pada tahun 1977 oleh Andreas Gruentzig di Switzerland (Korz 2022). Teknik ini adalah penting dalam mengurus penyakit CAD, penyakit arteri perifer, penyakit arteri karotid, dan penyakit buah pinggang kronik dengan memperbaiki aliran darah melalui arteri. PCI ialah prosedur bukan pembedahan yang digunakan untuk merawat penyempitan arteri koronari jantung yang terdapat dalam penyakit arteri koronari. PCI

ialah alternatif kepada CABG yang sering dirujuk sebagai "pembedahan pintasan" yang mana ia memintas arteri yang menyempit dengan mencantumkan saluran dari lokasi lain dalam badan. Prosedur PCI digunakan untuk meletakkan stent koronari iaitu tiub berjaring dawai kekal untuk membuka arteri koronari yang sempit (Ahmad et al. 2023).

Proses angioplasti melibatkan penyisipan sebuah kateter iaitu sebatang tiub yang panjang, nipis dan fleksibel ke dalam saluran darah pesakit yang kebiasaannya melalui pergelangan tangan atau pangkal paha. Kateter ini dilengkapi dengan belon kecil di hujungnya yang dinavigasikan ke bahagian arteri yang tersumbat. Setelah sampai di kawasan yang sempit, belon tersebut ditiup dengan perlahan-lahan. Pengembangan ini menolak plak yang terkumpul ke dinding arteri dan memampatkan plak ke dinding salur darah untuk membolehkan aliran darah lebih lancar.

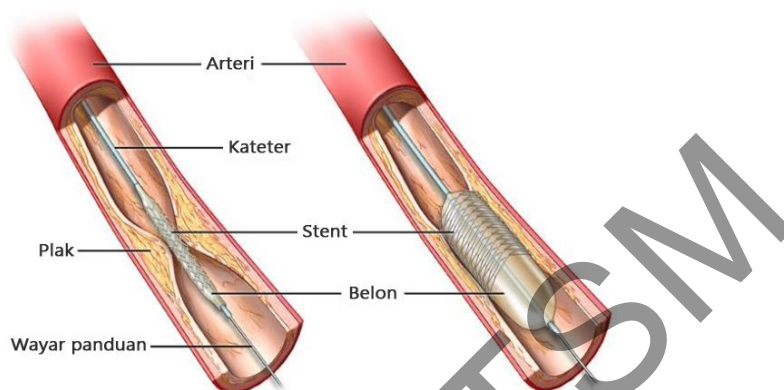


Rajah 1.1 Gambaran angioplasti

Dalam kes angioplasti koronari, prosedur ini khususnya menyasarkan arteri koronari yang menyediakan darah ke jantung. Ini adalah penting kerana apabila arteri koronari menjadi sempit atau tersumbat, ia akan menyebabkan berlakunya gejala seperti sakit dada atau serta serangan jantung kerana bekalan oksigen yang tidak mencukupi ke jantung. Angioplasti boleh dilakukan dalam keadaan kecemasan, seperti semasa serangan jantung, atau secara elektif iaitu setelah pakar perubatan mengesan penyakit jantung.

Selepas pengembangan arteri, angioplasti sering dilengkapi dengan penempatan stent iaitu sebuah tiub mesh wayar kecil yang diselitkan ke dalam arteri yang terjejas untuk menyokong pembukaan arteri. Kebanyakan stent mengeluarkan

ubat di mana stent ini dilapisi dengan ubat yang menghalang arteri daripada menyempit semula iaitu suatu keadaan yang disebut sebagai *restenosis*. Inovasi dalam teknologi stent ini telah mengurangkan keperluan untuk prosedur ulangan secara signifikan.



Rajah 1.2 Pemasangan stent

Dalam konteks ini, ML memainkan peranan penting dengan menganalisis data pesakit untuk mengenal pasti faktor risiko, memperbaiki diagnosis, dan memantau keberkesanan rawatan seperti angioplasti. Algoritma pembelajaran mendalam, sebagai contoh, dapat digunakan untuk menganalisis imej perubatan seperti angiogram, membantu pakar perubatan membuat keputusan yang lebih tepat. Selain itu, teknik ML juga boleh digunakan untuk meramal hasil rawatan, termasuk kejayaan angioplasti dan risiko komplikasi, berdasarkan data pesakit yang besar dan pelbagai.

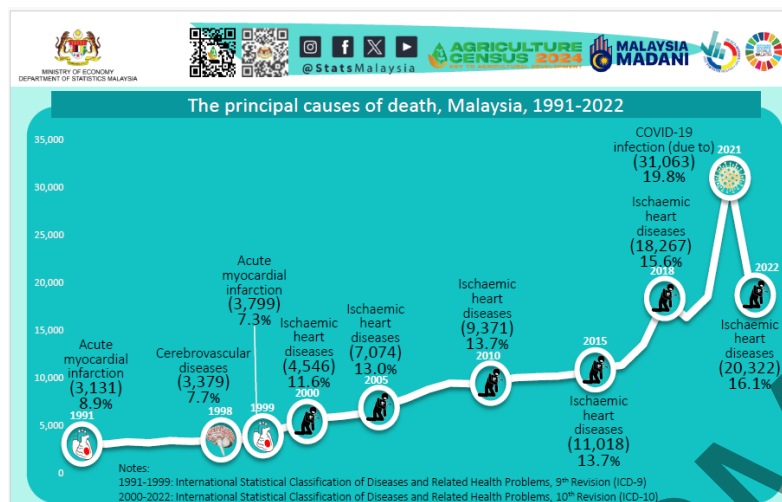
Gabungan ML dan prosedur perubatan seperti angioplasti membawa impak besar dalam pengurusan penyakit CAD. ML bukan sahaja menyokong diagnosis dan rawatan yang lebih tepat, tetapi juga membantu meramal dan mengurangkan risiko komplikasi melalui analisis data yang mendalam. Dalam masa yang sama, angioplasti kekal sebagai prosedur penting dalam rawatan CAD, dengan inovasi teknologi seperti stent berlapis ubat yang meningkatkan keberkesanan rawatan.

Kesinambungan antara ML, dan prosedur seperti angioplasti mencerminkan potensi besar teknologi ini untuk mentransformasi sektor kesihatan. Pendekatan berasaskan data yang diperoleh daripada ML dan aplikasi klinikal seperti angioplasti menjanjikan peningkatan kualiti hidup pesakit melalui rawatan yang lebih berkesan dan bersifat peribadi.

1.2 LATAR BELAKANG KAJIAN

Dalam landskap penjagaan kesihatan yang berkembang pesat, angioplasti koronari atau PCI telah muncul sebagai prosedur penting dalam membendung penyakit arteri koronari yang disenaraikan antara penyebab utama kematian di peringkat global. Ini amat relevan di Malaysia, di mana penyakit jantung kekal berada di kedudukan tertinggi berbanding punca kematian yang lain. Data terkini daripada Jabatan Perangkaan Malaysia (DOSM) mendedahkan bahawa penyakit kardiovaskular sebagai penyumbang utama kepada kematian, sekali gus meningkatkan keperluan terhadap metodologi rawatan dan pengoptimuman protokol penjagaan pesakit. Pengenalan pembelajaran mesin ke dalam sektor penjagaan kesihatan membuka satu dimensi baru untuk memperhalusi ramalan hasil PCI. Melalui penggunaan algoritma ML, kita mampu memproses dan menganalisis set data berskala besar untuk menganalisis ramalan di luar keupayaan manusia bagi tindak balas pesakit terhadap hasil prosedur PCI.

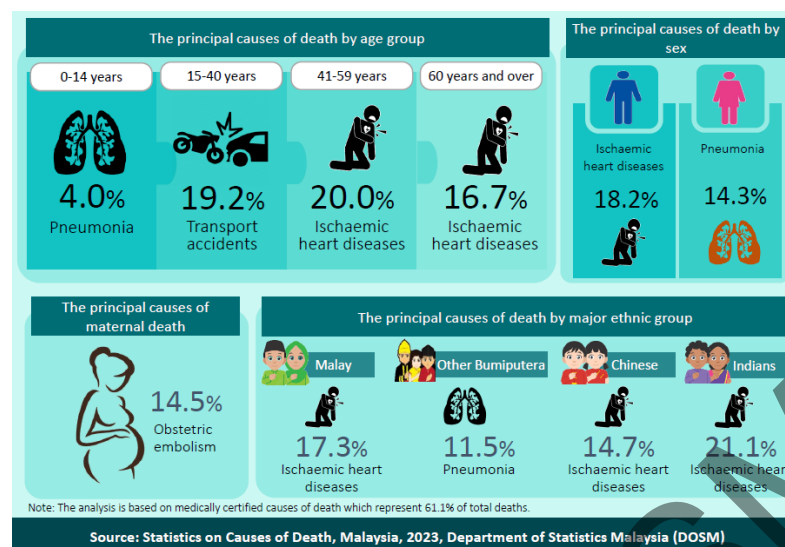
Menurut laporan Jabatan Perangkaan Malaysia, penyakit jantung iskemia (IHD) kekal menjadi pembunuh nombor satu di Malaysia sejak 2019 sehingga laporan terkini dibentangkan bagi tahun 2022. Ia menyumbang 17.0% dari 109,155 kematian yang disahkan secara perubatan pada tahun 2020 (DOSM 2024a). Kematian tertinggi bagi tahun 2021 adalah diakibatkan oleh wabak Covid-19 dan penyakit jantung iskemia kembali mencatatkan kematian tertinggi pada tahun 2022 iaitu sebanyak 16.1% bersamaan 20,322 jumlah kematian (DOSM 2024b). Sebanyak 13,817 (18.2%) kematian disebabkan oleh IHD adalah dalam kalangan lelaki, manakala sebab kematian utama bagi perempuan adalah pneumonia iaitu sebanyak 7,201 kes kematian (14.3%). IHD merupakan sebab kematian utama bagi semua kumpulan etnik utama di Malaysia yang mana melayu mencatatkan 17.3%, cina 14.7% dan india 21.1% kecuali bumiputera lain yang merekodkan pneumonia sebanyak 11.5% sebagai sebab kematian utama.



Rajah 1.3 Punca utama kematian di Malaysia bagi tahun 1991-2022

Sumber: DOSM (2024b)

Dari segi kumpulan umur, sebab kematian utama bagi penduduk berumur 41-59 tahun serta 60 tahun dan lebih adalah IHD yang mana masing-masing pada 20.0% dan 16.7%. Selain itu, kemalangan berkaitan kenderaan merupakan sebab kematian utama bagi penduduk berumur 15-40 tahun (19.2%), manakala pneumonia adalah sebab kematian utama bagi penduduk berumur 0-14 tahun (4.0%). Semua negeri di Malaysia merekodkan IHD sebagai sebab kematian utama pada tahun 2022 kecuali Sabah dan Sarawak yang merekodkan pneumonia sebagai sebab kematian utama pada tahun yang sama. Kelantan merekodkan peratusan tertinggi bagi kematian disebabkan oleh IHD iaitu 23.3% diikuti Perlis sebanyak 23.1% dan Melaka 20.0% (DOSM 2024b). Melalui laporan perangkaan ini, jelaslah bahawa pilihan rawatan yang berkesan dalam menangani isu ini amat menekan dan kritikal. Lazimnya, angioplasti koronari atau PCI menjadi antara pilihan utama bagi prosedur perubatan yang kerap digunakan dalam menguruskan kes IHD.



Rajah 1.4 Punca utama kematian mengikut kumpulan umur, jantina dan kumpulan etnik bagi tahun 2022

Sumber: DOSM (2024b)

Langkah untuk menerapkan pembelajaran mesin dalam bidang kesihatan di Malaysia penuh dengan cabaran termasuklah kebimbangan terhadap privasi data, keperluan penyediaan infrastruktur data yang mantap serta latihan komprehensif kepada pakar kesihatan dalam teknologi AI dan ML. Namun, sekiranya cabaran tersebut ditangani dengan baik, Malaysia mampu meningkatkan standard penjagaan jantung, mengurangkan kesan buruk berikutan PCI dan akhirnya menyumbang kepada pengurangan kadar kematian yang dikaitkan dengan penyakit kardiovaskular.

Penggunaan ML untuk meramalkan hasil PCI merupakan langkah awal ke arah mempraktikkan aplikasi AI dan ML ke dalam landskap penjagaan kesihatan Malaysia. Ringkasnya, aplikasi pembelajaran mesin untuk meramalkan hasil prosedur PCI menandakan kemajuan besar dalam pengurusan penyakit arteri koronari di Malaysia. Dengan memanfaatkan keupayaan dari ML, Malaysia bukan sahaja boleh memperbaiki hasil perawatan pesakit tetapi juga mampu menetapkan penanda aras bagi penyepaduan teknologi termaju dalam penjagaan kesihatan serta merangka masa depan yang lebih sihat dan sejahtera.

1.3 PERMASALAHAN KAJIAN

Pembangunan model pembelajaran mesin untuk mengklasifikasi dengan tepat hasil prosedur angioplasti, khususnya membezakan antara risiko kematian jangka masa pendek (kurang dari 30 hari dan antara 30 hari hingga 1 tahun) dan jangka panjang (antara 1 tahun hingga 2 tahun dan selepas 2 tahun), memberikan cabaran yang ketara dalam bidang perubatan. Tugas ini bukan sahaja melibatkan penggunaan teknik pembelajaran mesin tetapi juga pemahaman mendalam tentang fitur klinikal yang mempengaruhi hasil ini. Masalah utama terletak pada keupayaan untuk mereka bentuk dan melatih model yang boleh mengendalikan kerumitan dan kebolehubahan data pesakit, termasuk keadaan sedia ada, butiran prosedur dan faktor penjagaan selepas pembedahan. Pemilihan model yang sesuai adalah penting untuk mencapai ketepatan dan kebolehppercayaan yang tinggi dalam ramalan yang seterusnya boleh memberi kesan ketara kepada pengurusan pesakit dan perancangan rawatan.

Mengenal pasti fitur utama yang mempengaruhi prestasi model ramalan hasil angioplasti adalah salah satu isu yang kritikal. Fitur perubatan bagi pesakit yang menjalani prosedur PCI adalah sangat besar antaranya termasuk demografi pesakit, sejarah klinikal, spesifik prosedur dan penjagaan pasca prosedur. Cabarannya terletak pada penentuan fitur yang mempunyai kesan paling ketara terhadap keupayaan ramalan model. Ini memerlukan penggunaan teknik pemilihan fitur untuk menapis set data yang besar, mengurangkan dimensi dan memfokuskan pada maklumat yang paling relevan.

Selain itu, meningkatkan prestasi model ramalan hasil angioplasti melalui penalaan hiperparameter semasa proses latihan juga antara langkah penting dalam memastikan model yang dihasilkan adalah lebih relevan. Masalah sering timbul berkaitan hiperparameter di mana setiap kombinasi boleh menjejaskan ketepatan, kerumitan dan kebolehan menjana model dengan baik. Proses ini memerlukan pendekatan sistematik untuk mengenal pasti set hiperparameter optimum yang dapat menghasilkan prestasi ramalan terbaik. Cabaran ini ditambah lagi dengan sumber pengiraan serta masa yang diperlukan untuk latihan dan pengesahan model bagi mencapai hiperparameter yang optimum tanpa menjejaskan kualiti ramalan model.

1.4 OBJEKTIF KAJIAN

Objektif bagi kajian ini adalah:

1. Membangunkan model untuk mengelas hasil prosedur angioplasti.
2. Mengkaji faktor-faktor yang mempengaruhi pengelasan menggunakan kaedah *Shapley Additive Explanations* (SHAP).

1.5 SKOP KAJIAN

Skop kajian ini memfokuskan kepada data pesakit jantung dari Malaysia yang telah menjalani prosedur angioplasti sama ada angioplasti belon atau penempatan stent. Data kajian yang digunakan adalah diperoleh dari Institut Jantung Negara (IJN), Malaysia. Data kajian ini adalah berstatus sulit dan kebenaran penggunaan data telah dimohon terlebih dahulu daripada pihak IJN. Antara kategori data yang dikenalpasti untuk kegunaan kajian ini adalah:

Jadual 1.1 Penerangan bagi kategori data

Kategori Data	Penerangan
Demografi	Merangkumi maklumat asas tentang individu yang membantu penyelidik memahami ciri-ciri latar belakang populasi kajian dan menganalisis bagaimana ciri-ciri ini mempengaruhi pembolehubah lain.
Komorbidity	Merujuk kepada penyakit atau keadaan lain yang mungkin dimiliki pesakit selain dari keadaan utama yang berkaitan. Ianya penting kerana boleh mempengaruhi perkembangan penyakit, hasil rawatan, dan kesihatan keseluruhan pesakit.
Data Makmal	Merangkumi keputusan dari ujian yang dilakukan di dalam makmal. Data makmal adalah penting untuk diagnosis keadaan pesakit, pemantauan perkembangan penyakit, dan penilaian keberkesanan rawatan.
Data Klinikal	Data yang dikumpulkan melalui interaksi antara pesakit dan pakar kesihatan. Data klinikal memberikan pandangan menyeluruh tentang status kesihatan dan sejarah rawatan pesakit.
Hasil	Merujuk kepada hasil atau keputusan akhir penjagaan kesihatan yang diterima oleh pesakit yang boleh diukur untuk mencerminkan keberkesanan intervensi.

1.6 KEPENTINGAN KAJIAN

Penyelidikan ini boleh meningkatkan ketepatan ramalan hasil angioplasti dengan lebih baik yang membawa kepada penjagaan pesakit dan pengurusan sumber yang lebih

efisien khususnya kepada bidang kardiologi intervensi. Secara tidak langsung, ia sejajar dengan matlamat yang lebih luas untuk mengintegrasikan pembelajaran mesin dalam bidang perubatan untuk meningkatkan ketepatan menganalisa keputusan klinikal dan hasil prosedur pembedahan terhadap pesakit.

1.7 STRUKTUR TESIS

Laporan ini menerangkan keseluruhan proses pembangunan projek iaitu bermula dari proses perancangan hingga ke proses mengenalpasti model terbaik. Setiap tahap akan diulas dengan terperinci dan dipecahkan kepada lima bab. Setiap bab yang terdapat di dalam kajian ini secara ringkasnya adalah seperti berikut:

Bab II menerangkan berkenaan kajian-kajian yang telah dijalankan oleh penyelidik-penyelidik lain berkaitan dengan pengaplikasian model-model pembelajaran mesin bagi mendapatkan hasil ramalan terhadap prosedur angioplasti atau PCI.

Bab III menjelaskan proses-proses yang digunakan dalam kajian ini. Metodologi yang diterapkan dalam pembangunan model pengklasifikasian hasil angioplasti akan dibincangkan yang mencakupi perbandingan dan pengujian antara teknik pengklasifikasian. Bab ini juga akan menghuraikan tentang pemilihan fitur terbaik terhadap data berkaitan prosedur angioplasti.

Bab IV akan membentangkan hasil penelitian serta penilaian terhadap teknik pengklasifikasian yang digunakan dalam mengklasifikasikan hasil angioplasti.

Bab V merupakan bab penutupan. Bab ini akan merangkumkan keseluruhan kajian dan memberikan saranan penambahbaikan untuk pengkaji bagi bidang perubatan khususnya berkaitan angioplasti di masa yang akan datang.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Bab ini membicarakan tentang kajian-kajian terdahulu mengenai teknik-teknik pengelasan yang digunakan pada data pesakit yang menjalani prosedur angioplasti koronari dan pemasangan stent atau PCI. Terdapat banyak faktor risiko kematian selepas PCI yang telah dikenal pasti (Edward et al. 2023), termasuklah pembolehubah klinikal dan anatomi. Pembolehubah klinikal antaranya termasuklah umur, jantina, diabetes mellitus, penyakit paru-paru kronik, serangan sakit jantung sebelumnya, kegagalan fungsi ventrikel kiri, kegagalan buah pinggang, kejutan kardiogenik, arteri koronari utama kiri (LMCA), dan umur yang merupakan faktor jangkaan yang paling penting bagi kematian selepas pembedahan (Duggal et al. 2018). Oleh itu, kematian selepas pembedahan kekal sebagai isu utama dalam prosedur PCI.

2.2 KAJIAN MENGENAI PEMBELAJARAN MESIN MENGGUNAKAN DATA PROSEDUR *PERCUTANEOUS CORONARY INTERVENTION* (PCI)

Niimi et al. (2022) membicarakan keberkesanan model pembelajaran mesin (ML) dalam meramalkan risiko pasca-PCI berbanding skor risiko *National Cardiovascular Data Registry* (NCDR-CathPCI). Melalui analisis data dari 22,958 pesakit yang menjalani PCI dari 2008 hingga 2020, model XGBoost menunjukkan peningkatan diskriminasi yang sederhana untuk kejadian kecederaan buah pinggang akut (AKI) dan pendarahan. Walaubagaimana pun, risiko kematian di hospital pula menunjukkan hasil yang sebaliknya. Walaupun keberkesanan model dalam meramal risiko AKI dan pendarahan adalah tinggi, model XGBoost menghadapi masalah dalam meramal risiko bagi kematian di hospital bagi kalangan pesakit risiko rendah. Salah satu kelebihan yang dinyatakan melalui kajiannya adalah kemampuan model ML dalam memproses dan menganalisis jumlah data yang besar menggunakan pemboleh ubah yang kompleks.

Model pembelajaran mesin khususnya XGBoost, menunjukkan potensi dalam meningkatkan kualiti ramalan berbanding dengan pendekatan tradisional. Ini dapat dilihat melalui hasil kajiannya dalam meramal AKI dan pendarahan pasca-PCI. Untuk membuat model menjadi lebih tepat, beliau menyarankan peningkatan dalam pemilihan fitur dan teknik kolaborasi. Penggunaan set data yang lebih besar dan lebih pelbagai juga dapat membantu model mengurangkan *overfitting*. Selain itu, integrasi data klinikal yang lebih luas dan penggunaan teknik ML yang lebih canggih boleh meningkatkan ketepatan model dalam membuat ramalan.

Zhao et al. (2022) pula telah membangunkan model untuk meramal risiko kematian pesakit wanita di hospital (bagi semua punca kematian) yang disahkan mengalami STEMI dengan menggunakan teknik regresi logistik dan hutan rawak RF. Data dari *National Inpatient Sample* (NIS) dari tahun 2011 hingga 2013 telah digunakan pada kajian ini. Ia melibatkan pembangunan tiga model iaitu regresi logistik, hutan rawak penuh, dan hutan rawak tereduksi. Model regresi logistik yang merangkumi 11 fitur, menunjukkan indeks C yang setanding dengan model hutan rawak sekaligus menandakan kemampuan model untuk membezakan antara pesakit yang berisiko tinggi dan rendah terhadap kematian di hospital. Kelebihan kajian ini adalah kemampuannya untuk memproses dan menganalisis set data yang besar dengan struktur yang kompleks. Mereka juga menganalisis pola dan hubungan yang mungkin tidak dikenali melalui analisis statistik tradisional. Kajian ini mengenal pasti faktor risiko dan memperbaiki ketepatan ramalan bagi risiko kematian di hospital. Untuk meningkatkan ketepatan model, mereka menyarankan integrasi data klinikal yang lebih luas, termasuk penggunaan ubat-ubatan, faktor fizikal serta data makmal yang tidak terdapat pada pangkalan data NIS. Penggunaan fitur tambahan ini dapat membantu dalam mengenal pasti faktor risiko yang lebih spesifik dan relevan sehingga meningkatkan kemampuan ramalan model.

Kajian oleh Deng et al. (2022) telah menggunakan algoritma pembelajaran mesin (ML) untuk membangunkan model yang mampu meramal kejadian ketiadaan aliran semula (NR) dan kematian di hospital dalam kalangan pesakit yang mengalami STEMI dan pesakit yang menjalani intervensi koronari perkutaneus primer (pPCI). Dengan menggunakan empat algoritma ML iaitu hutan rawak (RAN), pohon keputusan

(CTREE), mesin sokongan vektor (SVM), dan algoritma jaringan saraf (NNET), mereka menemukan bahawa model RAN menunjukkan prestasi terbaik dalam meramalkan NR dengan $AUC = 0.7891$ dan kematian di hospital dengan $AUC = 0.9273$. Antara kelebihan pendekatan yang dijalankan ini adalah keupayaan algoritma ML dalam menangani data berskala besar dan kompleks, yang membolehkan penyelidik mengenal pasti faktor risiko yang relevan dan seterusnya berjaya membangunkan model yang lebih baik. Ketepatan yang tinggi pada model RAN menunjukkan potensi besar pembelajaran mesin dalam memperbaiki pengurusan data klinikal pesakit serta meningkatkan keupayaan ramalan terhadap komplikasi selepas pPCI. Untuk meningkatkan keupayaan model, mereka mencadangkan penambahan dan penyelidikan fitur baru yang mungkin mempengaruhi risiko NR dan kematian di hospital. Ini termasuklah faktor-faktor seperti penggunaan ubat-ubatan khusus selama pPCI, teknik intervensi, dan kesan fisiologi pesakit terhadap rawatan. Pemilihan fitur yang lebih berkesan dan penggunaan model hibrid yang menggabungkan beberapa algoritma ML dianggarkan mampu meningkatkan ketepatan dan generalisasi model dalam konteks klinikal yang lebih luas.

Kajian Jesús Sampedro-Gómez et al. (2020) memfokuskan pada pembangunan model pembelajaran mesin (ML) untuk meramal restenosis stent (SR) selepas prosedur PCI pada pesakit yang mengalami STEMI. Dengan menggunakan data pengujian GRACIA-3, kajian ini menganalisis ciri-ciri demografi harian, klinikal, dan angiografi. Enam model ML digunakan iaitu hutan rawak (RF), pohon rawak ekstrem (ERT), *gradient boosting* (GB), mesin sokongan vektor (SVC), *L2-regularized logistic regression* (LR) serta pengklasifikasi pohon rawak ekstrem (ERT). Model ERT menunjukkan prestasi terbaik dengan kawasan di bawah lengkungan (AUC-PR) sebanyak 0.46, melebihi skor risiko klinikal yang terdapat pada model ML lain dalam ramalan SR. Kelebihan kaedah ini adalah keupayaannya untuk mengintegrasikan dan menganalisis pelbagai jenis data serta menilai faktor-faktor yang mempengaruhi SR. Pendekatan menggunakan ERT meningkatkan potensi pengurusan pesakit pasca-PCI yang lebih baik dengan menyediakan ramalan yang lebih tepat dan perawatan secara khusus. Untuk meningkatkan kecekapan model, mereka menyarankan penambahbaikan dalam proses pemilihan fitur dan teknik validasi luaran. Penggunaan set data yang lebih besar dan lebih beragam dapat membantu meningkatkan generalisasi dan kejituan

model. Selain itu, kajian lebih lanjut pada pengembangan algoritma ML yang dapat menyesuaikan dengan perubahan dalam amalan klinikal dan teknologi stent dapat memungkinkan adaptasi yang lebih baik terhadap keperluan pesakit. Integrasi pemantauan pasca-PCI jangka panjang ke dalam model pembelajaran mesin juga dapat memberikan pemerhatian lebih lanjut untuk mengenal pasti lebih awal pesakit yang berisiko tinggi mengalami SR, memungkinkan pengelolaan awal untuk mencegah kejadian tersebut.

Kajian oleh Liu et al. (2021) menjelaskan bagaimana teknik pembelajaran mesin (ML) dapat digunakan untuk meramalkan kematian bagi jangka masa panjang terhadap pesakit arteri koronari yang menjalani PCI. Dengan menggunakan data dari 9,680 orang pesakit dan 87 faktor risiko yang dikenalpasti, enam model ML iaitu mesin sokongan vektor (SVM), pohon keputusan (DT), hutan rawak (RF), Pohon Keputusan dengan *Gradient Boosting* (GBDT), rangkaian neural (NN) dan regresi logistik (LR) dilatih dan dibandingkan. Model RF menunjukkan prestasi terbaik dengan AUC sebanyak 0.71 ± 0.04 , yang menandakan ketepatan yang sederhana dalam meramalkan kematian pesakit dalam jangka masa panjang. Ini menunjukkan peningkatan yang signifikan berbanding model risiko tradisional dan mereka menekankan potensi ML dalam meningkatkan stratifikasi risiko bagi pesakit yang menjalani PCI. Untuk meningkatkan keupayaan model, kajian ini mencadangkan penambahbaikan dalam pemilihan fitur dan teknik validasi luaran. Penggunaan algoritma ML yang lebih maju, seperti pembelajaran mendalam mungkin dapat membantu meningkatkan keupayaan model ramalan hasil prosedur PCI.

Kajian oleh Hamilton et al. (2024) membentangkan suatu pendekatan inovatif yang menggabungkan model pembelajaran mesin (ML) dengan input rujukan pesakit untuk memperbaiki ramalan risiko komplikasi pasca-PCI. Kajian ini menggunakan data dari 107,793 prosedur PCI dari 48 hospital di Michigan melalui rekod BMC2 dan validasi luaran pada pangkalan data COAP dengan 56,583 prosedur di Washington. Model XGBoost menunjukkan kemampuan model dipertingkatkan dalam meramal berbagai hasil prosedur termasuk kadar risiko kematian di hospital, AKI, dialisis baru, stroke, pendarahan major, dan transfusi. Salah satu kelebihan pendekatan ini adalah menggabungkan data rujukan pesakit ke dalam model ramalan di mana ia mampu

meramal risiko yang lebih relevan melalui maklumat pesakit dan menyediakan sokongan keputusan terhadap pengurusan pesakit. Model ini menunjukkan bahawa pembelajaran mesin khususnya algoritma XGBoost dapat memberikan ramalan yang lebih tepat jika dibandingkan dengan kaedah tradisional di mana ia memanfaatkan kompleksiti data klinikal dan data rujukan pesakit dengan lebih efektif. Untuk meningkatkan ketepatan model, kajian ini menyarankan penambahan lebih banyak pemboleh ubah klinikal dan karakteristik pesakit ke dalam model pembelajaran mesin, serta penerapan teknik ML yang lebih lanjut untuk mengkaji hubungan non-linear antara faktor-faktor risiko dan hasil prosedur.

Al'Aref et al. (2019) dalam kajiannya bertajuk "*Determinants of In-Hospital Mortality After Percutaneous Coronary Intervention: A Machine Learning Approach*", beliau menyelidik faktor jangkaan kematian di hospital di kalangan pesakit yang menjalani PCI menggunakan algoritma pembelajaran mesin. Dengan menggunakan data dari 479,804 pesakit dari New York Percutaneous Coronary Intervention Reporting System antara tahun 2004 hingga 2012, algoritma AdaBoost menunjukkan hasil terbaik dengan AUC sebanyak 0.927. Ini membuktikan kemampuan diskriminatif yang sangat baik untuk meramalkan kematian di hospital. Faktor penting yang digunakan termasuk usia, fraksi ejeksi, dan beberapa keadaan klinikal lain sebagai faktor ramalan yang signifikan terhadap kematian di hospital. Kelebihan dari penggunaan pembelajaran mesin dalam konteks ini termasuklah keupayaan untuk mengolah dan menganalisis data yang besar dan kompleks untuk mengenalpasti pola dan hubungan yang tidak dapat dikesan melalui analisis statistik tradisional. Ini membuka wawasan baru dalam meramal faktor risiko kematian di hospital dari pasca-PCI dan memungkinkan pembangunan model ramalan yang lebih tepat. Untuk menambahkan lagi ketepatan model, Al'Aref et al. (2019) menyarankan integrasi data yang lebih luas, termasuk maklumat tentang penggunaan ubat-ubatan dan faktor fisiologi yang lebih terperinci, yang mungkin mempengaruhi risiko kematian di hospital. Adaptasi model untuk memasukkan data baharu secara real-time dan penyesuaian terhadap perubahan dalam amalan klinikal juga dapat memberikan hasil jangkaan yang lebih tepat dan bermanfaat dalam pengambilan keputusan klinikal.

Kajian Zack et al. (2019) membentangkan penggunaan algoritma pembelajaran mesin, khususnya regresi hutan rawak, untuk meramalkan hasil bagi pasien yang menjalani intervensi koronari perkutaneus (PCI). Analisis dilakukan menggunakan 52 fitur yang diterima semasa kemasukan pesakit untuk meramalkan risiko kematian di hospital dan 358 fitur yang tersedia semasa pesakit keluar dari hospital sebagai data yang digunakan untuk meramal risiko pesakit kembali dirawat di hospital kerana kegagalan jantung kongestif (CHF). Model pembelajaran mesin, khususnya regresi hutan rawak, berhasil mengenal pasti kohort berisiko tinggi berlakunya kematian di hospital (2% pesakit dengan kadar kematian 45.5%) dan kembali di rawat di hospital kerana CHF dalam 30 hari pasca-pelepasan. Model ini menunjukkan prestasi yang sangat baik jika dibandingkan dengan model regresi logistik tradisional dalam meramalkan risiko pesakit dirawat kembali di hospital kerana CHF dan kematian kardiovaskular dalam tempoh 180 hari. Model regresi hutan rawak menunjukkan AUC yang lebih tinggi iaitu 0.90. Kelebihan utama dari pendekatan ini ialah kemampuannya untuk mengintegrasikan dan menganalisis berbagai pemboleh ubah secara komprehensif dan ini dapat membantu para doktor dalam mengenal pasti pesakit berisiko tinggi yang mungkin memerlukan intervensi atau pemantauan yang lebih intensif. Kajian ini menunjukkan potensi teknik pembelajaran mesin dalam meningkatkan stratifikasi risiko dan membezakan kelompok-kelompok pesakit berisiko tinggi pasca-PCI. Mereka juga menyarankan pendekatan yang lebih komprehensif dan berasaskan data dalam pengambilan keputusan klinikal.

Mortazavi et al. (2019) di dalam kajiannya yang bertujuan untuk menilai kemampuan teknik pembelajaran mesin (ML) dalam meramal risiko pendarahan major pasca-prosedur intervensi koronari perkutaneus (PCI), beliau telah menggunakan data dari National Cardiovascular Data Registry (NCDR) CathPCI Registry untuk membandingkan regresi logistik dengan regularisasi lasso dan boosting gradien (XGBoost) terhadap skor risiko sederhana dan model NCDR lengkap. Hasil kajian beliau menunjukkan bahawa XGBoost telah mencapai statistik C sehingga 0.82, meningkatkan pengelasan kes risiko tinggi sebanyak 3.7% dan kes risiko rendah sebanyak 1.0% jika dibandingkan dengan model NCDR. Antara kelebihan menggunakan pendekatan ML khususnya XGBoost adalah kemampuannya untuk menangani pola non-linear dan hubungan antara pemboleh ubah yang kompleks secara

efisien yang sukar dicapai melalui teknik statistik tradisional. Pendekatan ini membolehkan model untuk menganalisis dan meramal dengan lebih tepat berkenaan risiko pendarahan pasca-PCI, membantu pakar perubatan merancang pelan perawatan yang lebih efektif serta mengurangkan risiko pada pesakit. Untuk meningkatkan akurasi model lebih lanjut, mereka turut menyarankan peningkatan dalam strategi imputasi data dan penyesuaian model untuk menangani faktor prediktif tambahan yang mungkin belum sepenuhnya dieksplorasi dalam set data NCDR.

Burrello et al. (2022) mengkaji penggunaan algoritma pembelajaran mesin (ML) untuk meramalkan kebarangkalian kematian pasca-prosedur intervensi koronari perkutaneus (PCI) pada pesakit dengan lesi bifurkasi (Bifurcation Lesions) dengan menggunakan data dari 4,094 pesakit yang dirawat dengan stent yang sangat nipis untuk lesi bifurkasi dan menganalisis menggunakan model regresi hutan rawak, kajian ini berhasil membina model RAIN-ML yang menunjukkan kemampuan ramalan yang baik dengan nilai AUC yang tinggi dalam validasi internal dan eksternal. Ini menandakan model tersebut efektif dalam mengkategorikan pesakit ke dalam kumpulan risiko yang berbeza untuk pelbagai risiko kematian. Salah satu kelebihan utama dari pendekatan ini adalah kemampuannya untuk mengintegrasikan dan menganalisis secara komprehensif berbagai fitur klinikal, anatomi dan prosedural yang relevan dengan hasil prosedur PCI terhadap pesakit. Ia dapat menjangkakan keputusan atau kesan perawatan yang lebih tepat dan dapat membantu pengurusan pasca-PCI bagi pesakit lesi bifurkasi. Untuk meningkatkan ketepatan model, mereka menyarankan penambahan lebih banyak data atau fitur pesakit dan karakteristik prosedur ke dalam model ML, serta penerapan teknik ML yang lebih canggih dan diversifikasi model ML untuk mengkaji hubungan yang lebih kompleks antara faktor-faktor risiko dan hasil prosedur. Integrasi data longitudinal yang merangkumi penggunaan ubat-ubatan pasca-PCI, kepatuhan terhadap cadangan gaya hidup, dan pemantauan jangka panjang pasca-PCI dapat memberikan dimensi tambahan yang berguna untuk ramalan yang lebih baik. Pembaharuan model secara berkala dengan data baru dan validasi silang di antara populasi yang berbeza juga mampu meningkatkan kejituan dan kebolehan model terhadap amalan klinikal.

Kajian oleh Gldener et al. (2023) telah membentangkan potensi pembelajaran mesin, khususnya melalui penggunaan peta mandiri (*self-organizing maps*, SOMs),

dalam mengidentifikasi faktor risiko restenosis in-stent (ISR) pasca-intervensi koronari perkutaneus (pasca-PCI) dengan menganalisis data dari 10,004 pesakit. Kajian ini bertujuan untuk mengungkap pola dalam data yang mungkin terlepas pandang oleh penganalisis konvensional yang mana ianya mungkin dapat menambahbaik proses meramal ISR untuk 6 hingga 8 bulan selepas prosedur. Kelebihan dari pendekatan pembelajaran mesin ini termasuklah kemampuannya untuk mengintegrasikan dan memvisualisasikan data yang kompleks dari berbagai jenis, sehingga memungkinkan pengelasan faktor risiko baru yang tidak dapat ditemukan melalui analisis statistik tradisional. SOMs mampu memvisualisasikan data multidimensi dalam bentuk yang lebih mudah difahami yang membantu dalam mengenali pola kompleks berkaitan risiko restenosis. Untuk meningkatkan ketepatan ramalan model, mereka menyarankan penambahan lebih banyak data klinikal dan pemboleh ubah prosedural ke dalam analisis pembelajaran mesin. Eksplorasi lebih lanjut terhadap teknik pembelajaran mesin yang berbeza, seperti jaringan saraf tiruan (NN) dan pembelajaran mendalam (*deep learning*) dapat memberikan wawasan baru dan meningkatkan kemampuan ramalan model terhadap ISR. Selain itu, validasi luaran model di pelbagai populasi dan data klinikal akan membantu memastikan bahawa model tersebut dapat diaplikasikan dengan meluas berikutan ketepatannya yang tinggi. Kajian ini menunjukkan potensi signifikan pembelajaran mesin dalam memperkaya pemahaman mengenai faktor-faktor yang mempengaruhi risiko restenosis setelah pemasangan stent dan memudahkan pakar perubatan merancang kaedah rawatan perubatan yang lebih spesifik dan berasaskan data untuk mengurangkan insiden restenosis di masa hadapan.

Rios et al. (2022) pula mengkaji jumlah pemboleh ubah yang diperlukan secara manual dan pemboleh ubah pengimejan untuk mengekalkan ketepatan prognostik dalam meramalkan kejadian kardiovaskular utama (MACE) menggunakan model pembelajaran mesin (ML), khususnya pada pesakit yang menjalani pengimejan perfusi miokardium single-photon emission computed tomography (SPECT). Dengan menganalisis data dari 20,414 pesakit dalam daftar REFINE SPECT dan 2,984 pesakit dari Universiti Calgary untuk pengujian luaran, kajian ini membuktikan bahawa model ML yang dikurangkan pemboleh ubahnya mampu mencapai ketepatan prognostik yang hampir sama dengan model ML yang menggunakan semua pemboleh ubah. Mereka menggunakan hanya 12 dari 40 pemboleh ubah input manual dan 11 dari 58 pemboleh

ubah pengimejan. Antara kelebihan dari pendekatan ini termasuklah penggunaan data yang lebih efisien di mana model ML yang lebih sederhana dan memerlukan input yang lebih sedikit masih dapat memberikan ramalan yang tepat untuk MACE. Ini akan mengurangkan beban kerja dalam pengumpulan data dan memungkinkan aplikasi ML dalam keadaan kekurangan data klinikal dengan sumber yang terbatas. Untuk meningkatkan ketepatan model ML akan datang, kajian ini menyarankan eksplorasi pemboleh ubah tambahan yang mungkin memiliki nilai prognostik tetapi belum termasuk di dalam set pemboleh ubah minimum. Teknik ML yang lebih canggih seperti pembelajaran mendalam boleh diterapkan untuk mengkaji hubungan non-linear dan interaksi kompleks antara pemboleh ubah yang mungkin tidak terkesan oleh model ML yang lebih sederhana. Selain itu, validasi luaran model yang pelbagai beserta data klinikal akan membantu memastikan kejituan model ramala. Hal ini memungkinkan penggunaannya secara luas dalam amalan perubatan untuk menambah baik pengurusan pesakit arteri koronari.

Kajian oleh Galimzhanov et al. (2023) bertujuan membangunkan model pembelajaran mesin (ML) untuk meramalkan risiko kematian dalam pelbagai sebab, kejadian serebrovaskular iskemik (CVE), dan pendarahan major pada pesakit yang dirawat di hospital setelah menjalani prosedur intervensi koronari perkutaneus (PCI). Kajian ini menggunakan data dari National Inpatient Sample (NIS) bagi tahun 2016-2019. Mereka menggunakan lima algoritma ML umum iaitu regresi logistik, Mesin Sokongan Vektor (SVM), NB, hutan rawak (RF) dan XGBoost. Hasil kajian ini menunjukkan XGBoost memberikan keputusan terbaik dengan AUC 0.86 yang mana ia menandakan kebolehan ramalannya yang amat baik. Kelebihan pendekatan ML terutama melalui XGBoost di dalam kajian ini merangkumi kemampuannya untuk mengolah set data yang besar dan kompleks. Ia berkeupayaan mengesan pola prediktif yang lebih tepat untuk hasil klinikal pasca-PCI. Ini dapat membantu para doktor merancang rawatan yang terbaik dan mengurangkan risiko selepas prosedur.

Kajian oleh Huang, Yen-Chun et al. (2022) yang bertujuan untuk membangunkan model ramalan terhadap risiko kematian menggunakan kaedah pembelajaran mesin (ML) pada pesakit yang menjalani PCI tiga pembuluh. Dengan menggunakan set data *Taiwan National Health Insurance Research*, kajian ini

melibatkan pemilihan 42 faktor risiko yang mempengaruhi kelangsungan hidup, termasuk maklumat demografi, ciri-ciri klinikal, dan sejarah penyakit lain. Model pembelajaran gabungan yang diusulkan berhasil mengklasifikasikan pesakit dengan tahap ketepatan sebanyak 88.7%. Mereka mendapati bahawa fitur usia, gagal jantung kongestif, dan kegagalan ginjal kronis sebagai fitur paling penting untuk mengesan risiko berlakunya kematian. Kelebihan pendekatan ini terletak pada kemampuannya untuk meramal dengan ketepatan yang tinggi terhadap risiko kematian. Ia mengintegrasikan dan menganalisis pelbagai set data klinikal dan demografi secara komprehensif di mana ia memungkinkan pengesanan pesakit berisiko tinggi yang mungkin memerlukan rawatan khusus atau pemantauan yang lebih lanjut. Mereka mencadangkan peningkatan model pada masa akan datang dengan menyarankan peningkatan dalam proses pemilihan fitur dan penerapan model ML yang lebih canggih, seperti jaringan saraf tiruan atau pembelajaran mendalam, yang dapat mengesan hubungan non-linear dan interaksi kompleks antara fitur dengan lebih efektif.

Kajian Jun Ke et al. (2022) yang bertajuk "*Machine learning-based in-hospital mortality prediction models for patients with acute coronary syndrome*" memfokuskan kepada proses mengenal pasti faktor risiko kematian di hospital pada pesakit dengan sindrom koroner akut (ACS) dan pengujian perbezaan model ramalan tradisional dan pembelajaran mesin. Data dikumpulkan dari pesakit ACS yang dirawat di Hospital Fujian Provincial dari 1 Januari 2017 hingga 31 Mac 2020. Empat algoritma pembelajaran mesin (regresi logistik, pohon keputusan *gradient boosting*, hutan rawak, dan SVM) dibangunkan untuk model ramalan dengan pengujian model menggunakan sensitiviti, spesifisiti, dan lengkungan karakteristik operasi penerima (ROC). Model pembelajaran mesin, khususnya model *gradient boosting* dan hutan rawak, menunjukkan prestasi terbaik dalam meramalkan risiko kematian di hospital berbanding model regresi logistik tradisional, dengan model *gradient boosting* dan hutan rawak di mana ia menunjukkan AUC yang lebih tinggi. Faktor bebas yang disifatkan sebagai faktor risiko kematian di hospital antaranya adalah umur, jenis NSTEMI, kelas Killip III dan IV, D-dimer, cTnI, CK, NT-proBNP, kolesterol HDL, dan penggunaan statin. Kelebihan model pembelajaran mesin dalam kajian ini terletak pada kemampuan mereka untuk mengintegrasikan dan menganalisis pemboleh ubah yang luas dan kompleks untuk menghasilkan ramalan yang lebih tepat.

Ringkasan kajian literatur mengenai penggunaan pembelajaran mesin dalam meramal hasil intervensi koronari perkutaneus (PCI) boleh didapati dalam Jadual 2.1.

LIBRARY FTSM

Jadual 2.1 Ringkasan kajian literatur berkaitan aplikasi pembelajaran mesin dalam meramal hasil intervensi koronari perkutaneus (PCI)

No.	Penulis, Tahun	Objektif	Algoritma	Dapatan
1	Niimi et al. (2022)	Membincangkan keberkesanan model pembelajaran mesin dalam meramalkan risiko pasca-PCI berbanding skor risiko NCDR-CathPCI.	Regresi Logistik, XGBoost dan <i>Gradient Boosting</i> .	XGBoost menunjukkan peningkatan diskriminasi yang sederhana untuk kejadian AKI dan pendarahan. Namun, ia menghadapi masalah dalam meramal risiko kematian di hospital bagi pesakit risiko rendah. Model menunjukkan potensi dalam meningkatkan kualiti ramalan berbanding dengan pendekatan tradisional.
2	Zhao et al. (2022)	Membangunkan model untuk meramal risiko kematian di hospital bagi pesakit wanita yang mengalami STEMI.	Regresi Logistik, Hutan Rawak Penuh dan Tereduksi.	Model hutan rawak penuh menunjukkan AUC terbaik sebanyak 0.90, menandakan kemampuannya yang lebih unggul dalam meramal kematian di hospital jika dibandingkan dengan model regresi logistik dan hutan rawak tereduksi yang memiliki Indeks C setara.
3	Deng et al. (2022)	Membangunkan model yang mampu meramal kejadian ketiadaan aliran semula (NR) dan kematian di hospital dalam kalangan pesakit yang mengalami STEMI dan menjalani pPCI.	Hutan Rawak (RAN), Pohon Keputusan (CTREE), Mesin Sokongan Vektor (SVM) dan Algoritma Jaringan Saraf (NNET).	Model RAN menunjukkan prestasi terbaik dengan AUC 0.7891 untuk NR dan AUC 0.9273 untuk kematian di hospital. Keupayaan model ini dalam menangani data besar dan kompleks membolehkan pengesanan faktor risiko yang relevan yang berpotensi memperbaiki pengurusan data klinikal pesakit serta meningkatkan keupayaan ramalan terhadap komplikasi selepas pPCI.
4	Jesús Sampeder-Gómez et al. (2020)	Membangunkan model pembelajaran mesin untuk meramal restenosis stent selepas PCI pada pesakit STEMI.	Hutan Rawak (RF), Pohon Rawak Ekstrem (ERT), Gradient Boosting (GB), Mesin Sokongan Vektor (SVC), Regresi Logistik L2-regularized (LR), Pengklasifikasi Pohon Rawak Ekstrem (ERT)	Model ERT menunjukkan prestasi terbaik dengan Area Under the Curve (AUC-PR) sebanyak 0.46 yang lebih baik dari skor risiko klinikal yang terdapat pada model ML lain dalam ramalan restenosis stent.
5	Liu et al. (2021)	Meramalkan kematian jangka panjang bagi pesakit arteri koronari yang menjalani PCI.	Mesin Sokongan Vektor (SVM), Pohon Keputusan (DT), Hutan Rawak (RF), Pohon Keputusan dengan <i>Gradient Boosting</i> (GBDT), Neural Network (NN), Regresi Logistik (LR).	Model Hutan Rawak (RF) menunjukkan prestasi terbaik dengan AUC 0.71±0.04, menunjukkan ketepatan yang sederhana dalam meramalkan kematian pesakit dalam jangka masa panjang. Kajian ini menunjukkan peningkatan yang signifikan berbanding model risiko tradisional dan menekankan potensi ML dalam meningkatkan stratifikasi risiko bagi pesakit yang menjalani PCI.

bersambung...

...sambungan

- | | | | | |
|---|-------------------------|--|--|--|
| 6 | Hamilton et al. (2024) | Menggabungkan model pembelajaran mesin dengan input rujukan pesakit untuk memperbaiki ramalan risiko komplikasi pasca-PCI. | XGBoost. | Model XGBoost menunjukkan kemampuan dalam meramal berbagai hasil prosedur termasuk risiko kematian di hospital, AKI, strok, dan pendarahan. Pendekatan ini membolehkan penggabungan data rujukan pesakit ke dalam model ramalan yang membantu dalam meramal risiko yang lebih relevan. |
| 7 | Al'Aref et al. (2019) | Menyelidik faktor jangkaan kematian di hospital di kalangan pesakit yang menjalani PCI menggunakan algoritma pembelajaran mesin. | Adaptive Boosting (AdaBoost), XGBoost, dan Regresi Logistik. | Algoritma AdaBoost menunjukkan hasil terbaik dengan kawasan di bawah lengkungan (AUC) sebanyak 0.927, menandakan keupayaan diskriminatif yang tinggi untuk meramal kematian di hospital. |
| 8 | Zack et al. (2019) | Menentukan sama ada pembelajaran mesin boleh digunakan untuk mengenal pasti pesakit yang berisiko tinggi untuk kematian atau dimasukkan semula ke hospital akibat CHF selepas PCI. | Regresi hutan rawak dan regresi logistik (teknik statistik). | Model regresi hutan rawak (pembelajaran mesin) lebih berdaya ramal dan lebih diskriminatif berbanding kaedah regresi standard dalam mengenal pasti pesakit yang berisiko. |
| 9 | Mortazavi et al. (2019) | Menentukan sama ada teknik pembelajaran mesin lebih baik dalam meramalkan pendarahan major selepas PCI berbanding model sedia ada dalam Pendaftaran Data Kardiovaskular Nasional (NCDR). | Regresi logistik dengan regularisasi lasso (teknik statistik) dan XGBoost. | Penggunaan XGBoost dan rangkaian penuh pemboleh ubah terpilih mencapai statistik C sebanyak 0.82 (95% CI, 0.82-0.82), dengan skor F sebanyak 0.31 (95% CI, 0.30-0.31). XGBoost mengenal pasti tambahan 3.7% kes dengan betul. Penggunaan kaedah pembelajaran mesin secara strategik membolehkan penambahbaikan dalam prestasi model ramalan. |

bersambung...

...sambungan

- | | | | | |
|----|---------------------------|---|---|---|
| 10 | Burrello et al. (2022) | Membangun model penstratifikasian risiko berasaskan pembelajaran mesin yang dibina berdasarkan ciri klinikal, anatomi, dan prosedur untuk meramalkan kematian akibat pelbagai punca selepas PCI bifurkasi kontemporari. | Analisis diskriminan linear (LDA), regresor hutan rawak (RF), mesin sokongan vektor (SVM) dengan kernel berbeza, dan hutan pengasingan. | Model RAIN-ML yang menggunakan regresi hutan rawak mencapai nilai AUC yang tinggi, menunjukkan keberkesanan dalam mengkategorikan pesakit ke dalam kumpulan risiko yang berbeza untuk kematian pasca-PCI. |
| 11 | Guldener et al. (2023) | Mengkaji penggunaan algoritma pembelajaran mesin untuk mengidentifikasi faktor risiko restenosis in-stent pasca-PCI. | Peta Mandiri (Self-Organizing Maps, SOMs). | SOMs berhasil mengidentifikasi pola-pola kompleks dalam data yang mempengaruhi risiko restenosis, menunjukkan keberkesanan dalam memvisualisasikan dan mengintegrasikan data yang kompleks untuk analisis risiko restenosis. |
| 12 | Rios et al. (2022) | Mengkaji jumlah pemboleh ubah yang diperlukan untuk meramalkan kejadian kardiovaskular utama (MACE) menggunakan model ML pada pesakit yang menjalani SPECT. | Pembelajaran mesin (tidak dinyatakan). | Model yang dikurangkan pemboleh ubahnya mencapai ketepatan prognostik yang hampir sama dengan model yang menggunakan semua pemboleh ubah. Ia menunjukkan keberkesanan pendekatan ML dalam pengurangan data sambil memelihara ketepatan. |
| 13 | Galimzhanov et al. (2023) | Membangun model pembelajaran mesin untuk meramalkan risiko kematian, kejadian serebrovaskular iskemik, dan pendarahan major pada pesakit yang dirawat di hospital selepas menjalani PCI. | Regresi Logistik, Mesin Sokongan Vektor (SVM), NB, Hutan Rawak (RF) dan XGBoost. | XGBoost memberikan keputusan terbaik dengan AUC 0.86, menunjukkan kebolehan ramalannya yang amat baik dalam mengolah set data yang besar dan kompleks untuk mengesan pola peramalan yang tepat untuk hasil klinikal pasca-PCI. |

bersambung...

...sambungan

- | | | | | |
|----|----------------------|---|---|---|
| 14 | Huang et al. (2022) | Membangunkan model ramalan risiko kematian menggunakan kaedah pembelajaran mesin pada pesakit yang menjalani PCI tiga pembuluh darah. | Model pembelajaran gabungan (<i>ensemble learning models</i>). | Model ini berjaya mengklasifikasikan pesakit dengan ketepatan sebanyak 88.7% dan mendedahkan bahawa usia, kegagalan jantung kongestif, dan kegagalan ginjal kronik sebagai fitur penting. |
| 15 | Jun Ke et al. (2022) | Menganalisis faktor risiko kematian di hospital pada pesakit dengan sindrom koroner akut dan membandingkan model ramalan pembelajaran mesin dengan model tradisional. | Regresi Logistik, Pohon Keputusan dengan <i>Gradient Boosting</i> , Hutan Rawak (RF), Mesin Sokongan Vektor (SVM) | Model pembelajaran mesin, khususnya model gradien boosting dan hutan rawak, menunjukkan prestasi terbaik dalam meramalkan risiko kematian di hospital dengan AUC yang lebih tinggi berbanding model regresi logistik tradisional. |
-

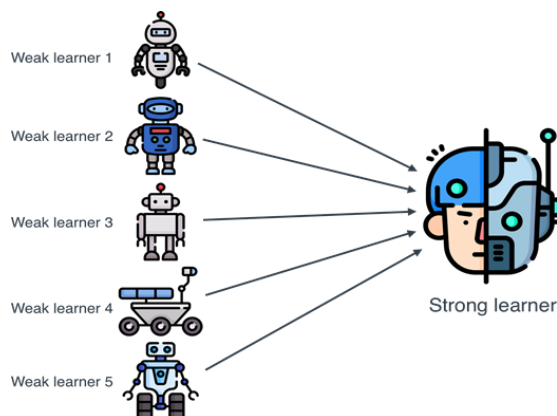
LIBRARY ETSM

Berdasarkan kajian terdahulu, SVM sering digunakan dalam beberapa kajian tetapi tidak menjadi model yang mempamerkan prestasi terbaik. Sebaliknya, model pembelajaran mesin bergabung secara konsisten mengatasi model bukan gabungan dalam meramal kesan buruk dan risiko kematian selepas PCI. Secara khusus, hutan rawak, *gradient boosting*, dan XGBoost sering muncul sebagai model berprestasi terbaik dari segi ketepatan. Dalam penerbitan terdahulu, pelbagai teknik pembelajaran gabungan telah dibincangkan, termasuk AdaBoost, XGBoost, LightGBM, dan CatBoost. Antara kaedah ini, XGBoost dan *gradient boosting* menunjukkan prestasi paling berkesan dalam membangunkan model ramalan bagi hasil PCI bersama-sama dengan hutan rawak. Oleh itu, kajian ini memberi tumpuan kepada tiga model utama iaitu hutan rawak, *gradient boosting*, dan XGBoost.

SVM merupakan algoritma pembelajaran mesin klasik dan membuat ramalan dengan cara mempelajari hiperplane sempadan dalam ruang ciri antara sampel positif dan negatif. Kedua-dua XGBoost dan RF pula merupakan kaedah ML gabungan yang berasaskan model pohon keputusan. XGBoost bertindak lebih pantas dan lebih cekap terutamanya model yang menggunakan set data yang besar dan kompleks. Pembelajaran mendalam (*deep learning*) tidak menunjukkan sebarang peningkatan dalam prestasi untuk data tabular berstruktur berbanding algoritma berasaskan pohon. Bukan itu sahaja, pembelajaran mendalam juga memerlukan pengiraan yang tinggi dan tempoh masa yang lama (Shwartz-Ziv et al. 2022, Grinsztajn et al. 2022). Justeru, kajian ini tidak menggunakan teknik pembelajaran mendalam.

2.3 MODEL PEMBELAJARAN MESIN BERGABUNG

Pembelajaran gabungan merupakan satu teknik yang mampu meningkat keberkesanan dalam pembelajaran mesin di mana sebilangan model yang sering dirujuk sebagai "*weak learners*" dilatih dan digabungkan untuk menyelesaikan masalah pengiraan tertentu dengan lebih berkesan daripada menggunakan satu model sahaja.



Rajah 2.1 Pembelajaran gabungan

Sumber: Encord (2023)

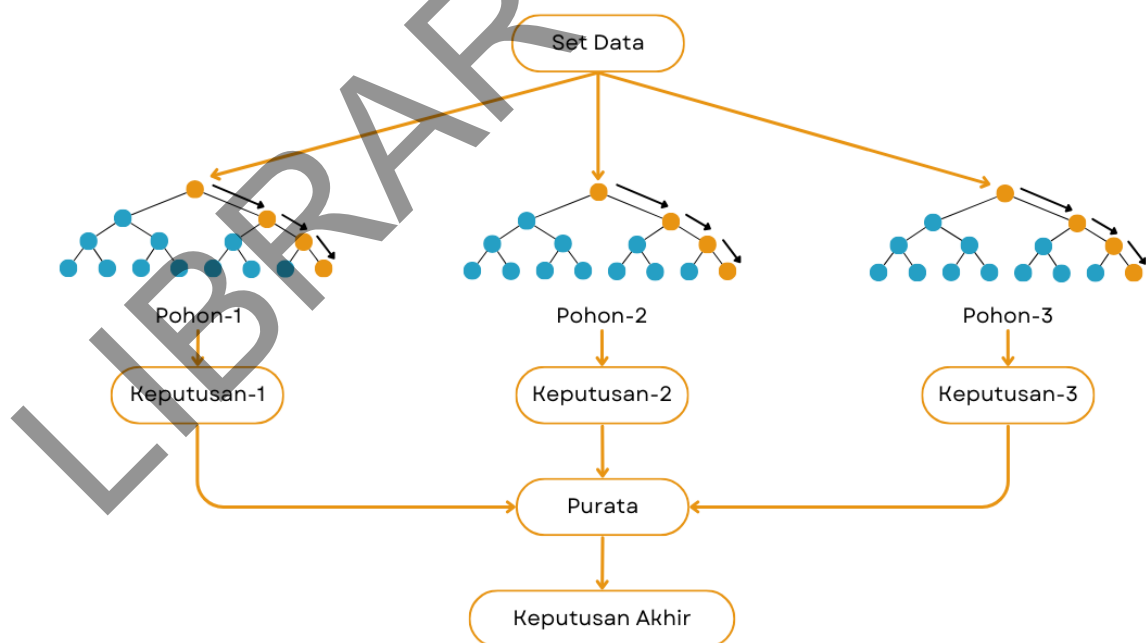
Pembelajaran gabungan meningkatkan prestasi model pembelajaran mesin dengan menggabungkan ramalan daripada berbilang model. Dengan memanfaatkan kekuatan algoritma yang pelbagai, kaedah gabungan bertujuan untuk mengurangkan kedua-dua bias dan varians, dengan itu menghasilkan ramalan yang lebih tepat pada set data yang kompleks serta meningkatkan kebolehpercayaan ramalan. Strategi ini amat berkesan kerana ia menggabungkan ramalan model yang dilatih pada perspektif masalah yang berbeza, meningkatkan keteguhan model kepada ralat dan ketidakpastian. Ini penting dalam bidang seperti penjagaan kesihatan atau kewangan, di mana ketepatan dan kebolehpercayaan adalah perkara yang amat penting.

Teknik seperti *bagging*, *boosting*, dan *stacking* adalah tunggak pembelajaran gabungan. *Bagging* membantu dalam mengurangkan varians, *boosting* digunakan untuk mengurangkan bias, dan tindanan (*stacking*) menggabungkan pelbagai model untuk meningkatkan ketepatan ramalan. Kaedah ini secara keseluruhannya menyumbang kepada prestasi dan kebolehpercayaan sistem pembelajaran mesin sekaligus menjadikannya ia satu keperluan kritikal untuk mereka yang ingin membangunkan penyelesaian ML yang mantap. Dengan menyepadukan pendekatan ini, pembelajaran gabungan bukan sahaja mencapai ketepatan ramalan yang unggul berbanding model tunggal tetapi juga menawarkan daya tahan yang lebih besar terhadap kerumitan yang wujud dalam data dunia sebenar.

2.3.1 Hutan Rawak

Hutan rawak (RF) ialah teknik pembelajaran mesin yang digunakan secara meluas merentasi pelbagai bidang untuk tugas klasifikasi dan regresi. Ia merupakan kaedah pembelajaran gabungan yang membina banyak pohon keputusan semasa latihan dan menghasilkan mod kelas (pengkelasan) atau ramalan min (regresi) bagi pohon individu. Setiap pohon dalam hutan rawak beroperasi pada bahagian data yang dipilih secara rawak dan menilai sebahagian fitur yang dipilih secara rawak bagi setiap pemisahan membuat keputusan. Kaedah rawak ini membantu dalam mencipta set model yang pelbagai serta mengurangkan *overfitting* dan meningkatkan keteguhan model.

Ini dilakukan secara penggantian yang dikenali sebagai *bootstrap*. Setiap pepohon keputusan dibina pada set data *bootstrap* ini dan semasa pembinaan setiap pepohon, subset fitur dipilih secara rawak untuk memisahkan nod, meningkatkan kepelbagaian model dan akhirnya membawa kepada model keseluruhan yang lebih mantap. Struktur asas hutan rawak boleh diperhatikan pada Rajah 2.2.



Rajah 2.2 Struktur hutan rawak

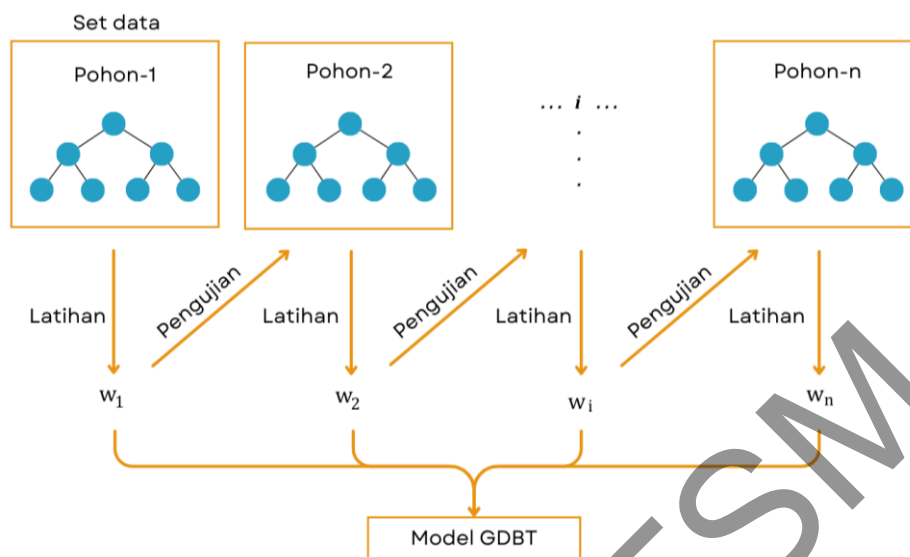
Selain daripada membina setiap pohon menggunakan subset data yang berbeza, hutan rawak berbeza dalam cara pembinaan pohon. Dalam pohon keputusan standard, setiap nod dipecahkan menggunakan keputusan optimum untuk pembahagian antara

semua pembolehubah untuk meminimumkan entropi akibat pembahagian set data yang diwakili oleh nod induk (Wang et al. 2020). Dalam hutan rawak, titik pemisahan setiap nod dipilih secara rawak daripada titik pemisahan terbaik di antara subset peramal. Hutan rawak dengan demikian mengelakkan masalah *overfitting* yang biasa terjadi pada satu pohon keputusan yang mendalam.

2.3.2 Gradient Boosting

Gradient boosting adalah teknik pembelajaran mesin yang popular kerana kemampuannya mencipta model ramalan yang sangat tepat. Dengan menggunakan kaedah ini, model dibina secara berurutan di mana setiap model baru membetulkan kesilapan yang dibuat oleh model sebelumnya, khususnya untuk mengurangkan bias. Pada dasarnya, *gradient boosting* mencipta satu model ramalan yang lemah, biasanya pohon keputusan yang ringkas. Model lemah ini digabungkan untuk membentuk model keseluruhan yang lebih kuat.

Teknik *gradient boosting* digunakan untuk tugas klasifikasi dan regresi, termasuk untuk klasifikasi pelbagai kelas seperti yang terdapat pada kajian ini iaitu untuk meramalkan pelbagai hasil angioplasti. Dalam teknik ini, setiap iterasi melibatkan penambahan pohon keputusan baru yang dibina untuk meramalkan sisa (ralat) daripada pohon sebelumnya. Ketepatan dan kelajuan pelaksanaan *gradient boosting* dapat ditingkatkan dengan membuat pensampelan data latihan secara rawak untuk mengelakkan *overfitting*. Parameter seperti kadar pembelajaran dan jumlah pohon (*n_estimators*) dikawal untuk memastikan model seimbang antara bias dan varian.

Rajah 2.3 Struktur *gradient boosting*

Pohon baru ditambahkan untuk meramal kesilapan pohon sebelumnya yang seterusnya digabungkan bersama ketika model dikemas kini secara berurutan untuk meningkatkan ketepatan. Algoritma ini berfungsi dengan mengoptimumkan fungsi kos (*cost function*) yang mana ia secara beransur-ansur meningkatkan ketepatan model. Ia menumpukan pada meminimumkan kesilapan dengan menyesuaikan pemberat setiap titik data yang diramalkan tidak tepat dalam iterasi sebelumnya. Parameter kadar pembelajaran mengawal keterlibatan setiap pohon bertujuan untuk mengelakkan model dari bertindak balas terlalu kuat terhadap pembetulan serta membantu mengekalkan keseimbangan antara bias dan varian.

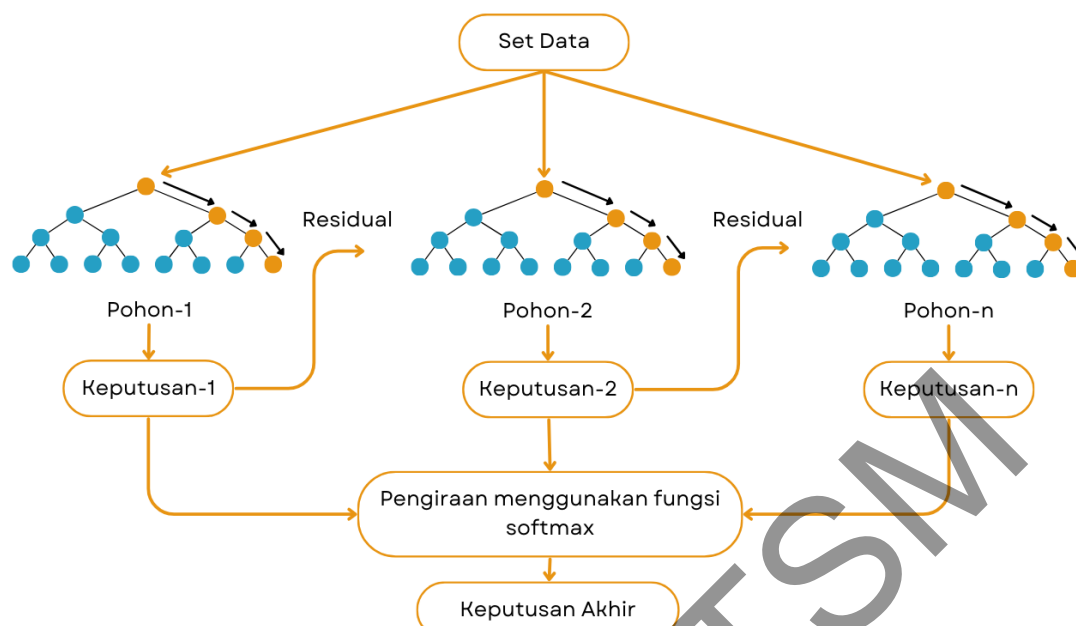
2.3.3 eXtreme Gradient-Boosting (XGBoost)

XGBoost atau *Extreme Gradient Boosting* adalah versi *gradient boosting* yang sangat efisien dan boleh diskala. Ia direka untuk kelajuan serta prestasi dengan mengintegrasikan regularisasi untuk mengelakkan *overfitting* bagi meningkatkan ketepatan ramalannya. XGBoost membina pohon secara berurutan dengan setiap pohon membetulkan kesilapan yang dibuat oleh pohon sebelumnya dengan menggunakan regularisasi lanjutan iaitu *Lasso Regression* dan *Ridge Regression* yang memperbaiki keupayaan generalisasi model. Ia berkesan menangani set data yang besar dan fitur-fitur

yang kompleks, menjadikannya pilihan utama untuk pelbagai masalah pembelajaran mesin yang kompetitif.

XGBoost digunakan secara meluas kerana mampu memberikan prestasi yang efisien dalam tugas-tugas seperti regresi dan klasifikasi serta menangani nilai yang hilang dalam data secara langsung dengan mengekalkan kelajuan dan ketepatan. Ini kerana ia menggunakan campuran pembelajaran gabungan yang dioptimumkan dan menggabungkan ramalan dari pelbagai model lemah untuk menghasilkan ramalan yang lebih kuat. Pemberat diberikan kepada pembolehubah bebas dan disesuaikan berdasarkan kualiti ramalan hasil oleh model. Proses iteratif ini menghasilkan model pembelajaran mesin yang tepat dan boleh dipercayai. XGBoost sesuai untuk beberapa aplikasi termasuk ramalan kadar klik (*click-through rate*), sistem cadangan dan pertandingan Kaggle (Spiceworks 2024).

XGBoost adalah perpustakaan algoritma *gradient boosting* teragih yang dioptimumkan (Gao et al. 2024). Ia melaksanakan algoritma pembelajaran mesin menggunakan rangka kerja pohon keputusan dengan *gradient boosting* (GBDT). Metodologinya melibatkan penciptaan pohon keputusan berurutan dengan setiap pohon meminimumkan residual dari model pohon sebelumnya. Tidak seperti GBDT tradisional yang hanya menggunakan derivatif pertama dari maklumat kesilapan, XGBoost melakukan pengembangan Taylor perintah kedua dari fungsi kos, menggunakan kedua-dua derivatif pertama dan kedua. Selain itu, XGBoost menyokong fungsi kos yang disesuaikan dan pemprosesan selari yang memungkinkan latihan yang efisien pada set data yang besar. Struktur asas XGBoost ditunjukkan dalam Rajah 2.4.



Rajah 2.4 Struktur XGBoost

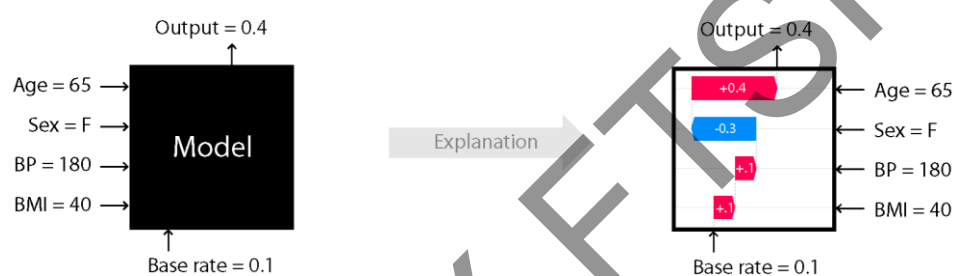
Residual dari pohon-1 dimasukkan ke pohon-2 untuk mengurangkan residual dan proses ini berterusan secara iteratif (Wang et al. 2020). Fungsi *softmax* digunakan untuk pengiraan kebarangkalian ramalan bagi setiap kelas dan kelas dengan kebarangkalian tertinggi dipilih sebagai ramalan akhir.

2.4 SHAPLEY ADDITIVE EXPLANATION (SHAP)

Shapley Additive Explanations atau singkatannya SHAP, telah diperkenalkan oleh Lundberg & Lee pada 2017 sebagai kaedah untuk meningkatkan kebolehdjelasan pasca pemodelan melalui pendekatan model-agnostik dan perkaitan fitur (Minh et al. 2022). SHAP berdasarkan nilai Shapley, yang berasal daripada teori permainan kooperatif. Dalam konteks pembelajaran mesin, SHAP digunakan untuk mengukur sumbangan setiap fitur terhadap ramalan model dengan mengambil kira semua kemungkinan kombinasi fitur. Pendekatan ini memberikan penjelasan yang konsisten dan adil tentang bagaimana setiap fitur mempengaruhi keputusan model (Lundberg & Lee 2017).

Salah satu alat visualisasi utama SHAP ialah plot beeswarm yang memaparkan taburan nilai SHAP bagi semua fitur dalam set data. Plot beeswarm membantu

penyelidik memahami fitur-fitur yang paling mempengaruhi keputusan model serta arah pengaruhnya (positif atau negatif). Dalam plot ini, setiap titik mewakili satu pemerhatian dan warna titik menggambarkan nilai sebenar fitur (rendah hingga tinggi). Susunan fitur adalah berdasarkan kepentingannya, yang mana fitur paling signifikan berada di bahagian atas (Molnar 2022). SHAP dan plot beeswarm banyak digunakan dalam pelbagai aplikasi pembelajaran mesin, terutamanya dalam sektor kesihatan untuk menjelaskan ramalan model yang kompleks seperti dalam diagnostik perubatan atau ramalan hasil klinikal (Lundberg et al. 2018).



Rajah 2.5 Tafsiran model menggunakan SHAP

Sumber: SHAP Documentation (2025)

Rajah 2.5 menunjukkan bagaimana SHAP menerangkan keputusan ramalan model. Model menerima input seperti Age = 65, Sex = F, BP = 180, dan BMI = 40, menghasilkan Output = 0.4 dengan Base rate = 0.1 sebagai nilai asas tanpa pengaruh fitur. SHAP menjelaskan sumbangan setiap fitur kepada keputusan: Age = 65 meningkatkan prediksi sebanyak +0.4, Sex = F mengurangkannya sebanyak -0.3, manakala BP = 180 dan BMI = 40 masing-masing menyumbang +0.1. Analisis ini membantu memahami secara telus bagaimana setiap fitur mempengaruhi keputusan model.

Kajian oleh Betts et al. (2021) adalah bertujuan untuk meramalkan sindrom gangguan pernafasan neonatal (RDS) dan hipoglikemia menggunakan data kesihatan pentadbiran dari Queensland, Australia. Pemodelan yang digunakan oleh penyelidikan tersebut adalah *gradient boosted trees* dengan menggunakan set data yang mengandungi data pesakit dalam tempoh kandungan sebelum 39 minggu dari 2009 hingga 2015. Model tersebut mencapai ketepatan yang tinggi dengan nilai kawasan di bawah lengkung (AUC) 0.923 bagi RDS dan 0.832 bagi hipoglikemia. Nilai SHAP

adalah penting dalam mentafsir model pembelajaran mesin ini dengan menonjolkan pengaruh fitur pada ramalan seperti umur kehamilan, intervensi kelahiran dan keadaan kesihatan ibu. Penggunaan SHAP memberikan pemahaman yang lebih mendalam tentang pengaruh sesuatu fitur dan seterusnya dapat mengukuhkan potensi pembelajaran mesin dalam meningkatkan penjagaan neonatal dengan meramalkan keadaan kritikal sebelum gejala menjadi lebih jelas. Pendekatan ini bukan sahaja membantu dalam diagnosis awal tetapi juga meningkatkan pelaksanaan lebih strategik terhadap amalan perubatan, selain berpotensi mengurangkan kesan buruk yang bakal berlaku.

Dalam kajian yang lain, Naegelin et al. (2023) menggunakan teknik pembelajaran mesin untuk mengenal pasti tahap tekanan menggunakan data fisiologi dan tingkah laku yang dikumpulkan dalam persekitaran pejabat simulasi. Kajian ini bertujuan untuk mengaitkan pelbagai titik data seperti kadar pembolehkan denyutan jantung, tahap aktiviti dan faktor persekitaran dengan hasil tekanan untuk membina model ramalan. Nilai SHAP digunakan untuk mentafsir ramalan model, menjelaskan faktor yang paling ketara menyumbang kepada tekanan. Kaedah penjelasan ini membantu dalam menentukan pencetus tekanan tertentu dan memahami impaknya, sekali gus memberikan cerapan berharga tentang keadaan tempat kerja yang boleh dioptimumkan untuk mengurangkan tekanan. Hasil kajian ini mencadangkan intervensi dan pengubahsuaian yang berpotensi dalam persekitaran pejabat yang boleh meningkatkan kesejahteraan dan produktiviti pekerja dengan mengurangkan tekanan.

Projek ViPal telah membangunkan rangka kerja ramalan untuk menilai virulensi virus influenza berdasarkan data genetik (Yin et al. 2023). Dengan menggunakan model pembelajaran mesin, kajian ini dijalankan bertujuan untuk memahami bagaimana konfigurasi genetik yang berbeza mempengaruhi virulensi strain influenza. Nilai SHAP sangat penting dalam mengenal pasti ciri genetik yang paling berkesan meramal kewujudan virulensi sekaligus boleh membimbing pembangunan vaksin dan penyelidikan antivirus. Pendekatan ini menekankan kepentingan genomik dalam pengurusan penyakit berjangkit dan menyediakan laluan untuk intervensi penjagaan kesihatan yang disasarkan berdasarkan pemahaman yang lebih mendalam tentang genetik patogen.

Penyelidikan oleh Timilsina et al. (2023) pula meneroka sinergi antara imputasi laluan genetik dengan analisis SHAP untuk meningkatkan pemahaman tentang kerentanan penyakit. Dengan menggabungkan data genetik dengan tafsiran pembelajaran mesin, kajian tersebut menjelaskan cara laluan genetik tertentu menyumbang kepada risiko penyakit. Analisis SHAP memberikan penjelasan terhadap model ramalan dan membolehkan penyelidik menentukan faktor genetik yang paling berpengaruh. Pengetahuan ini boleh membawa kepada pendekatan perubatan yang lebih tertumpu kepada kes tertentu serta menyesuaikan strategi pencegahan dan rawatan berdasarkan profil genetik individu. Kajian itu menunjukkan potensi menggabungkan pandangan genetik dengan teknik analisis lanjutan untuk memperhalusi pemahaman tentang mekanisme penyakit.

Juraev et al. (2022) meramalkan kadar kematian dalam unit rawatan rapi neonatal (ICU) dengan membangunkan model gabungan dinamik. Penyelidikan menggunakan set fitur klinikal yang komprehensif untuk meramalkan hasil dengan nilai SHAP memberikan kebolehtafsiran kepada model ramalan. Dengan menganalisis bagaimana pelbagai input klinikal seperti tanda-tanda vital dan intervensi perubatan mempengaruhi risiko kematian, model ini telah membantu dalam membuat keputusan penjagaan kritikal. Pendekatan ini bukan sahaja meningkatkan ketepatan prognosis tetapi juga menyokong pakar perubatan dalam mengutamakan penjagaan berdasarkan risiko yang diramalkan dan seterusnya berpotensi meningkatkan kadar kelangsungan hidup di kalangan bayi yang lahir dengan penyakit kritikal.

2.5 KESIMPULAN

Bab ini telah membincangkan kajian-kajian lepas yang dilakukan berkaitan dengan ramalan hasil angioplasti menggunakan kaedah pembelajaran mesin. Berdasarkan kajian yang dirujuk, di Malaysia, masih belum banyak kajian yang menyeluruh mengenai penggunaan pembelajaran mesin dalam meramalkan hasil prosedur angioplasti. Kajian yang ada sebelumnya kebanyakannya hanya tertumpu pada penggunaan pembelajaran mesin untuk meramalkan prestasi atau hasil klinikal terhadap pesakit di negara-negara luar khususnya dari Amerika Syarikat (USA). Oleh itu, dalam kajian ini, teknik pengelasan pembelajaran mesin khusus akan digunakan untuk

menganalisis data pesakit yang menjalani angioplasti di IJN. Tujuan utama kajian ini adalah untuk membangunkan model yang dapat meramalkan risiko kematian jangka pendek dan panjang pasca-angioplasti dengan menggunakan pendekatan pelombongan data dan algoritma pembelajaran mesin yang berbeza untuk mencapai ketepatan yang lebih tinggi dalam ramalan. Nilai SHAP turut digunakan pada kajian ini bagi mentafsir dengan lebih jelas terhadap fitur yang mempengaruhi ketepatan ramalan model yang dibangunkan.

LIBRARY FETSM

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

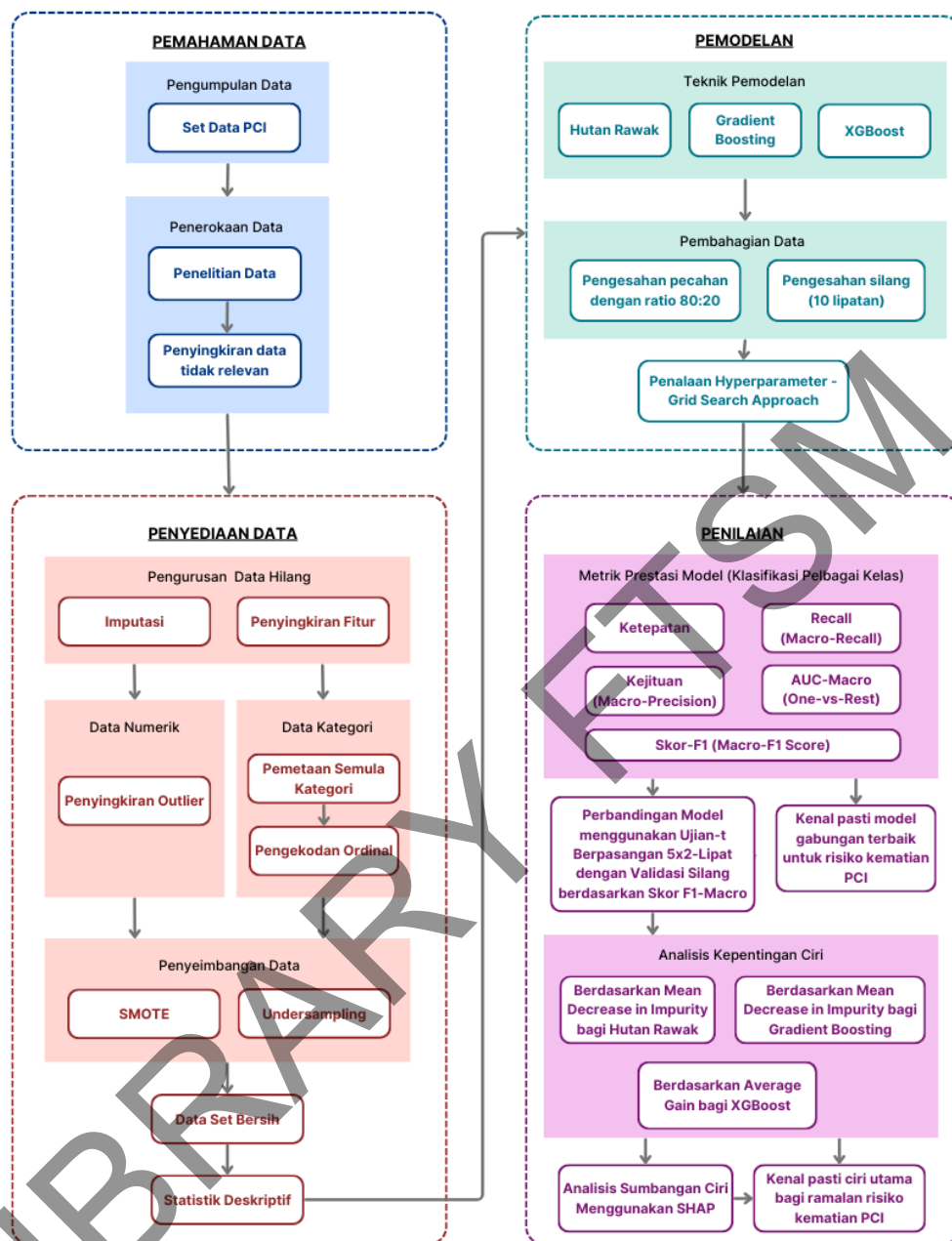
Bab ini menerangkan metodologi yang digunakan dalam kajian ini untuk membangunkan model ramalan risiko kematian pesakit yang akan menjalani prosedur PCI. Ianya termasuklah proses pengumpulan data, prapemprosesan data, pembangunan model, penilaian model, serta langkah-langkah untuk memastikan hasil kajian adalah sesuai untuk klasifikasi pelbagai kelas.

3.2 KERANGKA PENYELIDIKAN

Sains data merupakan disiplin untuk mendapatkan pandangan berharga daripada data melalui model dan aplikasi matematik serta analitikal. Projek sains data boleh mendapat manfaat daripada pengurusan projek dan metodologi proses. Metodologi tersebut berfungsi sebagai salah satu faktor kejayaan sesuatu projek (Christoph Schröer et al. 2021).

Bahagian ini membentangkan gambaran keseluruhan kaedah penyelidikan yang digunakan dalam kajian ini. Proses kajian bermula dari pemahaman data hingga ke tahap penilaian ditunjukkan pada Rajah 3.1. Pembangunan model untuk meramalkan kematian pesakit PCI bermula dengan pengumpulan data dan pemeriksaan awal untuk mengeluarkan data yang tidak relevan. Seterusnya, nilai hilang diimput dan beberapa fitur dibuang berdasarkan kadar nilai hilang. Data numerik digugurkan untuk mengeluarkan *outlier*, manakala data kategori dipetakan semula dan dikodkan. Sebelum pembangunan model, statistik deskriptif digunakan pada set data yang telah diproses untuk memahami taburan data bagi setiap fitur.

Seterusnya, model ramalan kematian dibangunkan menggunakan algoritma hutan rawak, *gradient boosting* dan XGBoost. Prestasi model yang dibangunkan dinilai menggunakan pelbagai metrik klasifikasi khusus untuk mengklasifikasikan pelbagai kelas seperti ketepatan, *recall (macro-recall)*, kejituan (*macro-precision*), skor F1 (*macro-F1 score*), dan *AUC-macro (one-vs-rest)*. Untuk mengenal pasti kaedah gabungan terbaik bagi ramalan kematian pesakit PCI, ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang (*5x2 cross-validated paired t-test*) berdasarkan skor F1-macro digunakan. Analisis kepentingan fitur dan SHAP digunakan untuk mengenal pasti fitur utama yang mempengaruhi ramalan. Penjelasan terperinci mengenai setiap langkah dalam kaedah penyelidikan disediakan dalam Seksyen 3.5 hingga 3.7.



Rajah 3.1 Carta alir keseluruhan kajian ini.

3.3 PEMAHAMAN DATA

Fasa pemahaman data dimulakan dengan pengumpulan data untuk tujuan perlombongan dan mengenal pasti isu kualiti data. Akses kepada data berkualiti tinggi memudahkan pengenalpastian masalah dari peringkat awal dan merancang langkah penyelesaian. Data berkualiti menyediakan justifikasi dan bukti yang diperlukan untuk membuat keputusan berinformasi. Pengumpulan data secara sistematik membantu menjimatkan masa dalam proses akses data. Selain itu, data yang dianalisis dengan baik

dapat mengarahkan kepada penyelesaian yang lebih tepat dan efisien serta meningkatkan kepercayaan terhadap hasil analisis.

Untuk kajian ini, set data yang digunakan adalah set data perubatan yang mengandungi hasil angioplasti. Set data yang digunakan untuk ramalan hasil angioplasti dalam kajian ini diperoleh melalui *IJN Research Ethics Committee* (IJNREC). Pihak IJN berperanan sebagai penyelidik bersama (*co-investigator*) di dalam projek ini. Surat kelulusan oleh IJN boleh dirujuk di Lampiran A.

Set data yang dibekalkan oleh pihak IJN terkandung di dalam dua set data berasingan iaitu set data berkaitan prosedur PCI (Set Data A) yang merangkumi 88 fitur dengan maklumat demografi pesakit, status klinikal, faktor risiko, serta keputusan prosedur dan tindak lanjut selepas rawatan. Set data ini melibatkan sejumlah 41,709 rekod pesakit yang menjalani prosedur angioplasti di IJN dari tahun 2007 sehingga tahun 2020. Manakala set data kedua (Set Data B) pula mengandungi maklumat terperinci berkaitan dengan fitur lesi yang dirawat semasa prosedur PCI. Set data ini mengandungi 38 fitur yang merangkumi maklumat mengenai bifurkasi, aliran TIMI, ukuran stent, serta penggunaan alat bantu semasa prosedur, dan melibatkan sejumlah 57,213 rekod lesi yang dirawat.

Data-data adalah terdiri daripada data angka dan nominal. Jadual 3.2 menunjukkan senarai fitur, jenis dan penerangan bagi fitur set data berkenaan prosedur PCI manakala Jadual 3.3 menunjukkan penerangan data bagi set data fitur lesi yang dirawat.

Jadual 3.1 Jenis dan penerangan fitur bagi Set Data A (prosedur PCI)

Bil.	Fitur	Jenis	Penerangan
1	PCINotifID	Nominal	Nombor unik yang digunakan untuk mengenal pasti prosedur PCI yang dilakukan pada pesakit.
2	PatientID	Nominal	Nombor unik yang digunakan untuk mengenal pasti pesakit.
3	DateofAdmissionddmmyyy	Nominal	Tarikh pesakit dimasukkan ke hospital sebelum prosedur PCI dilakukan.
4	Gender	Nominal	Jantina pesakit, sama ada lelaki atau perempuan.

bersambung...

...sambungan

5	Nationality	Nominal	Kewarganegaraan pesakit, seperti Malaysia atau negara lain.
6	Ageonadmission	Nominal	Umur pesakit ketika dimasukkan ke hospital untuk prosedur PCI.
7	EthnicGroup	Nominal	Kumpulan etnik pesakit.
8	OtherMalaysianspecify	Nominal	Jika pesakit adalah warganegara Malaysia tetapi tidak termasuk dalam kumpulan etnik utama, etnik lain akan dinyatakan.
9	OtherMalaysianspecify_A	Nominal	Sekiranya terdapat etnik Malaysia lain yang perlu dijelaskan lebih lanjut.
10	Foreignerspecifycountryoforigin	Nominal	Jika pesakit adalah warga asing, negara asal mereka akan dinyatakan.
11	Heightm	Numerik	Ketinggian pesakit dalam meter.
12	Weightkg	Numerik	Berat badan pesakit dalam kilogram.
13	BMI	Numerik	Indeks jisim badan (BMI) pesakit, dikira berdasarkan berat dan ketinggian.
14	Diabetes	Nominal	Menunjukkan sama ada pesakit menghidap diabetes atau tidak.
15	OHA	Nominal	Penggunaan agen hipoglisemik oral (OHA) untuk rawatan diabetes.
16	Insulin	Nominal	Sebahagian daripada rawatan diabetes.
17	Nonpharmacologytherapydiettherapy	Nominal	Terapi diet yang tidak melibatkan penggunaan ubat.
18	Hypertension	Nominal	Menunjukkan sama ada pesakit mempunyai hipertensi (tekanan darah tinggi).
19	PreviousPCI	Nominal	Menunjukkan sama ada pesakit pernah menjalani prosedur PCI sebelum ini.
20	Cerebrovascularisease	Nominal	Menunjukkan sama ada pesakit mempunyai sejarah penyakit serebrovaskular.
21	PreviousCABG	Nominal	Menunjukkan sama ada pesakit pernah menjalani pembedahan pintasan arteri koronari (CABG).
22	Peripheralvascularisease	Nominal	Menunjukkan sama ada pesakit mempunyai penyakit vaskular periferal.
23	Smokingstatus	Nominal	Status merokok pesakit.
24	Historyofheartfailure	Nominal	Menunjukkan sama ada pesakit mempunyai sejarah kegagalan jantung.
25	Systolic	Numerik	Bacaan tekanan darah sistolik pesakit pada masa prosedur PCI.
26	Diastolic	Numerik	Bacaan tekanan darah diastolik pesakit pada masa prosedur PCI.
27	Baselinecreatinine	Numerik	Tahap kreatinin serum pesakit sebelum prosedur PCI.
28	FastingBloodGlucose	Numerik	Tahap glukosa darah pesakit selepas berpuasa sebelum prosedur PCI.
29	Totalcholesterol	Numerik	Tahap kolesterol total pesakit sebelum prosedur.

bersambung...

...sambungan

30	LDLLevels	Numerik	Tahap lipoprotein berkepadatan rendah (LDL).
31	EFStatusattimeofPCIprocedure	Numerik	Status pecahan ejeksi jantung (EF) semasa prosedur PCI.
32	Dateofproceduredmmyyyy	Nominal	Tarikh prosedur PCI dijalankan pada pesakit.
33	Year	Numerik	Tahun prosedur PCI dijalankan.
34	Timeofprocedure	Nominal	Masa prosedur PCI dijalankan.
35	PCIstatus	Nominal	Status prosedur PCI.
36	Elective	Nominal	Menunjukkan sama ada prosedur PCI dijalankan sebagai prosedur elektif (terancang).
37	NSTEMIUA	Nominal	Menunjukkan sama ada pesakit mengalami angina tidak stabil (UA) atau infark miokardium tanpa peningkatan segmen ST (NSTEMI).
38	STEMI	Nominal	Menunjukkan sama ada pesakit mengalami infark miokardium dengan peningkatan segmen ST (STEMI).
39	CardiacArrest	Nominal	Menunjukkan sama ada pesakit mengalami serangan jantung.
40	Anginatype	Nominal	Jenis angina yang dialami oleh pesakit.
41	CanadianCardiovascularScoreCCS	Nominal	Skor CCS untuk angina, digunakan untuk mengukur tahap simptom.
42	NYHA	Nominal	Untuk menilai tahap kritikal kegagalan jantung.
43	KillipClassSTEMIampNSTEMI	Nominal	Untuk menilai kegagalan jantung pada pesakit STEMI atau NSTEMI.
44	CoronaryArteryDiseaseCADPresentation	Nominal	Gambaran klinikal penyakit arteri koronari yang dialami oleh pesakit.
45	DateSTEMIonset	Nominal	Tarikh mula simptom STEMI berlaku.
46	TimeSTEMIonset	Nominal	Masa mula simptom STEMI berlaku.
47	NotApplicableSTEMIonset	Nominal	Menunjukkan sama ada maklumat tentang simptom STEMI tidak terpakai untuk pesakit ini.
48	DateArrivalatfirsthospital	Nominal	Tarikh ketibaan pesakit di hospital pertama untuk rawatan kecemasan.
49	TimeArrivalatfirsthospital	Nominal	Masa ketibaan pesakit di hospital pertama.
50	NotApplicableArrivalatfirsthospital	Nominal	Menunjukkan sama ada maklumat ini tidak terpakai.
51	DateArrivalatPCIhospital	Nominal	Tarikh ketibaan pesakit di hospital yang menjalankan prosedur PCI.
52	TimeArrivalatPCIhospital	Nominal	Masa ketibaan pesakit di hospital yang menjalankan PCI.
53	NotApplicableArrivalatPCIhospital	Nominal	Menunjukkan sama ada maklumat ini tidak terpakai.
54	DateFirstballooninflationstentappiration	Nominal	Tarikh pengembangan belon atau penyedutan stent pertama semasa PCI.

bersambung...

...sambungan

55	TimeFirstballooninflationstentasp iration	Nominal	Masa pengembangan belon atau penyedutan stent pertama.
56	NotApplicableFirstballooninflat ionstentasp iration	Nominal	Menunjukkan sama ada maklumat ini tidak terpakai.
57	DateInhospitalSTEMI	Nominal	Tarikh pesakit mengalami STEMI semasa berada di hospital.
58	TimeInhospitalSTEMI	Nominal	Masa pesakit mengalami STEMI semasa berada di hospital.
59	NotApplicableInhospitalSTEMI	Nominal	Menunjukkan sama ada maklumat ini tidak terpakai.
60	DoorToBalloonTimeminute	Numerik	Masa diukur (minit) dari ketibaan pesakit di hospital hingga pengembangan belon semasa PCI.
61	TotalNo.oflesiontreated	Numerik	Jumlah lesi arteri yang dirawat semasa prosedur PCI.
62	Cardiogenicshock	Nominal	Menunjukkan sama ada pesakit mengalami kejutan kardiogenik.
63	ArrhythmiaVTVFBrady	Nominal	Kehadiran aritmia seperti fibrilasi ventrikular (VF) atau bradikardia.
64	TIAStroke	Nominal	Sejarah serangan iskemia sementara (TIA) atau strok.
65	Tamponade	Nominal	Menunjukkan sama ada pesakit mengalami tamponade jantung.
66	Contrastreaction	Nominal	Tindak balas alergi atau kesan terhadap bahan kontras yang digunakan semasa prosedur.
67	Worseningrenalimpairment	Nominal	Kemerosotan fungsi buah pinggang selepas prosedur PCI.
68	Bleeding	Nominal	Menunjukkan sama ada pesakit mengalami pendarahan selepas prosedur.
69	BleedingEpisodeCriteriaIfYes	Nominal	Kriteria pendarahan, jika berlaku.
70	Bleedingsite	Nominal	Lokasi pendarahan yang berlaku pada pesakit.
71	BleedingsiteIfothersspecify	Nominal	Sekiranya lokasi pendarahan lain, akan dinyatakan.
72	RBCWholeBloodTransfusion	Nominal	Menunjukkan sama ada pesakit menerima pemindahan darah penuh atau sel darah merah.
73	DischargeOutcome_A	Nominal	Hasil pelepasan pesakit selepas rawatan, sama ada masih hidup atau meninggal dunia.
74	DateOutcome	Nominal	Tarikh hasil rawatan terakhir pesakit selepas prosedur PCI dijalankan.
75	Primarycauseofdeath	Nominal	Punca utama kematian pesakit yang tercatat dalam rekod perubatan, jika pesakit meninggal dunia.
76	PrimarycauseofdeathOthersspeci fy	Nominal	Punca lain kematian pesakit, jika pesakit meninggal dunia.
77	Locationofdeath	Nominal	Lokasi di mana kematian pesakit berlaku.

bersambung...

...sambungan

78	EPTrackRef	Nominal	Rujukan kepada rekod pesakit dalam sistem EP Track.
79	DateLastFup	Nominal	Tarikh lawatan susulan terakhir pesakit untuk memantau perkembangan kesihatan selepas prosedur PCI.
80	Status	Nominal	Status terkini pesakit selepas lawatan susulan terakhir.
81	Combine DatelastFup/Datedeceased	Nominal	Tarikh lawatan susulan terakhir atau tarikh kematian.
82	Date deceased	Nominal	Tarikh kematian pesakit.
83	FupDuration(yr)	Numerik	Tempoh susulan dalam tahun selepas prosedur PCI.
84	FupDuration(days)	Numerik	Tempoh susulan dalam hari selepas prosedur PCI.
85	StatusfupOverall	Nominal	Status keseluruhan pesakit berdasarkan tempoh susulan.
86	Deathw30days	Nominal	Rekod kematian jika berlaku dalam tempoh 30 hari.
87	Deathw1yr	Nominal	Rekod kematian jika berlaku dalam tempoh 1 tahun.
88	Deathw2yr	Nominal	Rekod kematian jika berlaku dalam tempoh 2 tahun.

Jadual 3.2 Jenis dan penerangan fitur bagi Set Data B (fitur lesi yang dirawat)

Bil.	Fitur	Jenis	Penerangan
1	AutoLesionID	Nominal	Nombor unik automatik yang diberikan kepada setiap lesi yang dirawat semasa prosedur PCI.
2	PCINotifID	Nominal	Nombor unik yang digunakan untuk mengenal pasti prosedur PCI yang dilakukan pada pesakit.
3	PatientID	Nominal	Nombor unik yang digunakan untuk mengenal pasti pesakit.
4	Dateofprocedure	Nominal	Tarikh prosedur PCI dijalankan pada pesakit.
5	Year	Numerik	Tahun prosedur PCI dijalankan.
6	TreatedVessel	Nominal	Saluran darah atau vesel koronari yang dirawat semasa prosedur.
7	TIMIFlowpre(Main Branch)	Nominal	Aliran TIMI (<i>Thrombolysis In Myocardial Infarction</i>) sebelum prosedur dijalankan pada cabang utama arteri yang dirawat.
8	TIMIFlowpost(Main Branch)	Nominal	Aliran TIMI selepas prosedur di cabang utama arteri yang dirawat.
9	Stent#1Diameter	Numerik	Diameter stent pertama yang digunakan (milimeter) semasa prosedur PCI.

bersambung...

...sambungan

10	Stent#1Length	Numerik	Panjang stent pertama yang digunakan (milimeter) semasa prosedur PCI.
11	Stent#2Diameter	Numerik	Diameter stent kedua yang digunakan, jika ada.
12	Stent#2Length	Numerik	Panjang stent kedua yang digunakan, jika ada.
13	Stent#3Diameter	Numerik	Diameter stent ketiga yang digunakan, jika ada.
14	Stent#3Length	Numerik	Panjang stent ketiga yang digunakan, jika ada.
15	Stent#4Diameter	Numerik	Diameter stent keempat yang digunakan, jika ada.
16	Stent#4Length	Numerik	Panjang stent keempat yang digunakan, jika ada.
17	Stent#5Diameter	Numerik	Diameter stent kelima yang digunakan, jika ada.
18	Stent#5Length	Numerik	Panjang stent kelima yang digunakan, jika ada.
19	Stent#6Diameter	Numerik	Diameter stent keenam yang digunakan, jika ada.
20	Stent#6Length	Numerik	Panjang stent keenam yang digunakan, jika ada.
21	MaxBalloonPostdilatationSize	Numerik	Saiz maksimum belon yang digunakan untuk pengembangan semula selepas pemasangan stent.
22	Aspirationcatheter	Nominal	Penggunaan kateter aspirasi.
23	Ventilator	Nominal	Menunjukkan sama ada pesakit memerlukan bantuan pernafasan melalui ventilator semasa prosedur.
24	TemporaryCardiacPacingWire	Nominal	Penggunaan wayar pacing jantung sementara semasa prosedur.
25	Circulatorysupport	Nominal	Penggunaan peranti sokongan peredaran darah semasa prosedur.
26	IABP	Nominal	Penggunaan belon pam intra-aorta (IABP) untuk membantu peredaran darah semasa prosedur.
27	Impella	Nominal	Penggunaan peranti Impella untuk sokongan mekanikal jantung.
28	ECMO	Nominal	Penggunaan ECMO untuk sokongan pernafasan.
29	PCPS	Nominal	Penggunaan Sokongan Peredaran Mekanik Percutaneous (PCPS) semasa prosedur.
30	BifurcationID	Nominal	Pengenalan kepada lesi bifurkasi, iaitu lesi yang melibatkan cabang arteri koronari.
31	TIMIFlowpre(Side branch)	Nominal	Aliran TIMI sebelum prosedur di cabang sampingan arteri yang dirawat.
32	TIMIFlowpost(Side branch)	Nominal	Aliran TIMI selepas prosedur di cabang sampingan arteri yang dirawat.

bersambung...

...sambungan

33	SideStentDiameter1	Numerik	Diameter stent pertama yang digunakan di cabang sampingan arteri, jika ada.
34	SideStentLength1	Numerik	Panjang stent pertama yang digunakan di cabang sampingan arteri, jika ada.
35	SideStentDiameter2	Numerik	Diameter stent kedua yang digunakan di cabang sampingan arteri, jika ada.
36	SideStentLength2	Numerik	Panjang stent kedua yang digunakan di cabang sampingan arteri, jika ada.
37	SideStentDiameter3	Numerik	Diameter stent ketiga yang digunakan di cabang sampingan arteri, jika ada.
38	SideStentLength3	Numerik	Panjang stent ketiga yang digunakan di cabang sampingan arteri, jika ada.

Untuk pembangunan model, set data ini ditapis untuk memasukkan rekod dan fitur yang berkaitan dengan objektif penyelidikan sahaja. Sebelum penerokaan data, semua fitur dikaji berdasarkan pengetahuan awal mengenai prosedur angioplasti. Set Data A merupakan set data utama yang mengandungi kebanyakan data penting yang diperlukan untuk kajian ini iaitu data berkenaan prosedur PCI. Sementara Set Data B pula mempunyai data sokongan yang memperincikan maklumat dengan lebih mendalam berkaitan lesi yang dirawat semasa prosedur tersebut dijalankan. Sebarang fitur yang tidak relevan dengan kajian ini dikeluarkan dari pertimbangan. Senarai fitur yang dikeluarkan dalam fasa penyediaan data dan alasan untuk pengecualian boleh didapati di Lampiran B.

Kajian ini bertujuan untuk membangunkan model ramalan yang dapat meramalkan hasil prosedur angioplasti pada masa kemasukan pesakit. Untuk mencapai matlamat ini, fitur yang mengandungi maklumat yang diperoleh selepas prosedur perlu dikeluarkan. Fitur-fitur ini termasuklah DateLastFup, Status, dan Combine DatelastFup/Datedeceased. Skop kajian ini juga hanya tertumpu kepada rakyat Malaysia sahaja. Oleh itu, bagi fitur Nationality, data yang bertanda “Non-Malaysia” perlu disingkirkan terlebih dahulu bagi memastikan kajian ini hanya menggunakan data pesakit warganegara Malaysia sahaja. Setelah itu, fitur Nationality juga dikeluarkan kerana hanya memegang satu nilai yang sama bagi semua data. Selain itu, fitur-fitur yang tidak menyumbang kepada maklumat yang relevan bagi pembangunan model juga dikecualikan, seperti PCINotifID, PatientID, dan DateofAdmissionddmmyyy.

Untuk mengekalkan konsistensi data, beberapa fitur boleh menghasilkan fitur baharu yang berguna kepada kajian atau disebut sebagai data terbitan. Dengan menggunakan data terbitan yang baharu, fitur yang menghasilkan data terbitan berkenaan juga perlu dikeluarkan seperti Deathw30days, Deathw1yr, Deathw2yr dan DeathAfter2yr. Fitur ini dilabel dengan "Menggunakan fitur terbitan Death" dalam Lampiran B. Selain itu, fitur yang mempunyai maklumat berlebihan juga dikeluarkan. Sebagai contoh, Heightm dan Weightkg dikeluarkan kerana BMI sudah digunakan berikutan ia mempunyai tujuan yang sama. Dalam Lampiran B, fitur ini dinyatakan dengan alasan penghapusan kerana terdapat pada fitur yang lain.

Untuk menghasilkan satu set data yang menyeluruh, kedua-dua set data digabungkan berdasarkan PCINotifID yang terdapat pada kedua-dua set data. PatientID tidak digunakan sebagai kunci data kerana seorang pesakit mempunyai kemungkinan lebih dari satu prosedur PCI yang memegang maklumat berbeza bagi setiap prosedur. Selepas penyingkiran rekod dan fitur yang tidak relevan, jumlah set data menjadi 57,212 rekod dengan 27 fitur.

3.4 PENYEDIAAN DATA

Fasa pra-pemrosesan data amat penting untuk meningkatkan kualiti data, yang seterusnya meningkatkan kualiti pengetahuan yang boleh diekstrak. Fasa ini melibatkan beberapa langkah termasuklah pengendalian nilai yang hilang, penyingkiran outlier, pemetaan semula kategori, dan pengekodan sebelum pembangunan model. Langkah-langkah ini dilaksanakan sehingga mencapai output yang dikehendaki.

Dalam kajian ini, terdapat nama-nama fitur dalam bentuk singkatan dan terdapat juga nama fitur yang terlalu panjang seperti yang ditunjukkan dalam Jadual 3.1 dan Jadual 3.2. Untuk meningkatkan kejelasan dan pemahaman, fitur-fitur ini digantikan dengan istilah yang lebih sesuai dan sebahagiannya dikekalkan sekiranya perlu. Jadual 3.3 menunjukkan penggantian istilah-istilah 27 fitur yang telah terpilih.

Jadual 3.3 Penggantian istilah bagi fitur dalam set data

Bil.	Fitur	Penggantian Istilah
1	Gender	Gender
2	Ageonadmission	Age
3	EthnicGroup	Ethnicity
4	BMI	BMI
5	OHA	OHA
6	Insulin	Insulin
7	Nonpharmacologytherapydiettherapy	DietTherapy
8	Hypertension	Hypertension
9	PreviousPCI	PrevPCI
10	Cerebrovascularisease	CVAdisease
11	PreviousCABG	PrevCABG
12	Peripheralvascularisease	PeripheralVASCdisease
13	Smokingstatus	SmokingStatus
14	Historyofheartfailure	HeartFailureHist
15	Systolic	Systolic
16	Diastolic	Diastolic
17	Baselinecreatinine	BaselineCreatinine
18	Totalcholesterol	TotalCholesterol
19	PCIstatus	PCIstatus
20	CanadianCardiovascularScoreCCS	CCSscore
21	NYHA	NYHA
22	KillipClassSTEMIampNSTEMI	KillipClass
23	TotalNo.oflesiontreated	LesionsTreated
24	MaxStentDiameter	MaxStentDiameter
25	TotalStentLength	TotalStentLength
26	SideStent	SideStent
27	Death	Death

Proses penggantian istilah dalam Jupyter Notebook diperjelaskan dengan menamakan semula fitur menggunakan fungsi *rename* dalam *pandas*. Pendekatan ini memastikan set data siap untuk tugas perlombongan data seterusnya, yang membawa kepada hasil yang lebih boleh dipercayai dan boleh ditafsirkan.

3.4.1 Pengurusan Data Hilang

Proses penyediaan data dimulakan dengan memeriksa kewujudan nilai yang hilang dalam set data. Jadual 3.4 menunjukkan nisbah nilai yang hilang bagi setiap fitur.

Jadual 3.4 Peratusan nilai yang hilang bagi setiap fitur

Bil.	Fitur	Nilai Hilang (%)
1	Gender	0.000
2	Age	0.000
3	Ethnicity	0.002
4	BMI	9.624
5	OHA	0.000
6	Insulin	0.000
7	DietTherapy	13.574
8	Hypertension	0.435
9	PrevPCI	0.220
10	CVAdisease	0.357
11	PrevCABG	0.231
12	PeripheralVASCdisease	0.357
13	SmokingStatus	14.177
14	HeartFailureHist	0.413
15	Systolic	7.248
16	Diastolic	7.350
17	BaselineCreatinine	4.366
18	TotalCholesterol	18.451
19	PCIstatus	0.000
20	CCScore	6.371
21	NYHA	5.370
22	KillipClass	1.454
23	LesionsTreated	0.000
24	MaxStentDiameter	9.856
25	TotalStentLength	9.841
26	SideStent	0.000
27	Death	0.000

Di dalam Jadual 3.4, secara keseluruhannya, peratus kehilangan data adalah rendah. Bagi fitur yang mempunyai peratus kehilangan di bawah 5%, data bagi fitur ini disingkirkan agar ketepatan data dapat dikekalkan. Sementara 10 fitur yang mempunyai kehilangan data melebihi 5% (BMI, DietTherapy, SmokingStatus, Systolic, Diastolic, TotalCholesterol, CCScore, NYHA, MaxStentDiameter dan TotalStentLength), proses untuk mengisi nilai yang hilang (imputasi) adalah seperti berikut:

1. **BMI, Systolic, Diastolic, TotalCholesterol MaxStentDiameter dan TotalStentLength (Numerik).** Imputasi bagi lima fitur ini adalah dengan

menggunakan kaedah median. Median lebih popular berbanding purata (*mean*) kerana kurang sensitif kepada nilai terpencil (*outliers*). Ukuran tekanan darah (Systolic dan Diastolic) boleh berubah dengan ketara, dan median memberikan gambaran yang lebih meyakinkan untuk nilai-nilai tersebut. Dengan mengisi nilai yang hilang menggunakan median, ia dapat mengurangkan pengaruh nilai-nilai ekstrem dan mengekalkan integriti data.

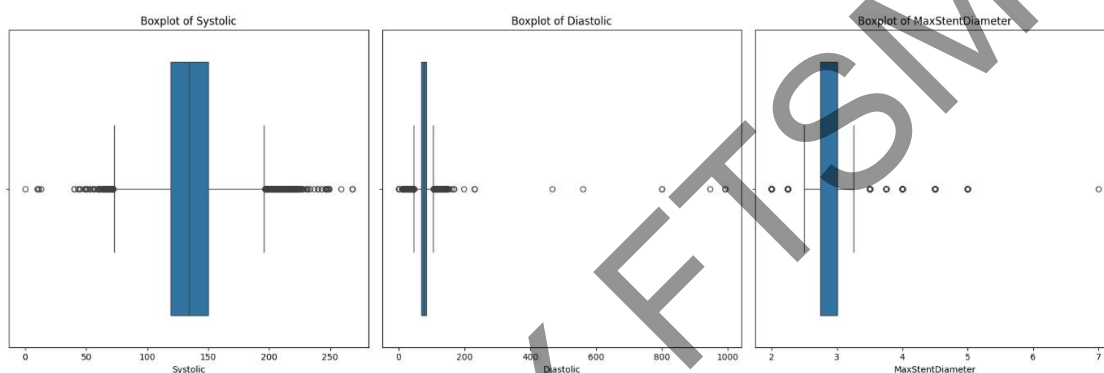
2. **DietTherapy, SmokingStatus, CCSscore, dan NYHA (Nominal).** Imputasi bagi empat fitur ini adalah dengan menggunakan kaedah mod. Kaedah mod membantu mengekalkan konsistensi dalam taburan kategori dan mengisi nilai yang hilang dengan kategori yang paling mungkin atau kerap. Contohnya, jika nilai bagi CCSscore adalah "CCS 0, CCS 1", "CCS 2", "CCS 3 dan CCS 4" dan kekerapan tertinggi adalah "CCS 2", maka mod bagi CCSscore ialah "CCS 2". Nilai hilang seterusnya digantikan dengan "CCS 2". Terdapat juga data yang bertanda "Not Applicable / Not available" yang memberi satu nilai yang bermakna dan bukannya bermaksud sebagai data hilang. Data seperti ini terdapat pada fitur KillipClass. Bagi pesakit yang menjalani PCI Elektif, Killip Class akan ditandakan sebagai "Not Applicable". Killip Class hanya untuk pesakit STEMI & NSTEMI semasa pendaftaran masuk wad.

3.4.2 Penyingkiran Outlier untuk Data Numerik

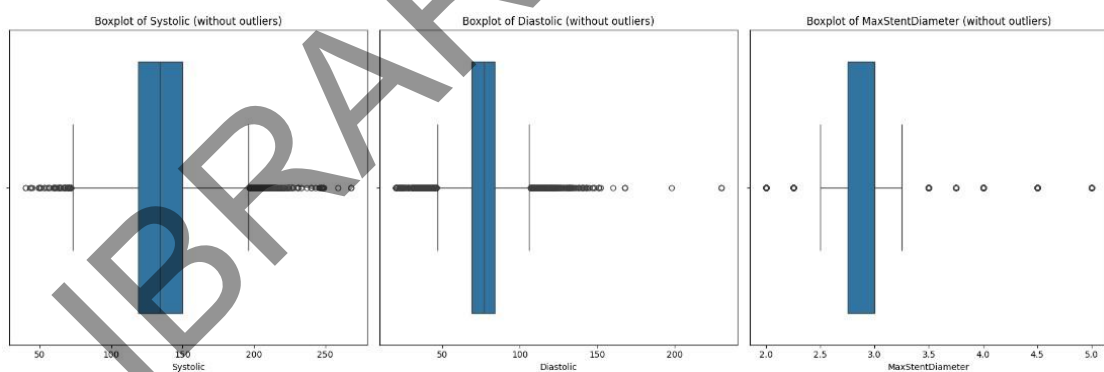
Setelah memastikan tiada pertindihan rekod, langkah mengenal pasti *outlier* dijalankan. *Outlier* merujuk kepada nilai-nilai yang sangat berbeza dan mencerminkan penyimpangan yang ketara daripada kebanyakan nilai dalam set data. Pengendalian *outlier* adalah proses penting kerana ia biasanya merupakan kesilapan yang boleh menjejaskan keputusan analisis (Deneshkumar et al. 2014). Dalam konteks data perubatan, nilai ekstrem tidak semestinya dianggap sebagai *outlier* kerana ia mungkin mewakili nilai yang sah dan wajar.

Dalam kajian ini, *outliers* terdapat di dalam fitur Systolic, Diastolic dan MaxStentDiameter. Nilai wajar Systolic bagi seorang pesakit sepatutnya tidak kurang dari 40 mmHg manakala nilai wajar Diastolic tidak kurang dari 20 mmHg dan tidak melebihi 250 mmHg. Nilai yang tidak berada dalam julat tersebut disingkirkan bagi

memelihara integriti data. Rajah 3.2 menunjukkan boxplot bagi fitur Systolic, Diastolic dan MaxStentDiameter yang mengandungi *outlier*. Dalam boxplot MaxStentDiameter, whiskers mewakili julat data yang berada dalam 1.5 kali julat antara kuartil (IQR). Mana-mana titik data yang berada di luar julat ini dianggap sebagai *outlier* dan ditunjukkan sebagai bulatan individu. Nilai 7 jelas ditandakan sebagai *outlier* yang ekstrem kerana terletak jauh daripada pengagihan data lain yang membuktikan ia adalah anomali.



Rajah 3.2 Boxplot bagi Systolic, Diastolic dan MaxStentDiameter dengan *outlier*



Rajah 3.3 Boxplot bagi Systolic, Diastolic dan MaxStentDiameter tanpa *outlier*

Nilai 2.8 pula terletak sedikit bawah pada whisker bawah yang menunjukkan ia juga dianggap sebagai *outlier*. Rajah 3.3 menunjukkan boxplot bagi fitur Systolic, Diastolic dan MaxStentDiameter selepas penyingkiran *outlier*.

3.4.3 Pemetaan Semula dan Pengekoden untuk Data Kategori

Selain daripada ciri-ciri numerikal, ciri-ciri kategori juga diperiksa dan diproses dengan teliti. Dalam projek ini, pelaksanaan hutan rawak, *gradient boosting*, dan XGBoost bergantung pada rangka kerja *scikit-learn* dan XGBoost di mana ia tidak menyokong fitur berjenis kategori sebagai input. Oleh itu, semua sampel input ditukar kepada format float64 (*scikit-learn* 2023b). Hasilnya, fitur-fitur kategori ditransformasikan kepada perwakilan numerikal sebelum fasa latihan model. Tambahan pula, normalisasi atau penyeragaman data tidak diterapkan kerana hutan rawak, *gradient boosting*, dan XGBoost adalah teknik gabungan berasaskan pohon yang beroperasi berdasarkan logik peraturan dan bukannya metrik jarak (Prakash 2022; Cui et al. 2022).

Pengekoden binari dipilih untuk menukar fitur-fitur jenis kategori yang hanya mempunyai dua klasifikasi kepada perwakilan numerikal, di mana setiap kategori dipetakan kepada nombor 0 atau 1. Tidak seperti pengekoden *one-hot* yang mengubah setiap kategori menjadi ciri binari dan boleh meningkatkan bilangan fitur dalam set data dengan ketara, pengekoden binari mengekalkan bilangan fitur asal. Pendekatan ini lebih cekap dari segi pengiraan dan mengelakkan penurunan prestasi, terutamanya bagi set data yang melibatkan banyak fitur kategori. Untuk kelas sasaran yang mempunyai lebih daripada dua klasifikasi, pengekoden label (*label encoding*) digunakan untuk menukar setiap kelas kepada nombor unik. Oleh itu, pengekoden binari dan pengekoden label digunakan untuk menukar fitur kategori kepada perwakilan numerikal untuk model hutan rawak, *gradient boosting*, dan XGBoost.

Jadual 3.5 Kelas label Death

Kategori	Taburan	Pengekoden
Kematian kurang dari 30 hari.	1.41%	0
Kematian antara 30 hari hingga 1 tahun.	6.4%	1
Kematian antara 1 tahun hingga 2 tahun.	4.04%	2
Kematian selepas 2 tahun.	88.15%	3

Dalam set data angioplasti ini, pesakit dibezakan berdasarkan pelbagai fitur perubatan dan demografi. Sebagai contoh, fitur Gender menunjukkan jantina pesakit sama ada lelaki atau perempuan. Nilai yang dilabel 'Male' dipetakan kepada 0, manakala

nilai 'Female' dipetakan kepada 1. Jadual 3.6 menunjukkan taburan kategori dalam fitur Gender selepas proses pemetaan semula dan pengekodan binari beserta taburan kelas label kematian (Death) untuk setiap jantina.

Jadual 3.6 Pengekodan binari untuk Gender

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Male	82.23%	0	3 (88.98%)
			1 (5.92%)
			2 (3.81%)
			0 (1.29%)
Female	17.77%	1	3 (84.3%)
			1 (8.61%)
			2 (5.12%)
			0 (1.97%)

Bagi fitur Ethnicity, adalah dijangkakan bahawa kumpulan etnik Melayu (Malay) mempunyai peratusan tertinggi kerana ia merupakan etnik terbesar di Malaysia iaitu sebanyak 58.01%. Manakala etnik India (Indian) pula mempunyai taburan kedua tertinggi iaitu sebanyak 24.94% dan seterusnya adalah Cina iaitu sebanyak 14.29%. Bagi etnik selebihnya, taburannya sangat kecil iaitu berada di bawah 5%. Kumpulan-kumpulan etnik yang mempunyai peratusan kurang daripada 5% telah digabungkan dalam kategori Other. Jumlah kategori Ethnicity telah dikurangkan kepada empat kategori sahaja Malay, Indian, Chinese, dan Other seperti yang ditunjukkan pada Jadual 3.7.

Bagi fitur OHA, nilai No memiliki taburan sebanyak 60.3%, manakala nilai Yes adalah 39.7%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 beserta taburan kelas label kematian untuk setiap nilai OHA seperti yang ditunjukkan dalam Jadual 3.8 di bawah.

Bagi fitur Insulin, nilai No memiliki taburan sebanyak 84.32%, manakala nilai Yes adalah 15.68%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 seperti yang ditunjukkan dalam Jadual 3.9.

Jadual 3.7 Pemetaan semula dan pengkodan untuk Ethnicity

Nilai Asal		Nilai Baharu			
Kategori	Taburan	Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Malay	58.01%	Malay	58.01%	0	3 (87.65%) 1 (6.54%) 2 (4.28%) 0 (1.53%)
Indian	24.94%	Indian	24.94%	1	3 (87.53%) 1 (6.86%) 2 (4.29%) 0 (1.32%)
Chinese	14.29%	Chinese	14.29%	2	3 (90.3%) 1 (5.36%) 2 (3.05%)
Foreigner	1.25%				0 (1.3%)
Punjabi	1.07%				
Other Malaysian	0.42%				
Iban	0.08%				
Kadazan Dusun	0.08%	Other	2.77%	3	3 (93.13%) 1 (4.52%) 2 (2.17%)
Orang Asli	0.04%				
Bidayuh	0.03%				
Bajau	0.02%				0 (0.18%)
Murut	0.01%				
Melanau	0.00%				

Jadual 3.8 Pengekodan binari untuk OHA

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	60.3%	0	3 (88.04%) 1 (6.39%) 2 (3.96%)

bersambung...

...sambungan

			0 (1.61%)
Yes	39.7%	1	3 (88.31%)
			1 (6.41%)
			2 (4.17%)
			0 (1.11%)

Jadual 3.9 Pengekodan binari untuk Insulin

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	82.79%	0	3 (90.6%)
			1 (5.09%)
			2 (3.27%)
			0 (1.04%)
Yes	17.21%	1	3 (76.38%)
			1 (12.7%)
			2 (7.76%)
			0 (3.17%)

Bagi fitur DietTherapy, nilai No memiliki taburan sebanyak 97.68%, manakala nilai Yes adalah 2.32%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 dan taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.10.

Jadual 3.10 Pengekodan binari untuk DietTherapy

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	97.68%	0	3 (88.27%)
			1 (6.36%)
			2 (4.01%)
			0 (1.36%)
Yes	2.32%	1	3 (83.19%)
			1 (7.76%)
			2 (5.6%)
			0 (3.45%)

Bagi fitur Hypertension, nilai Yes memiliki taburan sebanyak 74.48%, manakala nilai No adalah 25.52%. Nilai Yes dikodkan sebagai 0 dan No dikodkan

sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.11.

Jadual 3.11 Pengekodan binari untuk Hypertension

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Yes	74.48%	0	3 (86.3%)
			1 (7.42%)
			2 (4.63%)
			0 (1.66%)
No	25.52%	1	3 (93.55%)
			1 (3.41%)
			2 (2.35%)
			0 (0.69%)

Bagi fitur PrevPCI, nilai No memiliki taburan sebanyak 72.02%, manakala nilai Yes adalah 27.98%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.12.

Jadual 3.12 Pengekodan binari untuk PrevPCI

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	72.02%	0	3 (89.04%)
			1 (5.84%)
			2 (3.67%)
			0 (1.45%)
Yes	27.98%	1	3 (85.86%)
			1 (7.83%)
			2 (5%)
			0 (1.3%)

Bagi fitur CVAdisease, nilai No memiliki taburan sebanyak 96.86%, manakala nilai Yes adalah 3.14%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.13.

Jadual 3.13 Pengkodan binari untuk CVAdisease

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	96.86%	0	3 (88.42%) 1 (6.21%) 2 (3.99%) 0 (1.38%)
Yes	3.14%	1	3 (79.78%) 1 (12.1%) 2 (5.89%) 0 (2.23%)

Bagi fitur PrevCABG, nilai No memiliki taburan sebanyak 94.64%, manakala nilai Yes adalah 5.36%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.14.

Jadual 3.14 Pengkodan binari untuk PrevCABG

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	94.64%	0	3 (88.73%) 1 (6.17%) 2 (3.72%) 0 (1.37%)
Yes	5.36%	1	3 (77.89%) 1 (10.35%) 2 (9.7%) 0 (2.05%)

Bagi fitur PeripheralVASCdisease, nilai No memiliki taburan sebanyak 98.92%, manakala nilai Yes adalah 1.08%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.15.

Jadual 3.15 Pengkodan binari untuk PeripheralVASCdisease

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	98.92%	0	3 (88.37%) 1 (6.29%)

bersambung...

...sambungan

			2 (3.96%)
			0 (1.38%)
Yes	1.08%	1	3 (68.06%)
			1 (16.2%)
			2 (11.57%)
			0 (4.17%)

Bagi fitur SmokingStatus, nilai Never memiliki taburan sebanyak 55.72%, nilai Former (quit > 30 days) adalah sebanyak 22.27%, manakala nilai Current (any tobacco use within last 30 days) adalah 22.02%. Nilai Never dikodkan sebagai 0, Former (quit > 30 days) sebagai 1 dan Current (any tobacco use within last 30 days) dikodkan sebagai 2 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.16.

Jadual 3.16 Pengekodan label untuk SmokingStatus

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Never	55.72%	0	3 (86.35%)
			1 (7.34%)
			2 (4.61%)
			0 (1.7%)
Former (quit > 30 days)	22.27%	1	3 (92.66%)
			1 (3.59%)
			2 (2.78%)
			0 (0.97%)
Current (any tobacco use within last 30 days)	22.02%	2	3 (88.14%)
			1 (6.84%)
			2 (3.88%)
			0 (1.14%)

Bagi fitur HeartFailureHist, nilai No memiliki taburan sebanyak 95.01%, manakala nilai Yes adalah 4.99%. Nilai No dikodkan sebagai 0 dan Yes dikodkan sebagai 1 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.17.

Jadual 3.17 Pengekodan binari untuk HeartFailureHist

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	95.01%	0	3 (89.24%)
			1 (5.72%)
			2 (3.67%)
			0 (1.37%)
Yes	4.99%	1	3 (67.33%)
			1 (19.34%)
			2 (11.22%)
			0 (2.1%)

Bagi fitur PCIstatus, nilai Elective memiliki taburan sebanyak 87.71%, nilai NSTEMI/UA adalah sebanyak 6.37%, manakala nilai STEMI adalah 5.92%. Nilai Elective dikodkan sebagai 0, NSTEMI/UA sebagai 1 dan STEMI dikodkan sebagai 2 serta taburan kelas kematian adalah seperti yang ditunjukkan dalam Jadual 3.18.

Jadual 3.18 Pengekodan label untuk PCIstatus

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Elective	87.71%	0	3 (88.68%)
			1 (6.27%)
			2 (3.81%)
			0 (1.24%)
NSTEMI/UA	6.37%	1	3 (83.42%)
			1 (7.86%)
			2 (6.44%)
			0 (2.28%)
STEMI	5.92%	2	3 (85.4%)
			1 (6.67%)
			2 (4.89%)
			0 (3.04%)

Begitu juga taburan serta pengekodan bagi CCSscore, NYHA, KillipClass, dan SideStent boleh dilihat pada Jadual 3.19, Jadual 3.20, Jadual 3.21 dan Jadual 3.22 berikut:

Jadual 3.19 Pengekodan label untuk CCSscore

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
CCS 2	43.24%	0	3 (86.55%) 1 (7.32%) 2 (4.63%) 0 (1.5%)
CCS 1	34.45%	1	3 (89.56%) 1 (5.67%) 2 (3.49%) 0 (1.28%)
CCS 0 = Asymptomatic	15.54%	2	3 (93.64%) 1 (3.83%) 2 (2.08%) 0 (0.44%)
CCS 3	4.99%	3	3 (78.9%) 1 (9.91%) 2 (7.85%) 0 (3.34%)
CCS 4	1.79	4	3 (82.78%) 1 (8.1%) 2 (5.32%) 0 (3.8%)

Jadual 3.20 Pengekodan label untuk NYHA

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
NYHA I	61.32%	0	3 (90.42%) 1 (5.16%) 2 (3.4%) 0 (1.02%)
NYHA II	33.49%	1	3 (85.71%) 1 (7.84%) 2 (4.72%) 0 (1.73%)
NYHA III	3.79%	2	3 (4.08%) 1 (75.58%) 2 (11.9%) 0 (8.59%)
NYHA IV	1.11%	3	3 (82.51%) 1 (10.76%) 2 (2.69%) 0 (4.04%)

Jadual 3.21 Pengekodan label untuk KillipClass

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
Not Applicable / Not available	75.51%	0	3 (88.82%)
			1 (6.17%)
			2 (3.85%)
			0 (1.16%)
			0 (1.16%)
I No clinical signs of HF	15.2%	1	3 (86.32%)
			1 (6.54%)
			2 (5.03%)
			0 (2.1%)
			0 (2.1%)
II Left Heart Failure (LHF)	7.54%	2	3 (86.68%)
			1 (7.62%)
			2 (3.78%)
			0 (1.92%)
			0 (1.92%)
IV Cardiogenic Shock	1.12%	3	3 (82.67%)
			1 (9.33%)
			2 (3.56%)
			0 (4.44%)
			0 (4.44%)
III Acute Pulmonary Oedema (APO)	0.61 %	4	3 (78.86%)
			1 (9.76%)
			2 (8.13%)
			0 (3.25%)
			0 (3.25%)

Jadual 3.22 Pengekodan binari untuk SideStent

Kategori	Taburan	Pengekodan	Taburan Kelas Label Death
No	97.34%	0	3 (88.06%)
			1 (6.46%)
			2 (4.07%)
			0 (1.41%)
			0 (1.41%)
Yes	2.66%	1	3 (91.53 %)
			1 (3.95 %)
			2 (3.01 %)
			0 (1.51 %)
			0 (1.51 %)

Dalam projek ini, penggunaan model pembelajaran mesin seperti hutan rawak, *gradient boosting*, dan XGBoost tidak memerlukan proses normalisasi data. Ini kerana ketiga-tiga algoritma ini adalah berasaskan pohon keputusan (*decision trees*) yang tidak bergantung kepada skala atau unit ukuran data. Algoritma berasaskan pohon keputusan secara semula jadi membahagikan data berdasarkan pemisahan yang optimum tanpa

mengambil kira julat atau unit fitur (Breiman 2001; Friedman 2001; Chen 2016). Berbeza dengan model seperti regresi logistik atau SVM yang kepekaan terhadap skala data memerlukan normalisasi, hutan rawak, *gradient boosting* dan XGBoost boleh menangani ciri-ciri data yang mempunyai unit dan julat yang berbeza dengan cekap. Oleh itu, langkah normalisasi pada projek ini tidak diperlukan kerana ia tidak akan meningkatkan prestasi model yang dipilih. Selain itu, ia juga dapat mengelakkan daripada melaksanakan penambahan proses pembangunan model yang tidak memberikan impak signifikan terhadap analisis atau keputusan akhir dalam projek ini.

3.4.4 Penerangan Set Data Akhir

Selepas pra-pemprosesan data untuk kedua-dua data numerik dan kategori, set data terdiri daripada 53,850 rekod, 26 fitur, dan 1 pemboleh ubah sasaran iaitu Death. Jadual 3.23 menunjukkan statistik deskriptif bagi fitur-fitur numerikal dalam bentuk min (Mean), sisihan piawai (Std), nilai minimum (Min), persentil bawah (25%), median (50%), persentil atas (75%), dan nilai maksimum (Max). Jadual 3.24 membentangkan statistik deskriptif untuk fitur-fitur kategori, termasuk bilangan kategori unik bagi setiap fitur, kategori dengan jumlah tertinggi, bilangan kategori tertinggi dan perkadarannya dalam set data. Taburan kategori terperinci untuk setiap fitur boleh didapati dalam Jadual 3.6 hingga 3.22.

Histogram dalam Rajah 3.4 menunjukkan taburan setiap fitur selepas pengekodan binari. Dalam visualisasi ini, taburan setiap fitur ditunjukkan selepas langkah-langkah pra-pemprosesan, memberikan pandangan jelas mengenai taburan data dan sebarang potensi kewujudan *outlier* sekiranya ada. Histogram ini membantu dalam memahami taburan keseluruhan nilai dalam setiap fitur yang penting untuk analisis lanjut dan pembangunan model. Dari histogram ini, dapat dilihat bahawa set data ini menunjukkan keadaan data tidak seimbang yang sangat ketara. Oleh itu, sebelum proses pembangunan model dijalankan, proses menyeimbangkan data perlu dilaksanakan. Keterangan lanjut berkaitan proses ini diterangkan pada sub topik 3.7.2 Pembahagian dan Penyeimbangan Data.

Untuk memahami korelasi antara fitur dan pemboleh ubah sasaran, peta haba ditunjukkan dalam rajah di atas. Secara asasnya, peta haba (*heatmap*) digunakan untuk

menggambarkan hubungan antara dua atau lebih pemboleh ubah dalam bentuk matriks. Dalam konteks korelasi, nilai dalam peta haba berkisar antara -1 hingga 1. Nilai -1 menunjukkan korelasi negatif yang sempurna iaitu apabila satu pemboleh ubah meningkat, pemboleh ubah lain menurun secara berkadar. Nilai 0 menunjukkan tiada korelasi, bermaksud tiada hubungan antara kedua-dua pemboleh ubah. Nilai 1 pula menunjukkan korelasi positif yang sempurna iaitu apabila satu pemboleh ubah meningkat, pemboleh ubah lain juga meningkat secara berkadar.

Dalam analisis projek ini, korelasi berdasarkan pekali korelasi Pearson digunakan. Pekali korelasi ini menunjukkan darjah hubungan antara pemboleh ubah dalam julat dari -0.35 hingga 0.52. Secara amnya, kebanyakan fitur dalam set data ini menunjukkan hubungan yang lemah antara satu sama lain, kecuali beberapa korelasi yang lebih ketara.

Sebagai contoh, Systolic dan Diastolic menunjukkan korelasi positif yang agak tinggi sebanyak 0.52 kerana secara logiknya, kedua-duanya berkaitan dengan tekanan darah di mana kenaikan salah satunya mempengaruhi kenaikan fitur yang satu lagi. Terdapat juga korelasi yang sederhana seperti KillipClass dan PCIstatus, yang menunjukkan nilai korelasi 0.41. Korelasi ini mungkin menunjukkan hubungan antara keadaan klinikal pesakit dengan status PCI yang dijalankan. Keseluruhan korelasi ini membantu memberikan gambaran tentang hubungan antara fitur-fitur dalam dataset, dan dapat membantu dalam memahami bagaimana setiap fitur mungkin mempengaruhi pemboleh ubah sasaran iaitu Death.

Berikutan saiz set data yang telah diproses ini adalah terlalu besar iaitu sebanyak 53,850 rekod data, perkakasan sedia ada yang digunakan bagi kajian ini tidak mampu menampung keperluan bagi proses pembangunan model. Oleh itu, kajian ini telah mengecilkan sampel data kepada 20,000 rekod data serta mengekalkan 26 fitur dengan 1 kelas sasaran.

Jadual 3.23 Statistik deskriptif bagi data numerik setelah pra-pemrosesan

Fitur	Mean	Std	Min	25%	50%	75%	Max
Age	58.94	10.28	16.24	52.21	59.07	66.06	97.67
BMI	26.93	4.34	10.50	24.20	26.56	29.11	71.62
Systolic	135.78	24.51	40.00	119.00	134.00	150.00	268.00
Diastolic	76.38	13.41	20.00	68.00	77.00	84.00	230.00
BaselineCreatinine	132.01	146.20	25.00	80.00	95.00	116.00	1899.00
TotalCholesterol	4.24	1.10	1.10	3.60	4.10	4.70	24.20
LesionsTreated	1.68	0.83	1.00	1.00	1.00	2.00	7.00
MaxStentDiameter	3.00	0.44	2.00	2.75	3.00	3.00	5.00
TotalStentLength	30.58	16.70	8.00	18.00	26.00	36.00	169.00

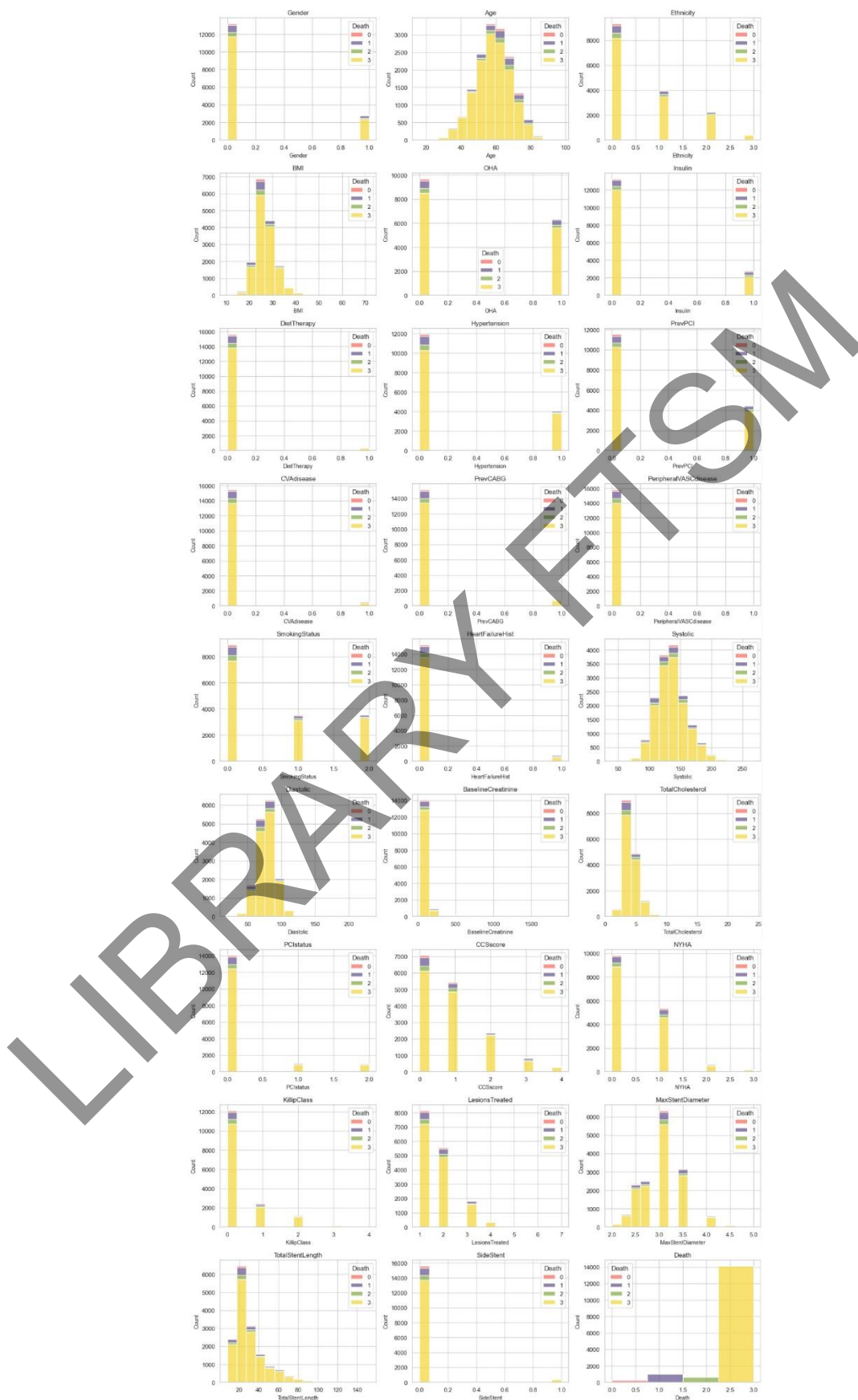
Jadual 3.24 Statistik deskriptif bagi data kategori setelah pra-pemrosesan

Fitur	Unik	Top	Frekuensi (Taburan)
Gender	2	0	16446 (82.23%)
Ethnicity	4	0	11602 (58.01%)
OHA	2	0	12059 (60.30%)
Insulin	2	0	16558 (82.79%)
DietTherapy	2	0	19536 (97.68%)
Hypertension	2	0	14897 (74.48%)
PrevPCI	2	0	14405 (72.02%)
CVAdisease	2	0	19372 (96.86%)
PrevCABG	2	0	18928 (94.64%)
PeripheralVASCdisease	2	0	19784 (98.92%)

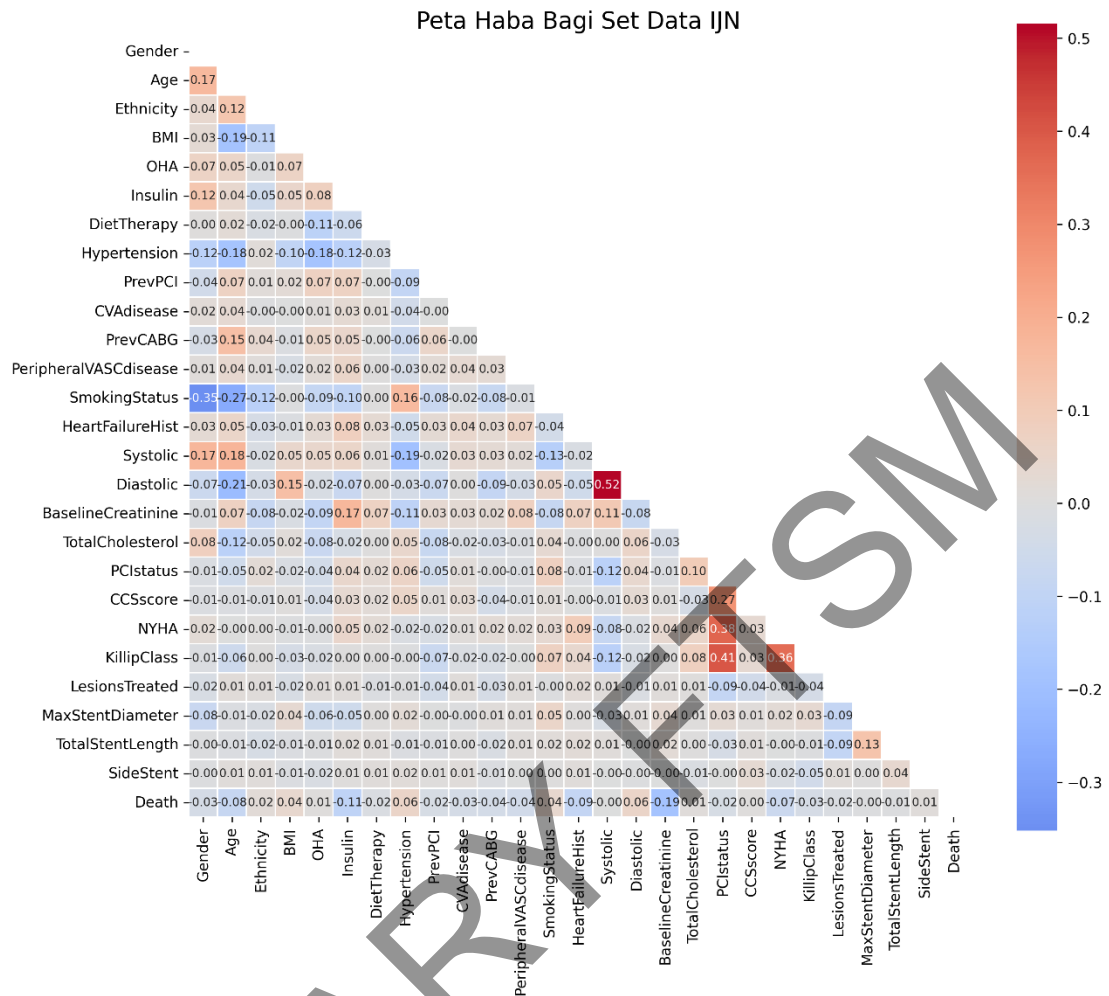
bersambung...

...sambungan			
SmokingStatus	3	0	11143 (55.72%)
HeartFailureHist	2	0	19002 (95.01%)
PCIstatus	3	0	17542 (87.71%)
CCSscore	5	0	8840 (44.20%)
NYHA	4	0	12264 (61.32%)
KillipClass	5	0	15102 (75.51%)
SideStent	2	0	19469 (97.34%)
Death	4	3	17630 (88.15%)

LIBRARY ETSU



Rajah 3.4 Taburan fitur bagi set data pra-pemprosesan menggunakan histogram



Rajah 3.5 Kolerasi peta haba Spearman antara fitur dengan pembolehubah sasaran bagi set data IJN

3.5 PEMODELAN

Dalam bahagian ini, seni bina yang dicadangkan untuk model ramalan hasil prosedur angioplasti menggunakan hutan rawak, *gradient boosting*, dan XGBoost dibincangkan dengan lebih terperinci. Pembahagian data untuk pembangunan model serta hiperparameter penting yang khusus untuk algoritma pengelasan ini turut dibincangkan dalam bahagian ini.

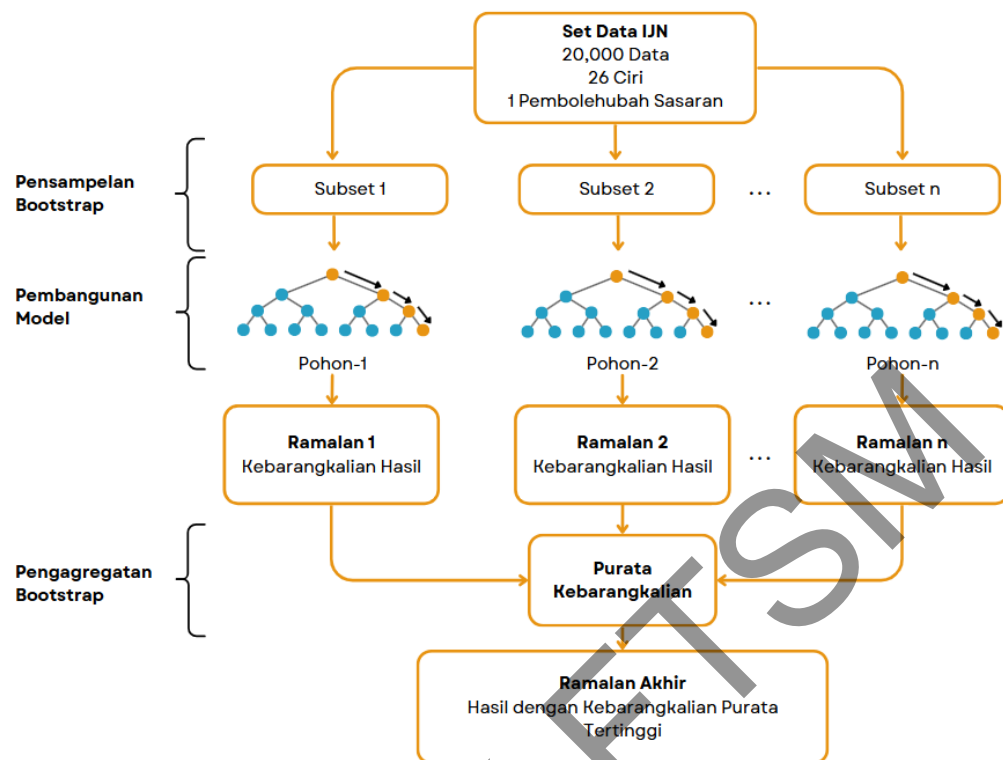
3.5.1 Pemilihan Teknik Pemodelan

Dalam Bab 2, ulasan kajian kesusasteraan menunjukkan bahawa kaedah gabungan sering melangkaui kedua-dua algoritma statistik dan pembelajaran mesin tradisional

dalam pemodelan hasil perubatan. Penyelidikan oleh Zhao et al. (2023), Deng et al. (2022), Liu et al. (2021) dan Zack et al. (2019) menunjukkan bahawa hutan rawak secara konsisten mengatasi teknik pembelajaran mesin lain, termasuk pendekatan gabungan yang lain. Begitu juga kajian oleh Niimi et al. (2022), Mortazavi BJ et al. (2019), Galimzhanov et al. (2023) dan Hamilton et al. (2024) menunjukkan bahawa XGBoost menunjukkan prestasi yang lebih unggul. Manakala kajian menggunakan kaedah boosting seperti *gradient boosting* oleh Jun Ke et al. (2022) membuktikan antara pilihan popular bagi mengkaji set data perubatan. Berdasarkan penemuan ini, kajian ini menggunakan algoritma hutan rawak, *gradient boosting*, dan XGBoost untuk membangunkan model bagi meramalkan hasil angioplasti. Perpustakaan scikit-learn digunakan untuk membangunkan model hutan rawak dan *gradient boosting*, manakala perpustakaan XGBoost untuk Python digunakan untuk membina model XGBoost. Bahagian-bahagian berikut menerangkan seni bina yang dicadangkan untuk model hutan rawak, *gradient boosting*, dan XGBoost.

a. **Arkitektur Hutan Rawak**

Untuk membangunkan model hutan rawak, set data IJN yang telah diproses digunakan di mana ianya merangkumi 20,000 rekod pesakit, 26 fitur, dan satu pembolehubah sasaran. Pada mulanya, set data ini dibahagikan kepada subset melalui pensampelan *bootstrap* dengan mengekalkan saiz yang sama seperti input asal kerana parameter *max_samples* adalah *None* secara lalai (scikit-learn 2024a). Setiap pohon dalam hutan rawak menghasilkan kebarangkalian untuk pelbagai hasil angioplasti. Jumlah pohon ditentukan oleh hiperparameter *n_estimators*. Hasil akhir model diklasifikasikan berdasarkan purata kebarangkalian tertinggi. Sebagai contoh, jika purata kebarangkalian untuk kategori "berjaya" lebih tinggi daripada kategori "tidak berjaya", maka hasil tersebut akan dikelaskan sebagai "berjaya". Seni bina model yang dicadangkan ini digambarkan di dalam Rajah 3.6.

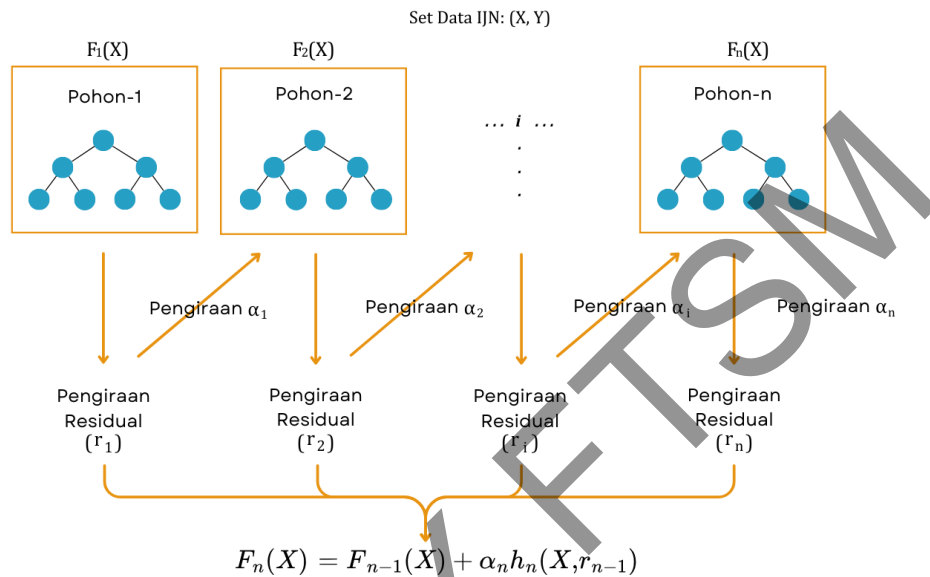


Rajah 3.6 Seni bina model hutan rawak yang dicadangkan

b. Arkitektur Gradient Boosting

Untuk membangunkan model *gradient boosting*, set data IJN yang telah diproses digunakan seperti yang digunakan pada model hutan rawak. Ianya yang merangkumi 20,000 rekod pesakit, 26 fitur, dan satu pemboleh ubah sasaran. Untuk kajian mengenai ramalan hasil angioplasti menggunakan set data IJN, *GradientBoostingClassifier* lebih sesuai digunakan berbanding *GradientBoostingRegressor* kerana pembolehubah sasarannya adalah berjenis kategori. Hasil kajian ini merupakan output kategori diskret seperti "Semasa Discaj," "3 Bulan," "1 Tahun," dan "Melebihi 1 Tahun". Pohon pertama dilatih dengan data asal, manakala setiap pohon berikutnya dilatih dengan sisa (ralat) dari pohon sebelumnya untuk meningkatkan prestasi model. Jumlah pohon ditentukan oleh hiperparameter $n_estimators$ dan kadar pembelajaran dikawal oleh hiperparameter $learning_rate$ untuk menyesuaikan sumbangan setiap pohon. Semua pohon pembelajaran mempunyai pemberat yang sama dalam kes *gradient boosting*. Pemberat biasanya ditetapkan sebagai kadar pembelajaran (α) yang kecil dalam magnitud. Ramalan akhir dalam *gradient boosting* adalah kombinasi berwajaran daripada ramalan semua pohon, di mana kadar pembelajaran menentukan sumbangan setiap pohon. Dengan menyesuaikan kadar pembelajaran, pengaruh setiap pohon individu pada model

keseluruhan dapat dikawal. Penyesuaian ini membolehkan model menjadi lebih fleksibel dan sering kali dapat meningkatkan prestasinya. Seni bina model yang dicadangkan ini digambarkan dalam Rajah 3.7.



Rajah 3.7 Seni bina model *gradient boosting* yang dicadangkan

Rajah 3.8 menunjukkan hasil akhir iaitu model *gradient boosting*, $F_n(X)$ di mana α_i dan r_i adalah parameter pengawalseliaan dan sisa yang dikira dengan pohon ke- i , dan h_i adalah fungsi yang dilatih untuk meramalkan sisa, r_i menggunakan X untuk pohon ke- i . Untuk mengira α_i kita menggunakan ralat yang dikira, r_i dan mengira seperti berikut:

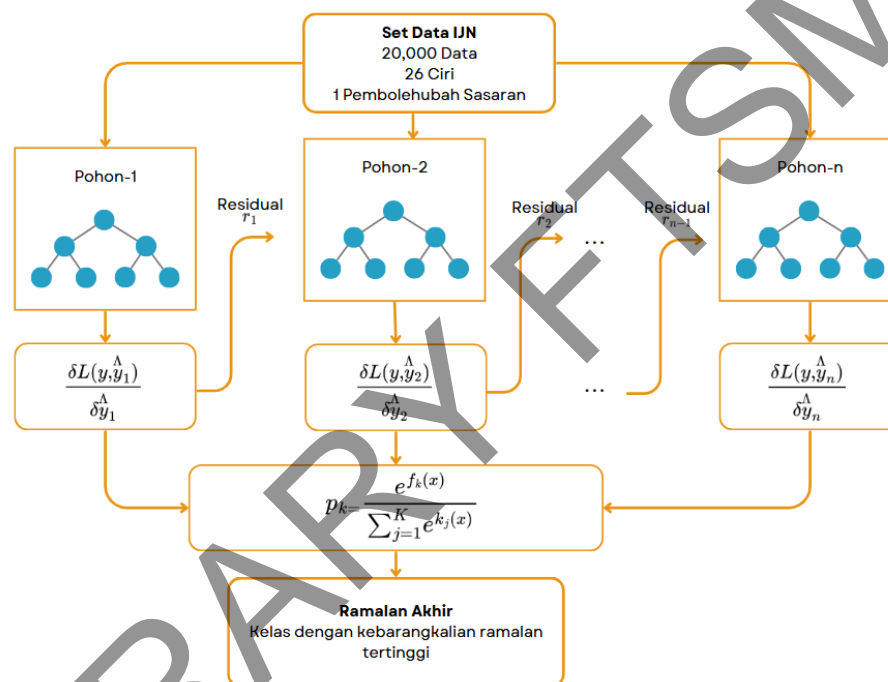
$$\arg \min_{\alpha} = \sum_{i=1}^n L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1})) \quad \dots(3.1)$$

di mana $L(Y, F(X))$ adalah fungsi kerugian (*loss function*) yang boleh dibezakan.

c. Arkitektur XGBoost

Seperti seni bina hutan rawak, model XGBoost menggunakan set data IJN yang telah diproses sebagai input. Berbeza dengan hutan rawak, tiada pensampelan data sebelum pembangunan pohon dilaksanakan kerana nilai lalai untuk hiperparameter subsample

ditetapkan kepada 1 (XGBoost 2024). Bilangan pohon dalam model XGBoost ditentukan oleh hiperparameter $n_estimator$. Untuk $n > 1$, setiap pohon dibangunkan menggunakan sisa dari pohon iterasi sebelumnya (Deng et al. 2022). Ramalan akhir adalah merupakan jumlah ramalan daripada semua pohon. Dengan hiperparameter $objective = multi:softmax$ untuk klasifikasi berbilang kelas, output model XGBoost ditentukan dengan kebarangkalian ramalan tertinggi. Seni bina model XGBoost yang dicadangkan untuk meramalkan hasil angioplasti ditunjukkan dalam Rajah 3.8.



Rajah 3.8 Seni bina model XGBoost yang dicadangkan

Keputusan atau ramalan akhir di tentukan melalui pengiraan menggunakan fungsi softmax, p_k di mana $f_k(x)$ adalah skor output bagi kelas k daripada model XGBoost dan K adalah jumlah keseluruhan kelas.

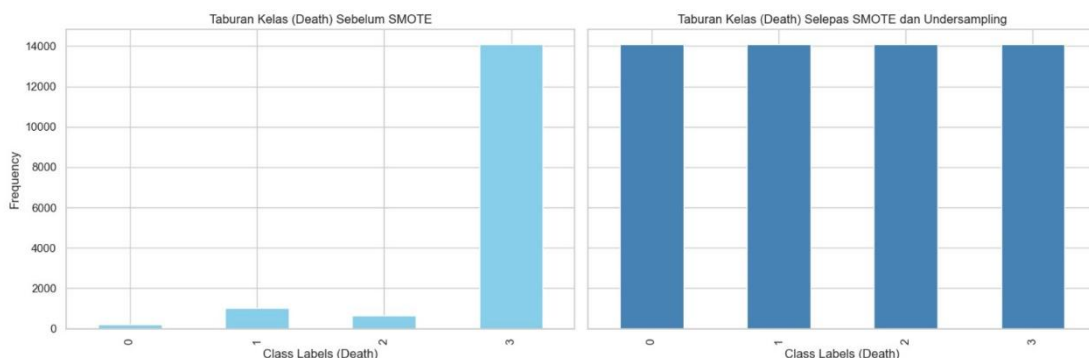
3.5.2 Pembahagian dan Penyeimbangan Data

Kajian ini menggabungkan SMOTE (*Synthetic Minority Over-sampling Technique*) dengan *undersampling* di mana ia merupakan satu teknik yang popular untuk menangani ketidakseimbangan kelas dalam set data. Pendekatan ini bertujuan untuk menyeimbangkan pengagihan kelas dengan menambah bilangan sampel sintetik dalam kelas minoriti melalui SMOTE serta mengurangkan bilangan sampel dalam kelas

majoriti melalui *undersampling*. Penggunaan SMOTE sahaja mungkin menghasilkan terlalu banyak sampel sintetik, terutamanya jika ketidakseimbangan kelas terlalu besar. Ini boleh menyebabkan *overfitting* pada kelas minoriti. Sebaliknya, jika hanya *undersampling* digunakan dalam menyeimbangkan data, ia boleh menyebabkan kehilangan maklumat penting daripada kelas majoriti kerana terlalu banyak sampel yang dibuang. Oleh itu, dengan menggabungkan kedua-dua teknik ini, SMOTE akan memastikan kelas minoriti diwakili dengan secukupnya, manakala *undersampling* menghalang kelas majoriti daripada mendominasi proses pembelajaran.

Berikut merupakan langkah-langkah menggabungkan SMOTE dengan *undersampling*. Rajah 3.9 menunjukkan perbezaan taburan data sebelum dan selepas teknik ini dilaksanakan:

1. **Pembahagian data awal:** Sebelum menerapkan SMOTE dan *undersampling*, data dibahagikan kepada set latihan (*train set*) dan set pengujian (*test set*). SMOTE dan *undersampling* hanya dilakukan pada data latihan untuk mengelakkan kebocoran data (*data leakage*). Data pengujian dikekalkan dalam keadaan asal dengan ketidakseimbangan kelas yang semula jadi, bagi menilai prestasi model dalam data dunia sebenar.
2. **Aplikasi SMOTE pada kelas minority:** Langkah pertama adalah menggunakan SMOTE untuk menghasilkan sampel sintetik bagi kelas minoriti pada data latihan. SMOTE mencipta data sintetik baharu antara data sedia ada menggunakan kaedah k-jiran terdekat (kNN). Langkah ini meningkatkan saiz kelas minoriti dengan menambah sampel sintetik untuk mengimbangi pengagihan kelas dengan lebih baik.



Rajah 3.9 Sebelum dan selepas penyeimbangan data

3. **Undersampling pada kelas majoriti:** Walaupun selepas *oversampling* dengan SMOTE, data mungkin masih didominasi oleh kelas majoriti. Untuk mengatasinya, *undersampling* rawak diterapkan pada kelas majoriti, di mana sampel daripada kelas majoriti dibuang secara rawak. Tujuannya adalah untuk mengurangkan bilangan sampel dalam kelas majoriti tanpa kehilangan terlalu banyak maklumat penting.
4. **Menggabungkan SMOTE dan *undersampling*:** Gabungan kedua-dua teknik ini melibatkan *oversampling* kelas minoriti terlebih dahulu menggunakan SMOTE. Kemudian diikuti dengan *undersampling* kelas majoriti untuk mengurangkan saiznya. Hasil akhir dari teknik ini adalah set data yang lebih seimbang, di mana kelas minoriti dan majoriti mempunyai saiz sampel yang lebih serupa, menjadikan model lebih berupaya mengenal pasti pola dari kedua-dua kelas dengan tepat.

3.5.3 Penalaan Hiperparameter

Hiperparameter adalah parameter model yang boleh ditetapkan sebelum proses latihan model dan mempengaruhi prestasi model (López et al. 2022). Mengoptimumkan hiperparameter adalah penting untuk meningkatkan ketepatan dan ketahanan model. Dalam kajian ini, penalaan hiperparameter dilakukan untuk mencari kombinasi terbaik nilai hiperparameter bagi mengoptimumkan ramalan luar sampel. Pendekatan *Grid Search* digunakan untuk penalaan hiperparameter di mana satu set nilai yang telah ditetapkan untuk setiap hiperparameter diterokai dan model dilatih dan dievaluasi merentasi semua kombinasi nilai tersebut.

a. Hiperparameter Hutan Rawak

Bagi model hutan rawak, beberapa hiperparameter utama ditala dalam kajian ini termasuklah *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, dan *max_features*. Penjelasan dan kepentingan setiap hiperparameter adalah seperti berikut:

1. ***n_estimators***: Menentukan bilangan pohon keputusan yang membentuk hutan rawak (Owen 2022). Secara umum, prestasi model meningkat dengan peningkatan bilangan pohon tetapi ia memerlukan masa pengiraan yang lebih lama. Walau bagaimanapun, bagi bilangan tertentu untuk *estimators*, peningkatan prestasi model menjadi marginal. Selain itu, *overfitting* lebih berkemungkinan berlaku apabila model mempunyai bilangan pohon yang lebih tinggi.
2. ***max_depth***: Kedalaman maksimum bagi pohon keputusan (Owen 2022). Kedalaman maksimum yang tinggi boleh menyebabkan *overfitting*, manakala kedalaman maksimum yang rendah boleh menyebabkan *underfitting*. Jika kedalaman maksimum ditetapkan sebagai *none*, pohon akan terus berpecah sehingga bilangan sampel dalam setiap daun lebih rendah daripada *min_samples_split*.
3. ***min_samples_split***: Menentukan bilangan minimum sampel yang diperlukan untuk membahagikan nod kepada nod anak (Owen 2022). Bilangan minimum sampel yang tinggi bagi pemisahan boleh mengelakkan *overfitting* model. Selain itu, bilangan minimum sampel yang tinggi bagi pemisahan boleh meningkatkan prestasi dengan kos pengiraan yang lebih tinggi (Krauß 2022).
4. ***min_samples_leaf***: Menunjukkan bilangan minimum sampel yang diperlukan dalam setiap daun (Owen 2022). Bilangan minimum sampel yang tinggi dalam daun boleh mengelakkan *overfitting* model. Sebaliknya, bilangan minimum sampel yang rendah dalam daun menjadikan model lebih kompleks (Krauß 2022).

5. ***max_features***: Bilangan fitur yang digunakan oleh pohon keputusan dalam hutan rawak untuk pemisahan nod (Owen 2022). Nilai maksimum fitur yang tinggi boleh mengurangkan varians dengan kos pengiraan yang lebih tinggi. Selain itu, nilai maksimum fitur yang tinggi boleh menyebabkan *overfitting* pada model (Krauß 2022). Jika *max_features* ditetapkan sebagai *none*, semua fitur akan dipertimbangkan pada setiap pemisahan (scikit-learn 2024a).

Definisi hiperparameter dan nilai lalai seperti yang digariskan dalam scikit-learn (scikit-learn 2024a) serta nilai yang digunakan untuk penalaan hiperparameter model hutan rawak boleh didapati dalam Jadual 3.25.

Jadual 3.25 Penalaan hiperparameter untuk hutan rawak

Hiperparameter	Definisi	Nilai Lalai	Nilai Penalaan
<i>n_estimators</i>	Bilangan pohon dalam hutan.	100	100, 200, 500
<i>max_depth</i>	Kedalaman maksimum setiap pohon	None	None, 20, 30
<i>min_samples_split</i>	Bilangan minimum sampel untuk memecahkan nod dalaman.	2	2, 4, 6
<i>min_samples_leaf</i>	Bilangan minimum sampel dalam nod daun.	1	1, 2, 5
<i>max_features</i>	Bilangan fitur yang dipertimbangkan untuk setiap pemecahan.	sqrt	sqrt, log2, None

b. Hiperparameter Gradient Boosting

Gradient boosting adalah teknik pembelajaran mesin yang popular kerana keupayaannya menghasilkan model ramalan yang sangat tepat dengan mengurangkan kesilapan model sebelumnya secara berulang. Dalam kaedah ini, model lemah dibina secara berturutan untuk memperbaiki kesilapan yang dilakukan oleh model-model sebelumnya, menjadikannya antara pilihan utama untuk tugas klasifikasi dan regresi. Hiperparameter utama untuk *gradient boosting* yang akan ditala adalah seperti berikut:

1. ***n_estimators***: Menentukan bilangan pohon keputusan yang digunakan dalam pembinaan model (AnalyticsVidhya 2024). Bilangan pohon yang lebih tinggi biasanya meningkatkan prestasi model, tetapi mungkin menyebabkan masa

pengiraan yang lebih lama dan risiko *overfitting*. Nilai yang biasa digunakan adalah antara 100 hingga 500.

2. ***learning_rate***: Kadar pembelajaran mengawal sumbangan setiap pohon terhadap model akhir (MachineLearningMastery 2024). Nilai yang lebih kecil seperti 0.01 atau 0.1 memberikan prestasi yang lebih baik dengan mengurangkan kesilapan tetapi memerlukan bilangan pohon yang lebih banyak. Kadar pembelajaran yang tinggi pula mungkin menghasilkan model yang cepat tetapi kurang tepat.
3. ***max_depth***: Kedalaman maksimum pohon keputusan (AnalyticsVidhya 2024). Pohon yang lebih dalam berpotensi menangkap lebih banyak corak dalam data, namun berisiko untuk *overfitting*. Nilai yang biasa dipertimbangkan adalah antara 3 hingga 10.
4. ***min_samples_split***: Jumlah minimum sampel yang diperlukan untuk memecahkan nod dalaman (AnalyticsVidhya 2024). Nilai yang lebih tinggi membantu mengawal *overfitting* dengan mengurangkan kerumitan pohon. Pilihan nilai biasanya adalah 2, 5, atau 10.
5. ***min_samples_leaf***: Bilangan minimum sampel dalam setiap daun (AnalyticsVidhya 2024). Nilai yang lebih tinggi menghasilkan pohon yang lebih ringkas dan mengurangkan kemungkinan *overfitting*. Nilai yang biasa digunakan adalah 1, 5, atau 10.
6. ***subsample***: Menentukan pecahan sampel latihan yang digunakan untuk setiap pohon (AnalyticsVidhya 2024). Nilai antara 0.5 hingga 1.0 membantu mengurangkan varians dan *overfitting* dalam model.

Nilai lalai dan julat penalaan untuk hiperparameter ini ditunjukkan dalam Jadual

3.26:

Jadual 3.26 Penalaan hiperparameter untuk *gradient boosting*

Hiperparameter	Definisi	Nilai Lalai	Nilai Penalaan
<i>n_estimators</i>	Bilangan pohon keputusan dalam model.	100	100, 200, 500
<i>learning_rate</i>	Kadar pembelajaran yang mengawal sumbangan setiap pohon.	0.1	0.01, 0.05, 0.1, 0.2
<i>max_depth</i>	Kedalaman maksimum setiap pohon keputusan.	3	3, 5, 10
<i>min_samples_split</i>	Bilangan minimum sampel untuk memecahkan nod dalaman.	2	2, 5, 10
<i>min_samples_leaf</i>	Bilangan minimum sampel di setiap daun.	1	1, 2, 5
<i>subsample</i>	Pecahan sampel latihan yang digunakan untuk setiap pohon.	1.0	0.5, 0.75, 1.0

c. Hiperparameter XGBoost

Untuk XGBoost, *n_estimators*, *max_depth*, *learning_rate*, *min_child_weight*, dan *colsample_bytree* adalah hiperparameter yang digunakan dalam penalaan model ini. Berikut adalah penjelasan bagi setiap hiperparameter yang ditala dalam kajian ini:

1. ***n_estimators***: Bilangan pohon keputusan yang digunakan untuk membina model XGBoost (Wade & Glynn 2020). Berbeza dengan hutan rawak, pohon keputusan dalam XGBoost dilatih berdasarkan sisa (*residual*) model sebelumnya. Untuk mengelakkan *overfitting*, kaedah *early stopping* boleh digunakan pada hiperparameter ini.
2. ***max_depth***: Kedalaman maksimum pohon keputusan dalam XGBoost (Krauß 2022). Peningkatan kedalaman pohon boleh menyebabkan *overfitting* dan meningkatkan kerumitan model. Justeru, kedalaman maksimum yang sesuai perlu ditetapkan bagi mengimbangi ketepatan dan kerumitan model.
3. ***learning_rate***: Menentukan saiz langkah bagi penyesuaian turunan separa fungsi kehilangan, juga dikenali sebagai *gradient* (Krauß 2022). Kadar pembelajaran ini mengawal pengecilan pemberat pohon dalam algoritma *boosting* dan nilai yang lebih rendah boleh membantu mengurangkan *overfitting* (Wade & Glynn 2020).

4. ***min_child_weight***: Menentukan jumlah minimum pemberat yang diperlukan untuk memecahkan satu nod menjadi anak nod (Krauß 2022). Jika jumlah pemberat berada di bawah nilai minimum yang ditetapkan, proses pemisahan akan dihentikan. Nilai yang lebih tinggi bagi *min_child_weight* dapat mengurangkan risiko *overfitting* (Wade & Glynn 2020).
5. ***colsample_bytree***: Pecahan fitur dalam set latihan yang digunakan untuk membina setiap pohon dalam model XGBoost (Krauß 2022). Hiperparameter ini serupa dengan *max_features* pada hutan rawak, tetapi diaplikasikan pada peringkat pohon dalam XGBoost. Nilai ini mengawal bilangan fitur yang diambil sampel untuk pembinaan pohon, dan nilai yang lebih rendah boleh membantu mengurangkan varians model.

Definisi hiperparameter XGBoost dan nilai lalai berdasarkan dokumentasi XGBoost (XGBoost 2024) boleh didapati dalam Jadual 3.27 berikut bersama dengan nilai penalaan bagi setiap hiperparameter.

Jadual 3.27. Penalaan hiperparameter untuk XGBoost

Hiperparameter	Definisi	Nilai Lalai	Nilai Penalaan
<i>n_estimators</i>	Bilangan iterasi <i>boosting</i> .	100	100, 200, 500
<i>max_depth</i>	Kedalaman maksimum pohon.	6	3, 6, 10
<i>learning_rate</i>	Saiz langkah untuk mengurangkan pemberat fitur selepas setiap iterasi.	0.3	0.01, 0.1, 0.2
<i>min_child_weight</i>	Jumlah minimum pemberat dalam nod untuk pemisahan.	1	0, 1, 3
<i>colsample_bytree</i>	Pecahan kolum yang diambil sampel untuk setiap pohon.	1	0.5, 0.8, 1

Grid Search memastikan bahawa model diuji merentasi pelbagai kombinasi hiperparameter untuk mencari konfigurasi terbaik bagi ramalan hasil angioplasti. Memandangkan kajian ini melibatkan set data perubatan dan telah menangani ketidakseimbangan kelas, penalaan hiperparameter akan membantu memastikan model dioptimumkan untuk menangani dataset tersebut.

3.6 PENILAIAN

Selepas pembangunan model, prestasi ketiga-tiga model dinilai dan dibandingkan menggunakan pelbagai metrik yang relevan untuk klasifikasi pelbagai kelas. Selain itu, ujian statistik dijalankan untuk menentukan sama ada terdapat perbezaan yang signifikan dalam prestasi antara model-model tersebut. Selain menilai metrik utama, kepentingan fitur diperhatikan untuk memahami bagaimana setiap fitur mempengaruhi ramalan model. Bagi mendapatkan pemahaman yang lebih jelas tentang hubungan antara fitur dan output model, SHAP digunakan untuk model hutan rawak, *gradient boosting*, dan XGBoost.

3.6.1 Metrik Utama Untuk Penilaian Prestasi Model

Model yang dibangunkan dinilai menggunakan ukuran yang sesuai untuk tugas klasifikasi pelbagai kelas, seperti ketepatan (*accuracy*), kejituan (*precision macro*), *recall (recall macro)*, skor F1 (*F1-score macro*), dan luas di bawah lengkung AUC (*AUC-Macro*). Metrik-metrik ini memastikan bahawa semua kelas diberikan kepentingan yang sama tanpa mengambil kira pengagihan kelas.

Ketepatan (*Accuracy*) ialah kadar keseluruhan sampel yang diklasifikasikan dengan betul oleh model. Dalam konteks klasifikasi pelbagai kelas, ketepatan memberikan gambaran umum mengenai sejauh mana model dapat mengklasifikasikan sampel ke dalam kelas yang betul. Ia boleh dikira menggunakan formula berikut (Han et al. 2012):

$$Accuracy = \frac{\sum_{i=1}^k TP_i}{\sum_{i=1}^k (TP_i + FN_i + FP_i)} \quad \dots(3.2)$$

Di mana k adalah bilangan kelas dan TP_i , FN_i serta FP_i masing-masing merujuk kepada bilangan *true positive*, *false negative*, dan *false positive* bagi setiap kelas i . Ketepatan digunakan sebagai satu ukuran global, tetapi ia boleh terkesan oleh ketidakseimbangan data kerana ia tidak mengambil kira bagaimana prestasi model terhadap setiap kelas secara individu.

Selain ketepatan, metrik kejutuan (*Precision Macro*) dan *Recall Macro* lebih sesuai untuk menilai prestasi model dalam klasifikasi pelbagai kelas. *Precision Macro* mengukur sejauh mana model dapat membuat ramalan yang tepat bagi setiap kelas tanpa mengambil kira saiz kelas tersebut, manakala *Recall Macro* menilai sensitiviti model dalam mengenal pasti sampel bagi setiap kelas. Formula bagi kedua-dua metrik ini adalah seperti berikut:

$$Precision\ Macro = \frac{1}{k} \sum_{i=1}^k \frac{TP_i}{TP_i + FP_i} \quad \dots(3.3)$$

$$Recall\ Macro = \frac{1}{k} \sum_{i=1}^k \frac{TP_i}{TP_i + FN_i} \quad \dots(3.4)$$

Di mana k adalah bilangan kelas dalam set data dan TP_i, FP_i serta FN_i masing-masing merujuk kepada bilangan *true positive*, *false positive* dan *false negative*, bagi setiap kelas i . *Precision Macro* dan *Recall Macro* memberikan penilaian yang lebih seimbang terhadap model kerana setiap kelas diberi kepentingan yang sama tanpa mengambil kira pengagihan data.

Salah satu metrik utama yang digunakan untuk menilai prestasi model dalam klasifikasi pelbagai kelas ialah Skor F1 Macro (*Macro-F1 Score*), yang merupakan purata harmonik antara *Precision Macro* dan *Recall Macro*. Skor F1 berguna dalam kes di mana terdapat ketidakseimbangan antara kelas kerana ia mempertimbangkan kedua-dua *precision* dan *recall* dalam satu ukuran bersepadu. Formula bagi Skor F1 Macro adalah seperti berikut :

$$F1 - Score\ Macro = \frac{1}{k} \sum_{i=1}^k \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad \dots(3.5)$$

Matriks kekeliruan (*Confusion Matrix*) digunakan untuk menggambarkan distribusi kelas sebenar berbanding dengan kelas yang diramalkan oleh model. Dalam konteks klasifikasi pelbagai kelas, matriks kekeliruan berbentuk matriks $k \times k$, di mana k adalah jumlah keseluruhan kelas. Setiap baris dalam matriks mewakili kelas sebenar, manakala setiap lajur mewakili kelas yang diramalkan oleh model. Matriks ini membolehkan kelemahan model dalam mengenal pasti kelas tertentu, terutamanya kelas minoriti, dapat dikenal pasti dengan lebih jelas. Penggunaan matriks kekeliruan juga membantu dalam analisis lebih mendalam tentang pola kesalahan model dalam mengklasifikasikan sampel.

Luas di bawah lengkung atau AUC juga merupakan ukuran lain untuk menilai prestasi model. AUC mempunyai nilai antara 0 dan 1, di mana pengklasifikasi dengan kemampuan tekaan rawak mempunyai AUC sebanyak 0,5 (Shen et al. 2021). Pengklasifikasi yang baik sepatutnya mempunyai nilai AUC yang tinggi.

Walaupun lengkungan AUC biasanya digunakan untuk klasifikasi binari, ia masih boleh digunakan untuk masalah berbilang kelas dengan beberapa pengubahsuaian (Saito et al. 2015). *Macro-Average AUC* ialah lanjutan daripada AUC untuk masalah klasifikasi berbilang kelas. Dalam klasifikasi binari, AUC mewakili kebarangkalian bahawa model meletakkan contoh positif secara rawak lebih tinggi daripada contoh negatif secara rawak. Untuk masalah berbilang kelas, konsep ini diadaptasi dengan mengira AUC bagi setiap kelas secara individu menggunakan pendekatan satu-lawan-semua atau *One-vs-Rest* (Hand et al. 2001).

Pendekatan satu-lawan-semua digunakan untuk mengira AUC bagi setiap kelas secara individu. Dalam pendekatan satu-lawan-semua, setiap kelas dianggap sebagai kelas "positif" sementara semua kelas lain dikumpulkan sebagai "negatif." Ini menghasilkan skor AUC untuk setiap kelas secara individu. Misalnya, terdapat k kelas dan setiap kelas mempunyai nilai AUC yang ditandakan sebagai AUC_i untuk $i = 1, 2, \dots, k$. Purata tanpa wajaran kemudian dikira bagi nilai AUC untuk semua kelas (Saito et al. 2015).

$$\text{Macro - Average AUC} = \frac{1}{k} \sum_{i=1}^k \text{AUC}_i \quad \dots(3.6)$$

Di mana k ialah jumlah keseluruhan kelas. AUC_i ialah skor AUC untuk setiap kelas i dalam setup satu-lawan-semua. Pendekatan ini memberikan kepentingan yang sama kepada setiap kelas tanpa mengambil kira pengagihan kelas apabila kelas di dalam kajian ini seimbang setelah menggunakan SMOTE (rujuk 3.5.2 Pembahagian dan Penyeimbangan Data).

3.6.2 Ujian Statistik Pada Prestasi Model

Selepas penilaian model, ujian statistik dijalankan untuk menentukan model mana yang menunjukkan prestasi yang lebih baik berbanding model-model lain yang dibangunkan. Ujian statistik adalah penting untuk memastikan sama ada sesuatu model berprestasi dengan signifikan secara statistik lebih baik daripada model-model yang lain (Han et al. 2012). Dalam kajian ini, purata prestasi model hutan rawak, *gradient boosting* dan XGBoost terbaik dengan kaedah *holdout* dan *cross-validation* dibandingkan menggunakan ujian-t (*t-test*). Untuk ujian statistik dalam kajian ini, ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang yang diperkenalkan oleh Dietterich (1998) digunakan.

Salah satu sebab pelaksanaan ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang adalah kerana kaedah ini menunjukkan kebarangkalian yang boleh diterima bagi ralat Jenis I (*Type I error*). Sebaliknya, ujian-t berpasangan dengan *10-fold cross-validation* menunjukkan kemungkinan lebih tinggi untuk ralat Jenis I (Dietterich 1998). Dengan ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang ini, purata prestasi ketiga-tiga model dibandingkan dengan melaksanakan *2-fold cross-validation* sebanyak lima kali ulangan. Berdasarkan eksperimen Dietterich (1998), ralat Jenis I boleh meningkat jika bilangan ulangan *cross-validation* kurang atau melebihi lima. Selain itu, ujian-t berpasangan mampu memberikan anggaran variasi yang lebih baik dengan *2-fold cross-validation* kerana set latihan adalah bebas antara satu sama lain tanpa sebarang pertindihan berbanding dengan *10-fold cross-validation*.

Dalam kajian ini, ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang diaplikasikan menggunakan pustaka *mlxtend* untuk Python (Raschka 2018). Pustaka ini digunakan untuk membandingkan ketepatan pengklasifikasi dan menentukan kadar signifikan statistik perbezaan prestasi model. Dengan pustaka ini, nilai statistik-t (*t-statistics*) dan *p-value* dua hala dihasilkan (Raschka 2023).

3.6.3 Pemilihan Fitur

Untuk menilai pemilihan fitur bagi model hutan rawak dan *gradient boosting*, *Mean Decrease in Impurity* (MDI) digunakan kerana pembahagian nod untuk membina pohon keputusan ditentukan melalui pemaksimuman penurunan kekotoran (Aydede 2023). Kekotoran Gini (*Gini impurity*) digunakan sebagai ukuran kekotoran untuk pemilihan fitur dalam hutan rawak kerana ia juga digunakan dalam pembangunan model. Penurunan kekotoran ialah pengurangan kekotoran antara nod induk dan purata kekotoran berwajaran bagi nod anak. Penurunan kekotoran bagi sesuatu fitur dalam pohon keputusan dikira pada setiap pembahagian di mana fitur tersebut digunakan. Penurunan kekotoran bagi sesuatu fitur dalam hutan rawak atau *gradient boosting* ialah jumlah penurunan kekotoran merentasi semua pohon. Untuk mengira penurunan purata kekotoran, jumlah penurunan kekotoran sesuatu fitur dalam hutan rawak atau *gradient boosting* dibahagikan dengan jumlah keseluruhan penurunan kekotoran bagi semua fitur.

Pemilihan fitur bagi XGBoost dinilai menggunakan *gain* fitur (*feature gain*) kerana XGBoost bertujuan untuk memaksimumkan *gain* dengan meminimumkan fungsi kehilangan (*loss function*) (Tu 2020). Semasa pembahagian daun (*leaf splitting*), *gain* bagi semua fitur dikira. Kemudian, segmen dengan *gain* tertinggi dipilih kerana *gain* yang lebih tinggi menunjukkan kehilangan yang lebih rendah selepas pembahagian. Dalam XGBoost (2023c), *gain* ditakrifkan sebagai purata *gain* yang diperhatikan merentasi semua kejadian di mana fitur tersebut digunakan dalam pembahagian.

Kaedah ini dipilih kerana ia memberikan gambaran yang lebih tepat mengenai sumbangan setiap fitur kepada ketepatan keseluruhan model dalam konteks pembelajaran gabungan serta memudahkan analisis dengan menggunakan satu ukuran

kepentingan yang seragam untuk semua model. Dengan menggunakan kaedah ini, fitur yang mempunyai sumbangan tertinggi kepada model dapat dikenal pasti dengan jelas dan membantu dalam menafsir hasil serta membuat keputusan yang lebih berinformasi berkaitan ramalan hasil angioplasti (Rozemberczki et al. 2022).

3.6.4 Shapley Additive Explanations (SHAP)

Walaupun pemilihan fitur berdasarkan *gain* dapat menunjukkan kepentingan setiap fitur dalam model, ia tidak dapat menjelaskan hubungan antara fitur dan hasil model secara terperinci. Untuk mengatasi kelemahan ini, SHAP diaplikasikan dalam kajian ini bagi meningkatkan kefahaman tentang bagaimana setiap fitur menyumbang kepada hasil model bagi setiap sampel. Konsep asas SHAP adalah nilai Shapley yang menjelaskan ramalan dengan menjumlahkan nilai sumbangan setiap fitur secara individu bagi setiap ramalan (Lundberg & Lee 2017).

Dalam kajian ini, pustaka SHAP untuk Python digunakan untuk mengira dan memvisualisasikan nilai SHAP. Nilai SHAP dihasilkan dalam unit kebarangkalian yang sesuai dengan hasil ramalan model hutan rawak, *gradient boosting*, dan XGBoost. SHAP memberikan interpretasi berdasarkan *feature gain*, yang membolehkan kita melihat bagaimana setiap fitur mempengaruhi hasil model dalam ketiga-tiga model ini secara konsisten (Molnar 2022).

Visualisasi nilai SHAP dijalankan melalui plot beeswarm untuk membantu memahami hubungan antara fitur dan hasil model. Plot beeswarm memaparkan koleksi nilai SHAP setiap fitur bagi setiap sampel. Setiap nilai SHAP ditunjukkan sebagai titik dalam plot, di mana kedudukan titik mencerminkan nilai SHAP yang sepadan, dan warna menunjukkan nilai fitur tersebut. Dengan cara ini, nilai SHAP memberikan pandangan yang lebih mendalam mengenai peranan setiap fitur dalam ramalan angioplasti, memudahkan interpretasi hasil yang lebih bermakna bagi kajian ini (Lundberg 2023).

3.7 KESIMPULAN

Sebagai kesimpulan, bab ini menggariskan kaedah penyelidikan yang digunakan untuk membangunkan model ramalan hasil angioplasti. Langkah-langkah atau kaedah penyelidikan yang digunakan adalah termasuk pemahaman situasi semasa melalui kajian kesusasteraan, pengumpulan data, dan penyediaan data yang melibatkan pengendalian nilai yang hilang, penyingkiran *outlier*, pemetaan semula, dan pengekodan. Pada peringkat seterusnya, model ramalan hasil angioplasti dibina menggunakan kaedah gabungan khususnya hutan rawak, *gradient boosting*, dan XGBoost. Selanjutnya, model-model ini dinilai dengan pelbagai metrik klasifikasi. Untuk mengenal pasti kaedah gabungan yang terbaik, ujian-t berpasangan menggunakan 5x2-lipat dengan validasi silang digunakan sebagai ujian statistik terhadap prestasi model. Selain itu, analisis kepentingan fitur dan SHAP diaplikasikan untuk menentukan fitur utama yang mempengaruhi model ramalan. Bab seterusnya akan membincangkan hasil model yang dibangunkan untuk ramalan hasil angioplasti.