PENGESANAN PENIPUAN DALAM TRANSAKSI KEWANGAN BERASASKAN TEKNIK PEMBELAJARAN MESIN

SYAZNI BINTI JASLAN

UNIVERSITI KEBANGSAAN MALAYSIA

PENGESANAN PENIPUAN DALAM URUSAN KEWANGAN BERASASKAN TEKNIK PEMBELAJARAN MESIN

SYAZNI BINTI JASLAN

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

28 Mac 2024

SYAZNI BINTI JASLAN
P107700

PENGHARGAAN

Dengan nama Allah Yang Maha Pemurah lagi Maha Mengasihani. Syukur ke hadratNya atas nikmat ilmu serta kekuatan dan kesabaran yang diberikan dalam menyempurnakan kajian ini.

Pertama sekali saya ingin mengucapkan jutaan terima kasih dan setinggi-tinggi penghargaan kepada Profesor Madya Dr. Masnizah di atas bantuan dan kesudian beliau selaku penyelia utama saya sepanjang usaha dalam menyiapkan kajian ini. Segala panduan, dorongan dan bimbingan yang diberikan sepanjang penyeliaan adalah sangat bermakna serta berharga buat saya.

Tidak lupa juga ucapan terima kasih kepada semua tenaga pengajar terlibat bagi Program Sarjana Sains Data di Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia atas segala ilmu dan tunjuk ajar yang diberikan. Penghargaan ini juga saya tujukan kepada pihak Yayasan Sarawak yang menaja pengajian dan perbelanjaan saya di peringkat Sarjana di UKM.

Buat keluarga tercinta terutamanya kedua ibu bapa dan adik-beradik, terima kasih tidak terhingga kerana sentiasa percaya, memahami, menyokong dan memberi sokongan moral terhadap saya sepanjang menyiapkan kajian ini.

Akhir sekali tidak dilupakan juga buat rakan-rakan seperjuangan Program Sarjana Sains Data, sahabat-sahabat yang banyak berkongsi idea dan menghulurkan bantuan serta individu-individu yang turut membantu sama ada secara langsung ataupun tidak dalam kajian ini, ribuan terima kasih ditujukan, jasa kalian semua amat saya hargai.

Sekian.

ABSTRAK

Pengesanan penipuan kewangan yang praktikal dan komprehensif boleh membantu melindungi institusi kewangan daripada kerugian yang besar akibat aktiviti penipuan. Kaedah pengesanan penipuan kewangan sedia ada yang menggunakan penilaian berasaskan peraturan dan analisis statistik tidak lagi berkesan dengan peningkatan saiz data dan jumlah transaksi yang bercorak dinamik. Pemilihan ciri dan juga pembelajaran bersama (ensemble) adalah alternatif yang boleh diterokai untuk pengesanan penipuan kewangan yang melibatkan set data yang besar dan kompleks. Kajian ini menggunakan set data sintetik yang dihasilkan menggunakan simulator yang dipanggil PaySim telah digunakan sebagai pendekatan kepada masalah penipuan dalam transaksi kewangan. Kajian ini memberi tumpuan kepada pemilihan ciri yang informatif, pembangunan dan pemilihan model pembelajaran yang secara automatik boleh meramalkan transaksi kewangan sebagai transaksi sah ataupun tidak. Bagi tujuan ini, tiga kaedah pemilihan ciri yang berbeza, iaitu Penghapusan Ciri Secara Berulang (RFE), Keuntungan Maklumat (IG), dan Kepentingan Hutan Rawak (RFI). Kami mengintegrasikan kaedah pemilihan ciri ini dengan empat jenis pembelajaran mesin, iaitu dua daripadanya adalah pengelas tunggal, Mesin Vektor Sokongan dan Regresi Logistik) dan dua lagi adalah jenis pengelas ensemble, Hutan Rawak dan XGBoost. Prestasi bagi setiap model gabungan dibandingkan berdasarkan ketepatan (accuracy), ketepatan (precision), kepekaan (recall), dan skor F1. Model XGBoost+RFE memberikan prestasi ketepatan tertinggi untuk set data sintetik Paysim dengan nilai ketepatan 0.9945. Ujian hipotesis menerima kedua-dua hipotesis dan menunjukkan terdapat impak pemilihan ciri dan pembelajaran bersama (ensemble) yang signifikan terhadap pembelajaran mesin dalam mengesan penipuan dalam transaksi kewangan. Secara keseluruhannya, model XGBoost+RFE adalah model terbaik untuk mengesan penipuana kewangan bagi set data dalam kajian ini.

DETECTION OF FINANCIAL FRAUD BASED ON MACHINE LEARNING TECHNIQUES

ABSTRACT

Practical and comprehensive financial fraud detection can help protect financial institutions from significant losses due to fraudulent activities. Existing financial fraud detection methods that rely on rule-based assessment and statistical analysis are no longer effective with the increasing volume of data and dynamic transaction patterns. Feature selection and ensemble learning are alternative approaches that can be explored for financial fraud detection involving large and complex datasets. This study utilizes synthetic data generated using a simulator called PaySim as an approach to the problem of fraud detection in financial transactions. The study focuses on selecting informative features, developing and selecting machine learning models that can automatically predict financial transactions as either legitimate or fraudulent. For this purpose, three different feature selection methods, namely Recursive Feature Elimination (RFE), Information Gain (IG), and Random Forest Importance (RFI), are integrated with four types of machine learning models, two of which are single classifiers (Support Vector Machine and Logistic Regression), and the other two are ensemble classifier types (Random Forest and XGBoost). The performance of each combined model is compared based on accuracy, precision, recall, and F1 score. The XGBoost+RFE model achieves the highest accuracy performance for the synthetic Paysim dataset with an accuracy value of 0.9945. Hypothesis testing accepts both hypotheses and indicates a significant impact of feature selection and ensemble learning on machine learning in detecting fraud in financial transactions. Overall, the XGBoost+RFE model is the best model for detecting financial fraud in the datasets used in this study.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENGENALAN

1.1 PENDAHULUAN

Dalam era kemajuan teknologi yang pesat, urusan kewangan telah mengalami transformasi yang signifikan, bergerak menuju platform digital yang menawarkan kemudahan dan kecekapan yang tiada tandingannya. Namun, bersamaan dengan peningkatan pesat transaksi dalam talian, risiko penipuan dalam urusan kewangan turut meningkat. Penipuan kewangan merujuk kepada aktiviti menipu atau haram yang bertujuan untuk memperoleh keuntungan kewangan secara tidak jujur. Ini merangkumi pelbagai aktiviti penipuan yang lazim dilakukan termasuk penipuan kad kredit, pencurian identiti, pencucian wang, dan juga perdagangan dalam (insider trading) (Bhattacharyya et al. 2011). Aktiviti penipuan ini tidak hanya menyebabkan kerugian kewangan yang besar bagi individu dan organisasi, tetapi juga merosakkan kepercayaan dalam sistem kewangan.

Pengesanan penipuan dalam urusan kewangan adalah isu yang kritikal dan memerlukan tindakan segera untuk melindungi keselamatan kewangan pelanggan dan organisasi. Oleh itu, para penyelidik dan profesional industri telah memberikan perhatian kepada teknik pembelajaran mesin sebagai pendekatan yang berpotensi memberikan penyelesaian berkesan untuk mengesan dan mencegah aktiviti penipuan terutamanya secara masa nyata. Dengan memanfaatkan kekuatan algoritma ini, institusi kewangan dan organisasi dapat meningkatkan keupayaan sistem pengesanan penipuan mereka yang sedia ada supaya berkebolehan untuk patuh dengan taktik jenayah penipuan kewangan yang sentiasa berubah pada masa kini.

1.2 LATAR BELAKANG

Pengesanan penipuan dalam urusan kewangan adalah proses kritikal untuk mengenal pasti dan mencegah aktiviti penipuan yang merugikan pelanggan dan organisasi. Dalam usaha untuk meningkatkan keberkesanan pengesanan penipuan, para penyelidik telah menjalankan kajian intensif dalam bidang ini. Teknik pembelajaran mesin telah dikenal pasti sebagai kaedah yang berpotensi memberikan pendekatan yang canggih dan berkesan untuk pengesanan penipuan dalam urusan kewangan.

Teknik pembelajaran mesin memanfaatkan kecerdasan buatan untuk menganalisis data kewangan dalam jumlah yang besar dan mengenal pasti corak serta ciri-ciri yang mencurigakan. Dengan menerapkan algoritma pembelajaran mesin yang tepat, sistem pengesanan penipuan dapat dikembangkan untuk mengesan transaksi yang mencurigakan dalam masa nyata dan mengambil tindakan sewajarnya untuk mencegah kerugian kewangan lebih lanjut (Dal Pozzolo et al. 2018). Salah satu kelebihan utama teknik pembelajaran mesin dalam pengesanan penipuan adalah keupayaannya untuk menyesuaikan dan belajar dari corak baru dan trend penipuan yang wujud berdasarkan data. Manakala sistem berdasarkan peraturan tradisional mungkin sukar untuk mengimbangi taktik penipu yang berubah-ubah, algoritma pembelajaran mesin, sebaliknya, dapat terus belajar dan mengemaskini model mereka berdasarkan data baharu, membolehkan mereka mengesan corak penipuan yang sebelum ini tidak dikenal pasti. Pelbagai teknik pembelajaran mesin telah digunakan dalam pengesanan penipuan dalam urusan kewangan. Teknik-teknik ini termasuk lah rangkaian neural buatan, pohon keputusan, mesin vektor sokongan, dan juga kaedah ensemble. Setiap teknik mempunyai kelebihan dan kelemahan sendiri, dan para penyelidik terus meneroka dan menyempurnakan kaedah-kaedah ini untuk meningkatkan ketepatan dan keberkesanannya dalam mengesan aktiviti penipuan.

Walau bagaimanapun, pengesanan penipuan menggunakan teknik pembelajaran mesin tidak terkecuali dari cabarannya. Salah satu cabaran utama adalah dimensi dataset yang besar, di mana kehadiran data dalam jumlah besar sering kali memunculkan cabaran dalam hal komputasi dan kecekapan. Dengan melakukan pemilihan ciri, kita dapat mengurangkan kompleksiti dataset yang besar, memilih ciri-ciri yang paling relevan dan informatif sahaja, serta membantu model untuk mengenal pasti pola

penipuan dengan lebih efisien. Justeru, keberkesanan teknik ini juga masih bergantung kepada kualiti dan ciri-ciri dataset yang digunakan dalam latihan dan pengujian.

Sebagai kesimpulan, kajian ini bertujuan untuk menyiasat dan menganalisis keupayaan pelbagai teknik pembelajaran mesin dalam pengesanan penipuan dalam urusan kewangan, secara khususnya pada pemilihan ciri untuk data berskala besar. Dengan mengambil kira perubahan dan cabaran semasa dalam landskap kewangan, hasil daripada kajian ini diharapkan dapat memberikan sumbangan yang signifikan dalam memperkuatkan sistem pengesanan penipuan dan melindungi integriti kewangan.

1.3 PERNYATAAN MASALAH

Penipuan kewangan kekal menjadi cabaran yang signifikan dalam industri perbankan dan kewangan, yang mengakibatkan kerugian kewangan yang besar dan kerosakan reputasi. Kaedah tradisional untuk mengesan penipuan yang bergantung pada sistem berasaskan peraturan dan pemeriksaan manual, sering kali memakan masa dan juga mudah terdedah kepada kesilapan (West et al. n.d.). Tambahan, kaedah tradisional juga tidak mampu untuk mengenal pasti skema penipuan yang kompleks, menyebabkan keperluan yang semakin meningkat untuk pendekatan canggih dan lebih bersifat automatik yang dapat mengesan dan mencegah penipuan dalam transaksi kewangan.

Walau bagaimanapun, terdapat keperluan untuk mengkaji dan mengembangkan pendekatan berdasarkan pembelajaran mesin yang baru dan berkesan dalam mengesan penipuan dalam transaksi kewangan dengan tepat dan cekap, sambil mengatasi cabaran seperti data yang tidak seimbang, ciri-ciri dataset yang berdimensi tinggi dan corak penipuan yang dinamik. Justeru, kajian ini akan mengeksplorasi pendekatan pemilihan ciri yang melibatkan mengenal pasti ciri-ciri yang paling relevan dan informatif dari dataset untuk meningkatkan prestasi dan kecekapan model-model pengesanan penipuan (Hamal & Senvar 2021). Walau bagaimanapun, pemilihan ciri yang sesuai untuk pengesanan penipuan dalam industri kewangan adalah tugas yang kompleks disebabkan oleh beberapa sebab. Pertama, dataset kewangan sering mengandungi jumlah ciri yang besar, termasuk nisbah kewangan, atribut perakaunan, dan petunjuk kewangan lain (Hamal & Senvar 2021). Memilih ciri-ciri yang paling relevan daripada jumlah ciri yang besar ini adalah penting untuk mengurangkan dimensi dataset dan memberi tumpuan

kepada atribut-atribut yang paling informatif. Kedua, sifat dataset yang tidak seimbang membawa cabaran dalam pemilihan ciri (Hamal & Senvar 2021a). Kebanyakan contoh dalam dataset adalah bukan penipuan, menjadikan sukar untuk mengenal pasti ciri-ciri yang membezakan yang dapat membezakan kes penipuan dengan berkesan daripada yang bukan penipuan. Oleh itu, teknik pemilihan ciri perlu mengambil kira masalah ketidakseimbangan kelas dan memastikan ciri-ciri yang dipilih mampu menangkap ciri-ciri unik penyata kewangan yang menipu. Selanjutnya, pemilihan ciri dalam pengesanan penipuan kewangan juga perlu mengambil kira interpretabiliti dan penjelasan model. Institusi kewangan memerlukan model yang telus dan dapat difahami untuk mendapatkan pandangan mengenai faktor-faktor yang menyumbang kepada penipuan dan membuat keputusan yang berinformasi. Kesimpulannya, masalah pemilihan ciri dalam pengesanan penipuan kewangan adalah amat penting (West & Bhattacharya 2016). Ia melibatkan menangani cabaran dimensi yang tinggi, ketidakseimbangan kelas, dan interpretabiliti untuk mengenal pasti ciri-ciri yang paling relevan yang dapat membezakan penyata kewangan yang menipu dengan berkesan (Hamal & Senvar 2021). Dengan memilih set ciri yang tepat, model pembelajaran mesin dapat dibangunkan untuk mengesan penipuan kewangan dengan tepat, mengurangkan risiko, dan melindungi integriti industri kewangan (Xiuguo & Shengyong 2022).

1.4 HIPOTESIS KAJIAN

Hipotesis kajian adalah prosedur formal untuk membuat pembuktian dalam kajian. Ia adalah pernyataan yang diajukan berdasarkan anggapan atau ramalan mengenai hubungan antara dua atau lebih atribut dalam kajian. Hipotesis kajian juga bertujuan untuk diuji dan menyediakan jawapan atau penjelasan terhadap persoalan penyelidikan. Kajian ini membentuk hipotesis seperti di bawah:

H0: Pemilihan ciri akan meningkatkan kecekapan dalam latihan model, tanpa mengurangkan prestasi, dan dengan itu menghasilkan model-model yang lebih dapat diinterpretasikan.

H1: Kaedah pembelajaran bersama (ensemble) adalah lebih unggul dalam meningkatkan prestasi model pembelajaran mesin untuk pengesanan penipuan

kewangan berbanding kaedah-kaedah klasik pembelajaran mesin yang popular di dalam bidang ini.

1.5 PERSOALAN KAJIAN

Persoalan kajian adalah untuk melihat perkara yang ingin diketahui dan dijawab dalam kajian ini. Kajian ini membentuk persoalan kajian seperti di bawah:

1. Apakah teknik-teknik klasifikasi yang berbeza untuk lebih memahami pendekatan yang lebih dan paling sesuai untuk diaplikasikan dalam bidang pengesanan penipuan?

2. Apakah pengaruh penerokaan pemilihan ciri terhadap peningkatan ketepatan model dan peranannya dalam pengesanan penipuan menggunakan teknik pembelajaran mesin?

3. Bagaimanakah prestasi model pembelajaran bersama (ensemble) yang dicadangkan berbanding kaedah-kaedah lain yang popular dalam kajian literatur dalam konteks pengesanan penipuan?

1.6 OBJEKTIF KAJIAN

1. Meneroka kaedah pemilihan ciri dan peranannya untuk pengesanan penipuan dalam transaksi kewangan bagi membantu sesebuah organisasi untuk mengoptimumkan keupayaan mereka untuk mengesan potensi penipuan, seterusnya meminimalkan risiko kerugian dan meningkatkan reputasi.

2. Membandingkan teknik-teknik klasifikasi yang berbeza untuk lebih memahami pendekatan yang lebih dan paling sesuai untuk diaplikasikan dalam bidang pengesanan penipuan.

3. Menilai prestasi model pembelajaran bersama (ensemble) yang dicadangkan berbanding kaedah-kaedah lain yang popular dalam kajian literatur dalam konteks pengesanan penipuan.

1.7 RASIONAL/KEPENTINGAN KAJIAN

(Lam et al. 2022) menunjukkan bukti langsung untuk pengesanan penipuan yang tepat menggunakan data mentah laporan kewangan dan algoritma pembelajaran gabungan, menekankan kepentingan pendekatan pembelajaran mesin dalam pengesanan penipuan. Selain itu, (Yinhe Chen 2023) mencadangkan rangka kerja pemilihan ciri terintegrasi untuk meningkatkan keupayaan mengesan penipuan penyata kewangan syarikat tersenarai, menekankan kepentingan pemilihan ciri dalam pengesanan penipuan. Tambahan pula, satu kajian terdahulu telah membincangkan kepentingan teknik pembelajaran mesin dalam mengesan penipuan kad kredit, menekankan kerelevanan mengaplikasikan algoritma canggih untuk pengesanan penipuan dalam transaksi kewangan (Sharma & Chalapathi 2022).

Rujukan-rujukan ini secara kolektif menyokong rasional untuk kajian ini, kerana mereka menekankan kepentingan teknik pembelajaran mesin dan pemilihan ciri dalam pengesanan pelbagai jenis penipuan kewangan. Kajian-kajian ini memberikan bukti keberkesanan pendekatan pembelajaran mesin dan kaedah pemilihan ciri dalam mengesan aktiviti penipuan, dengan demikian membuktikan keperluan untuk penyelidikan lanjut dalam bidang ini.

1.8 SKOP KAJIAN

Pemilihan ciri memainkan peranan penting dalam meningkatkan prestasi model pengesanan penipuan kewangan dengan mengeluarkan ciri-ciri yang tidak diperlukan atau yang memberikan maklumat prediktif yang sedikit. Tambahan pula, perlu diingat bahawa pemilihan ciri memberi impak besar kepada ketepatan model pembelajaran mesin dalam mengesan penipuan kewangan (Liu et al. 2021).

Selanjutnya, penggunaan teknik pembelajaran mesin seperti RF, SVM, LR, dan XGBOOST telah banyak diuji dalam konteks pengesanan penipuan kewangan (Hajek et al. 2023; Yao et al. 2019). Teknik-teknik ini telah digunakan untuk mengesan aktiviti penipuan penyata kewangan berdasarkan pelbagai atribut kewangan dan bukan kewangan, menunjukkan keberkesanan mereka dalam mengenal pasti aktiviti penipuan (Yao et al. 2019). Selain itu, penggunaan XGBOOST telah ditekankan sebagai

pendekatan yang berkesan untuk pengesanan penipuan kewangan (Lei et al. 2020).

Tambahan itu, adalah penting untuk mempertimbangkan konteks khusus pengesanan penipuan kewangan, seperti penipuan kad kredit dan penipuan penyata kewangan, kerana jenis penipuan yang berbeza mungkin memerlukan pendekatan yang disesuaikan (Sharma & Chalapathi 2022; West & Bhattacharya 2016). Kesusasteraan menekankan kepentingan pemilihan ciri yang relevan untuk pengesanan penipuan kad kredit dan pengesanan penipuan penyata kewangan (Sharma & Chalapathi 2022). Ini menonjolkan kepentingan teknik pemilihan ciri dalam konteks jenis penipuan kewangan yang berbeza.

Secara keseluruhannya, skop kajian adalah untuk pengesanan penipuan kewangan menggunakan dataset Paysim dan teknik pemilihan ciri, bersama dengan teknik pembelajaran mesin, melibatkan pemilihan ciri dengan berhati-hati untuk meningkatkan prestasi model pengesanan. Penggunaan teknik pembelajaran mesin seperti Hutan Rawak, Mesin Vektor Sokongan (SVM), Regresi Logistik dan XGBOOST telah menunjukkan potensi dalam mengenal pasti aktiviti penipuan dengan berkesan dalam konteks penipuan kewangan.

1.9 METODOLOGI KAJIAN

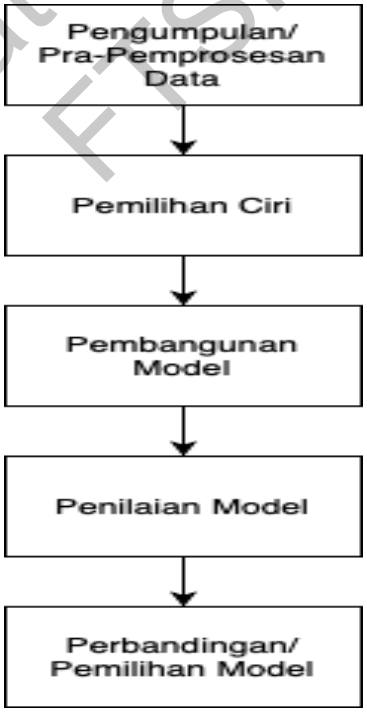
Metodologi kajian perlu dirangka berdasarkan objektif kajian dan dijadikan panduan bagi pelaksanaan kajian. Bersesuaian dengan objektif kajian ini, pendekatan yang digunakan dalam kajian ini tertumpu pada proses teliti dalam memilih ciri-ciri yang penting untuk meningkatkan keberkesanan pengesanan penipuan kewangan melalui teknik pembelajaran mesin. Objektif utama adalah untuk memahami dan mengasingkan atribut yang paling berkaitan yang memberikan sumbangan yang ketara kepada ketepatan model pengesanan penipuan.

Proses pemilihan ciri-ciri melibatkan penggunaan algoritma pembelajaran mesin, dengan penekanan khusus pada Kepentingan Maklumat, teknik Pengurangan Ciri secara Berulang (RFE) dan juga Kepentingan Hutan Rawak. Teknik-teknik pemilihan ciri dilakukan untuk meningkatkan keberkesanan model pengesanan penipuan kewangan menggunakan empat algoritma pembelajaran mesin, iaitu Regresi

Logistik, Hutan Rawak, Mesin Vector Sokongan, dan XGBoost. Gabungan ini akan menjadi satu kaedah pembelajaran gabungan yang kukuh, digunakan untuk menentukan kepentingan ciri, memberikan wawasan tentang keberkesanan setiap atribut.

Dataset ini dikenakan kepada teknik pemilihan ciri-ciri ini semasa fasa pra-pemprosesan, memastikan bahawa atribut yang dipilih merangkumi hakikat corak penipuan dalam data kewangan. Subset ciri yang disempurnakan, yang diperolehi, bertujuan untuk meningkatkan prestasi pengelasifikasi pembelajaran mesin seterusnya. Tambahan pula, mengambil pendekatan sistematik terhadap pemilihan ciri, dengan menggabungkan kelebihan pendekatan pembelajaran mesin dan pemilihan ciri, kajian ini bertujuan menyumbang wawasan berharga ke arah atribut yang mempengaruhi ketepatan model pengesanan penipuan kewangan.

Kerangka kerja bagi kajian ini merangkumi 5 fasa utama dan meliputi kesemua bab dalam kajian ini. Fasa dalam kajian ini adalah pengumpulan dan pra-pemprosesan, pemilihan ciri, pembangunan model, penilaian model, dan perbandingan model. Setiap fasa dalam kajian ini akan dibincangkan dengan lebih perinci dalam Bab 3.



Rajah 1.1 Fasa-fasa yang terlibat dalam kajian ini

1.10 ORGANISASI TESIS

Terdapat lima (5) bab utama dalam tesis ini yang akan menerangkan secara terperinci berkenaan kerja-kerja yang terlibat bagi menjayakan kajian ini. Berikut ialah ringkasan bagi setiap bab yang akan menerangkan keseluruhan perjalanan tesis:

Bab 1 membincangkan secara keseluruhan kajian yang ingin dilaksanakan termasuk latar belakang kajian, penyataan masalah, objektif kajian, rasional kajian serta skop kajian.

Bab 2 adalah berdasarkan kajian literatur yang lepas tentang isu-isu yang berkaitan dengan model pengesanan penipuan, memahami kajian literatur, memilih model pembelajaran yang sesuai untuk kajian dan rujukan penyelesaian untuk pemilihan ciri yang akan digunakan.

Bab 3 adalah berkisar tentang metodologi proses penyediaan data yang melalui proses analisis dan pra-pemprosesan seperti pembersihan, integrasi, transformasi dan sebagainya. Teknik pemilihan ciri dan pembangunan model pembelajaran, penilaian model pembelajaran yang dihasilkan diperincikan dengan lebih lanjut.

Bab 4 akan memberi maklumat berkaitan hasil dapatan kajian yang dijalankan. Hasil dapatan merangkumi hasil analisis deskriptif dan analisis prediktif. Pengukuran prestasi dijalankan ke atas pengelasan yang dijana oleh model – model pembelajaran.

Bab 5 merupakan rumusan keseluruhan untuk kajian ini. Selain itu, sumbangan kajian juga akan dijelaskan hasil daripada dapatan kajian. Akhir sekali penyelidikan pada masa hadapan bagi model pengesanan penipuan kewangan dibincangkan dan dicadangkan untuk kajian lanjut.

BAB II

SOROTAN KESUSASTERAAN

2.1 PENGENALAN

Bab ini membincangkan literatur berkaitan yang menjadi asas kepada kajian ini dalam konteks pengesanan penipuan dalam domain teknologi kewangan, khususnya transaksi kewangan, cabaran menggunakan dataset yang tersedia, pembangunan dan pemilihan model – model pembelajaran dan pemilihan fitur. Kajian literatur dijalankan berdasarkan penelitian dan mengenal pasti masalah jurang dalam bidang kajian. Perbincangan kajian literatur dimulakan dengan pemahaman mengenai istilah jenayah penipuan, jenis-jenis penipuan yang sering dilakukan dalam domain ini, model pembelajaran mesin dalam kajian lepas, pengenalan kepada model pembelajaran bersama yang kian popular, dan keberkesanan pemilihan ciri dalam set data berdimensi tinggi,

Bab ini terbahagi kepada lima (5) bahagian iaitu:

1. Pengesanan Penipuan Dalam Transaksi Kewangan;
2. Kompleksiti dan Cabaran Dataset;
3. Teknik Pembelajaran Mesin;
4. Penerokaan Pemilihan Ciri
5. Penilaian Model

2.2 PENGESANAN PENIPUAN DALAM TRANSAKSI KEWANGAN

Bahagian ini akan menerangkan definisi asas penipuan dalam domain teknologi kewangan. Teori menyatakan, penipuan boleh didefinisikan sebagai "tindakan penipuan yang sengaja dirancang untuk memberikan perolehan yang melanggar undang-undang

kepada pelaku atau untuk menafikan hak kepada mangsa" (“What Is Fraud? Definition, Types, and Consequences” n.d.). Namun, mengesan penipuan merupakan cabaran yang sangat besar kerana penipuan adalah jenayah yang bersifat adaptif dan berubah menyesuaikan diri (Chalapathy & Chawla 2019). Oleh itu, perlunya dataset kewangan dalam skala yang besar. Dataset – dataset yang sering digunakan dalam kajian melibatkan domain kewangan akan mewakili gabungan transaksi dari trafik rangkaian kewangan semasa dalam tempoh tertentu. Dalam dataset ini, penipuan adalah anomali yang berbeza dari rekod biasa. Untuk mengenal pasti corak-corak tersebut, teknik-teknik telah digunakan dalam pengecaman anomali dalam domain Pembelajaran Mesin (ML). Secara amnya, teknik-teknik pengecaman anomali membolehkan pengenalan penipuan daripada dataset besar. Mereka telah terbukti mencapai hasil yang menonjol dalam mengelaskan data dalam pengumpulan tersebut. Malah, kelebihan-kelebihan ini menjadikan teknik-teknik ini pilihan semulajadi untuk menangani cabaran pengecaman penipuan. Menurut satu kajian oleh (Stojanovi´cstojanovi´c et al. 2021), terdapat tiga jenis penipuan utama yang difokuskan dalam domain teknologi kewangan, iaitu penipuan dalam kad kredit, transaksi kewangan, dan transaksi blockchain. Kajian ini hanya akan memberi fokus kepada transaksi kewangan sahaja.

Jenayah transaksi kewangan adalah salah satu jenis penipuan dalam domain maya teknologi kewangan. Ini termasuk, sebagai contoh, pengubahan wang haram dan penipuan lelong dalam talian. Yang terakhir merangkumi transaksi palsu, pengembalian wang palsu, dan pinjaman palsu, ketidakbayaran dan pembelian yang tidak dibenarkan (J. S. Chang & Chang 2012). Tindakan jenayah siber yang dilakukan secara dalaman dalam syarikat dikenali sebagai penipuan pekerjaan. Biasanya, tindakan ini dipengaruhi oleh perolehan data pengguna yang sulit. Cabaran dalam mengesan penipuan sebegini adalah sukar kerana aktiviti teknologi kewangan sering kali dilakukan melalui rangkaian perdagangan yang interaktif. Dengan cara ini, penipuan boleh dikaitkan dengan mana-mana pengguna, item, atau masa (Webga & Lu 2015). Ini menjadi masalah terutama dalam zon perdagangan bebas, iaitu negara-negara dengan pasaran yang longgar atau tidak teratur. Untuk melengkapkan proses pengesanan penipuan, dasar seperti pencegahan pengubahan wang haram (AML) harus dilaksanakan dan dikuatkuasakan di peringkat negeri (Le Khac & Kechadi 2010).

Kajian-kajian berikut menangani masalah ini dengan menggunakan pendekatan berdasarkan pembelajaran mesin (ML) kepada dataset kewangan. (Magomedov et al. 2018) mencadangkan kaedah pengesanan anomali dalam pengurusan penipuan berasaskan ML dan pangkalan data graf. Sebuah kajian dengan objektif yang sama, yang memberi tumpuan kepada pengubahan wang haram, dibentangkan oleh (Huang et al. 2018).Mereka memperkenalkan rangka kerja pengesanan, yang dipanggil CoDetect, yang menganalisis rangkaian, iaitu entiti dan transaksi, dan seterusnya mengesan penipuan dan corak ciri. CoDetect menggunakan pendekatan perlombongan Graf untuk pelbagai senario penipuan dunia sebenar. Selain itu, (La & Kim 2018) juga mencadangkan rangka kerja komprehensif untuk menguruskan transaksi teknologi kewangan yang menggunakan kecerdasan berasaskan pembelajaran mesin dalam menghasilkan model pengesanan anomali dan penyediaan keselamatan teknologi kewangan yang bersifat adaptif.

(Le Khac & Kechadi 2010) menggunakan algoritma k-means untuk mengesan pengubahan wang haram manakala (W. H. Chang & Chang 2010) menggunakan kaedah yang sama untuk mengesan penipuan lelong dalam talian. Selain itu, (J. S. Chang & Chang 2012) mencadangkan kaedah untuk pengesanan penipuan awal dalam lelongan dalam talian. Mereka mengurangkan atribut yang digunakan untuk menghasilkan model pembelajaran melalui analisis utama dan menggunakan 20% terakhir sejarah transaksi dalam membina model untuk memaksimumkan kadar pengesanan sambil meminimumkan bebanan komputasi. Sesetengah pengarang menggunakan pendekatan hibrid untuk memaksimumkan prestasi pengesanan penipuan. Di sisi lain, (Glancy & Yadav 2011) dan (Torgo & Lopes 2011) menggunakan pengelompokan hierarki untuk pengesanan anomali dalam transaksi kewangan. (Yaram 2017)mencadangkan pengelompokan dokumen dan algoritma klasifikasi untuk mengenal pasti penipuan dalam tuntutan insurans.

(J. J. Xu et al. 2015) membincangkan jenis penipuan kewangan yang agak baru, iaitu dalam pinjaman wang dari seorang individu ke individu (P2P). Peminjaman P2P berlaku dalam pasaran dalam talian, di mana pinjaman diperolehi tanpa institusi kewangan di antaranya. Perbincangan ini penting kerana peminjaman P2P belum menarik banyak minat penyelidikan dalam pengesanan penipuan. Pertama, pengarang menjelaskan kaedah pengesanan sedia ada, termasuk berasaskan ML, dalam konteks ini.

Selepas itu, mereka menyediakan arah penyelidikan yang mungkin berkenaan dengan pengesanan penipuan dalam persekitaran seperti itu.

(Leite et al. 2018) menjalankan kajian menyeluruh terhadap pendekatan pengesanan penipuan visual yang sedia ada. Kerja ini memberi tumpuan kepada teknik visualisasi seperti plot garis, gambarajah node-link, plot sebar, dan lain-lain. Selain itu, pengarang menyediakan penilaian perbandingan untuk setiap pendekatan. Akhirnya, mereka menyimpulkan bahawa kebanyakan pendekatan yang rumit tidak mengintegrasikan kaedah automatik untuk pengesanan penipuan. Pendekatan lain yang membincangkan analitik visual untuk pengesanan penipuan secara langsung (Webga & Lu 2015)

(Wedge et al. 2019) mencadangkan pendekatan untuk kejuruteraan ciri automatik yang direka untuk mengurangkan bilangan positif palsu. Pengarang menyatakan bahawa maklumat yang boleh diakses mengenai kad dan pelanggan dapat meningkatkan saiz set ciri yang berpotensi secara drastik. Pengeluaran ciri secara manual memerlukan masa yang banyak, dan mungkin wujud keperluan untuk mengulangi prosedur tersebut beberapa kali, iaitu setiap kali bank baru ditambah dalam dataset. Untuk menangani isu ini, pengarang mencadangkan kaedah automatik untuk kejuruteraan ciri, iaitu Deep Feature Synthesis (DFS). Hasil menunjukkan penurunan positif palsu sebanyak 54% dalam dataset yang sebelum ini tidak dilihat yang terdiri daripada 1952 juta transaksi. (Long et al. 2019) juga menganggap kejuruteraan ciri menggunakan Pembelajaran Mendalam. Pengarang mencadangkan model hujung ke hujung untuk pengekstrakan ciri dari sampel deretan masa kewangan dan ramalan pergerakan harga, menggunakan neuron konvolusi dan ulangan skerta rangkaian neural multi-filter. (Baesens et al. 2021)menyatakan bahawa kejuruteraan data adalah penting untuk meningkatkan prestasi kebanyakan model pembelajaran mesin. Dalam kajian mereka, proses kejuruteraan data yang terdiri daripada beberapa langkah kejuruteraan ciri dan contoh dicadangkan dan ditunjukkan pada dataset transaksi pembayaran dari sebuah bank besar Eropah.

2.3 KOMPLEKSITI DAN CABARAN DATASET

Malangnya, satu masalah umum bagi penyelidikan dalam pengecaman anomali di domain kewangan adalah ketiadaan data ujian yang boleh diakses secara umum. Oleh itu, data yang paling dikenali dan banyak digunakan mewakili dataset Kaggle. Ini termasuk dataset untuk kad kredit, data transaksi bank, dan data sejarah blockchain. Dataset sintetik yang sedikit lebih lama boleh ditemui di Repositori ML UC Irvine, contohnya, UC Irvine. Selain itu, simulator seperti BankSim dan PaySim (Alonso Lopez-Rojas et al. 2016) digunakan untuk mengkaji domain ini. Dataset BankSim mewakili simulator berdasarkan ejen pembayaran bank, manakala Paysim mensimulasikan transaksi mudah alih dengan menghasilkan pelanggan dan melaksanakan transaksi. Kedua-dua simulator telah menghasilkan dataset yang menyerupai pengguna sebenar dan transaksi sekaligus.

Selain itu, kebanyakan penyelidikan awal cenderung untuk menghasilkan data sintetik yang disimulasikan berdasarkan ciri-ciri yang diperoleh dari penipuan dunia sebenar dan transaksi sah. Untuk melakukan ini, (Rieke et al. 2013) mengekstrak corak pengubahan pembayaran dari peristiwa dunia sebenar. Walau bagaimanapun, jumlah contoh yang ada tidak mencukupi untuk pengesanan penipuan yang efisien, seperti yang ditunjukkan oleh kadar negatif palsu (sah) yang rendah dalam kajian awal (Coppolino et al. 2015; Rieke et al. 2013). Kemajuan yang besar telah dicapai dengan memperkenalkan simulasi kewangan PaySim (Edgar Alonso Lopez-Rojas et al. 2018) yang menyerupai transaksi mudah alih biasa dan menyuntik kelakuan penipuan untuk menghasilkan lebih banyak penipuan kewangan. Simulasi berasaskan agen dan analisis statistik mengesahkan bahawa data yang disimulasikan serupa data sebenar yang asal yang digabungkan secara anonim, oleh itu, mewakili persekitaran kawalan optimum untuk pengesanan penipuan dalam transaksi pembayaran mudah alih. Dengan memanfaatkan data PaySim, (Edgar A. Lopez-Rojas & Barneaud 2019)memperlihatkan kelebihan mereka berbanding dengan dataset dunia sebenar yang agak kecil. Selain itu, data yang disimulasikan mengekalkan transaksi dan dinamik kausal data asal. Walau bagaimanapun, perlu diingat bahawa dengan mengekalkan sifat statistik data dunia sebenar, imbangan kelas yang tinggi berpihak kepada transaksi sah juga dikekalkan dalam dataset yang disimulasikan.

Selain kekurangan dataset yang boleh diakses secara umum, terdapat dua cabaran utama lain dalam bidang pengesanan penipuan: ketidakseimbangan kelas, iaitu terdapat lebih banyak transaksi yang sah daripada transaksi palsu, dan perubahan konsep, iaitu kebiasaan pelanggan dan penipu berkembang (Dal Pozzolo et al. 2015).Terdapat beberapa pendekatan yang memberi tumpuan kepada perubahan konsep. (Dal Pozzolo et al. 2015)mereka bentuk dua sistem pengecaman penipuan berdasarkan pendekatan pembelajaran bersama (ensemble) dan jendela geser untuk penyesuaian perubahan konsep. Dalam kerja terkini oleh kumpulan pengkaji yang sama, perubahan konsep juga diambil kira. Aspek perubahan konsep juga menjadi fokus dalam kerja oleh (Ma et al. 2019) di mana satu kaedah pembelajaran maya virtual bertahap untuk pengemaskinian rangkaian neural diusulkan. (Somasundaram & Reddy 2019) mencadangkan satu pembelajaran bersama dan bertahap untuk menangani perubahan konsep dan ketidakseimbangan data.

Terdapat beberapa kertas kajian yang merangkumi pengecaman anomali dalam teknologi kewangan, yang memberikan pandangan yang sangat baik ke atas tren semasa. Kajian menyeluruh awal mengenai penyelesaian pintar untuk pengesanan penipuan kewangan telah dijelaskan oleh (Ngai et al. 2011). Kajian yang dilakukan oleh (Ahmed et al. 2016) memberikan gambaran keseluruhan mengenai kaedah pengecaman anomali, khususnya algoritma pengelompokan, dalam domain kewangan. Selain itu, ia memberikan ulasan mengenai aplikasi kaedah pengecaman anomali pada data raya (Big Data) dalam pasaran kewangan. Seterusnya, (Ahmed et al. 2017) menetapkan andaian bagaimana mengesan anomali dan merangkum hasil kerja yang menggunakan algoritma pengelompokan berasaskan partisi dan hierarki. Selain itu, (Gai et al. 2018) mencadangkan satu kajian yang sangat komprehensif mengenai teknologi kewangan secara umum. Kemudian, (West & Bhattacharya 2016)mengetengahkan hasil kajian mengenai aplikasi algoritma pengelasan kepada pengesanan penipuan kewangan. Tambahan pula, mereka menganalisis kelebihan dan kelemahan pendekatan berdasarkan pengelasan kepada pengesanan penipuan kewangan dan mengelas karya sedia ada dari segi prestasi, algoritma yang digunakan, dan jenis penipuan. Gambaran umum mengenai kaedah pengecaman anomali berasaskan grafik disampaikan oleh (Pourhabibi et al. 2020). Teknik ensemble juga merupakan salah satu kaedah yang popular dan dikaji secara terkini dalam dalam pengesanan penipuan kewangan kerana

ia menggabungkan beberapa model pembelajaran mesin untuk meningkatkan ketepatan mengenal pasti transaksi penipuan (Khedr et al. 2021)

2.4 TEKNIK PEMBELAJARAN MESIN

Pengesanan penipuan dalam urusan kewangan adalah bidang penyelidikan yang penting dan telah menjadi subjek kajian yang meluas. Pelbagai kajian telah meneroka keberkesanan algoritma pembelajaran mesin dalam mengesan pelbagai jenis penipuan kewangan, termasuk penipuan kad kredit, penipuan perakaunan, dan transaksi palsu dalam sektor kewangan dan teknik pembelajaran mesin telah muncul dan terbukti sebagai alat yang berkesan untuk mengesan dan mencegah aktiviti penipuan. Seksyen ini bertujuan untuk memberikan gambaran tentang penyelidikan yang telah dijalankan dalam domain ini, menekankan aplikasi pelbagai teknik pembelajaran mesin untuk pengesanan penipuan dalam transaksi kewangan.

Sejumlah besar literatur telah diterbitkan mengenai pengesanan penipuan kewangan, seperti (West & Bhattacharya 2016) untuk ulasan dan (Hajek & Henriques 2017) untuk penilaian komprehensif mengenai kaedah pengesanan penipuan kewangan. Faktor risiko penipuan kewangan telah disiasat, menunjukkan bahawa tekanan/insentif untuk melakukan penipuan adalah faktor risiko yang paling penting (S. Y. Huang et al. 2017). Kajian - kajian berkaitan boleh dikategorikan mengikut jenis penipuan kewangan seperti berikut (Onwubiko 2020): (1) penipuan ambil alih akaun, (2) penipuan pembayaran, dan (3) penipuan aplikasi. (Onwubiko 2020) juga mengenal pasti empat saluran penipuan utama, iaitu fizikal, web, telefoni, dan mudah alih. Penipuan dalam transaksi pembayaran mudah alih semakin diakui sebagai kebimbangan utama dalam kewangan disebabkan oleh perkembangan terkini dalam perkhidmatan pembayaran mudah alih (Yanyu Chen et al. 2023). Oleh itu, keperluan keselamatan harus dipenuhi untuk menangani isu keselamatan yang berkaitan dengan transaksi pembayaran mudah alih, seperti perisian berbahaya mudah alih dan serangan berdasarkan SMS (J. Kang 2018). Pelbagai platform mudah alih perisian dan perkakasan membuatkan masalah keselamatan menjadi lebih mencabar (Li & Clark 2013).

(Albashrawi 2022) telah menjalankan kajian literatur yang komprehensif mengenai penyelidikan yang dijalankan antara tahun 2004 hingga 2015. Kajian adalah

berkaitan dengan pengesanan penipuan dalam urusan kewangan menggunakan teknik perlombongan data. Kajian ini merangkumi 65 artikel yang relevan dan telah menekankan penggunaan teknik perlombongan data dalam pengesanan penipuan kewangan seperti penipuan penyata kewangan dan penipuan bank. Kajian ini telah mengelaskan aplikasi penipuan kewangan dalam rangka kerja peringkat tinggi dan juga terperinci, dan mengenal pasti teknik perlombongan data yang paling signifikan untuk digunakan dalam domain ini. Kajian oleh (Albashrawi 2022) juga menekankan kepentingan pemilihan ciri dalam pengesanan penipuan kewangan. Kaedah pemilihan ciri telah digunakan untuk mengenal pasti nisbah kewangan dan petunjuk yang paling relevan dan memberikan impak signifikan terhadap penyata kewangan penipuan. Kajian ini menekankan keberkesanan algoritma pembelajaran mesin yang asas seperti regresi logistic (LR), pohon keputusan (DT) dan mesin vektor sokongan (SVM) dalam pengesanan penipuan ini. Dalam satu kajian lain oleh (Ashfaq et al. 2022), para penyelidik mencadangkan mekanisme pengesanan penipuan yang berdasarkan pembelajaran mesin dan teknologi blockchain. Kajian ini memberi tumpuan kepada pengesanan penipuan kad kredit dan menggunakan teknik pembelajaran mesin yang tidak dibimbing untuk mengesan anomali kewangan. Walaubagaimanapun, para penyelidik merumuskan bahawa teknik pembelajaran mesin yang dibimbing biasanya lebih berkesan dalam pengesanan penipuan. (Lam et al. 2022) menjalankan kajian mengenai pengesanan penipuan kewangan bagi syarikat-syarikat tersenarai di China menggunakan pendekatan pembelajaran mesin. Kajian ini membandingkan keberkesanan algoritma pembelajaran mesin klasifikasi tunggal dan algoritma pembelajaran bersama (ensemble). Keputusan menunjukkan bahawa algoritma pembelajaran bersama, terutamanya algoritma stacking, lebih berkesan daripada algoritma klasifikasi tunggal dalam pengesanan penipuan bagi syarikat-syarikat tersenarai di China.

Dalam tahun-tahun terkini, teknik-teknik pembelajaran bersama (ensemble) telah mula digunakan dalam kajian, kebanyakan melebihi prestasi pengelas tunggal. Pengelas ensemble mengintegrasikan ramalan daripada pelbagai model asas. Banyak penemuan empirikal dan teoretikal telah menunjukkan bahawa penggabungan model-model yang berbeza dapat meningkatkan ketepatan ramalan (Bertomeu et al. 2021). Tambahan pula, model ensemble terkenal dengan keupayaannya untuk mengurangkan

kecenderungan dan varians. Ramai penyelidik telah menunjukkan minat dalam mengkaji model ensemble yang menggabungkan teknik-teknik seperti boosting (Bao et al. 2020), bagging (Whiting et al. 2012), dan kaedah hibrid lain (H. Li & Wong 2015) pada data yang seimbang dan tidak seimbang. Keberkesanan model-model tersebut telah disimpulkan bergantung pada pemilihan pengelas asas. Walaupun penyelidikan terdahulu menunjukkan bahawa pengelas ensemble adalah yang terbaik dalam mengesan penipuan kewangan, terdapat kurang kajian mengenai mereka berbanding dengan pengelas tunggal (Al Ali et al. 2023). Kebanyakan kajian terdahulu telah menggunakan dataset yang tidak seimbang untuk penilaian, sebagaimana yang berlaku dalam data dunia nyata. Oleh itu, disebabkan oleh masalah bias kelas yang umum, model ensemble tradisional harus secara umumnya disertakan dengan teknik persampelan seperti peningkatan atau pengecilan untuk menyeimbangkan taburan kelas. Hanya beberapa kajian yang telah mempertimbangkan isu ketidakseimbangan semasa pemodelan. Selain itu, walaupun kebanyakan penyelidik telah menggunakan nisbah kewangan untuk ramalan, yang lain berhujah bahawa atribut mentah menghasilkan hasil yang lebih baik (Grove & Basilico 2008; Kanapickienė & Grundienė 2015). Pelbagai metrik boleh digunakan untuk menilai prestasi pengelas, dengan yang dominan adalah kepekaan atau recall, ketepatan, dan ketepatan (Gu et al. 2009). Dalam kajian ini, penilaian pengelas yang berbeza yang digunakan dalam pengesanan penipuan sambil mempertimbangkan isu ketidakseimbangan kelas dan juga data berdimensi tinggi. Kajian ini mengambil kira kedua-dua atribut kewangan mentah dan nisbah kewangan untuk isu yang kedua.

Secara keseluruhannya, kajian literatur ini menekankan kepentingan teknik pembelajaran mesin dalam pengesanan penipuan kewangan. Kajian-kajian ini telah membuktikan kepentingan pemilihan ciri, keberkesanan algoritma pembelajaran mesin yang berbeza, dan potensi algoritma pembelajaran bersama dalam meningkatkan ketepatan pengesanan penipuan. Aplikasi pelbagai teknik pembelajaran mesin mebuktikan kepentingan dan kebolehan mereka dalam menangani kompleksiti pengesanan penipuan dalam transaksi kewangan. Ini sekaligus memperlihatkan potensi mereka untuk menyumbang kepada pembangunan sistem pengesanan penipuan yang kukuh, meningkatkan keselamatan dan integriti dalam sektor kewangan, sekaligus membantu mengurangkan kerugian kewangan yang berkaitan dengan aktiviti penipuan.

2.5 PENEROKAAN PEMILIHAN CIRI

Beberapa masalah yang dihadapi ketika berurusan dengan set data transaksi kewangan termasuk dimensi yang tinggi dan ketidakseimbangan kelas (Ala'raj et al. 2022; Xiaoming Zhang et al. 2022) menjadikannya sukar bagi pengelas ML untuk belajar dan membuat ramalan yang tepat. Selain itu, data dimensi tinggi sering membuat proses pembelajaran menjadi kompleks dan mahal dari segi pemrosesan, menghasilkan model dengan keupayaan generalisasi yang lemah (Yang et al. 2022). Oleh itu, pemilihan ciri adalah penting dalam set data tersebut untuk mengurangkan beban komputasi dan meningkatkan keupayaan generalisasi model. Sebagai contoh, (Chaquet-Ulldemolins et al. 2022) mencatat peningkatan prestasi pengelasan pengelas ML selepas memperkenalkan pemilihan ciri. Secara umum, kaedah pemilihan ciri berguna dalam aplikasi di mana jumlah ciri mempengaruhi prestasi pengelas.

Pemilihan ciri dalam pengurangan jumlah data merujuk kepada proses memilih satu subset ciri yang relevan daripada set ciri yang lebih besar untuk mengurangkan dimensi data sambil mengekalkan ciri-ciri yang paling informatif dan membezakan. Proses ini penting dalam pengesanan penipuan kerana ia membantu meningkatkan kecekapan latihan model, mengurangkan sumber komputasi, dan meningkatkan interpretabiliti model. Dalam konteks pengesanan penipuan penyata kewangan, pemilihan ciri memainkan peranan penting. (Yinhe Chen 2023) mencadangkan satu kaedah pemilihan ciri terintegrasi untuk membina sistem ciri bagi mengesan penipuan penyata kewangan dalam syarikat-syarikat tersenarai. Kaedah ini bertujuan untuk menghapuskan ciri-ciri yang berlebihan atau ciri-ciri dengan maklumat ramalan yang sedikit, dengan itu mengurangkan dimensi dataset sambil mengekalkan prestasi. Dengan memilih ciri-ciri penting, model yang dihasilkan lebih mudah untuk diinterpretasikan dan kurang cenderung kepada overfitting.

2.5.1 Kaedah Pembungkus (Wrapper Method)

Kaedah pemilihan ciri bungkus (wrapper) telah digunakan secara meluas dalam pelbagai aplikasi (Al-Yaseen et al. 2022; Beheshti 2022). Ia adalah pendekatan popular untuk pemilihan ciri dalam pengesanan penipuan kewangan, di mana prestasi algoritma pembelajaran mesin tertentu digunakan sebagai kriteria untuk menilai sub-himpunan

ciri. Kaedah ini mengira kepentingan setiap ciri berdasarkan kegunaannya semasa melatih model pembelajaran mesin (ML). Komponen utama kaedah bungkus adalah pengelas pembelajaran dan strategi carian. Kaedah bungkus wujud sebagai lapisan luar pengelas pembelajaran dan menggunakan pengelas yang sama untuk memilih ciri yang paling relevan. Oleh itu, pengelas pembelajaran yang kukuh dapat meningkatkan pemilihan ciri berasaskan bungkus. Selain itu, strategi carian yang digunakan dalam kaedah bungkus dapat mempengaruhi pemilihan ciri, dan menggunakan strategi carian yang betul untuk sesuatu aplikasi adalah penting untuk mencapai prestasi yang baik.

(Kolli & Tatavarthi 2020) mencadangkan satu strategi pengesanan penipuan yang merangkumi fasa pemilihan ciri menggunakan model wrapper. Model wrapper ini memilih ciri-ciri penting dan sesuai daripada data yang telah diproses, meningkatkan ketepatan dan kecekapan proses pengesanan penipuan. Selain itu, (Hamal & Senvar 2021) juga telah menjalankan kajian mengenai pengesanan penipuan perakaunan kewangan dalam SME Turki. Mereka menggunakan kernel polinomial, yang biasa digunakan dengan Mesin Vector Sokongan (SVM), untuk pemilihan ciri bagi mengenal pasti penipuan perakaunan kewangan. Kajian tersebut membandingkan prestasi pelbagai pengelasifikasi pembelajaran mesin, termasuk SVM, Naive Bayes, rangkaian neural buatan, K-jiran terdekat, hutan rawak, regresi logistik, dan bagging (Hamal & Senvar 2021). Hasilnya menunjukkan bahawa model hutan rawak tanpa pemilihan ciri dibangunkan dengan kaedah oversampling mempunyai prestasi yang lebih baik daripada model lain. Dalam satu kajian lain oleh (Ghazikhani et al. 2012), pendekatan wrapper diperkenalkan sebagai kaedah pra pemilihan ciri. Pendekatan wrapper menggunakan hasil penilaian sistem (pengelas) untuk memacu fasa pra-pemprosesan dan mencari kawasan yang sesuai untuk pengambilan sampel. Mereka menggunakan algoritma genetik sebagai asas pendekatan pembungkus untuk menambahbaik kawasan optimal (Ghazikhani et al. 2012). Selain itu, (Kolli & Tatavarthi 2020) mencadangkan kaedah pengesanan penipuan menggunakan model wrapper dan rangkaian neural ulangan mendalam berdasarkan pengoptimuman air Harris. Model pembungkus digunakan untuk pemilihan ciri agar pengelas dapat mengesan aktiviti penipuan dengan lebih cekap. Model pengoptimuman berasaskan ukuran kecekapan digunakan untuk menilai hasil pengesanan yang tepat (Kolli & Tatavarthi 2020).

Secara ringkasnya, kaedah wrapper telah digunakan secara meluas dalam pengesanan penipuan kewangan untuk memilih ciri-ciri yang relevan dan meningkatkan prestasi model pembelajaran mesin. Ia mempertimbangkan prestasi pengelas tertentu sebagai kriteria pemilihan ciri. Kajian-kajian terdahulu kebanyakannya telah menunjukkan keberkesanan kaedah wrapper dalam pengesanan penipuan kewangan dan memperlihatkan potensinya dalam meningkatkan ketepatan dan kecekapan pengesanan penipuan.

2.5.2 Kaedah Penyaring (Filter Method)

Kaedah penyaringan dalam pemilihan ciri adalah aspek penting dalam pembelajaran mesin dan perlombongan data. Kaedah ini tidak bergantung kepada algoritma klasifikasi dan digunakan untuk pemilihan ciri berdasarkan kriteria tertentu. Kaedah penyaringan menjadi lebih berguna apabila kos komputasi yang rendah atau ketidakbergantungan dengan teknik pengelasan dan kesederhanaan adalah penting. Kaedah penyaringan menilai kerelevanan susunan ciri dengan menggunakan ciri-ciri asal ataupun semulajadi data (Zhu et al. 2007). Kaedah ini menghasilkan kedudukan ciri-ciri dengan menilai hubungan atau persamaan mereka berdasarkan teori maklumat dan statistik. Tambahan pula, kaedah penyaringan memberi skor kepada setiap ciri berdasarkan korelasi dan memilih ciri-ciri dengan menetapkan nilai had (threshold) (Nalluri & Kurra 2021). Selain itu, kaedah penyaringan sering menghasilkan peringkat ciri-ciri (rank) di mana ciri yang lebih penting atau relevan diberikan peringkat yang lebih tinggi berbanding dengan yang lain. Ciri-ciri dengan peringkat tertinggi pada akhirnya akan dipilih tanpa perlunya algoritma pembelajaran.

Pendekatan keuntungan maklumat adalah salah satu teknik yang popular untuk teknik penyaring (filter) dalam pemilihan ciri untuk pengesanan penipuan kewangan, di mana ia mengukur jumlah maklumat yang dimiliki oleh suatu ciri terhadap atribut sasaran. Sebagai contoh, (Albashrawi 2022) telah menjalankan satu kajian yang menyeluruh terhadap penyelidikan-penyelidikan yang telah dijalankan dari tahun 2004 hingga 2015 dalam mengesan penipuan kewangan menggunakan teknik perlombongan data. Tinjauan tersebut menekankan penggunaan pendekatan keuntungan maklumat sebagai kaedah pemilihan ciri dalam pelbagai aplikasi kewangan, termasuklah pengesanan penipuan insurans kesihatan dan juga kad kredit. Kajian tersebut

menekankan kepentingan pemilihan ciri informatif untuk meningkatkan ketepatan pengesanan penipuan.

Selain itu, dalam satu kajian oleh (H. Xu et al. 2022), satu kaedah pemilihan petunjuk utama yang baharu berdasarkan mod hibrid pembelajaran mesin telah dicadangkan untuk ramalan penipuan kewangan. Ukuran keuntungan maklumat digunakan untuk mengenal pasti tahap sumbangan algoritma dan model yang dipilih. Kajian tersebut menunjukkan bahawa model hibrid kaedah Lasso dan hutan rawak memberikan prestasi terbaik dari segi ujian kawasan di bawah lengkung (AUC), menunjukkan keberkesanan pendekatan keuntungan maklumat dalam pemilihan ciri untuk pengesanan penipuan kewangan. Kajian lain oleh (Sharma & Chalapathi 2022) tertumpu kepada pengesanan penipuan kad kredit menggunakan teknik pembelajaran mesin. Ukuran keuntungan maklumat digunakan untuk memilih ciri-ciri yang paling relevan dari data kewangan kad kredit. Kajian tersebut menekankan kepentingan pemilihan ciri dalam meningkatkan ketepatan model pengesanan penipuan. Akhir sekali, (Chen 2023) telahmencadangkan satu kaedah pemilihan ciri bersepadu untuk pengesanan penipuan penyata kewangan. Kaedah tersebut bertujuan untuk mengurangkan dimensi dataset dengan mengeluarkan ciri-ciri yang berlebihan ataupun ciri-ciri yang mempunyai maklumat ramalan yang sedikit. Ukuran keuntungan maklumat digunakan untuk menilai kerelevanan ciri-ciri dalam pengesanan penipuan penyata kewangan.

Secara keseluruhan, selain kaedah wrapper, pendekatan keuntungan maklumat juga telah digunakan secara meluas dalam pengesanan penipuan kewangan untuk memilih ciri-ciri yang relevan seterusnya meningkatkan prestasi model pembelajaran mesin. Kajian-kajian yang telah diterokai menunjukkan keberkesanan pendekatan keuntungan maklumat dalam mengenal pasti ciri-ciri yang informatif dan meningkatkan ketepatan pengesanan penipuan. Dengan memilih ciri-ciri yang paling relevan, model pembelajaran mesin boleh dibangunkan untuk mengesan penipuan kewangan dengan tepat, mengurangkan risiko, dan melindungi integriti industri kewangan.

2.5.3 Kaedah Terbenam (Embedded Method)

Kaedah terbenam melibatkan manfaat dari kaedah pembungkus (wrapper) dan penyaring (filter) dengan menyertakan interaksi fitur tetapi tetap menjaga kadar komputasi yang wajar. Teknik ini bersifat iteratif yang bermaksud bahwa ia menjaga setiap iterasi dari proses pelatihan model dan dengan hati-hati mengekstrak fitur-fitur yang memberikan kontribusi paling besar terhadap pelatihan untuk iterasi tertentu. Beberapa kajian memberikan pandangan yang berharga mengenai penggunaan teknik pemilihan ciri terbenam, khususnya dalam konteks model pembelajaran mesin berasaskan hutan rawak untuk pengesanan penipuan dalam transaksi kewangan. Sebagai contoh, (Xinwei Zhang et al. 2021) mengemukakan metodologi kejuruteraan ciri baru, HOBA, untuk pengesanan penipuan kad kredit dengan menggunakan senibina pembelajaran mendalam. Pendekatan ini menunjukkan potensi pemilihan ciri yang terbenam dalam meningkatkan keberkesanan model pengesanan penipuan dengan memanfaatkan teknik pembelajaran mendalam. Tambahan pula, (Hu et al. 2021) menekankan penggunaan penghapusan ciri berulang hutan rawak (RF-RFE) sebagai kaedah pemilihan ciri, yang dipilih untuk dibandingkan dengan kaedah pengurangan dimensi ciri lain, menunjukkan keberkesanannya dalam pembelajaran mesin. Selain itu, (Prasetiyowati et al. 2022) menunjukkan bahawa pemilihan ciri menggunakan Information Gain, FFT, dan SMOTE meningkatkan ketepatan prestasi Hutan Rawak, menunjukkan potensi teknik pemilihan ciri yang terbenam dalam meningkatkan ketepatan model pengesanan penipuan. Tambahan pula, (Geng & Yang 2021) menyimpulkan bahawa gabungan ciri-ciri pengenalan penipuan kewangan yang dibina oleh algoritma Relief dan model hutan rawak memberikan kesan pengenalan terbaik, membuktikan keberkesanan pemilihan ciri yang terbenam dalam pengesanan penipuan. Kesimpulannya, rujukan-rujukan ini memberikan pemehaman yang bernilai mengenai penggunaan teknik pemilihan ciri yang terbenam dalam konteks model pembelajaran mesin berasaskan hutan rawak untuk pengesanan penipuan dalam transaksi kewangan. Kajian-kajian ini secara kolektif menekankan kepentingan memanfaatkan kaedah kejuruteraan dan pemilihan ciri yang canggih untuk meningkatkan ketepatan dan keberkesanan sistem pengesanan penipuan.

Teknik pemilihan ciri juga boleh digabungkan dengan kaedah lain untuk meningkatkan model pengesanan penipuan. Sebagai contoh, (Ikeda et al. 2020)

mencadangkan satu rangka kerja baru dalam kejuruteraan ciri untuk pembelajaran mesin dalam pengesanan penipuan kewangan. Rangka kerja ini merangkumi penciptaan ciri, pengumpulan ciri, transformasi ciri, dan pemilihan ciri. Dengan menggabungkan teknik-teknik ini, para pengarang menunjukkan prestasi yang lebih baik dalam mengesan penipuan kewangan. Secara ringkasnya, pemilihan ciri dalam pengurangan jumlah data adalah langkah penting dalam pengesanan penipuan. Ia melibatkan pemilihan subset ciri yang relevan daripada set yang lebih besar untuk mengurangkan dimensi, meningkatkan kecekapan model, dan meningkatkan interpretabiliti. Pelbagai pendekatan seperti kaedah pemilihan ciri yang informatif, model wrapper, dan teknik ciri terbenam telah dicadangkan dalam kajian terdahulu untuk mengatasi cabaran dalam domain ini dan meningkatkan keberkesanan model pengesanan penipuan dalam transaksi kewangan.

2.6 PENILAIAN MODEL

Kebolehan untuk meramal bagi pengelas-pengelas berdasarkan pembelajaran mesin lazimnya diukur berdasarkan nilai ketepatan, precision, kepekaan,, dan skor F1. Ini adalah metrik-metrik penilaian umum dalam pembelajaran mesin yang biasanya digunakan untuk menilai prestasi model pembelajaran mesin, terutamanya dalam konteks pengesanan penipuan, di mana pertukaran antara mengenalpasti transaksi penipuan dengan betul (sensitiviti) dan mengelakkan positif palsu (ketepatan) adalah penting.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$Precision = \frac{TruePositive}{(TotalPredictedPositive)}$$

$$Recall = \frac{True\ Positive}{(Total\ Actual\ Positive)}$$

$$F1 - score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Rajah 2.1 Pengukur-pengukur prestasi yang digunakan dalam kajian ini

2.6.1 Ketepatan (Accuracy)

Ketepatan (Accuracy) ialah nisbah di antara contoh-contoh yang diklasifikasikan dengan betul (kedua-dua positif sebenar dan negatif sebenar) dengan jumlah keseluruhan. Ia memberikan ukuran keseluruhan seberapa dekat ramalan yang dibuat oleh model kepada nilai sebenar atau sebenar. Dengan kata lain, ketepatan mengukur seberapa dekat ramalan yang dibuat oleh model kepada nilai sebenar atau sebenar.

2.6.2 Ketepatan (Precision)

Ketepatan (Precision) ialah nisbah ramalan positif sebenar di antara jumlah contoh yang diramalkan sebagai positif. Metrik ini mengukur seberapa dekat ramalan yang dibuat oleh model kepada satu sama lain. Ia juga dikenali sebagai nilai ramalan positif dan memberi tumpuan kepada ketepatan ramalan positif sahaja.

2.6.3 Kepekaan (Recall)

Sensitiviti atau juga dikenali sebagai kepekaan, adalah metrik yang digunakan dalam pembelajaran mesin untuk menilai keupayaan sesuatu model untuk mengenalpasti semua contoh yang berkaitan dengan suatu kelas. Ia adalah nisbah ramalan positif sebenar di antara jumlah keseluruhan contoh positif sebenar. Dalam kata lain, ia mengukur keupayaan pengklasifikasi untuk mengenalpasti semua contoh positif, termasuk yang terlepas oleh model.

Kepekaan amat penting dalam pengesanan penipuan, di mana kehilangan transaksi penipuan boleh mempunyai akibat yang signifikan. Skor sensitiviti yang tinggi menunjukkan bahawa model adalah efektif dalam mengenalpasti semua contoh positif, manakala skor sensitiviti yang rendah menunjukkan bahawa model terlepas beberapa contoh positif.

2.6.4 Skor F1

Skor F1 ialah satu ukuran ketepatan model dalam tugas pengelasan binari, di mana contoh-contoh diklasifikasikan sebagai 'positif' atau 'negatif'. Ia merupakan cara untuk menggabungkan ketepatan dan sensitiviti model, dan didefinisikan sebagai min harmoni antara ketepatan dan sensitiviti model. Skor F1 secara lazim digunakan untuk menilai

sistem pengambilan maklumat seperti enjin carian dan pelbagai jenis model pembelajaran mesin, terutamanya dalam pemprosesan bahasa semula jadi.

Ia adalah satu metrik tunggal yang menyampaikan keseimbangan antara ketepatan dan sensitiviti model. Ia boleh diubahsuai untuk memberi lebih penekanan kepada ketepatan berbanding sensitiviti, atau sebaliknya, dengan menggunakan skor F yang disesuaikan seperti F0.5 dan F2. Skor F1 adalah sangat berguna apabila kelas-kelas tidak seimbang, dan ia merupakan metrik yang boleh dipercayai hanya jika dataset itu seimbang dari segi kelas.

2.7 KESIMPULAN

Seterusnya, kajian - kajian terdahulu yang dirujuk dalam penerokaan teknik pembelajaran mesin, khususnya mengaplikasikan pemilihan ciri telah dirumuskan dalam Jadual 2.1 seperti di bawah untuk membuat perbandingan dan memahami objektif serta hasil kajian secara lebih terperinci untuk pembangunan model pengesanan penipuankewangan.

Jadual 2.1 Rumusan kajian pembelajaran mesin dan pemilihan ciri dalam pengesanan penipuan kewangan

Bil	Kajian	Objektif	Algoritma	Dapatan Kajian
1	(Ghazikhani et al. 2012)	Kajian ini bertujuan untuk menangani masalah ketidakseimbangan kelas menggunakan pendekatan penambahan secara rawak berdasarkan kaedah pembungkusan.	Model Wrapper, Oversampling	Memperkenalkan pendekatan penambahan secara rawak berdasarkan pembungkusan berjaya mengurangkan kesan ketidakseimbangan kelas dengan berkesan, dengan demikian meningkatkan prestasi model pembelajaran mesin dalam menangani set data yang tidak seimbang.
2	(Kolli & Tatavarthi 2020)	Menggunakan model pembungkus dan rangkaian neural berulang mendalam berdasarkan pengoptimuman air Harris untuk meningkatkan ketepatan, kepekaan, dan spesifisiti pengesanan penipuan dalam transaksi bank.	Model Wrapper, RNN	Model pembungkus digunakan untuk memilih ciri-ciri yang lebih baik dalam mengesan aktiviti penipuan dan terbukti meningkatkan ketepatan model pengesanan penipuan.
3	(Hu et al. 2021)	Untuk membina pendekatan klasifikasi yang kukuh yang megaplikasikan pembelajaran mesin untuk membezakan dengan tepat.	Eliminasi Ciri Berulang (RFE), Hutan Rawak	Menunjukkan keberkesanan penggunaan eliminasi ciri berulang dengan hutan rawak (RF-RFE) sebagai kaedah pemilihan ciri.
4	(H. Xu et al. 2022)	Untuk membangunkan pendekatan inovatif dalam pemilihan atribut utama dalam konteks model ramalan penipuan kewangan.	Keuntungan Maklumat, Model Hibrid (Lasso + Hutan Rawak)	Pendekatan keuntungan maklumat yang menggunakan model hibrid menunjukkan prestasi terbaik dalam pengukuran AUC, menunjukkan kepentingan memilih ciri-ciri dalam pengesanan penipuan kewangan.
5	(Sharma & Chalapathi 2022)	Dengan menumpukan kepada ciri-ciri unik dan cabaran yang berkaitan dengan data kewangan kad kredit, objektifnya adalah untuk menyumbang kepada pembangunan model pengesanan penipuan yang khusus dan berkesan yang menangani	Keuntungan Maklumat, Pembelajaran Mesin	Menekankan kepentingan pemilihan ciri menggunakan ukuran keuntungan maklumat dalam meningkatkan ketepatan model pengesanan penipuan.

bersambung…

		kompleksiti penipuan kad kredit.		
6	(Prasetiyowati et al. 2022)	Untuk meningkatkan ketepatan dan keberkesanan Hutan Rawak dengan memanfaatkan teknik pemilihan ciri dan strategi penyeimbangan untuk mengoptimumkan kebolehan model dalam meramal.	Keuntungan Maklumat, FFT, SMOTE, Hutan Rawak	Kajian tersebut menunjukkan bahawa proses pemilihan ciri, dengan menggunakan kaedah seperti Keuntungan Maklumat, FFT, dan SMOTE, menyumbang kepada peningkatan ketepatan prestasi Hutan Rawak.
7	(Yinhe Chen 2023)	Tujuan kajian ini adalah untuk menyiasat dan menunjukkan peningkatan dalam pengesanan penipuan penyata kewangan melalui integrasi teknik pemilihan ciri dan pembelajaran ketidakseimbangan.	Keuntungan Maklumat, SMOTE	Menunjukkan bahawa ukuran keuntungan maklumat untuk menilai kerelevanan ciri-ciri yang dicadangkan dalam penyelidikan itu telah meningkatkan masalah generalisasi yang lemah dalam kaedah pemilihan ciri tunggal. Selain itu, kajian tersebut menunjukkan bahawa Teknik Pengimbangan Minoriti Tiruan (SMOTE) secara berkesan memperkuat dan meningkatkan keupayaan model untuk mengesan penipuan penyata kewangan dalam syarikat yang disenaraikan.

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

Bab ini menerangkan metodologi kajian yang digunakan untuk menjalankan penyelidikan ini. Metodologi merujuk kepada kaedah serta tatacara pelaksanaan kajian bagi mencapai matlamat kajian. Pendekatan kajian ini berasaskan kepada penyelidikan eksperimental yang dibahagikan kepada lima fasa utama iaitu fasa pengumpulan dan pra-pemprosesan data, fasa pemilihan ciri, fasa pembangunan model, fasa penilaian dan juga perbandingan model. Bab ini bermula dengan penerangan berkaitan dengan penerangan data dan kerangka metodologi kajian yang menjadi tunjang kepada kajian ini. Lima fasa utama dalam kajian ini dibahagikan kepada beberapa sub fasa yang akan dijelaskan secara lebih terperinci di dalam bab ini. Secara ringkasnya, untuk fasa pra-pemprosesan dalam kajian ini, penskalaan ciri kategorikal dan juga pengurangan sampel telah dilaksanakan. Ciri-ciri yang paling relevan akan dipilih dari dataset yang telah dipilih dalam fasa pemilihan ciri dan digunakan untuk membangunkan model klasifikasi (fasa pembangunan/latihan model). Seterusnya, hasil gabungan model-model yang telah dibangunkan akan menjalani fasa penilaian untuk pemilihan model yang terbaik. Hasil eksperimen ini akan membincangkan bagaimana ciri-ciri yang dipilih memainkan peranan dalam pengesanan penipuan dalam urusan kewangan. Ini akan memberikan gambaran yang lebih jelas mengenai mengapa ciri-ciri tertentu adalah kriteria penting dalam mengenal pasti aktiviti penipuan dalam data kewangan.

3.2 SET DATA

Terdapat kekurangan dataset yang boleh didapati secara awam mengenai perkhidmatan kewangan, terutamanya dalam domain transaksi wang mudah alih yang sedang berkembang. Dataset kewangan adalah penting untuk banyak penyelidik dan khususnya bagi yang menjalankan penyelidikan dalam domain pengesanan penipuan. Sebahagian daripada masalah ini adalah kerana sifat peribadi intrinsik transaksi kewangan, yang membawa kepada ketiadaan dataset yang boleh diakses oleh awam.

Untuk kajian semasa, set data sintetik yang dihasilkan menggunakan simulator yang dipanggil PaySim telah digunakan sebagai pendekatan kepada masalah penipuan dalam transaksi kewangan. PaySim menggunakan data yang diagregat daripada dataset peribadi untuk menghasilkan satu dataset sintetik yang menyerupai operasi normal transaksi dan menyuntik tingkah laku berbahaya untuk menilai prestasi kaedah pengesanan penipuan. PaySim mensimulasikan transaksi wang mudah alih berdasarkan sampel transaksi sebenar yang diekstrak dari sebulan log kewangan dari satu perkhidmatan wang mudah alih yang dilaksanakan di sebuah negara Afrika. Log asal disediakan oleh sebuah syarikat multinasional, yang merupakan pembekal perkhidmatan kewangan mudah alih yang kini beroperasi di lebih daripada 14 negara di seluruh dunia. Dataset sintetik ini dikurangkan saiznya sebanyak 1/4 daripada dataset asal, dibuat khas untuk Kaggle dan boleh dimuat turun daripada https://www.kaggle.com/datasets/ealaxi/paysim1. Salah satu cabaran yang terbesar dalam masalah pengesanan penipuan adalah data yang sangat tidak seimbang. Justeru, set data ini telah terpilih untuk menjadi set data eksperimen memandangkan ia mengandungi lebih banyak transaksi yang sah berbanding transaksi penipuan. Ini adalah sampel yang bagus untuk mewakili senario dunia yang sebenar.

Berdasarkan ketersediaan data, terdapat 6,363,620 transaksi kewangan mudah alih termasuk sama ada setiap transaksi adalah penipuan ataupun sah, dengan jumlah akhir 8213 transaksi penipuan dan 6354407 transaksi sah, akan dijadikan sampel untuk melatih dan menilai model. Butiran 11 atribut, termasuk penerangan diberikan dalam Jadual 1. Kebanyakan atribut adalah sedia ada numerikal dan sesuai untuk analisis, tanpa memerlukan proses penukaran jenis data. Terdapat satu atribut yang dinamakan

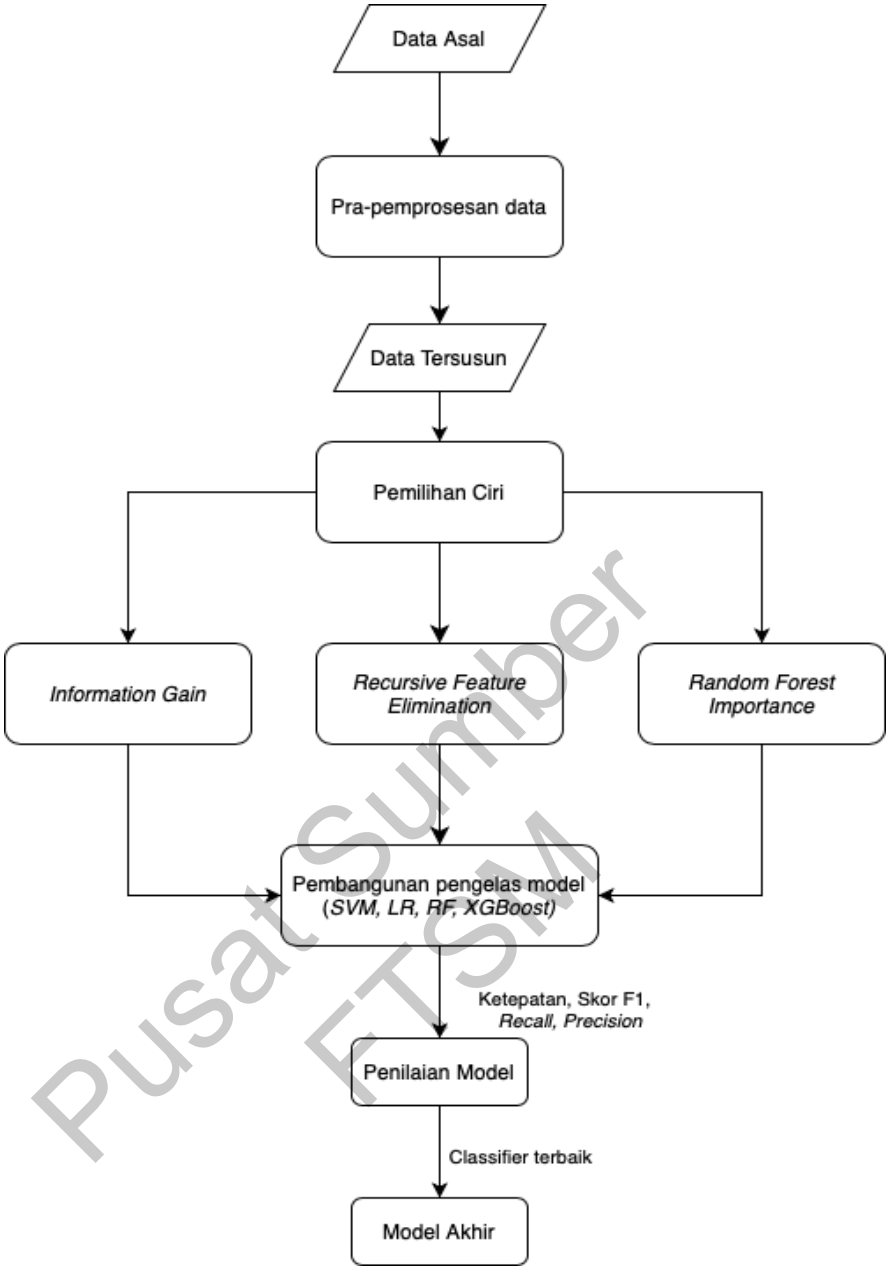
'isFraud' yang menunjukkan status penipuan sebenar bagi transaksi itu. Ini adalah atribut kelas untuk analisis projek ini.

Jadual 3.1 Penerangan atribut

Atribut	Penerangan	Jenis Data
Step	Memetakan satu unit masa dalam dunia nyata. 1 langkah adalah 1 jam masa. Jumlah keseluruhan adalah 744 langkah (30 hari simulasi)	Numerikal
Type	Menunjukkan jenis transaksi. Ia boleh menjadi CASH-IN, CASH-OUT, DEBIT, PAYMENT atau TRANSFER	Kategorikal
Amount	Jumlah transaksi dalam mata wang tempatan.	Numerikal
NameOrig	Pelanggan yang memulakan transaksi.	Kategorikal
OldBalanceOrg	Baki awal pemberi transaksi sebelum transaksi.	Numerikal
NewBalanceOrig	Baki akhir pemberi transaksi selepas transaksi.	Numerikal
NameDest	Penerima yang menerima transaksi.	Kategorikal
OldBalanceDest	Baki awal penerima sebelum transaksi.	Numerikal
NewBalanceDest	Baki akhir penerima selepas transaksi.	Numerikal
isFlaggedFraud	Model perniagaan yang mengenalpasti percubaan haram untuk memindahkan lebih dari 200,000 dalam satu transaksi.	Kategorikal
isFraud	Menunjukkan sama ada transaksi itu sebenarnya penipuan atau tidak. Nilai 1 = Ya dan 0 = Tidak	Binari

3.3 KERANGKA METODOLOGI KAJIAN

Satu representasi skematik mengenai peringkat-peringkat yang terlibat dalam bahagian seterusnya kajian ini ditunjukkan dalam Rajah 3.1. Fasa pertama ialah fasa pengumpulan dan juga pra pemprosesan data, diikuti oleh fasa pemilihan ciri, fasa pembangunan pengelas (classifier), fasa penilaian dan akhirnya fasa perbandingan/pemilihan model terbaik.



Rajah 3.1 Fasa-fasa yang terlibat dalam kajian ini

3.3.1 Fasa Pengumpulan dan Pra-Pemprosesan Data

Dalam pengesanan penipuan kewangan berdasarkan teknik pembelajaran mesin, pra-pemprosesan data memainkan peranan penting dalam menyediakan data untuk analisis dan meningkatkan prestasi model pengesanan penipuan. Beberapa kajian telah menekankan kepentingan kaedah pra-pemprosesan data dalam mengatasi cabaran

seperti kelas yang tidak seimbang, dataset yang bersakala besar, seterusnya boleh meningkatkan prestasi model ramalan penipuan.

a. Keseimbangan Kelas: Pengurangan Sampel secara Rawak

Salah satu cabaran biasa dalam pengesanan penipuan kewangan adalah ketidakseimbangan antara kes penipuan dan kes bukan penipuan. Dalam kajian ini pengurangan sampel secara rawak telah dilakukan dengan mengimbangi taburan kelas dalam dataset dengan mengeluarkan contoh-contoh dari kelas majoriti. Kaedah ini berguna apabila dataset tidak seimbang, dan kelas minoriti (dalam kes ini, transaksi penipuan) kurang diwakili. Dengan mengurangkan bilangan contoh dalam kelas majoriti, pengurangan sampel dapat meningkatkan prestasi model pembelajaran mesin dalam mengesan transaksi penipuan.

b. Kejuruteraan Ciri Kategorikal: Penskalaan Label

Untuk menggunakan atribut kategorikal dalam pelbagai algoritma pembelajaran mesin, langkah pra-pemprosesan untuk menukar nilai-nilai tersebut kepada nilai-nilai numerikal adalah diperlukan. Menukarkan data kategorikal ke dalam format nombor adalah penting bagi algoritma pembelajaran mesin, kerana algoritma-algoritma ini beroperasi pada data berangka. Data kategorikal, seperti label atau nama, tidak boleh diproses secara langsung oleh model pembelajaran mesin. Terdapat pelbagai teknik untuk penukaran data kategorikal, termasuk penskalaan label, penskalaan satu-panas, dan penskalaan ordinal. Setiap teknik mempunyai kelebihan tersendiri dan dipilih berdasarkan sifat data kategorikal dan keperluan model pembelajaran mesin.

Kajian ini akan menggunakan teknik penskalaan label untuk menukarkan data kategorikal menjadi data numerikal dimana setiap nilai kategorikal yang terdapat di dalam atribut ‘type’ akan diberi label nombor yang unik. Dalam penskalaan label, setiap kategori atau label unik diberi nilai integer yang unik. Teknik ini sesuai untuk data kategorikal nominal tanpa sebarang susunan atau ranking intrinsic/semulajadi di antara kategorinya. Walau bagaimanapun, ia mungkin tidak sesuai untuk data kategorikal ordinal di mana kategori mempunyai susunan yang jelas, semula jadi, dan intrinsik. Antara langkah lain yang diambil dalam mentransformasikan data adalah dengan

mengeluarkan atribut kategorikal yang lain iaitu ‘nameOrig’ dan ‘nameDest’. Atribut-atribut ini tidak memberikan sebarang kepentingan untuk proses klasifikasi. Rujuk Jadual 3.2 untuk ringkasan atribut-atribut yang berkenaan dengan proses transformasi ciri dan Jadual 3.3 pula untuk pemetaan nilai atribut ‘type’ sebelum dan selepas menjalani proses penskalaan label.

Jadual 3.2 Atribut - atribut kategorikal yang menjalani proses transformasi

Atribut	**Penerangan**	**Jenis Data**	**Proses Transformasi**
Type	Menunjukkan jenis transaksi. Ia boleh menjadi CASH-IN, CASH-OUT, DEBIT, PAYMENT atau TRANSFER	Kategorikal	Penskalaan Label
NameOrig	Pelanggan yang memulakan transaksi.	Kategorikal	Dikecualikan
NameDest	Penerima yang menerima transaksi.	Kategorikal	Dikecualikan

Jadual 3.3 Nilai atribut ‘type’ sebelum dan selepas penskalaan label

Nilai Asal	**Nilai Kod**
CASH-IN	0
CASH-OUT	1
DEBIT	2
PAYMENT	3
TRANSFER	4

3.3.2 Fasa Pemilihan Ciri

Selain itu, pemilihan ciri adalah aspek penting dalam pra-pemprosesan data untuk pengesanan penipuan kewangan. Ia membantu menyusun semula dataset dengan mengurangkan dimensi dan memilih ciri-ciri yang paling relevan dan informatif sahaja untuk pengesanan penipuan. Dengan mengurangkan jumlah data dan mengekalkan ciri-ciri yang paling penting, pemilihan ciri meningkatkan kecekapan dan keberkesanan pengesanan penipuan. Ini membantu model pembelajaran mesin dalam mengenal pasti aktiviti penipuan dengan lebih baik, walaupun menggunakan dataset yang lebih ringkas. Seperti yang telah diterangkan dalam Bab 2, terdapat 3 kaedah utama pemilihan ciri yang telah digunakan dalam kajian ini. Untuk setiap kaedah tersebut, satu sub-kaedah telah dipilih untuk melaksanakan pemilihan ciri dalam dataset Paysim.

Jadual 3.4 Kaedah dan subkaedah pemilihan ciri dalam pembelajaran mesin

Kaedah Pemilihan Ciri	Sub-Kaedah
Pembungkus (Wrapper Method)	Penghapusan fitur secara berulang (Recursive Feature Elimination)
Penyaring (Filter Method)	Keuntungan maklumat (Information Gain)
Terbenam (Embedded Method)	Kepentingan hutan rawak (Random Forest Importance)

a. Penghapusan Fitur Secara Berulang (Recursive Feature Elimination)

Penghapusan Ciri Secara Berulang (RFE) ialah teknik pemilihan ciri yang lazim digunakan dalam pembelajaran mesin. Ia berfungsi dengan menghapuskan secara berulang ciri yang paling tidak penting sehingga jumlah ciri yang ditetapkan dicapai. RFE menyusun ciri-ciri mengikut model "coef" atau sifat "kepentingan ciri" dan kemudian menghapuskan ciri-ciri yang paling tidak penting satu per satu. Proses ini mengurangkan kompleksiti model dengan memilih ciri-ciri yang paling signifikan, yang boleh membawa kepada peningkatan prestasi model. Dalam kajian ini, teknik Penghapusan Fitur Secara Berulang (RFE) telah digunakan bersama pengklasifikasi Hutan Rawak kerana model pembelajaran mesin ini cenderung untuk memberikan nilai kepentingan yang baik untuk setiap ciri. Oleh itu, gabungan ini akan cenderung untuk mengenal pasti subset fitur yang paling memberi kesan terhadap hasil klasifikasi.

Jadual 3.5 Penjelasan untuk setiap langkah untuk teknik RFE

Langkah		Penjelasan/Pelaksanaan
Inisialisasi algoritma Hutan Rawak	:	100 pohon keputusan (n_estimators=100)
Inisialisasi RFE	:	Parameter n_features_to_select menentukan jumlah ciri untuk dipilih, yang ditetapkan kepada 5 dalam kes ini.
Pemilihan ciri pada data latihan	:	Pemilih RFE (rfe_selector) disesuaikan dengan data latihan (X_train, y_train) menggunakan kaedah fit_transform. Kaedah ini mengenal pasti dan memilih n_features_to_select 5 ciri teratas dari data latihan berdasarkan kepentingannya seperti yang ditentukan oleh algoritma Random Forest. Ciri yang dipilih kemudian diubah dan diberikan kepada X_train_rfe.
Pemilihan ciri pada data ujian	:	Pemilihan ciri yang sama diterapkan pada data ujian (X_test) menggunakan kaedah transform untuk memastikan keseragaman dalam dimensi ciri antara dataset latihan dan ujian. Data ujian yang diubah tersebut

bersambung…

…sambungan

		diberikan kepada X_test_rfe.
Pengekstrakan ciri terpilih	:	Nama-nama atribut yang dipilih disimpan dalam pemboleh ubah selected_features_rfe.

b. Keuntungan Maklumat (Information Gain)

Ukuran keuntungan maklumat atau maklumat bersama membantu untuk mengira pengurangan entropi dari transformasi suatu dataset. Kaedah ini telah digunakan dalam kajian ini untuk menilai kebolehberhubungan atribut bukan sasaran berbanding atribut sasaran (penipuan atau bukan penipuan). Ukuran ini mengukur jumlah maklumat yang disediakan oleh satu ciri tentang atribut sasaran. Kaedah pemilihan ciri ini tidak berasaskan kepada model-model pembelajaran mesin dan hanya bergantung pada ciri tunggal. Ianya tidak selalu memberikan hasil yang optimal apabila digunakan bersama dengan model pembelajaran mesin seperti Hutan Rawak.

Jadual 3.6 Penjelasan untuk setiap langkah untuk teknik IG

Langkah		**Penjelasan/Pelaksanaan**
Pengiraan keuntungan maklumat	:	Fungsi mutual_info_classif dari scikit-learn digunakan untuk mengira keuntungan maklumat bagi setiap ciri dalam data latihan (X_train) berbanding atribut sasaran (y_train). Keuntungan maklumat mengukur pengurangan entropi atau ketidakpastian atribut sasaran yang diperoleh dengan mengetahui nilai suatu ciri.
Pemilihan ciri	:	Ciri-ciri dengan nilai keuntungan maklumat lebih besar daripada 0.1 dan kurang daripada 0.3 dipilih berdasarkan julat ambang (threshold) yang ditetapkan. Ambang ini telah dipilih secara empirikal dan berdasarkan pengetahuan domain. Nama-nama atribut ciri yang dipilih disimpan dalam pembolehubah selected_features_ig.
Subset data latihan dan ujian (X_train_ig, X_test_ig)	:	Dataset latihan dan ujian disubsetkan untuk hanya merangkumi ciri-ciri yang dipilih yang diperoleh dari pengiraan keuntungan maklumat. Ini memastikan hanya ciri-ciri yang relevan disimpan untuk latihan model dan penilaian.
Pengekstrakan ciri terpilih	:	Skrip mencetak nama ciri-ciri yang dipilih (selected_features_ig) untuk memberi gambaran tentang ciri mana yang dianggap penting berdasarkan skor keuntungan maklumat mereka.

c. Kepentingan Hutan Rawak (Random Forest Importance)

Hutan Rawak merupakan algoritma pembelajaran mesin yang sering digunakan dalam pengesanan penipuan. Ia merangkumi beberapa pokok keputusan untuk meningkatkan ketepatan dan kebolehtahan model. Dengan menggunakan Hutan Rawak, kita boleh menilai kepentingan setiap ciri dalam dataset, membantu kita mengenal pasti ciri-ciri yang paling relevan untuk mengesan penipuan. Secara keseluruhan, Hutan Rawak adalah algoritma yang berkuasa untuk mengesan penipuan dan mengenal pasti ciri-ciri penting dalam proses tersebut.

Jadual 3.7 Penjelasan untuk setiap langkah untuk teknik RFI

Langkah		Penjelasan/Pelaksanaan
Inisialisasi dan latihan algoritma Hutan Rawak	:	100 pohon keputusan (n_estimators=100) dan penetapan rawak untuk kebolehulangan (random_state=42). Model kemudian dilatih dengan data latihan (X_train dan y_train) menggunakan kaedah fit.
Pengiraan kepentingan ciri	:	Selepas latihan model Hutan Rawak, kepentingan ciri diperoleh menggunakan atribut feature_importances_ dari model yang dilatih. DataFrame yang dinamakan allfeature_importance_rf dicipta untuk menyimpan nama ciri dan kepentingan mereka yang sepadan. DataFrame disusun mengikut turutan menurun berdasarkan kepentingan ciri.
Pemilihan ciri	:	Ciri-ciri dengan skor kepentingan yang terletak dalam julat yang ditetapkan dari 0.1 hingga 0.3 telah dipilih. Julat ambang (threshold) ini dipilih secara empirikal atau berdasarkan pengetahuan domain. Ciri-ciri yang dipilih disimpan dalam pembolehubah selected_features_rf.
Subset data latihan dan ujian (X_train_rf, X_test_rf)	:	Akhirnya, dataset latihan dan ujian disubsetkan untuk hanya merangkumi ciri-ciri yang dipilih yang diperoleh dari analisis kepentingan Hutan Rawak.
Pengekstrakan ciri terpilih	:	Nama-nama atribut yang dipilih disimpan dalam pemboleh ubah selected_features_rf.

bersambung…

3.3.3 Fasa Pembangunan Model

Ciri-ciri yang telah dipilih dalam fasa pemilihan ciri digunakan untuk melatih beberapa model pembelajaran mesin yang telah disebutkan dalam bab 1, termasuk mesin vektor sokongan, regresi logistik, hutan rawak, dan XGBoost. Dalam penetapan awal algoritma untuk latihan model, empat algoritma telah ditetapkan: Mesin Vektor Sokongan, Regresi Logistik, Hutan Rawak, dan juga XGBoost. Regresi Logistik dan Mesin Vektor

Sokongan mewakili pendekatan pembelajaran tradisional yang sering digunakan untuk membangunkan model pengesanan penipuan kewangan, sementara Hutan Rawak dan XGBoost adalah contoh pembelajaran bersama (ensemble). Setiap algoritma dilengkapi dengan modelnya masing-masing yang telah disesuaikan dengan parameter tertentu, seperti n_estimators = 100 untuk hutan rawak, dengan penetapan rawak untuk kebolehulangan (random_state=42). Ini membolehkan perbandingan prestasi setiap model dalam kajian. Model-model ini disimpan dalam kamus Python bernama "classifiers". Setiap pasangan kunci-nilai dalam kamus tersebut mewakili sebuah pengelas, di mana kunci adalah nama pengelas (misalnya, 'Regresi Logistik', 'Hutan Rawak', dan lain-lain), dan nilai adalah objek pengelas yang sesuai.

a. Mesin Vektor Sokongan (Support Vector Machine)

Mesin Vektor Sokongan (SVM) merupakan klasifier diskriminatif, biasanya dikenali sebagai hiperplane pemisah. Dengan kata lain, algoritma ini menghasilkan hyperplane optimal yang mengklasifikasikan contoh-contoh baru dari data latihan berlabel (pembelajaran terbimbing). Titik-titik data atau vektor yang terdekat dengan hyperplane, yang mempengaruhi arah hyperplane, disebut sebagai Vektor Sokongan, kerana vektor-vektor ini menyokong hyperplane.

b. Regresi Logistik (Logistic Regression)

Regresi Logistik adalah model statistik dan dijalankan apabila atribut bergantung adalah binari. Ia adalah pengelas diskriminatif yang linear dalam parameter-parameternya, dan digunakan untuk menerangkan hubungan antara satu atribut binari bergantung dengan satu atau lebih atribut bebas. Algoritma ini boleh mengendalikan data nominal dan data berangka.

c. Hutan Rawak (Random Forest)

Hutan Rawak (RF) adalah satu teknik pembelajaran gabungan yang diusulkan untuk pohon keputusan. Model RF terdiri daripada banyak pohon keputusan yang digabungkan untuk mengurangkan varians tinggi. Setiap pohon keputusan dilatih pada subset data yang berbeza, dan hasilnya digabungkan untuk membentuk model yang

lebih kukuh. RF juga menggunakan subset ciri-ciri secara rawak pada setiap langkah dalam pembinaan pohon keputusan, memberikan kepelbagaian kepada struktur model. Ini membantu mengurangkan risiko overfitting dan menghasilkan model yang lebih berkesan secara keseluruhan.

d. Peningkatan Gradien Ekstrim (XGBoost)

XGBoost, singkatan bagi eXtreme Gradient Boosting, ialah perisian pembelajaran mesin yang menggunakan pohon keputusan yang diperkukuhkan secara berperingkat atau lebih dikenali sebagai gradient-boosted decision tree (GBDT). Ia direka untuk menjadi cekap, fleksibel, dan mudah diaplikasikan, melaksanakan algoritma pembelajaran mesin di bawah kerangka kerja Gradient Boosting. XGBoost menyediakan pemantapan model secara serentak (juga dikenali sebagai GBDT, GBM) yang dapat menyelesaikan pelbagai masalah sains data dengan cepat dan tepat. Kaedah pembelajaran ini digunakan secara meluas dalam pengesanan penipuan kewangan kerana prestasinya yang tinggi dan kecekapan. Ia telah berjaya mengesan transaksi penipuan dengan cemerlang, sering kali melampaui prestasi algoritma pembelajaran mesin lain.

3.3.4 Fasa Penilaian Model

Dalam fasa penilaian, kajian ini akan menilai prestasi model-model yang telah dilatih menggunakan beberapa metrik penilaian seperti ketepatan (precision), ketepatan (accuracy), kepekaan (recall), dan juga skor F1. Dengan kata lain, setiap model pengelas akan dinilai menggunakan ciri-ciri yang telah dipilih menggunakan tiga kaedah pemilihan ciri yang berbeza: Penghapusan Ciri Secara Berulang (RFE), Keuntungan Informasi (Information Gain), dan Kepentingan Hutan Rawak (RFI). Jadual 3.8 di bawah menunjukkan parameter yang digunakan dalam setiap pengukuran model. Pengukuran akan dibuat untuk gabungan setiap kaedah pemilihan ciri dan juga setiap model pengelas. Model-model pengelas ini telah disimpan di dalam sebuah kamus Python yang diberi nama "classifiers".

Jadual 3.8 Parameter penilaian model untuk kaedah pemilihan ciri dan pengelas

Metriks		**Penghapusan Ciri Secara Berulang (RFE)**	**Keuntungan Maklumat (IG)**	**Kepentingan Hutan Rawak (RFI)**
Ketepatan (accuracy)	:	(X_train_rfe): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui RFE. (X_test_rfe): Hasil ramalan dibuat pada data ujian accuracy_score(): Pengukuran ketepatan	(X_train_ig): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui Keuntungan Maklumat. (X_test_ig): Hasil ramalan dibuat pada data ujian accuracy_score(): Pengukuran ketepatan	(X_train_rf): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui kepentingan Hutan Rawak. (X_test_rf): Hasil ramalan dibuat pada data ujian accuracy_score(): Pengukuran ketepatan
Ketepatan (precision)	:	(X_train_rfe): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui RFE. (X_test_rfe): Hasil ramalan dibuat pada data ujian precision_score(): Pengukuran ketepatan	(X_train_ig): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui Keuntungan Maklumat. (X_test_ig): Hasil ramalan dibuat pada data ujian precision_score(): Pengukuran ketepatan	(X_train_rf): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui kepentingan Hutan Rawak. (X_test_rf): Hasil ramalan dibuat pada data ujian precision_score(): Pengukuran ketepatan
Kepekaan (recall)	:	(X_train_rfe): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui RFE. (X_test_rfe): Hasil ramalan dibuat pada data ujian recall_score(): Pengukuran kepekaan	(X_train_ig): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui Keuntungan Maklumat. (X_test_ig): Hasil ramalan dibuat pada data ujian recall_score(): Pengukuran kepekaan	X_train_rf): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui kepentingan Hutan Rawak. (X_test_rf): Hasil ramalan dibuat pada data ujian recall_score(): Pengukuran kepekaan
Skor F1	:	(X_train_rfe): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui RFE. (X_test_rfe): Hasil ramalan dibuat pada data ujian F1_score(): Pengukuran skor F1	(X_train_ig): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui Keuntungan Maklumat. (X_test_ig): Hasil ramalan dibuat pada data ujian F1_score(): Pengukuran skor F1	X_train_rf): Model pengelas di latih menggunakan ciri-ciri yang dipilih melalui kepentingan Hutan Rawak. (X_test_rf): Hasil ramalan dibuat pada data ujian F1_score(): Pengukuran skor F1

Dengan melakukan langkah-langkah ini untuk setiap kaedah pemilihan ciri dan setiap model pengelas, kajian dapat menilai prestasi setiap gabungan pengelas dan kaedah pemilihan ciri dalam konteks pengesanan penipuan kewangan.

3.4 KESIMPULAN

Bab ini menerangkan mengenai metodologi kajian yang merangkumi penyediaan data, pendekatan kajian, pemilihan ciri, pembangunan model, penilaian model dan pemilihan model. Metodologi kajian yang dirancang dijadikan asas kepada fasa pelaksanaan aktiviti kajian bagi memastikan kajian berjaya mencapai matlamat dan objektif yang ditetapkan. Aktiviti utama dalam bab ini adalah penyediaan data yang menerangkan proses pra-pemprosesan data, pembersihan data, transformasi data, pengelasan data dan juga pemilihan model (subset data). Setiap aktiviti ini diperlukan untuk memastikan kualiti data yang akan diuji dengan model pembelajaran adalah tersedia. Analisis deskriptif yang akan diterangkan di bab seterusnya pula membantu untuk memahami data dengan lebih terperinci untuk membuat pembangunan model dan menganalisis hasil dapat daripada pengujian.

BAB IV

DAPATAN KAJIAN DAN ANALISIS

4.1 PENGENALAN

Bab ini membincangkan dapatan kajian dan analisis terhadap dapatan kajian pengesanan penipuan dalam urusan kewangan berdasarkan teknik pembelajaran mesin Mesin Vektor Sokongan (SVM), Regresi Logistik, XGBoost dan Hutan Rawak yang dibangunkan. Hasil dapatan kajian adalah berdasarkan input daripada metodologi yang dibincangkan di Bab 3. Dalam bab ini, penetapan eksperimen melalui penggunaan set data, pengendalian data kategorikal dalam kaedah pemilihan ciri, dan arkitektur model diterangkan secara lebih terperinci. Hasil dapatan kajian dan hasil analisis deskriptif dibincangkan secara analitikal.

4.2 PENGUMPULAN DATA DAN PRA-PEMPROSESAN DATA

4.2.1 Persediaan Eksperimen

Eksperimen dalam kajian ini dijalankan dengan menggunakan bahasa pengaturcaraan Python versi 3.7.3, platform distribusi Python; Anaconda, dan aplikasi web interaktif Jupyter Notebook. Penggunaan Python sebagai bahasa pengaturcaraan yang merangkumi proses muat naik data dalam format '.csv' dan berfungsi penting untuk menjalankan eksperimen, pemprosesan data, pemilihan ciri, pengelasan, dan evaluasi model.

4.2.2 Huraian Set Data

Set data yang digunakan untuk analisis ini adalah dataset transaksi digital yang dihasilkan secara sintetik menggunakan simulator yang dikenali sebagai PaySim. PaySim mensimulasikan transaksi wang mudah alih berdasarkan contoh transaksi nyata yang diekstrak dari satu bulan catatan kewangan servis wang mudah alih dari sebuah negara di Afrika. Ini mengumpulkan data yang telah dianonimkan dari dataset peribadi untuk menghasilkan dataset sintetik yang kemudiannya dimasukkan transaksi palsu. Pelbagai jenis transaksi penipuan dimasukkan, termasuk masukan tunai (meningkatkan baki akaun), keluar tunai (mengeluarkan tunai), pembayaran (membayar barang atau perkhidmatan), pemindahan (ke pengguna lain) dan debit (menghantar wang ke akaun bank). Dataset ini mengandungi lebih dari 6 juta transaksi dan 11 atribut. Terdapat satu atribut yang bernama 'isFraud' yang menunjukkan status penipuan yang sebenar dari transaksi tersebut. Ini adalah kelas sasaran untuk kajian ini. Fitur ini membantu pengkaji untuk membezakan perilaku pelanggan yang sah dan perilaku penipuan, sekaligus memperkaya penyelidikan dalam pengesanan penipuan transaksi kewangan.

Berikut adalah atribut-atribut yang terdapat dalam dataset:

Jadual 4.8 Atribut dalam dataset

Atribut	Penerangan
Step	Memetakan satu unit masa dalam dunia nyata. 1 langkah adalah 1 jam masa. Jumlah keseluruhan adalah 744 langkah (30 hari simulasi)
Type	Menunjukkan jenis transaksi. Ia boleh menjadi CASH-IN, CASH-OUT, DEBIT, PAYMENT atau TRANSFER
Amount	Jumlah transaksi dalam mata wang tempatan.
NameOrig	Pelanggan yang memulakan transaksi.
OldBalanceOrg	Baki awal pemberi transaksi sebelum transaksi.
NewBalanceOrig	Baki akhir pemberi transaksi selepas transaksi.
NameDest	Penerima yang menerima transaksi.
OldBalanceDest	Baki awal penerima sebelum transaksi.
NewBalanceDest	Baki akhir penerima selepas transaksi.
isFlaggedFraud	Model perniagaan yang mengenalpasti percubaan haram untuk memindahkan lebih dari 200,000 dalam satu transaksi.
isFraud	Menunjukkan sama ada transaksi itu sebenarnya penipuan atau tidak. Nilai 1 = Ya dan 0 = Tidak

4.2.3 Statistik Ringkas

Sebelum meneruskan dengan analisis, disajikan statistik ringkas atribut yang terdapat dalam set data ini. Untuk atribut numerikal, nilai purata, nilai minimum dan maksimum, sisihan piawai (standard deviation) dan julat nilai pada peratusan yang berbeza telah dikira. Bagi atribut kategorikal pula, dinilai hanya jumlah kategori unik, kategori yang paling kerap, dan kekerapan kategorinya.

Out[42]:

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
count	6362620.00	6362620.00	6362620.00	6362620.00	6.362620e+06	6.362620e+06	6362620.00	6362620.0
mean	243.40	179861.90	833883.10	855113.67	1.100702e+06	1.224996e+06	0.00	0.0
std	142.33	603858.23	2888242.67	2924048.50	3.399180e+06	3.674129e+06	0.04	0.0
min	1.00	0.00	0.00	0.00	0.000000e+00	0.000000e+00	0.00	0.0
25%	156.00	13389.57	0.00	0.00	0.000000e+00	0.000000e+00	0.00	0.0
50%	239.00	74871.94	14208.00	0.00	1.327057e+05	2.146614e+05	0.00	0.0
75%	335.00	208721.48	107315.18	144258.41	9.430367e+05	1.111909e+06	0.00	0.0
max	743.00	92445516.64	59585040.37	49585040.37	3.560159e+08	3.561793e+08	1.00	1.0

Rajah 4.1 Ringkasan statistik atribut numerikal

Out[9]:

	type	nameOrig	nameDest
count	6362620	6362620	6362620
unique	5	6353307	2722362
top	CASH_OUT	C1902386530	C1286084959
freq	2237500	3	113

Rajah 4.2 Ringkasan statistik atribut kategorikal

4.2.4 Kejuruteraan Ciri Kategorikal: Penskalaan Label

Adalah penting untuk memastikan bahawa semua atribut dalam set data adalah jenis yang sesuai untuk analisis, semakan menyeluruh telah dijalankan untuk mengenalpasti keperluan penukaran jenis data. Di bawah ini adalah output dari Python yang menunjukkan jenis asal (original) dari setiap atribut dalam dataset Paysim.

```
In [40]: # checking the overall info about the data
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
 10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 534.0+ MB
```

Rajah 4.3 Jenis data asal bagi setiap atribut

Berdasarkan output diatas, atribut 'type' adalah objek. Jenis data objek tidak sesuai untuk pembelajaran mesin kerana algoritma pembelajaran mesin umumnya memerlukan data berbentuk nombor untuk latihan dan membuat ramalan. Jenis data objek biasanya digunakan untuk mewakili atribut kategorikal atau data string, yang mungkin tidak boleh diinterpretasikan secara langsung oleh model pembelajaran mesin. Model pembelajaran mesin beroperasi pada representasi matematik data, dan ia memerlukan ciri-ciri dalam format nombor. Atribut kategorikal, seperti label atau kelas, perlu diubahsuai menjadi nilai nombor melalui teknik seperti penukaran satu-panas atau penskalaan label (label encoding). Jika atribut dibiarkan dalam jenis data objek tanpa penukaran yang sesuai, ia mungkin menyebabkan ralat atau prestasi model yang tidak optimum. Selain itu, kebanyakan algoritma pembelajaran mesin berasaskan pengiraan matematik dan optimasi, dan bekerja dengan data numerikal membolehkan latihan model yang lebih efisien dan tepat. Oleh itu, adalah penting untuk memproses dan menukarkan data kepada format numerikal yang sesuai sebelum menggunakan algoritma pembelajaran mesin.

```
In [43]: # checking type column categories
         df["type"].unique()

Out[43]: array(['PAYMENT', 'TRANSFER', 'CASH_OUT', 'DEBIT', 'CASH_IN'],
               dtype=object)

In [44]: # type_counts stores the count of each category in the "type" column using value_counts()
         type_counts = df["type"].value_counts()

         # transaction_categories stores the unique categories in the "type" column
         # using the index attribute of the type_counts Series.
         transaction_categories = type_counts.index

         # Extracting the counts (values) associated with each category
         quantity = type_counts.values
         quantity

Out[44]: array([2237500, 2151495, 1399284,  532909,   41432])
```

Rajah 4.4 Nilai-nilai yang berbeza untuk fitur 'type'

```
In [27]: from sklearn.preprocessing import LabelEncoder

         # Create a LabelEncoder instance
         label_encoder = LabelEncoder()

         # Apply label encoding to the 'type' column
         df['type'] = label_encoder.fit_transform(df['type'])
```

Rajah 4.5 Penskalaan label untuk fitur 'type'

```
In [5]: # Accessing the mapping of original categories to encoded values
        mapping = dict(zip(label_encoder.classes_, label_encoder.transform(label_encoder.classes_)))

        # Printing the mapping
        print(mapping)

        {'CASH_IN': 0, 'CASH_OUT': 1, 'DEBIT': 2, 'PAYMENT': 3, 'TRANSFER': 4}
```

Rajah 4.6 Pemetaan kategori asal kepada nilai-nilai yang telah dikodkan

Atribut "nameOrig" dan "nameDest" dalam dataset PaySim telah dikeluarkan kerana mengandungi terlalu banyak tahap unik, yang akan menjadikannya sukar untuk digunakan sebagai ciri dalam model pengesanan penipuan. Fitur-fitur alfanumerik ini mengandungi ID pelanggan dan nombor akaun, yang tidak relevan untuk tujuan pengesanan penipuan (Rube & Wirgen Isak 2021). Penghapusan fitur ini tidak memberi kesan terhadap integriti dataset, kerana mereka tidak diperlukan untuk analisis. Walaupun mereka mungkin berguna untuk jenis analisis tertentu, seperti analisis rangkaian atau penyesuaian data, mereka tidak biasanya dianggap sebagai ciri yang bernilai untuk pengesanan penipuan dalam konteks model pembelajaran mesin. Dalam

pengesanan penipuan, ciri-ciri seperti jumlah transaksi, jenis transaksi, masa, dan baki akaun lebih biasa digunakan untuk melatih model dan mengenal pasti aktiviti penipuan.

```
In [9]: # taking the required columns for further analysis
        # Dropping specified columns from X
        columns_to_drop = ['nameDest', 'nameOrig','isFraud']
        X = df.drop(columns=columns_to_drop, axis=1)

        # Displaying the first few rows of the updated X
        print(X.head())
```

```
   step  type    amount  oldbalanceOrg  newbalanceOrig  oldbalanceDest  \
0     1     2   9839.64       170136.0       160296.36             0.0
1     1     2   1864.28        21249.0        19384.72             0.0
2     1     4    181.00          181.0            0.00             0.0
3     1     1    181.00          181.0            0.00         21182.0
4     1     2  11668.14        41554.0        29885.86             0.0

   newbalanceDest  isFlaggedFraud
0             0.0               0
1             0.0               0
2             0.0               0
3             0.0               0
4             0.0               0
```

Rajah 4.7 Penyingkiran fitur-fitur kategorikal ‘nameDest’ dan ‘nameOrig’

4.2.5 Semakan Nilai Hilang

```
In [7]: # checking for null values
        df.isna().sum()
```

```
Out[7]: step              0
        type              0
        amount            0
        nameOrig          0
        oldbalanceOrg     0
        newbalanceOrig    0
        nameDest          0
        oldbalanceDest    0
        newbalanceDest    0
        isFraud           0
        isFlaggedFraud    0
        dtype: int64
```

Rajah 4.8 Semakan nilai hilang

Data yang hilang penting untuk dipertimbangkan kerana ia boleh mengakibatkan keputusan yang berat sebelah dan mengurangkan kekuatan statistik satu kajian, yang mungkin membawa kepada kesimpulan yang tidak sah. Penyebab data yang hilang adalah penting untuk menentukan jenis data yang hilang dan cara untuk mengatasinya. Terdapat tiga jenis utama data yang hilang: hilang sepenuhnya secara rawak (MCAR),

hilang secara rawak (MAR), dan tidak hilang secara rawak (MNAR). Memahami sebab-sebab data yang hilang adalah penting untuk mengendalikan data yang tinggal dengan betul. Disyorkan untuk secara berkala melakukan pemeriksaan data semasa kajian untuk mengenal pasti masalah data yang hilang dan mengambil tindakan segera untuk menanganinya. Secara umumnya, kadar data yang hilang yang lebih rendah adalah lebih baik, dengan kadar kehilangan kurang daripada atau sama dengan 5% dianggap remeh, dan analisis mungkin condong kepada berat sebelah jika 10% atau lebih data hilang (H. Kang 2013). Oleh itu, penting untuk mengendalikan data yang hilang dengan teliti untuk memastikan kesahan dan kebolehpercayaan keputusan kajian.

Dalam fasa ini, pemeriksaan jika terdapat nilai-nilai yang hilang dalam set data juga dilakukan. Walaubagaimanapun, set data kajian, yang merupakan dataset sintetik yang dihasilkan oleh simulator PaySim, tidak mengandungi nilai yang hilang atau nilai yang tidak betul. Ini disebabkan oleh fakta bahawa dataset ini dicipta untuk mencerminkan rekod kewangan dalam kehidupan nyata, dan sebagai hasilnya, tidak memerlukan pembersihan data yang mendalam. Seperti rajah 4.7 di atas, kod dan output menunjukkan jumlah keseluruhan nilai yang hilang dalam semua atribut, adalah sifar.

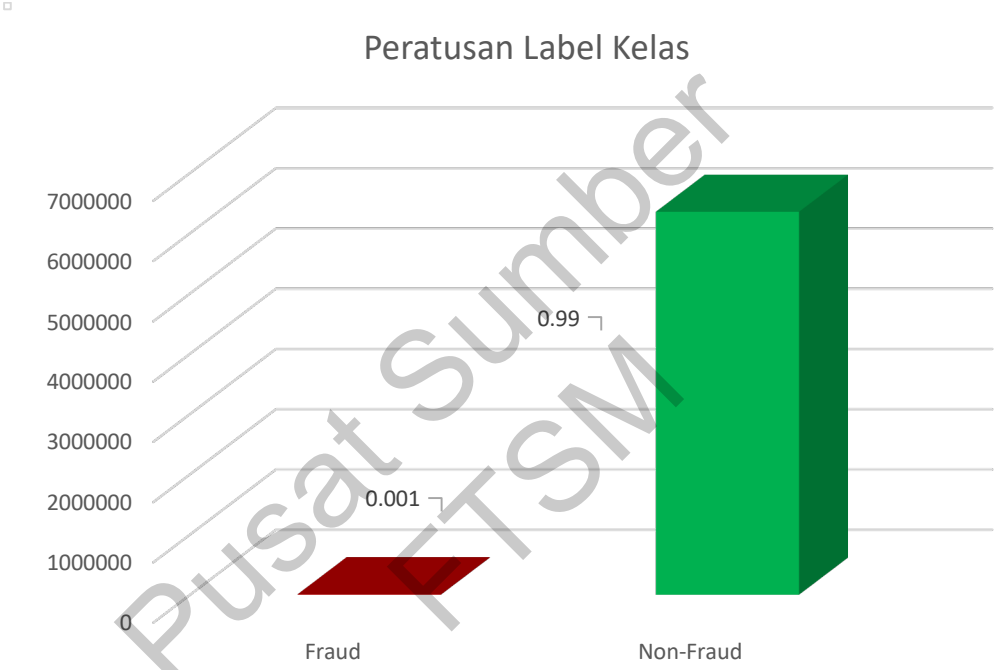
4.2.6 Keseimbangan Kelas: Pengurangan Sampel secara Rawak

Dalam analisis penerokaan ini, kita menilai ketidakseimbangan kelas dalam dataset. Ketidakseimbangan kelas ditakrifkan sebagai peratus daripada jumlah keseluruhan transaksi yang terdapat dalam atribut ‘isFraud’. Keluaran peratus frekuensi untuk atribut kelas ‘isFraud’ ditunjukkan di bawah:

```
In [13]: # Count the number of occurrences of each value in the 'isFraud' column
         fraud_counts = df['isFraud'].value_counts()

         # Calculate the percentage of fraud and non-fraud cases
         fraud_percentage = (fraud_counts[1] / len(df)) * 100
         non_fraud_percentage = (fraud_counts[0] / len(df)) * 100

         # Print the fraud and non-fraud percentages
         print("Fraud percentage:", fraud_percentage)
         print("Non-fraud percentage:", non_fraud_percentage)

Fraud percentage: 0.12908204481801522
Non-fraud percentage: 99.87091795518198
```

Rajah 4.9 Ketidakseimbangan kelas sasaran

Seperti yang dapat kita lihat dari Rajah 4.10 di bawah, terdapat perbezaan yang besar antara peratus transaksi kewangan yang sah dan penipuan. Hanya 0.13% (8,213) transaksi dalam dataset kajian adalah penipuan, menunjukkan ketidakseimbangan kelas yang tinggi dalam dataset. Ini adalah salah faktor penting yang harus dipertimbangkan dalam fasa pra-pemprosesan model kerana jika kita membina model pembelajaran mesin berdasarkan data yang sangat condong ini, transaksi bukan penipuan akan mempengaruhi latihan model hampir sepenuhnya, dengan itu mempengaruhi prestasi model kelak.



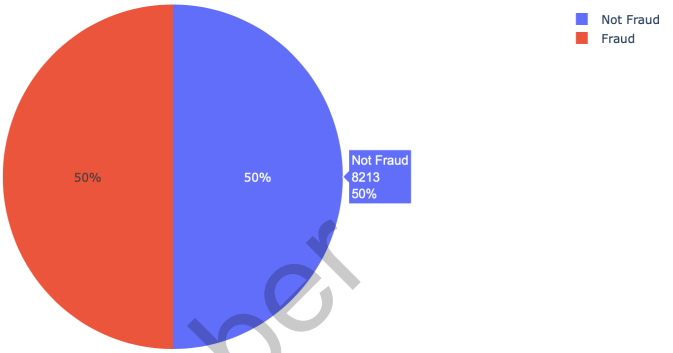
Rajah 4.10 Visualisasi ketidakseimbangan kelas sasaran

Pertimbangan terhadap ketidakseimbangan kelas dalam pengesanan penipuan dalam transaksi kewangan adalah penting untuk pembangunan model pengesanan yang berkesan Ketidakseimbangan kelas berlaku apabila satu kelas, seperti transaksi penipuan, jauh kurang berbanding kelas lain, seperti transaksi sah. Ketidakseimbangan ini boleh mengakibatkan model yang condong dan bias kepada kelas majoriti, menyebabkan pengesanan yang lemah terhadap kelas minoriti, yang sering kali merupakan kelas yang diinginkan dalam pengesanan penipuan. Selain itu, apabila terdapat ketidakseimbangan yang tinggi antara jumlah transaksi sah dan penipuan dalam dataset, ia juga boleh menyebabkan beberapa isu, seperti di bawah:

1. Bias Model: Oleh kerana majoriti data terdiri daripada transaksi yang sah, model pembelajaran mesin mungkin akan belajar untuk mengelaskan majoriti kelas dengan lebih tepat, manakala mengakibatkan prestasi buruk pada kelas minoriti (transaksi penipuan)
2. Kadar Positif Sebenar Berkurang: Disebabkan oleh ketidakseimbangan kelas, kadar positif sebenar, yang dikenali sebagai sensitiviti model mungkin lebih rendah, menyebabkan kurangnya transaksi penipuan yang dikesan.
3. Kadar Positif Palsu Meningkat: Akibat daripada ketidakseimbangan kelas, kadar positif palsu, yang dikenali sebagai spesifisiti model mungkin lebih tinggi, menyebabkan lebih banyak transaksi sah dilabel secara tidak betul sebagai transaksi penipuan
4. Pembelajaran Model yang Berlebihan (Overfitting): Ketidakseimbangan kelas boleh menyebabkan model pembelajaran mesin belajar secara berlebihan terhadap data, mengakibatkan generalisasi yang buruk dan prestasi yang berkurang pada data yang tidak dilihat sebelum ini ataupun dalam kata lain, data yang tidak termasuk dalam set data latihan.

Menangani ketidakseimbangan kelas adalah penting untuk pengesanan penipuan yang berkesan dalam transaksi kewangan, kerana ia membantu meningkatkan ketepatan dan kecekapan model pembelajaran mesin, akhirnya memberikan perlindungan yang lebih baik terhadap aktiviti penipuan. Untuk menangani masalah dalam pengesanan penipuan ini, kajian ini akan menggunakan teknik penyusunan semula data iaitu pengurangan data secara rawak (random undersampling). Ia boleh digunakan untuk menyeimbangkan kelas dalam dataset, walaupun kaedah ini tidak selalu berkesan atau efisien. Teknik pengurangan data secara rawak (RUS) adalah satu kaedah yang digunakan untuk mengatasi ketidakseimbangan kelas dalam pembelajaran mesin, terutamanya dalam konteks pengesanan penipuan dalam transaksi kewangan. Ia melibatkan pengurangan bilangan contoh dalam kelas majoriti (transaksi sah, di dalam dataset ini) untuk mencapai taburan yang lebih seimbang antara kelas majoriti dan minoriti. Dengan melakukan penurunan di bawah kelas majoriti, dataset latihan menjadi lebih seimbang, membolehkan model pembelajaran mesin untuk belajar dengan lebih berkesan dari kedua-dua kelas dan meningkatkan kebolehannya untuk mengesan aktiviti penipuan.

```
In [11]: # Undersampling:
undersampler = RandomUnderSampler(random_state=42)

# Contain the undersampled feature set and target variable, respectively.
X_res, y_res = undersampler.fit_resample(X, y)
print("X_res shape:", X_res.shape)

X_res shape: (16426, 8)

In [12]: #Checking for balance in target column
fig = go.Figure(data=[go.Pie(labels=['Not Fraud','Fraud'], values=y_res['isFraud'].value_counts())])
fig.show()
```



Rajah 4.11 Pecahan kelas sasaran selepas dikenakan Teknik RUS

4.3 PEMILIHAN CIRI

Jadual 4.9 Ciri-ciri terpilih untuk setiap kaedah pemilihan ciri

Kaedah Pemilihan Ciri	Ciri-Ciri Terpilih (0.1<=Nilai Informasi<=0.3)
Penghapusan Fitur Secara Berulang (Recursive Feature Elimination)	Type, oldBalanceOrg, Amount, newBalanceOrig, newBalanceDest
Keuntungan Maklumat (Information Gain)	Step, Type, oldBalanceOrg, Amount, newBalanceOrig
Kepentingan Hutan Rawak (Random Forest Importance)	Type, oldBalanceOrg, Amount, newBalanceOrig

Berdasarkan Jadual 4.2 di atas, melalui kesemua tiga kaedah pemilihan ciri, ciri-ciri 'Type', 'oldBalanceOrg', 'Amount', dan 'newBalanceOrig' secara konsisten muncul sebagai ciri-ciri yang paling penting untuk mengesan penipuan kewangan. 'Type' (jenis-jenis transaksi), 'oldBalanceOrg' (baki awal dalam akaun asal), 'Amount' ‘jumlah transaksi’, dan 'newBalanceOrig' (baki baru dalam akaun asal) diidentifikasi secara konsisten sebagai petunjuk penting bagi mengenal pasti aktiviti penipuan. Dapatan ini menunjukkan bahawa transaksi yang melibatkan jenis tertentu, jumlah tertentu, dan perubahan tertentu dalam baki asal adalah lebih cenderung berkaitan dengan tingkah laku penipuan. Konsistensi dalam pemilihan ciri melalui kaedah yang berbeza mengukuhkan kebolehpercayaan ciri-ciri ini dalam membezakan transaksi penipuan daripada transaksi sah.

'Jenis' transaksi adalah penting kerana jenis transaksi yang berbeza mungkin menunjukkan tahap risiko yang berbeza untuk aktiviti penipuan. Jenis transaksi tertentu mungkin lebih penting kepada manipulasi atau penyalahgunaan Sebagai contoh, dalam dataset kajian ini, transaksi TRANSFER dan CASH_OUT terbukti merupakan transaksi penipuan (rujuk Rajah 4.12). Selain itu, atribut 'oldBalanceOrg' dan 'newBalanceOrig' juga adalah penting untuk mengesan anomali dalam baki akaun asal sebelum dan selepas transaksi. Perubahan tiba-tiba atau ketidaksempurnaan dalam baki ini boleh menandakan tingkah laku penipuan. Akhir sekali, atribut 'Amount' adalah signifikan kerana jumlah transaksi besar atau tidak biasa mungkin menunjukkan aktiviti penipuan yang berpotensi, seperti penggubahan wang atau pemindahan dana tanpa kebenaran. Atribut-atribut yang dipilih selaras dengan corak dan ciri-ciri umum yang berkaitan dengan penipuan kewangan, memberikan pandangan berharga untuk membina model pengesanan penipuan yang berkesan. Konsistensi dalam pemilihan ciri melalui pelbagai kaedah meningkatkan keyakinan terhadap kepentingan ciri-ciri ini dalam mengesan penipuan kewangan, mengukuhkan relevansi dan kebolehpercayaan mereka dalam aplikasi pengesanan penipuan.


```
# Print the types of fraudulent transactions
# Provides a list of unique 'type' values for rows where 'isFraud' == 1 in the DataFrame
fraudulent_types = df.loc[df.isFraud == 1, 'type'].drop_duplicates().values
print('\nThe types of fraudulent transactions are: {}'.format(list(fraudulent_types)))

# Filter for fraudulent TRANSFERs
dfFraudTransfer = df.loc[(df.isFraud == 1) & (df.type == 'TRANSFER')]
print('\nThe number of fraudulent TRANSFERs = {}'.format(len(dfFraudTransfer)))

# Filter for fraudulent CASH_OUTs
dfFraudCashout = df.loc[(df.isFraud == 1) & (df.type == 'CASH_OUT')]
print('\nThe number of fraudulent CASH_OUTs = {}'.format(len(dfFraudCashout)))
```

```
The types of fraudulent transactions are: ['TRANSFER', 'CASH_OUT']

The number of fraudulent TRANSFERs = 4097

The number of fraudulent CASH_OUTs = 4116
```

Rajah 4.12 Jenis-jenis transaksi penipuan dalam dataset Paysim

Rajah 4.12 di atas menunjukkan analisis transaksi penipuan dalam dataset. Terlebih dahulu, ia mencetak jenis-jenis transaksi penipuan yang unik dengan mengekstrak nilai-nilai 'type' di mana 'isFraud' sama dengan 1 dalam DataFrame. Kemudian, ia menyaring transaksi TRANSFER dan CASH_OUT yang merupakan transaksi penipuan dengan menentukan syarat 'isFraud' sama dengan 1 dan 'type' sama dengan 'TRANSFER' atau 'CASH_OUT'. Jumlah transaksi penipuan TRANSFER dan CASH_OUT kemudian dicetak untuk tujuan pemeriksaan dan analisis selanjutnya.

4.4 PENILAIAN DAN PERBANDINGAN MODEL

Jadual 4.10 di bawah menyenaraikan model pengelas yang digunakan dalam ujian prestasi, termasuk Regresi Logistik, Mesin Vektor Sokongan (SVM), Hutan Rawak, dan XGBoost. Metrik prestasi utama yang diukur termasuk ketepatan (accuracy), ketepatan (precision), kepekaan (recall), dan skor F1 (F1 scores). Setiap model pengelas diuji menggunakan tiga kaedah pemilihan ciri yang berbeza, iaitu Penghapusan Ciri Secara Berulang (RFE), Keuntungan Informasi (IG), dan Kepentingan Hutan Rawak (RFI). Jadual ini menunjukkan skor prestasi untuk setiap kombinasi model pengelas dan kaedah pemilihan ciri. Skor untuk setiap metrik dinyatakan dalam nilai berkisar antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan prestasi yang lebih baik.

Jadual 4.10 Hasil dapatan yang mengandungi pengukuran untuk setiap model pengelas

Metriks **Model Pengelas**	**Ketepatan (accuracy)**	**Ketepatan (precision)**	**Kepekaan (recall)**	**Skor F1 (F1 scores)**
Regresi Logistik + RFE	0.8947	0.9067	0.8947	0.8938
Regresi Logistik + IG	0.9040	0.9044	0.9040	0.9040
Regresi Logistik + RFI	0.9042	0.9179	0.9042	0.9033
Mesin Vektor Sokongan + RFE	0.8389	0.8731	0.8389	0.8355
Mesin Vektor Sokongan + IG	0.8557	0.8799	0.8557	0.8537
Mesin Vektor Sokongan + RFI	0.8559	0.8800	0.8559	0.8539
Hutan Rawak + RFE	0.9925	0.9925	0.9925	0.9925
Hutan Rawak + IG	0.9933	0.9933	0.9933	0.9933
Hutan Rawak + RFI	0.9931	0.9931	0.9931	0.9931
XGBoost + RFE	0.9945	0.9945	0.9945	0.9945
XGBoost + IG	0.9935	0.9935	0.9935	0.9935
XGBoost + RFI	0.9943	0.9943	0.9943	0.9943

Prestasi keseluruhan model diukur melalui beberapa metrik, termasuk ketepatan (accuracy), ketepatan (precision), kepekaan (recall), dan skor F1. Dari perspektif ini, hasil menunjukkan bahawa semua model menunjukkan prestasi yang tinggi, dengan skor ketepatan melebihi 0.83 untuk semua kombinasi model pengelas dan kaedah pemilihan ciri. Walau bagaimanapun, penting untuk diperhatikan bahawa meskipun ketepatan adalah penting, ia mungkin tidak mencerminkan keseluruhan prestasi model, terutamanya jika dataset mempunyai ketidakseimbangan kelas. Selain itu, kaedah pemilihan ciri yang berbeza dapat memberikan kesan yang signifikan terhadap prestasi

model. Misalnya, untuk Regresi Logistik, prestasi ketepatan yang diperoleh dengan menggunakan Penghapusan Ciri Secara Berulang (RFE) adalah lebih rendah daripada ketepatan yang diperoleh dengan menggunakan Keuntungan Maklumat (IG) atau Kepentingan Hutan Rawak (RFI). Pengurangan ketepatan yang diperoleh dengan menggunakan Penghapusan Ciri Secara Berulang (RFE) berbanding dengan Keuntungan Maklumat (IG) atau Kepentingan Hutan Rawak (RFI) dalam Regresi Logistik disebabkan oleh beberapa faktor:

1. Pemilihan Ciri: RFE menggunakan proses iteratif untuk mengecilkan set ciri sehingga hanya ciri-ciri yang paling penting sahaja yang kekal. Walau bagaimanapun, proses ini mungkin tidak memberi tumpuan kepada keseluruhan maklumat yang terkandung dalam dataset, yang boleh mengurangkan ketepatan model. Ini disebabkan RFE mungkin membuang ciri yang sebenarnya penting untuk prestasi model Regresi Logistik.

2. Kompleksiti Model: Regresi Logistik mungkin lebih sensitif terhadap kehilangan maklumat yang diwakili oleh ciri tertentu. Jika RFE secara tidak sengaja mengeluarkan ciri penting, ia boleh menyebabkan penurunan ketepatan model.

3. Kebergantungan pada interaksi ciri: Regresi Logistik bergantung pada interaksi antara ciri untuk menghasilkan ramalan yang tepat. Jika RFE membuang ciri yang penting untuk interaksi ini, ia boleh menyebabkan penurunan ketepatan.

Kesimpulannya, perbezaan dalam ketepatan antara kaedah pemilihan ciri mungkin disebabkan oleh perbezaan dalam cara mereka menilai dan memilih ciri, serta bagaimana model pengelas merespon terhadap perubahan dalam set ciri. Ini menunjukkan bahawa penting untuk memilih kaedah pemilihan ciri yang sesuai dengan dataset dan tujuan kajian.

Tambahan pula, jenis model yang digunakan dalam kajian juga memainkan peranan penting dalam prestasi keseluruhan. Terdapat variasi dalam prestasi antara model-model yang berbeza, dengan beberapa model menunjukkan prestasi yang lebih baik daripada yang lain. Sebagai contoh, pengelas bersama (ensemble), iaitu Hutan Rawak dan juga XGBoost menunjukkan prestasi yang amat tinggi dengan skor

ketepatan melebihi 0.99 untuk semua kaedah pemilihan ciri, sementara Mesin Vektor Sokongan (SVM) cenderung menunjukkan prestasi yang lebih rendah. Ini adalah disebabkan keupayaan pembelajaran ensemble dalam mengendalikan data berskala besar. XGBoost dan Hutan Rawak cenderung lebih baik dalam menangani dataset besar berbanding SVM. Mereka dapat menyusun banyak pohon keputusan secara serentak dan secara efisien memproses maklumat dalam jumlah yang besar. Selain itu, pembelajaran ensemble menggabungkan ramalan dari beberapa model (pohon keputusan dalam kes ini) untuk membuat keputusan akhir. Pendekatan ini sering kali menghasilkan model yang lebih kuat daripada model tunggal seperti SVM dan juga Logistik Regresi. Model XGBoost dan Hutan Rawak juga cenderung lebih fleksibel dalam menangani pelbagai jenis data dan dapat menyesuaikan dengan baik dengan keadaan data yang berbeza. Ini dapat membantu mereka mengenali corak yang lebih rumit dalam data yang mungkin sulit untuk ditangani oleh kaedah-kaedah tradisional yang lain. Selanjutnya, tuning parameter yang mudah boleh dilakukan apabila menggunakan pembelajaran ensemble. XGBoost dan Hutan Rawak sering kali mempunyai parameter yang lebih mudah ditetapkan dan disesuaikan untuk mencapai prestasi yang optimal. Manakala, SVM pula memerlukan penyesuaian yang lebih teliti dan pemilihan parameter yang tepat untuk mencapai hasil yang optimum. Dengan mengambil kira faktor-faktor ini, XGBoost dan Hutan Rawak menjadi pilihan yang lebih unggul berbanding SVM dan juga Regresi Logistik untuk mengatasi masalah pengesanan penipuan dalam kajian ini.

Akhir sekali, satu aspek yang perlu diperhatikan adalah kestabilan dalam prestasi model menggunakan kaedah pemilihan ciri yang berbeza. Meskipun terdapat variasi dalam prestasi antara model-model yang berbeza, beberapa model menunjukkan kestabilan dalam prestasi mereka. Contohnya, XGBoost menunjukkan prestasi yang konsisten tinggi, dengan skor ketepatan, ketepatan, kepekaan, dan skor F1 yang hampir sama tinggi untuk semua kaedah pemilihan ciri. Dalam kesimpulannya, analisis hasil menunjukkan bahawa pemilihan model pengelas dan kaedah pemilihan ciri memainkan peranan penting dalam prestasi model untuk pengelasan data. Oleh itu, pemilihan model dan kaedah pemilihan ciri yang sesuai adalah kunci untuk mencapai prestasi yang optimum dalam pengelasan data. Peneliti perlu mempertimbangkan kelebihan dan

kelemahan setiap model dan kaedah pemilihan ciri sebelum membuat keputusan yang bermakna dalam konteks kajian mereka.

4.5 KESIMPULAN

Bab ini membentangkan keputusan kajian yang diperoleh hasil daripada eksperimen yang telah dijalankan berdasarkan metodologi kajian. Hasil analisis deskriptif memberikan panduan untuk perancangan pembangunan model. Penetapan eksperimen ditentukan berdasarkan kajian literatur dan dibuktikan melalui pengujian. Kaedah dan algoritma pembangunan arkitektur model dibincangkan secara terperinci. Model pengelasan pengesanan penipuan menggunakan pendekatan pemilihan ciri dan pembelajaran bersama (ensemble) telah berjaya dijalankan menggunakan set data sintetik Paysim. Hasil dapatan dibincangkan dengan membuat perbandingan antara hasil gabungan tiga (3) kaedah pemilihan ciri dengan model SVM, LR, RF dan XGBoost, dan juga peratusan label kelas melalui penilaian prestasi setiap gabungan-gabungan model ini.

Selain itu, model juga menunjukkan kestabilan dalam prestasi mereka. Contohnya, XGBoost menunjukkan prestasi yang konsisten tinggi, dengan skor ketepatan, ketepatan, kepekaan, dan skor F1 yang hampir sama tinggi untuk semua kaedah pemilihan ciri. Model terbaik dan paling menunjukkan kestabilan dalam prestasi mereka adalah model RFE-XGBoost dengan markah ketetapan (accuracy), ketepatan (precision), kepekaan (recall) dan skor F1 semuanya sebanyak 0.9945. Prestasi model pengelas XGBoost juga sangat memuaskan pada kedua kaedah pemilihan ciri yang lain iaitu IG dan RFI, dengan nilai ketepatan (accuracy), ketepatan (precision), kepekaan (recall), dan skor F1 masing-masing adalah 0.9935 dan 0.9943.

BAB V

RUMUSAN DAN CADANGAN

5.1 PENGENALAN

Bab ini merumuskan keseluruhan kajian yang telah dijalankan. Perkara yang dibincangkan di dalam bab ini adalah berkaitan dengan rumusan kajian, penerangan berkenaan sumbangan kajian; dan kajian di masa hadapan untuk proses penambahbaikan bagi hasil penyelidikan pengesanan penipuan menggunakan pendekatan pembelajaran mesin, termasuklah kaedah-kaedah pemilihan ciri, pembangunan model pengesanan penipuan, analisis keseimbangan data, dan model perbandingan pengesanan penipuan bagi kaedah pembelajaran bersama (ensemble) berbanding kaedah-kaedah lain yang popular dalam kajian literatur dalam konteks pengesanan penipuan.

5.2 RUMUSAN PENEMUAN DAN PENCAPAIAN OBJEKTIF KAJIAN

Secara keseluruhannya, kajian ini bertujuan untuk meneroka pemilihan ciri menggunakan keuntungan maklumat, hutan rawak dan kaedah pembungkus-RFE, seterusnya membandingkan teknik klasifikasi yang berbeza termasuk Mesin Vektor Sokongan (SVM), Regresi Logistik (LR), Hutan Rawak (RF), dan juga XGBoost. Kesimpulan kajian telah menekankan keberkesanan teknik pembelajaran mesin dalam mengesan penipuan kewangan. Perbandingan teknik klasifikasi menunjukkan bahawa hasil gabungan model pengelas XGBoost, serta kaedah pembungkus-RFE menunjukkan hasil yang paling unggul dalam pemilihan ciri untuk pengesanan penipuan dalam urusan kewangan. Kajian mendapati bahawa algoritma XGBoost menunjukkan prestasi yang kukuh dalam mengenal pasti aktiviti penipuan dengan tepat dalam data kewangan. Selain itu, kaedah pembungkus-RFE dalam pemilihan ciri yang relevan hanya terbukti berkesan untuk sesetengah model pengelas. Ini adalah disebabkan kerana

kebergantungan kaedah pemilihan ciri terhadap model pengelas yang dipilih. Semua ini menyumbang kepada kejayaan keseluruhan model pengesanan penipuan.

Selanjutnya, perbandingan teknik klasifikasi menunjukkan kelebihan Mesin Vektor Sokongan (SVM), Regresi Logistik (LR), Hutan Rawak (RF) dan XGBoost dalam mengesan penipuan kewangan. Kajian menunjukkan bahawa pembelajaran ensemble, XGBoost cemerlang dalam mengendalikan data yang kompleks dan berdimensi tinggi, menjadikannya sesuai untuk pengesanan penipuan dalam dataset kewangan. Ia menunjukkan prestasi yang sangat kukuh dalam mengendalikan hubungan kompleks dan bukan linear dalam data. Ia menunjukkan kebolehterjemahan dan kelogikan pelaksanaan, memberikan pandangan berharga tentang faktor-faktor yang menyumbang kepada aktiviti penipuan. Ia terbukti berkesan, terutamanya dalam senario di mana interpretabiliti model adalah penting. Pendekatan pembelajaran ensemble menunjukkan ketepatan tinggi dan kekuatan dalam mengenal pasti transaksi penipuan, memberikan keseimbangan antara ketepatan dan generalisasi. Pemilihan teknik yang paling sesuai bergantung pada pelbagai faktor, termasuk ciri-ciri khusus dataset dan keperluan interpretabiliti aplikasi. Penyelidik dan pengamal perlu mempertimbangkan dengan teliti faktor-faktor ini ketika melaksanakan sistem pengesanan penipuan.

Secara ringkasnya, kajian ini telah menekankan kepentingan pembelajaran mesin terutamanya ensemble dan kaedah ciri yang bersesuaian untuk model pengelas yang dipilih dalam mengesan penipuan kewangan secara efektif. Walaubagaimanapun, model pengelas XGBoost menunjukkan prestasi yang konsisten tinggi, dengan skor ketepatan, ketepatan, kepekaan, dan skor F1 yang hampir sama tinggi untuk semua kaedah pemilihan ciri yang digunakan dalam kajian ini. Penemuan ini menekankan potensi XGBoost dan kaedah pembungkus-RFE untuk pemilihan ciri, serta kelebihan setiap odel pengelas dalam mengklasifikasikan aktiviti penipuan. Hasil kajian telah menawarkan asas untuk pembangunan sistem pengesanan penipuan yang kukuh dan efisien, membantu usaha berterusan dalam memerangi salah laku kewangan dalam persekitaran yang pelbagai dan dinamik. Akhir sekali, kajian ini telah memberikan sumbangan kepada kemajuan metodologi pengesanan penipuan kewangan, memberikan pandangan berharga kepada pengamal dan penyelidik dalam bidang ini.

5.3 KAJIAN MASA HADAPAN

Dalam merangka kajian masa depan berkaitan pengesanan penipuan kewangan menggunakan pembelajaran mesin, adalah penting untuk meneroka bidang yang boleh meningkatkan keberkesanan dan kecekapan model pengesanan penipuan. Salah satu bidang kajian yang berpotensi melibatkan integrasi teknik pembelajaran mendalam, seperti rangkaian neural konvolusional (CNN) dan rangkaian neural berulang (RNN), untuk menganalisis data kewangan yang kompleks bagi mengesan aktiviti penipuan. Model pembelajaran mendalam telah menunjukkan potensi dalam pelbagai domain dan boleh menawarkan keupayaan yang lebih baik dalam mengenal pasti corak rumit yang menunjukkan penipuan dalam transaksi kewangan.

Selain itu, kajian masa depan boleh memberi tumpuan kepada pembangunan model hibrid yang menggabungkan kelebihan pelbagai algoritma pembelajaran mesin. Sebagai contoh, teknik pembelajaran ensemble, seperti stacking atau boosting, boleh digunakan untuk memanfaatkan sifat saling melengkapkan pelbagai pengelas, dengan demikian meningkatkan prestasi ramalan keseluruhan sistem pengesanan penipuan.

Selanjutnya, pengintegrasian metodologi kecerdasan buatan yang boleh dijelaskan (XAI) dalam model pengesanan penipuan kewangan merupakan bidang kajian yang menarik untuk penyelidikan masa depan. Teknik-teknik XAI bertujuan untuk memberikan output yang jelas dan dapat diinterpretasikan dari model pembelajaran mesin, membolehkan pihak berkepentingan memahami rasional di sebalik ramalan penipuan. Ini boleh menjadi sangat berharga dalam sektor kewangan, di mana interpretabiliti dan akauntabiliti adalah penting.

Akhir sekali, penerokaan sistem pengesanan penipuan secara waktu sebenar yang boleh menyesuaikan diri dengan corak dan trend penipuan yang berkembang adalah bidang yang sesuai untuk kajian masa depan. Dengan memanfaatkan pemprosesan data secara langsung dan mekanisme pembelajaran berterusan, sistem-sistem ini dapat menyesuaikan strategi pengesanan penipuan mereka secara dinamik untuk berkesan melawan taktik penipuan yang muncul.

5.4 KESIMPULAN

Bab ini merangkumi keseluruhan kajian yang telah dilakukan, termasuklah rumusan kajian, penerangan sumbangan kajian, dan kajian di masa hadapan untuk meningkatkan proses pengesanan penipuan menggunakan pendekatan pembelajaran mesin. Kajian ini bertujuan untuk meneroka pemilihan ciri menggunakan tiga kaedah pemilihan ciri yang berbeza dan membandingkan pelbagai teknik klasifikasi untuk mengesan penipuan dalam urusan kewangan. Hasil kajian menunjukkan bahawa model pengelas XGBoost dan kaedah pembungkus-RFE menunjukkan prestasi yang paling unggul dalam pemilihan ciri dan pengesanan penipuan. Model XGBoost terutamanya menonjol dengan prestasi yang tinggi dalam mengenal pasti aktiviti penipuan dalam data kewangan. Walau bagaimanapun, penekanan perlu diberi bahawa pemilihan teknik yang sesuai bergantung kepada ciri-ciri dataset dan keperluan aplikasi. Kajian ini memberikan landasan untuk pembangunan sistem pengesanan penipuan yang kukuh dan efisien, serta menyumbang kepada kemajuan dalam metodologi pengesanan penipuan kewangan.

Untuk kajian masa hadapan, adalah penting untuk meneroka integrasi teknik pembelajaran mendalam seperti CNN dan RNN untuk menganalisis data kewangan dengan lebih baik. Pembangunan model hibrid yang menggabungkan pelbagai algoritma pembelajaran mesin juga merupakan bidang yang menarik untuk diterokai, bersama dengan penggunaan metodologi XAI untuk meningkatkan interpretabiliti model. Selain itu, penelitian terhadap sistem pengesanan penipuan secara waktu sebenar yang dapat menyesuaikan diri dengan corak penipuan yang berkembang juga merupakan bidang yang berpotensi untuk kajian masa hadapan. Dengan demikian, kajian ini menyediakan landasan yang kukuh untuk penyelidikan lanjutan dalam bidang pengesanan penipuan kewangan menggunakan pembelajaran mesin.

RUJUKAN

Ahmed, M., Choudhury, N. & Uddin, S. 2017. Anomaly detection on big data in financial markets. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* 998–1001. https://www.researchgate.net/publication/321068299_Anomaly_Detection_on_Big_Data_in_Financial_Markets [14 January 2024].

Ahmed, M., Mahmood, A.N. & Islam, M.R. 2016. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems 55*: 278–288. https://www.researchgate.net/publication/270968895_A_survey_of_anomaly_detection_techniques_in_financial_domain [15 January 2024].

Al Ali, A., Khedr, A.M., El-Bannany, M. & Kanakkayil, S. 2023. Citation: A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique https://doi.org/10.3390/app13042272.

Ala'raj, M., Abbod, M.F., Majdalawieh, M. & Jum'a, L. 2022. A deep learning model for behavioural credit scoring in banks. *Neural Computing and Applications 34*(8): 5839–5866. https://www.researchgate.net/publication/357834732_A_deep_learning_model_for_behavioural_credit_scoring_in_banks [15 January 2024].

Albashrawi, M. 2022. Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. *Journal of Data Science 14*(3): 553–570. https://jds-online.org/journal/JDS/article/341/text [23 July 2023].

Alonso Lopez-Rojas, E., Axelsson, S. & Elmir, A. 2016. PAYSIM: A FINANCIAL MOBILE MONEY SIMULATOR FOR FRAUD DETECTION. https://www.researchgate.net/publication/313138956.

Al-Yaseen, W.L., Idrees, A.K. & Almasoudy, F.H. 2022. Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system. *Pattern Recognition 132* https://www.researchgate.net/publication/362154386_Wrapper_Feature_Selection_Method_based_Differential_Evolution_and_Extreme_Learning_Machine_for_Intrusion_Detection_System [15 January 2024].

Ashfaq, T., Khalid, R., Yahaya, A.S., Aslam, S., Azar, A.T., Alsafari, S. & Hameed, I.A. 2022. A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism. *Sensors 2022, Vol. 22, Page 7162 22*(19): 7162. https://www.mdpi.com/1424-8220/22/19/7162/htm [23 July 2023].

Baesens, B., Höppner, S. & Verdonck, T. 2021. Data engineering for fraud detection. *Decision Support Systems 150* [14 January 2024].

Bao, Y., Ke, B., Li, B., Yu, Y.J. & Zhang, J. 2020. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research 58*(1): 199–235. https://www.researchgate.net/publication/337468608_Detecting_Accounting_Fraud_in_Publicly_Traded_US_Firms_Using_a_Machine_Learning_Approach [15 January 2024].

Beheshti, Z. 2022. BMPA-TVSinV: A Binary Marine Predators Algorithm using time-varying sinus and V-shaped transfer functions for wrapper-based feature selection. *Knowledge-Based Systems 252* [15 January 2024].

Bertomeu, J., Cheynel, E., Floyd, E. & Pan, W. 2021. Using machine learning to detect misstatements. *Review of Accounting Studies 26*(2): 468–519. https://www.researchgate.net/publication/346049867_Using_machine_learning_to_detect_misstatements [15 January 2024].

Bhattacharyya, S., Jha, S., Tharakunnel, K. & Westland, J.C. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems 50*(3): 602–613. [14 February 2024].

Chalapathy, R. & Chawla, S. 2019. Deep Learning for Anomaly Detection: A Survey DEEP LEARNING FOR ANOMALY DETECTION: A SURVEY A PREPRINT. https://www.researchgate.net/publication/330357393.

Chang, J.S. & Chang, W.H. 2012. A cost-effective method for early fraud detection in online auctions. *International Conference on ICT and Knowledge Engineering* 182–188. https://www.researchgate.net/publication/261281480_A_cost-effective_method_for_early_fraud_detection_in_online_auctions [14 January 2024].

Chang, W.H. & Chang, J.S. 2010. Using clustering techniques to analyze fraudulent behavior changes in online auctions. *ICNIT 2010 - 2010 International Conference on Networking and Information Technology* 34–38. https://www.researchgate.net/publication/251935925_Using_clustering_techniques_to_analyze_fraudulent_behavior_changes_in_online_auctions [14 January 2024].

Chaquet-Ulldemolins, J., Gimeno-Blanes, F.J., Moral-Rubio, S., Muñoz-Romero, S. & Rojo-álvarez, J.L. 2022. On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders. *Applied Sciences (Switzerland) 12*(8) [15 January 2024].

Chen, Yanyu, kumara, E.K. & Sivakumar, V. 2023. RETRACTED ARTICLE: Investigation of finance industry on risk awareness model and digital economic growth. *Annals of Operations Research 326*: 15. https://www.researchgate.net/publication/355771078_Investigation_of_fi

nance_industry_on_risk_awareness_model_and_digital_economic_growth [15 January 2024].

Chen, Yinhe. 2023. Financial Statement Fraud Detection based on Integrated Feature Selection and Imbalance Learning. *Frontiers in Business, Economics and Management 8*(3): 46–48. https://doi.org/10.54097/fbem.v8i3.7557 [31 August 2023].

Coppolino, L., D'Antonio, S., Formicola, V., Massei, C. & Romano, L. 2015. Use of the Dempster–Shafer theory to detect account takeovers in mobile money transfer services. *Journal of Ambient Intelligence and Humanized Computing 6*(6): 753–762. https://www.researchgate.net/publication/277645521_Use_of_the_Dempster-Shafer_theory_to_detect_account_takeovers_in_mobile_money_transfer_services [15 January 2024].

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. & Bontempi, G. 2015. Credit card fraud detection and concept-drift adaptation with delayed supervised information. *Proceedings of the International Joint Conference on Neural Networks 2015-September* [14 January 2024].

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. & Bontempi, G. 2018. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems 29*(8): 3784–3797.

Gai, K., Qiu, M. & Sun, X. 2018. A survey on FinTech. *Journal of Network and Computer Applications 103*: 262–273. https://www.researchgate.net/publication/320446285_A_survey_on_FinTech [14 January 2024].

Geng, X. & Yang, D. 2021. Intelligent Prediction Mathematical Model of Industrial Financial Fraud Based on Data Mining. *Mathematical Problems in Engineering 2021* [15 January 2024].

Ghazikhani, A., Yazdi, H.S. & Monsefi, R. 2012. Class imbalance handling using wrapper-based random oversampling. *ICEE 2012 - 20th Iranian Conference on Electrical Engineering* 611–616. [3 September 2023].

Glancy, F.H. & Yadav, S.B. 2011. A computational model for financial reporting fraud detection. *Decision Support Systems 50*(3): 595–601. [14 January 2024].

Grove, H. & Basilico, E. 2008. Fraudulent Financial Reporting Detection: Key Ratios Plus Corporate Governance Factors. *International Studies of Management & Organization 38*(3): 10–42. https://www.researchgate.net/publication/240312194_Fraudulent_Financial_Reporting_Detection_Key_Ratios_Plus_Corporate_Governance_Factors [15 January 2024].

Gu, Q., Zhu, L. & Cai, Z. 2009. Evaluation measures of the classification performance of imbalanced data sets. *Communications in Computer and Information Science 51*: 461–471. https://www.researchgate.net/publication/226823368_Evaluation_Measures_of_the_Classification_Performance_of_Imbalanced_Data_Sets [15 January 2024].

Hajek, P. & Henriques, R. 2017. Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems 128*: 139–152. https://www.researchgate.net/publication/316770005_Mining_Corporate_Annual_Reports_for_Intelligent_Detection_of_Financial_Statement_Fraud_-_A_Comparative_Study_of_Machine_Learning_Methods [15 January 2024].

Hajek, P., Mohammad, ·, Abedin, Z. & Sivarajah, · Uthayasankar. 2023. Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework *25*: 1985–2003. https://doi.org/10.1007/s10796-022-10346-6 [13 January 2024].

Hamal, S. & Senvar, O. 2021. Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *International Journal of Computational Intelligence Systems 14*(1): 769–782. https://www.atlantis-press.com/journals/ijcis/125952602 [4 September 2023].

Hu, Y., Xu, L., Huang, P., Luo, X., Wang, P. & Kang, Z. 2021. Reliable Identification of Oolong Tea Species: Nondestructive Testing Classification Based on Fluorescence Hyperspectral Technology and Machine Learning. *Agriculture 2021, Vol. 11, Page 1106 11*(11): 1106. https://www.mdpi.com/2077-0472/11/11/1106/htm [15 January 2024].

Huang, D., Mu, D., Yang, L. & Cai, X. 2018. CoDetect: Financial Fraud Detection with Anomaly Feature Detection. *IEEE Access 6*: 19161–19174. [14 January 2024].

Huang, S.Y., Lin, C.C., Chiu, A.A. & Yen, D.C. 2017. Fraud detection using fraud triangle risk factors. *Information Systems Frontiers 19*(6): 1343–1356. https://www.researchgate.net/publication/301312579_Fraud_detection_using_fraud_triangle_risk_factors [15 January 2024].

Ikeda, C., Ouazzane, K. & Yu, Q. 2020. A NEW FRAMEWORK OF FEATURE ENGINEERING FOR MACHINE LEARNING IN FINANCIAL FRAUD DETECTION 205–220. [31 August 2023].

Kanapickienė, R. & Grundienė, Ž. 2015. The Model of Fraud Detection in Financial Statements by Means of Financial Ratios. *Procedia - Social and Behavioral Sciences 213*: 321–327. https://www.researchgate.net/publication/286542581_The_Model_of_Fra

ud_Detection_in_Financial_Statements_by_Means_of_Financial_Ratios [15 January 2024].

Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology 64*(5): 402. /pmc/articles/PMC3668100/ [15 January 2024].

Kang, J. 2018. Mobile payment in Fintech environment: trends, security challenges, and services. *Human-centric Computing and Information Sciences* *8*(1) https://www.researchgate.net/publication/328612693_Mobile_payment_in_Fintech_environment_trends_security_challenges_and_services [15 January 2024].

Khedr, A.M., El Bannany, M., Kanakkayil, S. & Bannany, M. El. 2021. PREPRINT An Ensemble Model for Financial Statement Fraud Detection An Ensemble Model for Financial Statement Fraud Detection https://doi.org/10.3897/arphapreprints.e69590.

Kolli, C.S. & Tatavarthi, U.D. 2020. Fraud detection in bank transaction with wrapper model and Harris water optimization-based deep recurrent neural network. *Kybernetes 50*(6): 1731–1750. [31 August 2023].

La, H.J. & Kim, S.D. 2018. A machine learning framework for adaptive FinTech security provisioning. *Journal of Internet Technology 19*(5): 1545–1553. https://www.researchgate.net/publication/328627152_A_machine_learning_framework_for_adaptive_FinTech_security_provisioning [14 January 2024].

Lam, A.Y., Geng, Y., Chen, Y. & Wu, Z. 2022. Citation: Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach https://doi.org/10.3390/su15010105 [13 January 2024].

Le Khac, N.A. & Kechadi, M.T. 2010. Application of data mining for anti-money laundering detection: A case study. *Proceedings - IEEE International Conference on Data Mining, ICDM* 577–584. https://www.researchgate.net/publication/220765959_Application_of_Data_Mining_for_Anti-money_Laundering_Detection_A_Case_Study [14 January 2024].

Lei, S., Xu, K., Huang, Y. & Sha, X. 2020. An Xgboost based system for financial fraud detection. *E3S Web of Conferences 214*: 02042. https://www.e3s-conferences.org/articles/e3sconf/abs/2020/74/e3sconf_ebldm2020_02042/e3sconf_ebldm2020_02042.html [13 January 2024].

Leite, R.A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E. & Kuntner, J. 2018. EVA: Visual Analytics to Identify Fraudulent Events. *IEEE Transactions on Visualization and Computer Graphics 24*(1): 330–339. [14 January 2024].

Li, H. & Wong, M.L. 2015. Financial Fraud Detection by using Grammar-based Multi-objective Genetic Programming with ensemble learning. *2015 IEEE Congress on Evolutionary Computation, CEC 2015 - Proceedings* 1113–1120. https://www.researchgate.net/publication/277562720_Financial_Fraud_Detection_by_using_Grammar-based_Multi-objective_Genetic_Programming_with_ensemble_learning [15 January 2024].

Li, Q. & Clark, G. 2013. Mobile security: A look ahead. *IEEE Security and Privacy 11*(1): 78–81. [15 January 2024].

Liu, Z., Ye, R. & Ye, R. 2021. Detecting Financial Statement Fraud with Interpretable Machine Learning https://www.researchsquare.com [13 January 2024].

Long, W., Lu, Z. & Cui, L. 2019. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems 164*: 163–173. [14 January 2024].

Lopez-Rojas, Edgar A. & Barneaud, C. 2019. Advantages of the PaySim Simulator for Improving Financial Fraud Controls. *Advances in Intelligent Systems and Computing 998*: 727–736. https://www.researchgate.net/publication/334301755_Advantages_of_the_PaySim_Simulator_for_Improving_Financial_Fraud_Controls [15 January 2024].

Lopez-Rojas, Edgar Alonso, Axelsson, S. & Baca, D. 2018. Analysis of fraud controls using the PaySim financial simulator. *International Journal of Simulation and Process Modelling 13*(4): 377–386. [15 January 2024].

Ma, T., Qian, S., Cao, J., Xue, G., Yu, J., Zhu, Y. & Li, M. 2019. An unsupervised incremental virtual learning method for financial fraud detection. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2019-November* https://www.researchgate.net/publication/339980210_An_Unsupervised_Incremental_Virtual_Learning_Method_for_Financial_Fraud_Detection [14 January 2024].

Magomedov, S., Pavelyev, S., Ivanova, I., Dobrotvorsky, A., Khrestina, M. & Yusubaliev, T. 2018. Anomaly detection with machine learning and graph databases in fraud management. *International Journal of Advanced Computer Science and Applications 9*(11): 33–38. [14 January 2024].

Nalluri, S.P. & Kurra, R.R. 2021. UNSUPERVISED FEATURE SELECTION FOR TEXT CLUSTERING USING DIFFERENTIAL INVERSE DOCUMENT FREQUENCY. *Indian Journal of Computer Science and Engineering 12*(4): 790–797. [17 February 2024].

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y. & Sun, X. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems 50*(3): 559–569. https://www.researchgate.net/publication/222578477_The_application_of_data_mining_techniques_in_financial_fraud_detection_A_classification_framework_and_an_academic_review_of_literature [15 January 2024].

Onwubiko, C. 2020. Fraud matrix: A morphological and analysis-based classification and taxonomy of fraud. *Computers and Security 96* https://www.researchgate.net/publication/341772309_Fraud_matrix_A_morphological_and_analysis-based_classification_and_taxonomy_of_fraud [15 January 2024].

Pourhabibi, T., Ong, K.L., Kam, B.H. & Boo, Y.L. 2020. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems 133* https://www.researchgate.net/publication/340691343_Fraud_detection_A_systematic_literature_review_of_graph-based_anomaly_detection_approaches [14 January 2024].

Prasetiyowati, M.I., Maulidevi, N.U. & Surendro, K. 2022. The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy. *PeerJ Computer Science 8*: e1041. https://peerj.com/articles/cs-1041 [15 January 2024].

Rieke, R., Zhdanova, M., Repp, J., Giot, R. & Gaber, C. 2013. Fraud detection in mobile payments utilizing process behavior analysis. *Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013* 662–669. [15 January 2024].

Rube, D. & Wirgen Isak. 2021. Supervised fraud detection of mobile money transactions on different distributions of imbalanced data [15 January 2024].

Sharma, N. & Chalapathi, V. 2022. A Novel Machine Learning Technique for Fraud Detection on Credit Card Financial Data. *International Journal of Engineering Technology and Management Sciences Website: ijetms.in Issue 4*(6) [3 September 2023].

Somasundaram, A. & Reddy, S. 2019. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Computing and Applications 31*: 3–14. https://www.researchgate.net/publication/326443739_Parallel_and_incremental_credit_card_fraud_detection_model_to_handle_concept_drift_and_data_imbalance [14 January 2024].

Song, X.P., Hu, Z.H., Du, J.G. & Sheng, Z.H. 2014. Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud:

Evidence from China. *Journal of Forecasting 33*(8): 611–626. https://onlinelibrary.wiley.com/doi/full/10.1002/for.2294 [23 July 2023].

Stojanovi´cstojanovi´c, B., Boži´cboži´c, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E., Runevic, J. & Bernal Bernabe, J. 2021. Follow the Trail: Machine Learning for Fraud Detection in Fintech Applications https://doi.org/10.3390/s21051594 [3 January 2024].

Torgo, L. & Lopes, E. 2011. Utility-based fraud detection. *IJCAI International Joint Conference on Artificial Intelligence* 1517–1522. https://www.researchgate.net/publication/220816231_Utility-Based_Fraud_Detection [14 January 2024].

Webga, K. & Lu, A. 2015. Discovery of rating fraud with real-time streaming visual analytics. *2015 IEEE Symposium on Visualization for Cyber Security, VizSec 2015* https://www.researchgate.net/publication/308302574_Discovery_of_rating_fraud_with_real-time_streaming_visual_analytics [14 January 2024].

Wedge, R., Max Kanter, J., Veeramachaneni, K., Moral Rubio, S. & Iglesias Perez, S. 2019. Solving the "false positives" problem in fraud prediction https://www.researchgate.net/publication/320582472_Solving_the_false_positives_problem_in_fraud_prediction [14 January 2024].

West, J. & Bhattacharya, M. 2016. Some Experimental Issues in Financial Fraud Mining [13 January 2024].

West, J., Bhattacharya, M. & Islam, R. (t.th.). Intelligent Financial Fraud Detection Practices: An Investigation [8 February 2024].

What Is Fraud? Definition, Types, and Consequences. (t.th.). https://www.investopedia.com/terms/f/fraud.asp [14 January 2024].

Whiting, D.G., Hansen, J. V., McDonald, J.B., Albrecht, C. & Albrecht, W.S. 2012. Machine learning methods for detecting patterns of management fraud. *Computational Intelligence 28*(4): 505–527. [15 January 2024].

Xiuguo, W. & Shengyong, D. 2022. An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. *IEEE Access 10*: 22516–22532. [31 August 2023].

Xu, H., Fan, G. & Song, Y. 2022. Novel Key Indicators Selection Method of Financial Fraud Prediction Model Based on Machine Learning Hybrid Mode. *Mobile Information Systems 2022* [3 September 2023].

Xu, J.J., Lu, Y. & Chau, M. 2015. LNCS 9074 - P2P Lending Fraud Detection: A Big Data Approach. *LNCS 9074*: 71–81. [14 January 2024].

Yang, Y., Fan, C.J., Chen, L. & Xiong, H.L. 2022. IPMOD: An efficient outlier detection model for high-dimensional medical data streams. *Expert Systems*

with Applications 191 https://www.researchgate.net/publication/356654568_IPMOD_An_efficient_outlier_detection_model_for_high-dimensional_medical_data_streams [15 January 2024].

Yao, J., Pan, Y., Yang, S., Chen, Y. & Li, Y. 2019. Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach. *Sustainability 2019, Vol. 11, Page 1579 11*(6): 1579. https://www.mdpi.com/2071-1050/11/6/1579/htm [13 January 2024].

Yaram, S. 2017. Machine learning algorithms for document clustering and fraud detection. *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016* [14 January 2024].

Zhang, Xiaoming, Yu, L., Yin, H. & Lai, K.K. 2022. Integrating data augmentation and hybrid feature selection for small sample credit risk assessment with high dimensionality. *Computers and Operations Research 146* https://www.researchgate.net/publication/361713971_Integrating_data_augmentation_and_hybrid_feature_selection_for_small_sample_credit_risk_assessment_with_high_dimensionality [15 January 2024].

Zhang, Xinwei, Han, Y., Xu, W. & Wang, Q. 2021. HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences 557*: 302–316. [15 January 2024].

Zhu, Z., Ong, Y.S. & Dash, M. 2007. Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37*(1): 70–76. [17 February 2024].