MODEL PEMBELAJARAN MENDALAM UNTUK PENGESANAN PENGELAKAN CUKAI TIDAK LANGSUNG DI MALAYSIA

NORHASLIZA BINTI HASHIM

UNIVERSITI KEBANGSAAN MALAYSIA

MODEL PEMBELAJARAN MENDALAM UNTUK PENGESANAN PENGELAKAN CUKAI TIDAK LANGSUNG DI MALAYSIA

NORHASLIZA BINTI HASHIM

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2022

PENGAKUAN

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

19 Mei 2022

NORHASLIZA BINTI HASHIM
P106801

PENGHARGAAN

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

Setinggi-tinggi penghargaan dan ucapan terima kasih kepada penyelia utama saya, Profesor Madya Dr. Shahnorbanun Sahran di atas segala panduan, bimbingan dan motivasi yang sangat diperlukan sepanjang usaha dalam menyiapkan kajian ini. Tidak lupa juga ucapan terima kasih kepada semua tenaga pengajar di Fakulti Teknologi dan Sains Maklumat (FTSM) atas segala ilmu dan tunjuk ajar yang diberikan.

Penghargaan ini juga saya tujukan kepada pengurusan tertinggi dan rakan sejawatan di Jabatan Kastam Diraja Malaysia (JKDM) yang sentiasa menyokong usaha saya mempertingkatkan ilmu pengetahuan untuk memberikan perkhidmatan yang lebih cemerlang. Ucapan terima kasih juga saya panjangkan kepada pihak Jabatan Perkhidmatan Awam (JPA) yang menaja pengajian dan perbelanjaan saya di peringkat Sarjana di UKM.

Akhir sekali kepada ibu bapa, keluarga, Muhamad Faiz, Syifa Ajwa dan Ajwa Anisa, terima kasih tidak terhingga di atas segala sokongan dan dorongan, serta tolak ansur yang diberikan bagi memudahkan segala urusan pengajian saya selama ini.

Semoga Allah merahmati kita semua.

ABSTRAK

Pengesanan pengelakan cukai yang praktikal dan komprehensif boleh membantu mengurangkan kehilangan hasil cukai negara. Kaedah pengesanan pengelakan cukai sedia ada yang menggunakan penilaian berasaskan peraturan, analisis statistik dan verifikasi pegawai cukai tidak lagi berkesan dengan peningkatan saiz data dan jumlah transaksi harian yang tinggi. Pembelajaran mendalam adalah alternatif yang boleh diterokai untuk pengesanan pengelakan cukai yang melibatkan set data yang besar dan kompleks. Kajian ini menggunakan set data sebenar pendaftaran pembayar cukai dan penyata cukai untuk cukai tidak langsung di Malaysia, iaitu Cukai Barangan dan Perkhidmatan (GST). Kajian ini memberi tumpuan kepada pembangunan dan pemilihan model pembelajaran mendalam yang secara automatik boleh meramalkan penyata cukai sebagai berisiko rendah dan berisiko tinggi terhadap pengelakan cukai. Bagi tujuan ini, Rangkaian Neural Pelingkaran (CNN) dan Memori Jangka Masa Panjang dan Pendek (LSTM) dibangunkan dan diuji untuk memilih model terbaik bagi pengesanan pengelakan cukai berdasarkan penyata cukai. Pengujian dijalankan pada set data penyata cukai dan empat (4) set data berkategori berdasarkan kekerapan pemfailan daripada set data penyata cukai. Prestasi bagi setiap model dibandingkan berdasarkan ketepatan, kejituan, dapatan semula, dan skor F1, serta ujian statistik untuk hipotesis kajian. Model CNN memberikan prestasi ketepatan tertinggi untuk set data penyata cukai dan set data berkategori dwibulanan dan pelbagai dengan ketepatan 0.793, 0.945 dan 0.914. Sementara itu, model LSTM memberikan ketepatan prestasi yang terbaik pada set data berkategori bulanan dan suku tahun, dengan skor ketepatan 0.794 dan 0.711. Ujian hipotesis menolak hipotesis nol dan menunjukkan terdapat impak yang signifikan terhadap model pembelajaran mendalam dalam mengesan pengelakan cukai tidak langsung di Malaysia. Secara keseluruhannya, model CNN adalah model terbaik untuk mengesan pengelakan cukai bagi set data dalam kajian ini. Keputusan pengelasan yang dijana oleh model pembelajaran mendalam CNN dinilai dan dipersetujui oleh pakar cukai di Jabatan Kastam Diraja Malaysia (JKDM).

DEEP LEARNING MODEL FOR THE DETECTION OF INDIRECT TAX EVASION IN MALAYSIA

ABSTRACT

Practical and comprehensive tax evasion detection can help reduce national tax revenue loss. Existing fraud detection methods using rule-based assessment, statistical analysis, and tax officer verification are no longer effective with the increasing data sizes and daily transaction volumes. Deep learning is an explorable alternative to tax evasion detection concerning large and complex datasets. This study uses the actual datasets of taxpayer registration and tax returns for indirect tax in Malaysia, the Goods, and Services Tax (GST). This study focuses on developing and determining deep learning models that can automatically predict tax returns as low risk and high risk to the tax evasion. For this purpose, the Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) are developed and experimented with to select the best model for detecting tax evasion based on tax returns statements. The experiments are conducted on the tax returns dataset and four (4) categorized datasets based on filing frequency from the tax returns dataset. The performance for each model is compared based on accuracy, precision, recall, and F1 score, as well as statistical tests for the study hypotheses. CNN model has the highest accuracy for tax returns dataset and categorized datasets of bimonthly and varied with the accuracy of 0.793, 0.945, and 0.914. Meanwhile the LSTM model gives better performance on categorized datasets of monthly and quarterly, with an accuracy score of 0.794 and 0.711. Hypothesis testing rejected the null hypothesis and showed a significant effect on the deep learning model detecting indirect tax evasion in Malaysia. Overall, the CNN model is the best model for detecting tax evasion for the datasets in the study. The classification results generated by CNN deep learning model are evaluated and approved by the tax experts at Royal Malaysian Customs Department (RMCD).

KANDUNGAN

Halaman

PENGAKUAN		ii
PENGHARGAAN		iii
ABSTRAK		iv
ABSTRACT		v
KANDUNGAN		vi
SENARAI JADUAL		ix
SENARAI ILUSTRASI		x
SENARAI SINGKATAN		xiii
SENARAI ISTILAH		xiv
BAB I	**PENGENALAN**	
1.1	Pendahuluan	1
1.2	Latar Belakang Kajian	2
1.3	Penyataan Masalah	5
1.4	Hipotesis Kajian	6
1.5	Persoalan Kajian	6
1.6	Objektif Kajian	7
1.7	Skop Kajian	7
1.8	Metodologi Kajian	8
1.9	Kepentingan Kajian	9
1.10	Organisasi Tesis	10
BAB II	**KAJIAN LITERATUR**	
2.1	Pengenalan	11
2.2	Pengelakan cukai	11
2.3	Pengelakan cukai di malaysia	13
2.4	Sistem Penilaian Risiko Fraud Cukai	14
2.5	Model Pembelajaran Mesin Untuk Pengesanan Fraud	16
2.6	Penerokaan Model Pembelajaran Mendalam	18
	2.6.1 Rangkaian Neural Pelingkaran (CNN)	23

	2.6.2 Memori Jangka Masa Panjang dan Pendek (LSTM)	25
BAB III	**METODOLOGI**	
3.1	Pengenalan	28
3.2	Pendekatan Kajian	28
	3.2.1 Kajian Data Sains Melalui Pendekatan CRIPS-DM	28
	3.2.2 Alatan Kajian	30
3.3	Kes Bisnes Dan Penerokaan Data	31
3.4	Penyediaan Data	31
	3.4.1 Perolehan Data	32
	3.4.2 Integrasi Data	34
	3.4.3 Pembersihan Data	34
	3.4.4 Padanan dan tapisan data	35
	3.4.5 Pengelasan Data	36
	3.4.6 Analisis Deskriptif Bagi Set Data Penyata Cukai	37
	3.4.7 Set Data Berkategori	42
	3.4.8 Data Tanda Aras (*Benchmark Data*)	48
3.5	Pembangunan model	50
	3.5.1 Set Latihan, Pengesahan dan Ujian	51
	3.5.2 Pengendalian Data Tidak Seimbang	51
	3.5.3 Rangkaian Neural Pelingkaran (CNN)	52
	3.5.4 Memori Jangka Masa Panjang dan Pendek (LSTM)	52
3.6	Penilaian Model	52
	3.6.1 Pemilihan Model Terbaik untuk Pengesanan Pengelakan Cukai	53
	3.6.2 Penilaian Pakar	54
3.7	Pengaturan Model	54
3.8	Kesimpulan	55
BAB IV	**DAPATAN KAJIAN DAN ANALISIS**	
4.1	Pengenalan	56
4.2	Trend Dan Hasil Dapatan Deskriptif	56
	4.2.1 Visualisasi Data Melalui Pengurangan Dimensi	56
	4.2.2 Taburan Data Nominal	58
	4.2.3 Taburan Data Numerik	60
4.3	Hasil Dapatan Pengujian Model Pengesanan Pengelakan Cukai	63

	4.3.1	Model Pengelasan CNN	63
	4.3.2	Model Pengelasan LSTM	64
	4.3.3	Pengujian Bagi Set Data Tanda Aras dan Set Data Penyata Cukai	65
	4.3.4	Pemilihan Model Pembelajaran Mendalam Terbaik	69
	4.3.5	Perbandingan Dua Model Pembelajaran Mendalam	74
	4.3.6	Pengujian Statistik untuk Hipotesis Kajian	75
4.4	Penilaian Pakar		77
4.5	Kesimpulan		79
BAB V	**KESIMPULAN DAN CADANGAN**		
5.1	Pengenalan		80
5.2	Perbincangan Kajian		80
5.3	Sumbangan Kajian		82
5.4	Kajian Masa Hadapan		83
	5.4.1	Teknik Pengimbangan Data	83
	5.4.2	Penalaan parameter (*parameter tuning*)	83
RUJUKAN			84
LAMPIRAN			
Lampiran A	KELULUSAN MENDAPATKAN DATA		89
Lampiran B	ARKITEKTUR MODEL CNN DAN LSTM		91
Lampiran C	CONTOH BORANG PENILAIAN PAKAR		92
Lampiran D	LATAR BELAKANG PAKAR		96

SENARAI JADUAL

No. Jadual		Halaman
Jadual 2.1	Corak pengelakan cukai dan tipologi fraud GST (Othman et al. 2019)	13
Jadual 2.2	Rumusan kajian pembelajaran mendalam untuk pengesanan pengelakan cukai	21
Jadual 3.1	Fasa kajian CRIPS-DM	29
Jadual 3.2	Perisian dan pengaturcaraan yang digunakan untuk kajian ini	30
Jadual 3.3	Atribut dan diskripsi medan data	33
Jadual 3.4	Jenis data dalam set data penyata cukai	36
Jadual 3.5	Medan bagi set data penyata cukai	37
Jadual 3.6	Jumlah data MyGST mengikut tahun bercukai	38
Jadual 3.7	Set data berkategori mengikut kekerapan pemfailan GST	42
Jadual 3.8	Peratusan label kelas untuk set data	51
Jadual 4.1	Model dan lapisan CNN	63
Jadual 4.2	Model dan lapisan LSTM	64
Jadual 4.3	Keputusan pengujian menggunakan set data tanda aras	65
Jadual 4.4	Keputusan prestasi ketepatan CNN dan LSTM berdasarkan bilangan epoc	66
Jadual 4.5	Keputusan pengujian set data berkategori	67
Jadual 4.6	Keputusan pengujian menggunakan set data penyata cukai cukai	69
Jadual 4.7	Pencapaian ketepatan model pembelajaran mendalam	70
Jadual 4.8	Pencapaian kejituan model pembelajaran mendalam	71
Jadual 4.9	Pencapaian dapatan semula model pembelajaran mendalam	72
Jadual 4.10	Pencapaian skor F1 model pembelajaran mendalam	73
Jadual 4.11	Keputusan pengujian statistik untuk perbandingan model CNN dan LSTM	75
Jadual 4.12	Keputusan pengujian statistik untuk hipotesis kajian	76

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 1.1	GST sebagai cukai tidak langsung	3
Rajah 1.1	Proses CRIPS-DM dalam kajian data sains	8
Rajah 1.1	Perbandingan model sedia ada dan model pembelajaran mendalam	9
Rajah 2.1	Contoh penjualan dan pembelian dalam pedagangan pelingkaran	13
Rajah 2.2	Prestasi pembelajaran mendalam berbanding pembelajaran mesin (Alom et al. 2019)	19
Rajah 2.3	Struktur lapisan CNN untuk pengelasan fraud cukai	24
Rajah 2.4	Struktur sel memori LSTM	26
Rajah 3.1	Intepretasi fasa kajian	29
Rajah 3.2	Penjanaan set data mentah bagi penyata GST-03 melalui sistem MyGST	32
Rajah 3.3	Contoh set data mentah daripada borang GST-03	32
Rajah 3.4	Proses integrasi data menggunakan pengaturcaraan python	34
Rajah 3.5	Proses pembersihan data menggunakan pengaturcaraan python	35
Rajah 3.6	Proses integrasi data menggunakan pengaturcaraan python	35
Rajah 3.7	Proses pelabelan data kepada 0 – risiko rendah, 1 – risiko tinggi	36
Rajah 3.8	Jumlah data MyGST mengikut tahun bercukai	38
Rajah 3.9	Pecahan atribut ‘Class’ berdasarkan label kelas 0,1	39
Rajah 3.10	Taburan rekod penyata GST mengikut tempoh penyata cukai	39
Rajah 3.11	Jumlah penyata cukai bagi set data penyata cukai mengikut stesen mengawal	40
Rajah 3.12	Taburan rekod penyata GST mengikut tempoh penyata cukai	40
Rajah 3.13	Matriks korelasi antara atribut dan label kelas bagi set data penyata cukai	41
Rajah 3.14	Jumlah penyata cukai mengikut kategori pemfailan	43

Rajah 3.15	Peratusan label kelas berdasarkan set data penyata cukai dan set data berkategori	43
Rajah 3.16	Matriks korelasi antara atribut dan label kelas bagi set data bulanan	45
Rajah 3.17	Matriks korelasi antara atribut dan label kelas bagi set data suku tahun	46
Rajah 3.18	Matriks korelasi antara atribut dan label kelas bagi set data dwibulanan	47
Rajah 3.19	Matriks korelasi antara atribut dan label kelas bagi set data pelbagai	48
Rajah 3.20	Peratusan label kelas bagi set data tanda aras	49
Rajah 3.21	Aliran Proses Eksperiment Pemodelan Pembelajaran Mendalam	50
Rajah 3.22	Matriks ralat bagi pengelasan kelas positif (risiko tinggi) dan kelas negatif (risiko rendah)	53
Rajah 4.1	Pengelompokan menggunakan pengurangan dimensi	57
Rajah 4.2	Plot taburan untuk atribut “V18”	58
Rajah 4.3	Plot taburan untuk atribut “V19”	59
Rajah 4.4	Plot taburan untuk atribut “V20”	59
Rajah 4.5	Plot taburan atribut “V21”	59
Rajah 4.6	Plot taburan atribut “V3”	60
Rajah 4.7	Plot taburan atribut “V4”	60
Rajah 4.8	Plot taburan atribut “V5”	61
Rajah 4.9	Plot taburan atribut “V6”	61
Rajah 4.10	Plot taburan atribut “V7”	61
Rajah 4.11	Plot taburan atribut “V8”	62
Rajah 4.12	Plot taburan atribut “V22”	62
Rajah 4.13	Prestasi model CNN dan LSTM untuk set data tanda aras	65
Rajah 4.14	Peningkatan prestasi ketepatan dengan peningkatan kadar *epocs*	67
Rajah 4.15	Prestasi model CNN dan LSTM untuk set data berkategori	68

Rajah 4.16	Prestasi model CNN dan LSTM untuk set data penyata cukai	69
Rajah 4.17	Prestasi ketepatan model ke atas peratusan peratusan label kelas	70
Rajah 4.18	Prestasi kejituan model ke atas peratusan peratusan label kelas	71
Rajah 4.19	Prestasi dapatan semula ke atas peratusan peratusan label kelas	72
Rajah 4.20	Prestasi skor F1 ke atas peratusan label kelas	73
Rajah 4.21	Jadual Kontigensi 2 x 2 Ujian Statistik McNemar	74
Rajah 4.22	Penilaian pakar terhadap kualiti set data yang digunakan	77
Rajah 4.23	Penilaian pakar terhadap kelas pengelasan yang dijana	78

SENARAI SINGKATAN

ANN	Rangkaian Neural Buatan, *Artifical Neural Network*
CNN	Rangkaian Neural Perlingkaran, *Convolutional Neural Network*
GST	Cukai Barangan dan Perkhidmatan, *Goods and Services Tax*
JKDM	Jabatan Kastam Diraja Malaysia
LSTM	Memori Jangka masa Panjang dan Pendek, *Long Short Term Memory*
AUC	Kawasan di bawah lengkung, *Area Under Curve*
CRIPS-DM	*Cross-Industry Standard Process for Data Mining*
DBM	*Deep Boltzmann Machine*
DNN	Rangkaian Neural Dalaman, *Deep Neural Network*
H1	Hipotesis alternatif
Hθ	Hipotesis nol
MLP	Lapisan Perceptron Berbilang Baris, *Multilayer Perceptron Layer*
RBM	*Restricted Boltzmann Machine*
RMCD	*Royal Malaysian Customs Department*
ROC	*Receiver Operating Curve*
SME	Perusahaan kecil sederhana, *Small Medium Enteprise*
SMOTE	*Synthetic Minority Oversampling Technique*
SVM	*Support Vector Machine*
VAT	*Value Added Tax*

SENARAI ISTILAH

Dapatan semula	*Recall*
Data tanda aras	*Benchmark data*
Epoc	*Epoch*
Jeda keyakinan	*Confidence interval*
Jeda keyakinan	*Confidence interval*
Kejituan	*Precision*
Ketepatan	*Accuracy*
Khi kuasa dua	*Chi-squared*
Lapisan tersembunyi	*Hidden layers*
Matriks korelasi	*Correlation matrix*
Matriks ralat	*Confusion matrix*
Nilai-p	*p-value*
Pagar input	*Input gate*
Pagar lupa	*Forget gate*
Pagar output	*Output gate*
Paras keertian	*Significance level*
Penalaan parameter	*Parameter tuning*
Skor F1	*Score F1*

BAB I

PENGENALAN

1.1 PENDAHULUAN

Jabatan Kastam Diraja Malaysia (JKDM) adalah sebuah agensi di bawah Kementerian Kewangan yang bertanggungjawab memungut cukai tidak langsung seperti duti eksport, duti import, duti eksais, Cukai Barang dan Perkhidmatan (*Goods and Services Tax*, GST), Cukai Jualan dan Cukai Perkhidmatan (SST), Cukai Perkhidmatan mengenai Perkhidmatan Digital (SToDS), Cukai Pelancongan (TTx), levi dan lain-lain hasil bukan cukai. Sebagai pemungut hasil kedua tertinggi di Malaysia, JKDM juga bertanggungjawab melaksanakan penguatkuasaan sempadan dan kesalahan narkotik.

Secara ringkasnya tiga peranan utama JKDM adalah;

1. Memungut hasil negara melalui Duti Import, Duti Eksport, Duti Eksais, Cukai Jualan, Cukai Perkhidmatan, Levi Keuntungan Luar Biasa, Levi Pelepasan, Levi Kenderaan, hasil bukan cukai, hasil negeri dan wang amanah dan Cukai Pelancongan;

2. Memberi fasiliti kepada sektor perdagangan dan perindustrian menerusi fasilitasi perkastaman selari dengan dasar semasa kerajaan, kemudahan pengecualian cukai kepada pengimport, pengecualian cukai terhadap bahan mentah dan mesin pengilang, kemudahan bayaran balik duti dan cukai, dan tuntutan pulang balik duti dan cukai, serta kemudahan pelepasan dagangan import, eksport dan pelepasan penumpang; dan

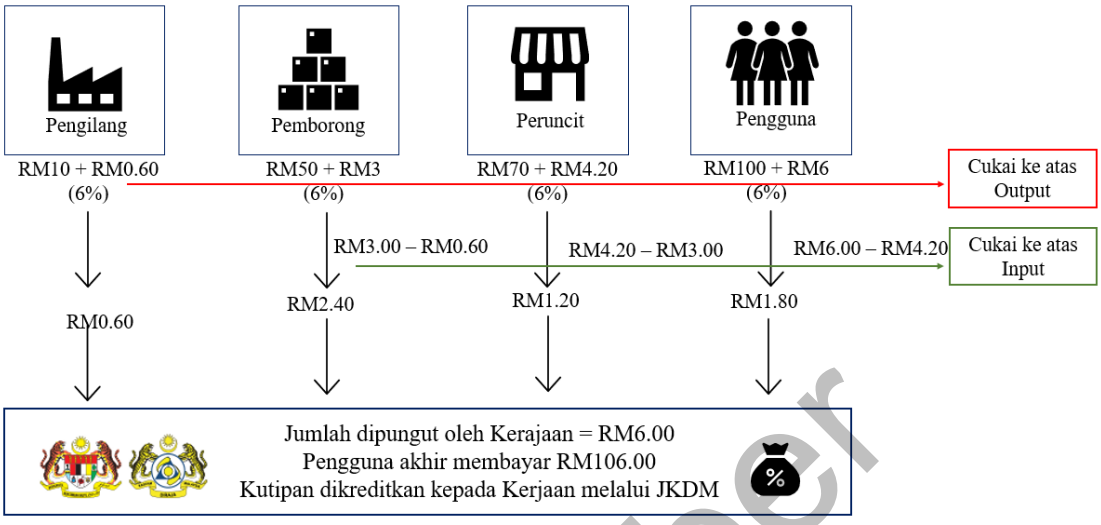
3. Menguatkuasakan undang-undang melalui penguatkuasaan akta-akta dan perundangan subsidiari.

1.2 LATAR BELAKANG KAJIAN

Malaysia dan negara lain di dunia memungut cukai langsung dan cukai tidak langsung sebagai salah satu daripada sumber untuk menjana ekonomi negara. Cukai tidak langsung dikenakan ke atas barangan dan perkhidmatan yang digunakan dan dikenakan oleh perniagaan atau individu bagi pihak kerajaan dan diselia oleh pentadbir cukai iaitu JKDM di Malaysia. JKDM berjaya menyumbangkan hasil kepada negara sebanyak RM41.454 bilion pada tahun 2018, RM42 bilion untuk tahun 2019, RM38.66 bilion pada tahun 2020 dan hasil kutipan terkini berjumlah RM42.5 bilion untuk tahun 2021. Manakala, cukai langsung adalah cukai yang dikenakan ke atas individu dan syarikat; sebagai contoh cukai pendapatan.

Kajian ini memfokuskan kepada cukai tidak langsung GST. GST diperkenalkan di Malaysia pada 1 April 2015 sebagai cukai kepenggunaan baharu dan juga dikenali secara meluas sebagai Cukai Nilai Tambah (*Value Added Tax,* VAT) di negara lain bagi menggantikan Cukai Jualan di bawah Akta Cukai Jualan 1972 dan Cukai Perkhidmatan di bawah Akta Cukai Perkhidmatan 1975. GST adalah cukai kepenggunaan berasaskan konsep nilai tambah secara berperingkat berbeza dengan Cukai Jualan dan Cukai Perkhidmatan yang merupakan cukai seperingkat.

Pungutan cukai dibuat di setiap peringkat oleh perantara dalam proses pengeluaran dan pengedaran. Cukai itu sendiri bukan merupakan suatu kos kepada perantara kerana mereka boleh menuntut kembali GST yang telah dikenakan ke atas input perniagaan mereka. GST dikenakan ke atas barang dan perkhidmatan di setiap peringkat pengeluaran dan pengedaran dalam rantaian pembekalan termasuk pengimportan barang dan perkhidmatan. Rajah 1.1 menunjukkan konsep percukaian GST sebagai cukai berperingkat.



Rajah 1.1 GST sebagai cukai tidak langsung

Kos dan perolehan dalam menjalankan perniagaan untuk syarikat juga dapat dikurangkan kerana GST yang dikenakan ke atas input perniagaan mereka dapat dituntut semula semasa mengemukakan penyata GST. Walaupun GST dikenakan ke atas harga jualan barang atau perkhidmatan, jumlah yang akan dibayar kepada Kerajaan hanyalah atas nilai tambah barang atau perkhidmatan di setiap peringkat rantaian pengedaran atau bekalan. Nilai tambah adalah nilai yang ditambah oleh pengeluar (sama ada pengilang atau pengedar dan lain-lain) kepada bahan mentah atau pembelian sebelum menjual produk atau perkhidmatan baru atau yang telah diperbaiki.

GST menggunakan mekanisme penolakan (*offset*) kredit cukai di mana GST yang dikenakan ke atas output perniagaan ditolak daripada GST yang dibayar ke atas barang atau perkhidmatan yang diperoleh sebagai oleh perniagaan. GST yang dikenakan ke atas output dipanggil cukai output. Manakala, GST yang dikenakan ke atas perolehan dipanggil cukai input. Mekanisme secara *offsetting* ini adalah untuk memastikan GST yang dibayar oleh perniagaan boleh diperoleh semula dan dengan itu membantu mengurangkan kos menjalankan perniagaan.

Tugas JKDM sebagai pentadbir cukai adalah untuk menjana kutipan hasil secara berkesan dengan mengekang ketirisan hasil negara melalui penguatkuasaan dan pematuhan undang-undang yang ditetapkan di bawah beberapa akta seperti Akta Kastam 1967, Akta Cukai dan Barangan 2014, Akta Cukai Jualan 2018 dan Akta Cukai

perkhidmatan 2018. Penguatkuasaan dan pematuhan terhadap pembayar cukai ini meliputi kempen kesedaran cukai, pemeriksaan dan lawatan audit, serta surat dan notis pemakluman cukai. Terdapat risiko besar kehilangan cukai yang dipungut bagi pihak pentadbir cukai disebabkan oleh penipuan penyata cukai dan pengelakan cukai seperti mengurangkan jumlah cukai kena bayar dengan sengaja.

Model pengesanan fraud cukai merupakan salah satu komponen pengurusan risiko untuk mengesan pengelakan cukai yang digunakan oleh JKDM dan kebanyakan agensi cukai yang lain. Model pengesanan fraud berasaskan peraturan (*rule*-based) menetapkan beberapa peraturan dan skor markah berdasarkan penilaian yang dibuat oleh pegawai-pegawai yang berpengalaman dengan menilai dan memahami corak pengisytiharan cukai yang berpotensi untuk mengelak dan mengeksploitasi penyata cukai. Kaedah ini kemudiannya memberikan markah berdasarkan peraturan yang telah ditetapkan. Pengelasan akan dibuat berdasarkan markah yang diberikan sebagai risiko rendah dan risiko tinggi. Penetapan peraturan ini juga hanya berdasarkan corak-corak pengelakan dan penipuan cukai yang telah dikesan sebelum ini dan perlu sentiasa dikemas kini oleh pegawai bergantung kepada penemuan trend pengelakan cukai.

Model pengesanan fraud secara tradisional ini perlu ditambah baik kerana ia terlalu bergantung kepada kepakaran pegawai dan perlu sentiasa dikemas kini. Selain itu, jumlah transaksi penyata cukai yang banyak dan semakin meningkat memerlukan model pengesanan fraud secara automatik dan lebih berkesan dengan pendekatan sains data seperti kaedah pembelajaran mendalam. Data percukaian yang bersaiz besar dan kompleks mampu dikendalikan oleh pembelajaran mendalam dengan lebih bersistematik.

Kajian ini memfokuskan kepada pengesanan pengelakan cukai tidak langsung oleh pembayar cukai untuk mempertingkatkan pungutan hasil melalui pendekatan data sains seperti perlombongan data, kaedah pembelajaran mendalam. Kajian ini juga selaras dengan arah tuju JDKM seperti dalam Pelan Tindakan Strategik JKDM untuk tahun 2020 – 2024 iaitu; Teras 2: Mempertingkatkan pungutan hasil dan memantapkan fasilitasi, dan Teras 4: Mempertingkatkan keupayaan modal insan dan memperkasakan penggunaan teknologi.

1.3 PENYATAAN MASALAH

Perkembangan teknologi yang mengutamakan sistem digital dan automatik telah memberi manfaat kepada pembayar cukai dan pentadbir cukai dengan pengisytiharan kendiri dan penghantaran penyata cukai secara elektronik untuk pemprosesan penyata cukai yang lebih mudah dan pantas. Walau bagaimanapun, ia menjadi satu cabaran kepada pentadbir cukai termasuk JKDM untuk memastikan pengisytiharan penyata cukai yang dikemukakan adalah benar, tepat dan patuh dalam tempoh masa yang munasabah dan verifikasi yang cepat.

Sistem berasaskan peraturan dan analitik data tradisional diamalkan secara meluas oleh pentadbir cukai termasuk di JKDM sebagai salah satu kaedah pengesanan pengelakan cukai. Sistem berasaskan peraturan sangat bergantung kepada peraturan yang ditetapkan dan terdedah kepada kesilapan teknikal dan kekhilafan daripada peraturan yang ditetapkan kerana peraturan-peraturan yang ditetapkan adalah bersifat subjektif dan kompleks (de Roux et al. 2018). Ini berbeza dengan penggunaan pembelajaran mesin dan pembelajaran mendalam yang dapat mempelajari corak dan paten penipuan cukai sendiri. Atas sebab-sebab ini juga, pentadbir cukai dan pengkaji mula meneroka kaedah pembelajaran mesin dan pembelajaran mendalam sebagai satu alternatif.

Pendekatan pembelajaran mesin dalam pengesanan fraud adalah salah satu bidang kajian sains data yang popular. Kajian-kajian terdahulu membuktikan bahawa pembelajaran mesin mampu membuat pengesanan fraud dengan lebih pantas (Alzubaidi et al. 2021; Deng et al. 2021; Janiesch et al. 2021; Shukla et al. 2018). Walau bagaimanapun, kajian-kajian tersebut menyatakan model pembelajaran mesin memerlukan banyak fasa dengan pra-pemprosesan yang menyeluruh dan prestasi model sangat bergantung kepada pengekstrakan ciri fitur yang jelas untuk memberikan. Oleh itu, kaedah pembelajaran mendalam di cadangkan sebagai eksplorasi model yang lebih mudah dan pantas.

Penerokaan pembelajaran mendalam untuk pengesanan fraud dalam kajian terdahulu kebanyakannya menjuruskan kepada ramalan pengelasan fraud cukai menggunakan set data awam yang terhad (Vanhoeyveld et al. 2020; F. Zhang et al.

2020) menyebabkan penerokaan pembelajaran mendalam dengan set data percukaian sebenar, bersaiz besar dan berlabel terhad kerana kesukaran mendapatkan data (Kleanthous & Chatzis 2020). Justeru itu, kajian ini menggunakan kelebihan set data yang besar dan kompleks untuk pembangunan model pembelajaran mendalam.

Bagi tujuan kajian ini, data yang digunakan merupakan data percukaian sebenar GST di Malaysia yang tidak pernah dijalankan kajian oleh mana-mana penyelidik lagi. Data percukaian juga adalah unik bergantung kepada rejim cukai sesebuah negara dan polisi pentadbir cukai itu sendiri. Data percukaian juga bersaiz besar kerana pembayar cukai melibatkan jumlah transaksi yang amat besar. Corak dan paten yang tersembunyi berpotensi untuk dapat dikenal pasti dengan menggunakan pembelajaran mendalam yang lebih komprehensif dan algoritma pembelajaran dalam dapat mengendalikan data yang besar (Alghofaili et al. 2020).

1.4 HIPOTESIS KAJIAN

Hipotesis kajian adalah prosedur formal untuk membuat pembuktian dalam kajian. Kajian ini membentuk hipotesis seperti di bawah:

H_0: Prestasi pengelasan pengelakan cukai menggunakan model pembelajaran mendalam pada set data penyata cukai dan set data berkategori adalah sama dengan prestasi model sedia ada.

H_1: Prestasi pengelasan pengelakan cukai menggunakan model pembelajaran mendalam pada set data penyata cukai dan set data berkategori adalah berbeza dengan prestasi model sedia ada.

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 taburan label kelas risiko rendah dan risiko tinggi pada set data penyata cukai?

2. Adakah pecahan set data mengikut kategori pemfailan mempengaruhi pengujian dan prestasi model pembelajaran mendalam yang dibangunakan?

3. Apakah model pembelajaran mendalam yang terbaik untuk membuat pengesanan pengelakan cukai tidak langsung di Malaysia?

1.6 OBJEKTIF KAJIAN

Objektif kajian ini adalah untuk membangunkan model pembelajaran mendalam untuk pengesanan pengelakan cukai tidak langsung bagi membantu JKDM untuk mengoptimumkan kutipan hasil cukai negara.

1. Membangunkan model deskriptif untuk menganalisis taburan data penyata cukai yang berisiko rendah dan berisiko tinggi.

2. Mengenal pasti keberkesanan pecahan set data untuk data transaksi percukaian yang besar dan kompleks.

3. Membangunkan model pembelajaran mendalam yang terbaik bagi pengesanan pengelakan cukai tidak langsung di Malaysia.

1.7 SKOP KAJIAN

Kajian ini melibatkan skop yang ditetapkan di bawah:

1. Kajian menggunakan dua jenis data transaksi kewangan iaitu data kad kredit dan data percukaian GST. Data kad kredit digunakan sebagai data tanda aras yang diperoleh daripada repositori *Kaggle* yang popular digunakan dalam kajian pengesanan fraud dan data percukaian GST di Malaysia daripada tahun 2014 sehingga tahun 2018 sebagai data mentah yang dikaji dalam penyelidikan ini.

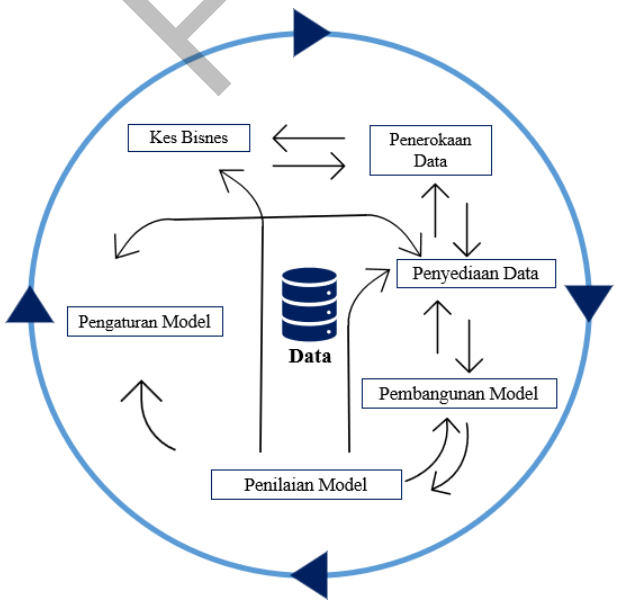
2. Kajian ini membangunkan model pembelajaran mendalam menggunakan pendekatan Rangkaian Neural Pelingkaran (*Convolutional Neural Network,* CNN) dan Memori Jangka Masa Panjang dan Pendek (*Long Short Term Memory,* LSTM) bagi pengelasan pengelakan cukai tidak langsung untuk data percukaian GST.

3. Kajian ini mengukur prestasi model dengan menggunakan pengukuran prestasi ketepatan, kejituan, dapatan semula dan skor F1 terhadap pengelasan yang dibuat oleh model pembelajaran mendalam yang dipilih.

1.8 METODOLOGI KAJIAN

Metodologi kajian perlu dirangka berdasarkan objektif kajian dan dijadikan panduan bagi pelaksanaan kajian. Bersesuaian dengan objektif kajian ini, pendekatan *Cross Industry Standard Proses for Data Mining* (CRIPS-DM) di adaptasikan sebagai kerangka kerja dalam kajian ini. CRIPS-DM mempunyai kerangka kerja dan struktur amalan terbaik bagi menghasilkan perlombongan data yang lebih berkesan dan pantas. Ia juga fleksibel dan boleh di ubah suai mengikut keperluan sesuatu projek atau kajian dan bergantung kepada keputusan yang dicapai dalam setiap fasa.

Bagi kajian ini, urutan fasa metodologi kajian yang telah dirangka adalah seperti dalam Rajah 1.1. Kerangka kerja CRIPS-DM bagi kajian ini merangkumi enam (6) fasa utama dan meliputi kesemua bab dalam kajian ini. Fasa dalam kajian ini adalah pemahaman bisnes, pemahaman data, penyediaan data, pembangunan model, penilaian model dan pengaturan model. Setiap fasa dalam CRISP-DM akan dibincangkan dengan lebih perinci dalam Bab 3.

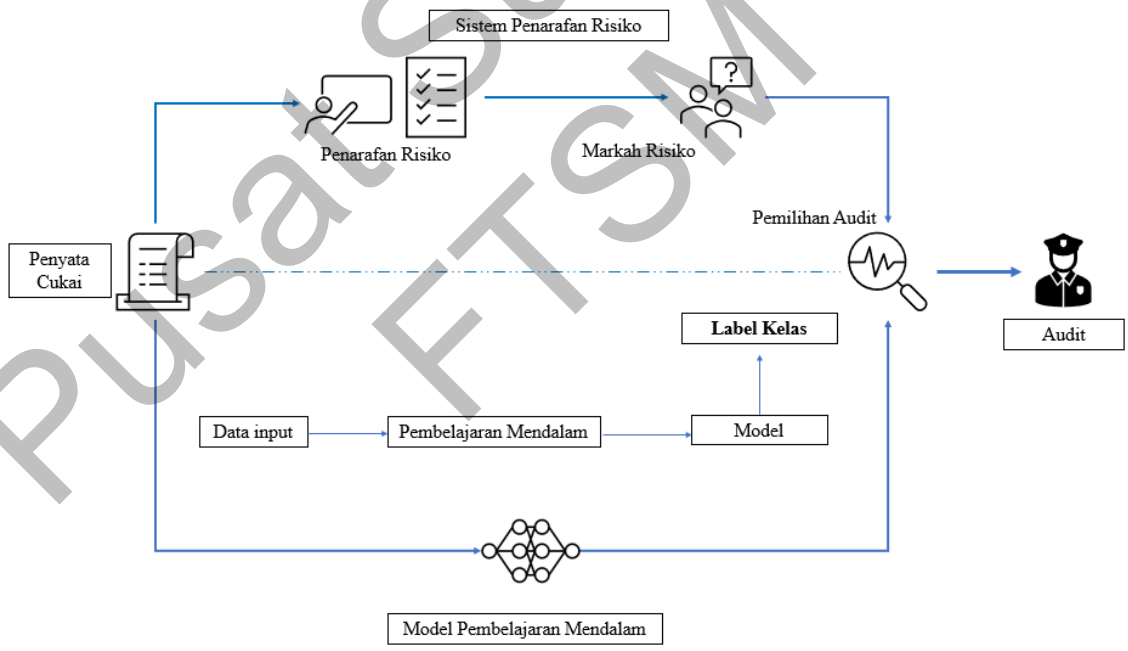


Rajah 1.1 Proses CRIPS-DM dalam kajian data sains

1.9 KEPENTINGAN KAJIAN

Pentadbir cukai sentiasa dalam tekanan untuk mengesan dan mencegah pengelakan cukai sebelum ia berlaku, tetapi ia juga penting untuk memerangi perkara ini tanpa menjejaskan pematuhan pembayar cukai. Hasil kajian dapat mengukuhkan kutipan hasil percukaian oleh JKDM kepada negara. Ia juga dapat mengubah budaya kerja sedia ada kepada budaya kerja berteknologi tinggi selaras dengan Pelan Tindakan Strategik JKDM untuk tahun 2020 – 2024.

Oleh itu, terdapat keperluan untuk rangka kerja pengesanan pengelakan cukai yang lebih komprehensif dan boleh mempelajari penipuan cukai dengan tepat. Pengenalan terhadap kaedah pembelajaran mendalam dapat membantu membuat pengelasan pengelakan cukai secara automatik dan lebih tepat berbanding sistem penarafan risiko sedia ada seperti Rajah 1.1.



Rajah 1.1 Perbandingan model sedia ada dan model pembelajaran mendalam

Organisation for Economic Co-operation and Development (OECD) menyokong perhubungan analitis data untuk tujuan pengurusan risiko yang lebih luas dan membantu memacu tadbir urus data, infrastruktur data yang lebih berkesan kepada sesuatu organisasi, khususnya JKDM. Penggunaan data analitik dan pembelajaran mesin untuk pengesanan fraud cukai berisiko boleh melengkapkan metodologi

kualitatif sedia ada yang dapat mengurangkan positif palsu dan negatif palsu (OECD 2019). Kajian data sains dalam domain percukaian dan kewangan amat popular bagi negara membangun. Walau bagaimanapun terdapat kekurangan kajian bagi domain tersebut di Malaysia. Kajian terdahulu adalah berkaitan dengan cukai pendapatan dan cukai korporat (Rahman et al. 2019 2020) dan faktor kepada pengelakan cukai (Othman et al. 2019; Tabandeh et al. 2012).

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 II adalah berdasarkan kajian literatur yang lepas tentang isu-isu yang berkaitan dengan model pengesanan fraud, memahami kajian literatur, memilih model pembelajaran mendalam yang sesuai untuk kajian dan rujukan penyelesaian untuk pemilihan kaedah yang akan digunakan.

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

Bab IV 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 pembelajaran mendalam. Ujian statistik diguna untuk membandingkan ketepatan prestasi kedua-dua model yang dikaji dan seterusnya dinilai oleh pakar percukaian.

Bab V 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 pengelakan cukai dibincangkan dan dicadangkan untuk kajian lanjut.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Bab ini membincangkan literatur berkaitan yang menjadi asas kepada kajian ini dalam konteks pematuhan percukaian, pembangunan dan pemilihan model pembelajaran mendalam dan pengekstrakan fitur melalui pembelajaran mendalam. Kajian literatur dijalankan berdasarkan penelitian dan mengenal pasti masalah jurang dalam bidang kajian. Perbincangan kajian literatur dimulakan dengan pemahaman mengenai pematuhan cukai, sistem penilaian risiko cukai yang digunakan oleh agensi pentadbir cukai dan lain-lain agensi pengurusan kewangan, model pembelajaran mesin dalam kajian lepas dan akhir sekali pengenalan kepada model pembelajaran mendalam yang kian popular.

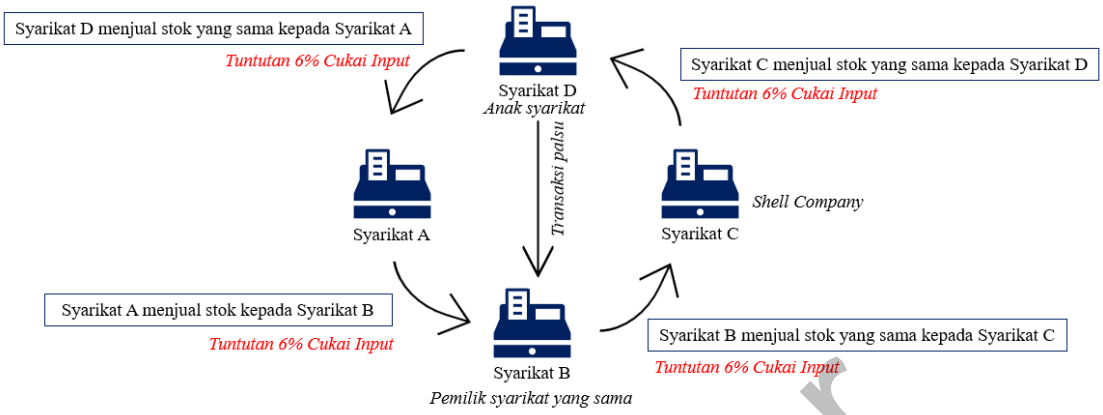
2.2 PENGELAKAN CUKAI

Konsep percukaian dan tahap pematuhan percukaian adalah saling berkait. Sistem percukaian yang mengamalkan pengikraran kendiri terdedah kepada unsur fraud dan dipantau dengan sistem pengesanan fraud dalaman dan tradisional. Pembelajaran mesin dan pembelajaran mendalam dapat menjadi alternatif kepada sistem pengesanan fraud yang lebih berkesan. Penyelidikan semasa dan lampau menjadi panduan dalam membangunkan model pengelakan cukai yang lebih berkesan dengan kawalan terhadap masalah set data yang tidak seimbang dan set data yang besar. Pengekstrakan fitur melalui pembelajaran mendalam juga dapat membantu pegawai audit untuk mengecilkan skop audit dengan hanya memberi fokus kepada fitur yang penting sahaja semasa buat semakan audit.

Konsep percukaian masa kini kebanyakannya mengamalkan pengisytiharan kendiri secara digital. Menurut Kertas Kerja Tabung Mata Wang Antarabangsa (*International Monetary Fund,* IMF) WP/20/245, pentadbir cukai semakin beralih kepada teknologi digital yang membolehkan pengikraran cukai secara elektronik dan pengesahan digital yang lebih pantas. Ini boleh meningkatkan pematuhan dan penguatkuasaan cukai dengan menyelaraskan perbezaan pembayaran, memantau kutipan hasil secara masa nyata, melaksanakan audit dan menggunakan data raya untuk menilai risiko pembayar cukai (Korauš et al. 2021). Walau bagaimanapun, risiko untuk membuat penipuan fraud cukai dan pengelakan cukai juga adalah sangat tinggi dengan teknologi digital dan automatik yang membuka peluang kepada pembayar cukai untuk melakukan fraud dan pengelakan cukai melalui penghantaran penyata cukai (Adamov 2020; Vanhoeyveld et al. 2020).

Fraud dan pengelakan cukai adalah terma yang berkait rapat. (Adekoya et al. 2020) mendefinisikan fraud cukai sebagai tindakan yang disengajakan untuk memperdaya, menindas dan membuat penipuan cukai. Tindakan ini melibatkan pengelakan cukai dan ketidakpatuhan terhadap garis panduan atau undang-undang yang ditetapkan. Ia dilakukan dengan membuat pengisytiharan palsu dan menyembunyikan maklumat cukai untuk memperdaya pihak berkuasa untuk mengurangkan liabiliti dan pembayaran cukai.

Kajian (Mehta, Mathews, Kumar, Suryamukhi, et al. 2019) menyatakan terdapat dua jenis pengelakan cukai yang popular untuk GST dan VAT iaitu perdagangan pelingkaran (*circular trading*) dan penipuan karusel. Melalui perdagangan pelingkaran, syarikat mengeluarkan invois kepada syarikat lain sama ada syarikat sendiri atau *shell company* tanpa membuat pembekalan barangan dan perkhidmatan dengan tujuan untuk menuntut pulang balik cukai input. Manakala penipuan karusel dilakukan dengan mengelak bayaran cukai kepada kerajaan berdasarkan pelepasan yang diberi tetapi mengenakan cukai di peringkat seterusnya. Rajah 2.1 menunjukkan salah satu contoh pengelakan cukai dan fraud melalui perdagangan pelingkaran.



Rajah 2.1 Contoh penjualan dan pembelian dalam pedagangan pelingkaran

2.3 PENGELAKAN CUKAI DI MALAYSIA

Penguatkuasaan ke atas pematuhan cukai adalah satu cabaran besar buat pentadbir cukai termasuk di Malaysia. Faktor pengelakan dan penipuan cukai ini juga bergantung kepada keadaan ekonomi dan rejim cukai itu sendiri yang memberi ruang kepada unsur-unsur penipuan. Kajian berlatar belakangkan fraud GST di Malaysia menyenaraikan tipologi fraud GST yang di kesan hasil daripada penyelidikan dan temu ramah dengan pegawai JKDM (Othman et al. 2019) seperti dalam Jadual 2.1.

Jadual 2.1 Corak pengelakan cukai dan tipologi fraud GST (Othman et al. 2019)

Bil	Corak Pengelakan Cukai	Tipologi Fraud GST
1	Tuntutan cukai input yang tidak munasabah	Pemalsuan tuntutan cukai input
		Manipulasi jualan
2	Ketidakpatuhan	Tidak mengemukakan penyata GST
		Kegagalan untuk mendaftar GST
		Mengelak membuat pembayaran GST
3	Penipuan GST tegar	Penipuan karusel

Kajian (Tabandeh et al. 2012) mengkaji faktor yang menyumbang kepada pengelakan cukai pendapatan dan kepentingan setiap faktor tersebut terhadap penguatkuasaan cukai di Malaysia menggunakan data daripada petunjuk pembangunan dunia (*world development indicator*, WDI) daripada Laporan Ekonomi tahunan. Kajian mengenal pasti beberapa faktor yang mendorong kepada pengelakan cukai di Malaysia iaitu bebanan cukai, pendapatan pembayar cukai, saiz kerjaan, kadar inflasi dan polisi perdagangan. Manakala kajian terdahulu bagi kepatuhan cukai di Malaysia

menumpukan kepada tahap kepatuhan cukai oleh perusahaan kecil dan sederhana (SME). Kajian (Mohamad et al. 2016) menyenaraikan beberapa pemboleh ubah yang menyumbang kepada pengelakan cukai seperti lokasi, ejen cukai, transaksi tunai, jenis industri, saiz perniagaan dan jenis transaksi seperti penjualan atau pembelian. Terdapat beberapa kajian lain dalam domain percukaian di Malaysia dan kebanyakannya adalah berkaitan dengan cukai pendapatan dan cukai korporat (Rahman et al. 2019; 2020) serta faktor kepada pengelakan cukai (Othman et al. 2019; Tabandeh et al. 2012).

Terdapat jurang dalam kajian pematuhan dan pengelakan cukai di Malaysia, justeru kajian ini memfokuskan pengelasan pengelakan cukai bercirikan perdagangan pelingkaran dan penyata cukai yang kurang mengisytiharkan (*under-declare*) penyata cukai GST dengan set data penyata cukai GST di Malaysia. Pengelakan cukai ini diklasifikasikan sebagai pembayar cukai yang berisiko dan tidak berisiko untuk membuat pengelakan cukai berdasarkan penyata pengisytiharan cukai yang dihantar. Hasil daripada pelabelan ini dihantar kepada pegawai audit untuk membuat audit meja dan audit lapangan.

Pematuhan cukai juga adalah berbeza mengikut sistem percukaian yang di kuat kuasakan dan bergantung kepada pentadbiran cukai. Kajian lepas (Ordonez et al. 2020) memfokuskan kepada ramalan kebarangkalian pembayar cukai yang berisiko tidak untuk tidak membuat bayar cukai dengan kerangka perlombongan data dan rangkaian neural untuk pengurusan cukai di Ecuador. Kajian pematuhan cukai yang popular dalam domain data sains adalah model ramalan dan klasifikasi pemilihan fail cukai untuk tujuan audit. Kajian ini bertujuan untuk mengenal pasti pembayar cukai yang mempunyai risiko tinggi untuk melakukan pengelakan cukai untuk dihantar kepada siasatan auditan.

2.4 SISTEM PENILAIAN RISIKO FRAUD CUKAI

Latar belakang ringkas mengenai sistem pengesanan fraud cukai dibincangkan untuk memahami konsep pengesanan fraud cukai dengan lebih mendalam. Sistem pengesanan fraud cukai yang digunakan di kebanyakan negara adalah berkonsepkan perlombongan data dan penilaian berasaskan peraturan (*rule-based*) menerusi penilaian risiko fraud (de Roux et al. 2018; Kleanthous & Chatzis 2020). Penilaian risiko dapat membantu

pegawai cukai mengenal pasti transaksi berisiko dan membuat aktiviti dan langkah kawalan termasuk meramalkan transaksi risiko tinggi. Pegawai hendaklah sentiasa memastikan peraturan yang ditetapkan adalah relevan dan memerlukan input daripada hasil audit terdahulu untuk mengemas kini peraturan risiko yang baru

Melalui model penilaian risiko fraud cukai, pensampelan dijalankan berdasarkan beberapa peraturan dan had ambang yang telah dikenal pasti oleh pegawai cukai. Analisis statistik dan analitik data juga digunakan secara meluas untuk mengenal pasti akaun cukai yang akan di audit. Pemilihan kes secara manual, maklumat daripada pemberi maklumat dan pemilihan berasaskan komputer adalah beberapa kaedah yang disenaraikan sebagai sistem pengesanan fraud cukai (Wu et al. 2019). Teknik yang digunakan ini memakan masa yang banyak dan memerlukan tenaga kerja yang ramai. Walaupun dengan sistem pengurusan dan penilaian risiko berdasarkan peraturan ini mampu mengurangkan beban pemilihan kes audit, ia mempunyai banyak batasan. Antara beberapa kekangan adalah seperti sumber tenaga dan masa untuk memeriksa kesemua pembayar cukai dan mengenal pasti pembayar cukai yang berisiko untuk melalukan fraud (Assylbekov et al. 2016) seperti yang dilaporkan oleh pentadbir cukai di Kazakhstan.

Kajian oleh (Carrasco & Sicilia-Urbán 2020) menyatakan teknik pengesanan penipuan cukai dengan bantuan komputer tidak dapat menampung jumlah transaksi dan pengikraran cukai yang besar dan selalunya memerlukan bantuan pegawai untuk menyemak semula transaksi dan penyata tersebut. Sistem berasaskan peraturan ini sangat bergantung kepada peraturan yang ditetapkan dan terdedah kepada kesilapan teknikal dan kekhilafan daripada peraturan yang ditetapkan kerana peraturan-peraturan yang ditetapkan adalah bersifat subjektif dan kompleks. Pegawai cukai hendaklah sentiasa memastikan peraturan yang ditetapkan adalah relevan dan memerlukan input daripada hasil audit terdahulu untuk mengemas kini peraturan risiko yang baru (Kleanthous & Chatzis 2020).

Kajian (Adamov 2019) membincangkan tentang kemunculan analitik data telah meningkatkan keupayaan pengesanan fraud dengan keupayaan perlombongan data yang tepat dan pendekatan algoritma yang berkesan dalam menganalisis trendd

pengelakan cukai. Walau bagaimanapun, data percukaian melibatkan jumlah data yang besar mewakili berjuta-juta transaksi dan pengisytiharan oleh pembayar cukai. Penggunaan sistem dan perisian tradisional seperti *spreadsheet* adalah sudah tidak praktikal untuk menampung keperluan data dan pengesanan fraud cukai kini memerlukan teknik dan model yang lebih pantas dan berkesan (González et al. 2021).

Kajian ini mengambil kira kekurangan data percukaian dalam industri dengan rekod pengelasan pengelakan cukai yang sebenar. Kajian (González et al. 2021) untuk mengenal pasti fraud cukai pendapatan di Spain menggunakan Lapisan Perceptron Berbilang Baris (*Multilayer Perceptron Layer,* MLP) menggunakan data pembayar cukai pendapatan di Sepanyol menyatakan model yang dibina mempunyai kekurangan kerana kekurangan data sebenar rekod fraud.

2.5 MODEL PEMBELAJARAN MESIN UNTUK PENGESANAN FRAUD

Pendekatan pembelajaran mesin untuk pengesanan fraud cukai dapat menambah baik model pengesanan fraud cukai yang sedia ada dengan prestasi yang lebih tinggi. Model pembelajaran mesin juga tidak terlalu bergantung kepada kepakaran pentadbir cukai sahaja. Malah ia dapat belajar dengan hanya menggunakan data-data penyata dan transaksi cukai untuk menghasilkan model pengesanan fraud cukai. Pembelajaran mesin mampu untuk membuat generalisasi terhadap paten dan corak tersembunyi

Pembelajaran mesin juga kian popular sebagai alternatif untuk pengesanan fraud sama ada dalam konteks percukaian mahupun kewangan. (Ozbayoglu et al. 2020) menerangkan fraud kewangan merangkumi fraud kad kredit, pengubahan wang haram, fraud kredit pengguna, pengelakan cukai, penipuan bank dan penipuan tuntutan insurans. Model pengesanan fraud kewangan popular dalam penyelidikan dan pembangunan model pengesanan melalui pembelajaran mesin.

Pengelasan, pengelompokan dan regresi antara teknik yang digunakan dalam pembelajaran mesin untuk pengesanan fraud (Didimo et al. 2020). Transaksi fraud dapat dipelajari melalui rekod sejarah transaksi yang berlabel dengan menggunakan algoritma seperti Hutan Rawak (*Random Forest*), Pokok Keputusan (*Decision Tree*)

dan Regresi Logistik (*Logistic Regression*) yang dilaporkan menghasilkan keputusan yang lebih tepat mengikut penyelidikan terkini (Carrasco & Sicilia-Urbán 2020).

(Deng et al. 2021) mencadangkan sistem pengesanan pembelajaran mesin untuk fraud kewangan berdasarkan Hutan Rawak dan pengesanan secara manual menggunakan data fraud IEEE CIS. Kajian menunjukkan model Hutan Rawak mampu mencapai ketepatan yang tinggi dan mencadangkan ia untuk diaplikasikan ke atas data fraud yang sebenar. Walau bagaimanapun, hasil daripada pengelasan untuk label fraud oleh Hutan Rawak masih perlu dinilai oleh pakar cukai untuk menentukan risiko penipuan cukai sama ada risiko tinggi atau risiko rendah.

Pembangunan model pembelajaran mesin merangkumi beberapa langkah antaranya pra-pemprosesan, pemilihan fitur, pengelompokan dan pengelasan seperti yang dijalankan dalam kajian (Shukla et al. 2018). Kajian tersebut mencadangkan model pengesanan pengelakan cukai dengan mengenal pasti ciri-ciri pembayar cukai dan kemudiannya menggunakan Rangkain Neural Suap Depan (*Multilevel Feed Forward Neural Network*, MFFNN) untuk mengenal pasti fraud cukai. Walaupun model yang dibangunkan adalah berjaya namun kajian mencadangkan eksplorasi kaedah yang lebih mudah dan pantas.

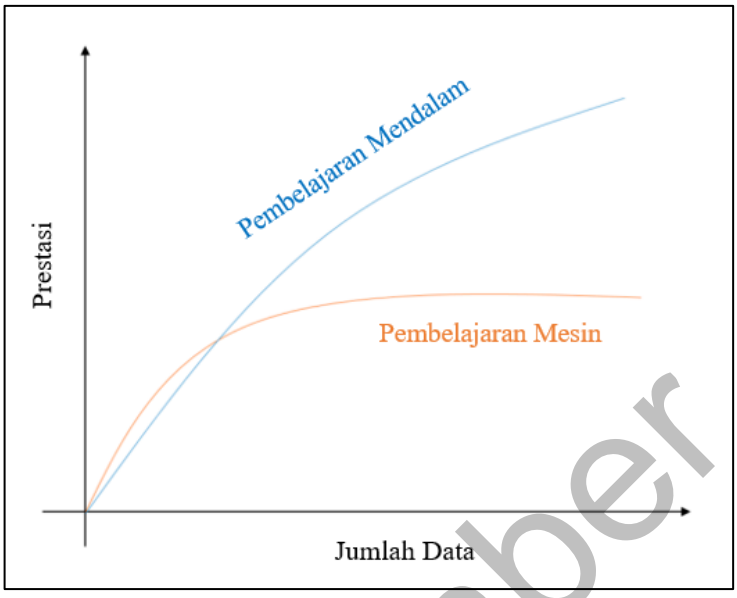
Antara kekangan model pembelajaran mesin adalah ia memerlukan pra pemprosesan yang menyeluruh, pemilihan fitur dan pengurangan dimensi. Kajian (Alzubaidi et al. 2021) merumuskan pembelajaran mesin dalam tugasan klasifikasi memerlukan banyak fasa dan setiap fasa akan mempengaruhi prestasi algoritma pembelajaran mesin. Kajian konseptual (Janiesch et al. 2021) menyatakan pembelajaran mesin sangat bergantung kepada ciri fitur yang jelas dan proses pengekstrakan ciri sangat mempengaruhi prestasi pembelajarannya. Pemilihan fitur yang tidak seimbang juga boleh menyebabkan pengelasan kelas yang kurang berkesan. Ini berbeza dengan kaedah pembelajaran mendalam yang mampu untuk membuat pembelajaran set fitur secara automatik untuk pelbagai aplikasi.

2.6 PENEROKAAN MODEL PEMBELAJARAN MENDALAM

Kajian-kajian terdahulu menjuruskan kepada pengunaan pembelajaran mesin dengan menggunakan data-data yang terhad. Kesukaran mendapatkan data daripada agensi cukai serta lain-lain data berkaitan transaksi kewangan menjadi batasan kajian. Data-data daripada agensi cukai adalah diklasifikasi sebagai sulit kerana melibatkan maklumat perniagaan dan individu (Kleanthous & Chatzis 2020). Walau bagaimanapun, kajian-kajian terdahulu bagi domain percukaian kebanyakkannya memfokuskan kepada model pengesanan frod cukai menggunakan pembelajaran mesin.

Terdapat keperluan untuk meneroka keupayaan pembelajaran mendalam yang lebih terperinci menggunakan kelebihan data yang besar dan algoritma yang lebih berkesan sebagai pendekatan alternatif. Data pembayar cukai melibatkan jumlah data yang besar, corak dan paten dapat dikenal pasti dengan menggunakan pembelajaran mendalam yang lebih komprehensif. Data percukaian adalah dalam saiz yang besar kerana ia melibatkan transaksi yang tinggi seperti data GST di India yang terdiri daripada berjuta-juta baris, dan saiz pangkalan data sebanyak 1.5 Terabait (Mehta, Mathews, Kumar, & Suryamukhi 2019).

Algoritma pembelajaran mendalam seperti LTSM dan CNN dapat mengendalikan data yang besar dengan pengelasan dan ramalan yang lebih tepat. Kemampuan model pembelajaran adalah daripada rangkaian neural dan mampu merekodkan fitur dan mengawal isu kompleks jika dibandingkan dengan pembelajaran mesin (Alom et al. 2019). Pembelajaran mendalam mampu mengesan fraud kewangan secara efektif. Rajah 2.2 menunjukkan prestasi pembelajaran mesin dan pembelajaran mendalam terhadap jumlah data yang digunakan.



Rajah 2.2 Prestasi pembelajaran mendalam berbanding pembelajaran mesin (Alom et al. 2019)

Kajian oleh (Alghofaili et al. 2020) memperkenalkan kaedah pengesanan fraud kewangan untuk set data kad kredit menggunakan teknik pembelajaran mendalam LSTM yang menghasilkan ketepatan latihan sebanyak 99.96% berbanding model Autoencoders dan model pembelajaran mesin seperti Hutan Rawak, SVM dan Regresi Logistik. Model LSTM dapat mengesan dan meramal fraud kewangan dengan berkesan berdasarkan penggunaan sel LSTM.

Hasil eksperimen oleh (Raghavan & Gayar 2019) dengan set data kad kredit bagi pengesanan fraud menunjukkan CNN adalah kaedah pembelajaran mendalam yang terbaik untuk set data yang besar mengatasi penggunaan RBM, DBM dan Autoencoders dengan pencapaian prestasi Pekali Korelasi Mathews (MCC) dan kawasan di bawah lengkung (AUC). Dalam kajian tersebut, model CNN dengan tujuh lapisan yang mempunyai tiga lapisan Pelingkaran, dua lapisan pengumpulan, satu lapisan penyambungan penuh dan 1 lapisan *softmax* telah dicadangkan.

Kajian lain yang berkaitan dengan pengelakan dan fraud percukaian kebanyakannya mencadangkan penggunaan model pembelajaran mendalam yang tidak diselia dan separuh diselia atas faktor kekangan data yang mempunyai label kelas. Dalam kajian pemilihan kes audit bagi VAT di Eropah, model pembelajaran mendalam separuh diselia, Autoencoders dapat membantu membuat pemilihan kes audit yang lebih berkesan berbanding dengan model pembelajaran mendalam diselia dengan

ketepatan yang negatif benar yang tinggi (Kleanthous & Chatzis 2020). Walau bagaimanapun, pada peringkat ini, kajian akan menumpukan kepada model pembelajaran mendalam diselia dengan menggunakan data yang mempunyai label kelas.

Kajian (Zumaya et al. 2021a) untuk mengenal pasti pengelakan cukai di Mexico menerangkan struktur LSTM untuk mengklasifikasikan pembayar cukai yang menggunakan resit digital untuk meningkatkan potongan cukai. Klasifikasi pembayar cukai sebagai berpotensi melalukan pengelakan cukai ini menggunakan tiga lapisan sel memori LSTM tersembunyi dengan setiap satu dengan 256 neuron yang menghubungkan setiap neuron dalam satu lapisan neuron yang lain ke lapisan berikutnya.

Seterusnya, kajian terdahulu yang dirujuk dirumuskan dalam Jadual 2.2 untuk membuat perbandingan dan memahami objektif serta hasil kajian secara lebih terperinci untuk pembangunan model pembelajaran mendalam bagi pengesanan pengelakan cukai.

Jadual 2.2 Rumusan kajian pembelajaran mendalam untuk pengesanan pengelakan cukai

Bil	Kajian	Objektif	Algoritma	Dapatan Kajian
1	(Alghofaili et al. 2020)	Kajian memperkenalkan model pembelajaran mendalam untuk fraud kewangan menggunakan kaedah LSTM mengguna kan set data fraud kewangan sebenar dan menambahbaikkan model seiring dengan penggunaan data raya.	LSTM, Autoencoders	Kajian membuat perbandingan dengan kaedah Autoencoders dan beberapa teknik pembelajaran mesin yang lain dan LSTM menunjukkan prestasi yang terbaik dalam mengesan fraud kewangan untuk set data mentah sebenar daripada industri.
2	(Babu & Pratap 2020)	Kajian mencadangkan penggunaan lapisan Pelingkaran untuk model pengesanan fraud kewangan yang bagi set data yang tidak seimbang.	CNN dengan 6 lapisan Perlingkaran	Arkitektur CNN yang digunakan dalam kajian dapat memberikan prestasi yang tinggi tanpa memerlukan fitur input dimensi tinggi. Walau bagaimanapun, pengumpulan maksimum mengurangkan prestasi CNN.
3	(Benchaji et al. 2021)	Membangunkan model pengesanan fraud kad kredit berdasarkan pemodelan data berurutan, mekanisme perhatian dan rangkaian saraf berulang dalam LSTM. Kajian juga mengoptimumkan proses mengelas pembelajaran melalui pemilihan fitur dan pengurangan dimensi.	LSTM	Model yang dicadangkan mencapai prestasi yang baik dalam pengelasan fraud melalui beberapa pendekatan seperti kecerdasan Swarm, kaedah *Uniform Manifold Approximation and Projection* (UMAP) untuk pengurangan dimensi dan pensampelan SMOTE untuk data yang tidak seimbang.
4	(Fang et al. 2021)	Meneroka keupayaan DNN untuk model pengesanan fraud kewangan dengan memberi fokus kepada pra pemprosesan data dan pengimbangan data.	DNN, SVM, Hutan Rawak, Pokok Keputusan dan Regresi Logistik	DNN mencapai prestasi yang lebih baik berbanding kaedah pembelajaran mesin untuk data kewangan daripada industri. DNN dengan pengoptimum Adam menghasilkan prestasi ketepatan yang lebih tinggi.

bersambung…

…sambungan

Bil	Kajian	Objektif	Algoritma	Dapatan Kajian
6	(Raghavan & Gayar 2019)	Menjalankan kajian penandaarasan untuk kaedah pembelajaran mesin dan pembelajaran mendalam yang terbaik untuk model pengesanan fraud kewangan menggunakan 3 set data fraud kewangan.	KNN, SVM, Hutan Rawak, Autoencoders, CNN, RBM, DBM.	Kajian melaporkan SVM dan CNN adalah kaedah yang terbaik untuk set data yang besar, manakala Hutan Rawak, SVM, KNN lebih sesuai untuk set data kecil. Prestasi CNN juga mengatasi kesemua kaedah pembelajaran mendalam yang lain.
7	(Zumaya et al. 2021)	Mengkaji penggunaan sains rangkaian dan pembelajaran mesin serta pembelajaran mendalam untuk mengenal pasti corak penipuan cukai secara automatik sama seperti yang dikesan oleh manusia.	DNN, ANN, LSTM, Hutan Rawak	Kajian membuktikan pengesanan corak penipuan cukai mampu dijalankan melalui sains rangkaian dan pembelajaran mesin tetapi corak dan paten yang dikenal pasti adalah berdasarkan corak yang diketahui oleh manusia dan bergantung pada paten statistik sahaja.
8	(Gupta et al. 2021)	Mengenal pasti model pengelasan fraud melalui pembelajaran mesin dan pembelajaran mendalam dan teknik pengimbangan data yang berkesan. Kajian menggunakan set data sektor kesihatan bagi pengesanan fraud insurans kesihatan.	Hutan Rawak, Pokok Keputusan, XGBoost, LightGBM, GBM, Rangkaian Neural	Kajian membuat pengimbangan data menggunakan set data asal, teknik *weighted*, *undersample*, SMOTE, ADASYN dan TGANs. Model rangkaian dengan teknik *weighted* menunjukkan prestasi yang tinggi. Pokok Keputusan menggunakan set data asal mempunyai ketepatan paling tinggi untuk pembelajaran masin,

Penerokaan pembelajaran mendalam untuk pengesanan pengelakan cukai dan pengesanan fraud dengan perbandingan terhadap model pembelajaran mesin menunjukkan algoritma seperti LSTM, CNN, Rangkaian Neural Mendalam (DNN), Rangkaian Neural Buatan (ANN) dan Autoencoders banyak digunakan dalam kajian lepas. Kajian (Alghofaili et al. 2020) membangunkan model LSTM dan Autoencoders untuk pengesanan fraud dan menunjukkan LSTM mempunyai prestasi yang terbaik dalam kajian itu. Model pembelajaran mendalam LSTM juga dibangunkan dalam kajian (Benchaji et al. 2021) dengan teknik seperti pengurangan dimensi dan pensampelan SMOTE yang telah meningkatkan prestasi model tersebut. Manakala kajian (Nguyen et al. 2020b; Raghavan & Gayar 2019) pula membangunkan model pembelajaran mesin dam mendalam lain bersama model CNN dan prestasi model CNN mengatasi kesemua model yang dibangunkan.

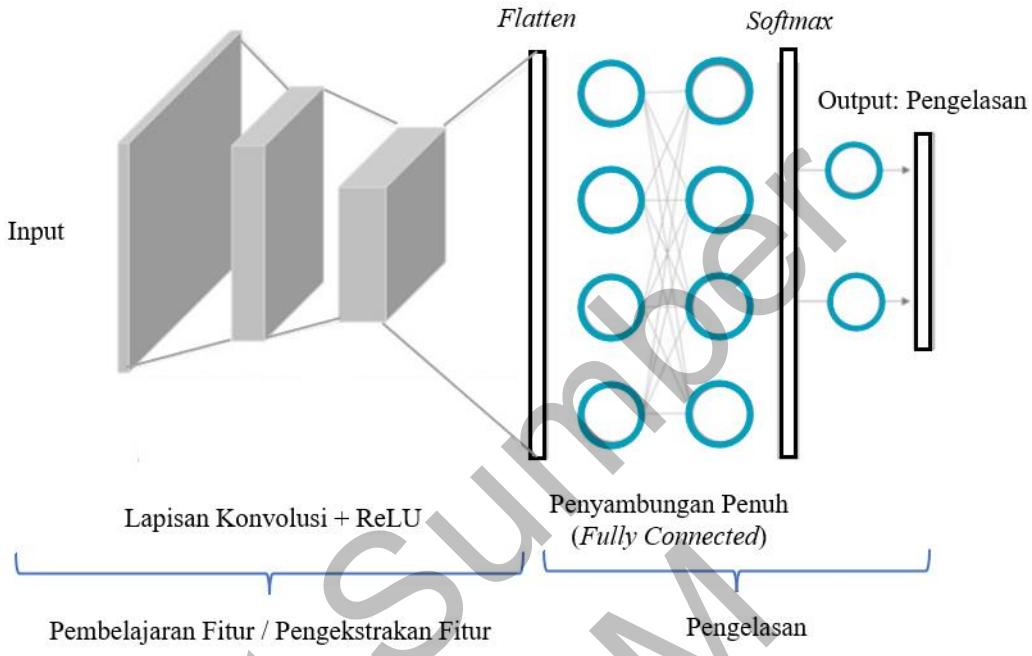
Hasil dapatan kajian-kajian terdahulu juga menunjukkan prestasi lebih baik bagi model LSTM dan CNN apabila dibandingkan dengan model pembelajaran mesin yang popular seperti Hutan Rawak, Regresi Logistik dan Pokok Keputusan (Fang et al. 2021; Gupta et al. 2021; Zumaya et al. 2021). Oleh itu, kajian ini mengambil pendekatan pembangunan model pembelajaran mendalam menggunakan kaedah algoritma LSTM dan CNN untuk mengenal pasti model yang terbaik bagi pengesanan pengelakan cukai menggunakan set data penyata cukai yang besar dan kompleks.

2.6.1 Rangkaian Neural Pelingkaran (CNN)

Rangkaian Neural Pelingkaran (*Convolutional Neural Network*, CNN) adalah satu kaedah dalam rangkaian neural mendalam yang kerap digunakan dalam kajian pembelajaran mendalam berkaitan dengan visual dan imej. CNN mampu mengesan dan menganalisis corak dengan mengaplikasikan fungsi korteks visual yang terdapat dalam otak manusia. Fungsi korteks visual mempunyai keupayaan untuk mengesan corak objek secara berkala dan hierarki di mana ia mengekstrak fitur imej secara automatik tanpa memerlukan pengekstrakan fitur secara manual seperti dalam pembelajaran mesin.

Terdapat dua bahagian utama dalam CNN iaitu pengekstrakan fitur dan pengelasan seperti dalam Rajah 2.3. Bahagian pengekstrakan fitur terdiri daripada

lapisan Pelingkaran dan unit linear diperbetulkan (*rectified linear unit*) atau ReLu yang mengekstrak output untuk dijadikan input kepada pengelasan (Carrasco & Sicilia-Urbán 2020). Manakala bahagian pengelasan mengandungi penyambungan penuh.



Rajah 2.3 Struktur lapisan CNN untuk pengelasan fraud cukai

Tiga (3) konsep dalam CNN ialah medan penerimaan tempatan (*local receptive fields*), berat dan bias yang dikongsi (*shared weights and biases*), serta pengaktifan dan pengumpulan (*activation and pooling*). Konsep medan penerimaan tempatan dalam rangkaian neural adalah setiap neuron dalam lapisan input disambungkan kepada neuron dalam lapisan tersembunyi (*hidden layers*). Walau bagaimanapun, untuk model CNN hanya kawasan kecil dalam neuron lapisan input yang bersambung ke neuron dalam lapisan tersembunyi yang dirujuk sebagai medan penerimaan tempatan. Medan penerimaan tempatan diterjemahkan merentasi input untuk mencipta peta ciri daripada lapisan input kepada neuron lapisan tersembunyi. Proses ini dipanggil lilitan atau Perlingkaran.

CNN mempunyai neuron dengan berat dan bias. Ia mempelajari nilai berat dan bias semasa proses latihan dan sentiasa mengemas kini setiap contoh latihan yang baharu. Dalam pengaktifan dan pengumpulan, langkah pengaktifan menggunakan transformasi pada output setiap neuron dengan menggunakan fungsi pengaktifan menggunakan ReLu. Ia mengambil output dan memetakannya kepada nilai positif tertinggi atau nilai negatif kepada sifar. Pengumpulan pula mengurangkan dimensi peta ciri dengan memadatkan output daripada kawasan kecil neuron kepada output tunggal. Ini membantu untuk meringkaskan lapisan seterusnya dan mengurangkan bilangan parameter yang perlu dipelajari oleh model.

Prestasi CNN dalam pengelasan imej telah menarik minat pengkaji untuk menggunakan model CNN dalam pengelasan transaksi fraud (Heryadi & Warnars 2018). Kelebihan CNN adalah keupayaannya untuk mengenal pasti kebergantungan citi dalam data input

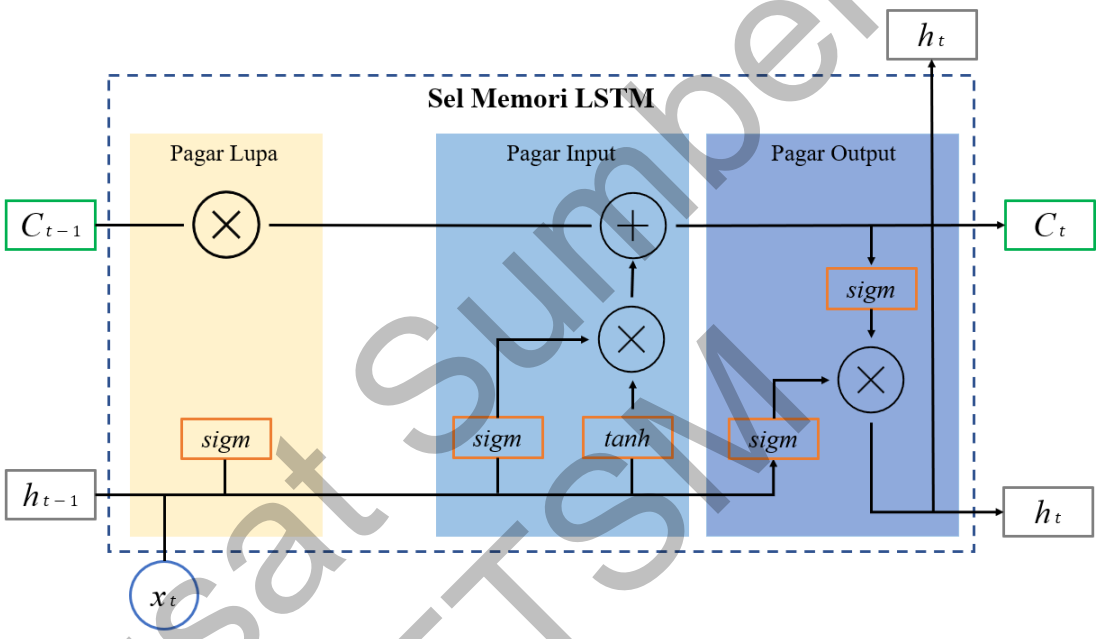
2.6.2 Memori Jangka Masa Panjang dan Pendek (LSTM)

Memori Jangka Masa Panjang dan Pendek (*Long Short Term Memory*, LSTM) merupakan sebahagian daripada Rangkaian Neural Berulang (RNN) yang digunakan untuk mengatasi masalah kecerunan dalam RNN. Algoritma ini membuat pengelasan dengan mempelajari corak urutan sampel.

LSTM memiliki tiga jenis pagar yang menguruskan memori secara berkesan dengan menyimpan maklumat dari sel memori pada pagar iaitu pagar input, pagar lupa (*forget gate*) dan pagar output (Jan 2021). Sel memori ini mampu menyimpan nilai pada selang masa dan ketiga tiga pagar yang mengawal aliran keluar masuk maklumat dari sel memori. Setiap pagar dalam sel memori LSTM adalah terdiri daripada rangkaian neural yang berasingan. Ia bertindak dengan menentukan maklumat yang akan dibawa masuk ke dalam sel memori dan disimpan sebagai memori. Semasa proses ini, pagar akan mempelajari dan menyimpan maklumat yang penting dan mengabaikan maklumat yang tidak penting.

Struktur sel memori LSTM dalam Rajah 2.4 menunjukkan blok memori dengan pagar dan tahap yang berlainan. Tahap sel ialah pautan paling penting dalam rangkaian

aliran maklumat. Tahap sel membenarkan maklumat asal yang tidak diubah suai untuk diteruskan. Pagar lupa (ft) memutuskan sama ada data perlu digugurkan atau tidak. Fungsi *sigmoid* digunakan untuk menyampaikan data di tahap sembunyi sebelumnya (h_{t-1}) dengan data input semasa (x_t). Fungsi *sigmoid* (σ)menentukan nilai antara 0 dan 1 dan jika nilai lebih dekat dengan 0, ia akan menandakan sebagai lupa. Manakala jika nilai lebih dekat dengan 1, ia akan menyimpan data tersebut. Selain itu, tahap vektor sel c_{t-1} menentukan komponen mana yang akan dilupakan (Alghofaili et al. 2020).



Rajah 2.4 Struktur sel memori LSTM

a. Pagar Lupa

Pagar ini berfungsi untuk memproses input dan membuat keputusan sama ada nilai tersembunyi pada sel memori terdahulu $h_t - 1$ akan digunakan dalam sel memori terkini ataupun tidak. Pagar ini juga akan menentukan informasi mana yang perlu dihapuskan daripada sel memori. Pagar lupa juga menentukan nilai output yang diperlukan untuk mengemas kini nilai sel memori.

b. Pagar Input

Pagar input berfungsi untuk memilih dan membuat keputusan nilai maklumat yang akan diimbas oleh pagar output untuk digunakan dalam sel memori. Ia menentukan nilai yang

akan dikemas kini dah mengubah nilai kepada 0 untuk nilai yang tidak penting dan 1 untuk nilai yang penting. Pada pagar input keluaran *sigmoid* akan menentukan maklumat yang penting untuk disimpan daripada keluaran *tanh*.

c. Pagar Output

Pagar output menentukan maklumat daripada pagar input sebelumnya. Pagar output berperanan untuk membaca nilai sel $C_t - 1$ dan membuat keputusan nilai maklumat untuk mengekstrak maklumat berguna daripada sel semasa sebagai output.

BAB III

METODOLOGI

3.1 PENGENALAN

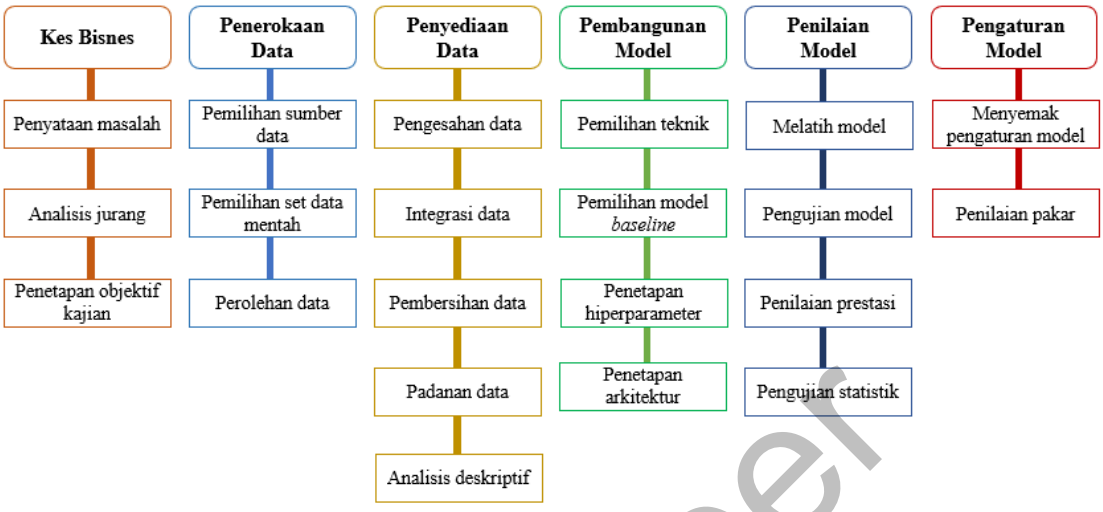
Bab ini menerangkan pendekatan metodologi yang digunakan dalam kajian ini. Kajian ini mengaplikasikan pendekatan *Cross-Industry Standard Process for Data Mining* (CRIPS-DM) dalam metodologi kajian. Bab ini juga memperincikan alatan kajian yang digunakan untuk membangunkan dan menguji model pembelajaran mendalam seperti perisian dan pengaturcaraan bagi perlombongan data dan algoritma pembelajaran mendalam yang terlibat.

3.2 PENDEKATAN KAJIAN

Pendekatan kajian CRIPS-DM merangkumi proses pemahaman bisnes, pemahaman data, penyediaan data, pembangunan model, penilaian model dan aplikasi model.

3.2.1 Kajian Data Sains Melalui Pendekatan CRIPS-DM

Kerangka metodologi bagi kajian ini digariskan berdasarkan pendekatan CRIPS-DM seperti yang telah dibincangkan pada para 1.8 dan Jadual 1.1. Kitaran CRIPS-DM bagi kajian ini mempunyai enam (6) fasa seperti dalam Rajah 3.1. Setiap fasa adalah fleksibel dan boleh diubah suai berdasarkan keputusan yang dicapai di setiap fasa menjadikan ia bersesuaian dengan objektif dan pelaksanaan projek perlombongan data dan kajian data sains (Paula et al. 2017).



Rajah 3.1 Intepretasi fasa kajian

Proses yang terlibat dalam Rajah 3.1 adalah pemahaman kes bisnes, penerokaan data, penyediaan data, pembangunan model, penilaian prestasi model dan pengaturan model dalam persekitaran bisnes. Setiap proses yang diterangkan secara berperingkat dalam bab ini dan bab seterusnya seperti dalam Jadual 2.1 di bawah.

Jadual 3.1 Fasa kajian CRIPS-DM

Fasa	**Deskripsi Aktiviti**	**Perbincangan**
Pemahaman Bisnes	Perbincangan berkenaan permasalahan dan jurang dalam bidang kajian.	Bab 1 dan 2
Pemahaman Data	Proses dapatan data yang tersedia untuk kajian dan pemahaman mengenai set data mentah daripada industri.	Bab 1 dan 2
Penyediaan Data	Penerokaan hubungan data dan proses pembersihan dan pra-pemprosesan data untuk pembangunan model pembelajaran mendalam.	Bab 3
Pembangunan Model	Mendapatkan model pengelasan melalui pembangunan algoritma model pengelasan dan penambahbaikan prestasi pengelasan model daripada fitur dan hiperparamater.	Bab 3 dan 4
Penilaian Model	Analisis terhadap model pengelasan pengesanan yang dibangunkan melalui kaedah penilaian metrik dan pengesahan pakar.	Bab 4
Pengaturan Model	Model dan kajian dibentangkan kepada pihak pemegang taruh sebagai garis panduan dan model prototaip. Penambahbaikan model berdasarkan data input semasa dan penilaian model.	Bab 4 dan 5

3.2.2 Alatan Kajian

Kajian ini melibatkan pengguna pelbagai perisian dan aplikasi di setiap fasa metodologi bagi tujuan pembangunan model pengesanan pengelak cukai. Bagi proses perolehan data, *MySQL* digunakan untuk mendapatkan data mentah dalam Sistem MyGST yang digunakan oleh JKDM. Set data mentah dijana ke dalam bentuk tabular oleh *Mircosoft Excel*. Bagi tujuan fasa penyediaan data, pembangunan model dan eksperimen kajian dilaksanakan dengan bahasa pengaturcaraan Python 3.7.13 melalui pelbagai perpustakaan digital seperti *Numpy*, *Pandas*, *Keras 2.8*, *Scikit-Learn* dan *Tensorflow*.

Pengaturcaraan Python berjalan pada platform *Google Colabotary Pro* (Google Colab) yang menawarkan sokongan GPU dan TPU dalam talian melalui persekitaran *Jupyter Notebook*. GPU membolehkan pemprosesan dan pembangunan model pembelajaran mendalam yang lebih pantas bersesuaian dengan keperluan model pembelajaran mendalam yang melibatkan jumlah data yang besar dan boleh mengambil masa yang lebih lama untuk belajar. Visualisasi set data untuk analisis deskriptif dijalankan menggunakan perisian Tableau yang mampu menjana plot dan graf bagi data bersaiz besar dengan mudah.

Ringkasan perisian dan aplikasi yang digunakan dalam kajian ini adalah seperti penerangan dalam Jadual 3.3 di bawah.

Jadual 3.2 Perisian dan pengaturcaraan yang digunakan untuk kajian ini

Bil	Perisian/Pengaturcaraan	Kegunaan
1	*MySQL*	Aktiviti capaian data daripada sistem MyGST
2	*Microsoft Excel*	Muat turun dan pengumpulan data
3	Tableau 2021.1	Visualisasi set
4	Google Colab Pro	Pembersihan data, pemprosesan data, pembangunan model pembelajaran mendalam dan pengujian model
5	Python	Pengaturcaraan untuk analisa data

3.3 KES BISNES DAN PENEROKAAN DATA

Fasa permulaan adalah untuk mengenal pasti kes bisnes dan memahami keperluan kes bisnes bagi tujuan penetapan matlamat dan objektif kajian. Analisis jurang dijalankan untuk memahami dengan lebih lanjut penetapan kajian dan metodologi yang ditetapkan pada kajian melalui kajian literatur dan penerokaan data. Kes bisnes yang dikenal pasti adalah keperluan untuk menambah baik dan memperkasakan kaedah pengesanan pengelakan fraud sedia ada secara automatik dan lebih berkesan untuk membantu JKDM mengoptimumkan kutipan hasil cukai.

Fasa penerokaan data bermula dengan mengenal pasti data yang tersedia untuk kajian dan seterusnya merancang untuk mendapatkan data. Dalam kajian ini, set data mentah daripada industri memerlukan penyelidik untuk mengenal pasti apakah set data yang tersedia untuk dikaji dan apakah proses-proses yang akan terlibat daripada pemerolehan data sehingga penyediaan data untuk tujuan pembangunan model. Data juga perlu disahkan oleh pakar sebelum proses seterusnya dijalankan.

Keperluan kajian ini diselaraskan berdasarkan pemahaman bisnes dan keperluan data yang sesuai. Berdasarkan objektif kajian, pemilihan data yang bersesuaian adalah penting. Data yang diperlukan ialah data penyata cukai, transaksi penyata cukai dan maklumat pembayar cukai. Bagi tujuan kajian ini, data daripada JKDM sebagai pemungut cukai GST adalah bersesuaian dan medan atribut yang dipilih hendaklah mewakili transaksi cukai oleh pembayar cukai.

3.4 PENYEDIAAN DATA

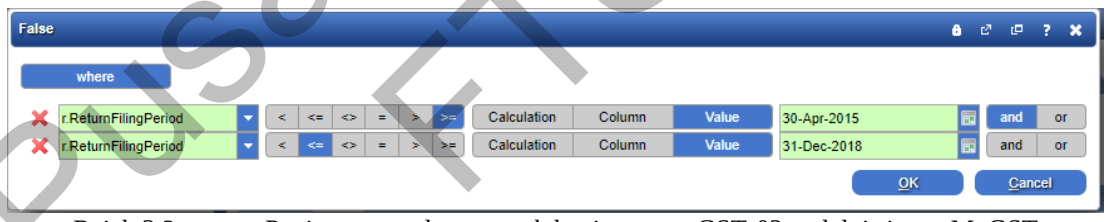
Penyediaan data bagi model pengesan pengelak cukai melibatkan beberapa proses bermula daripada perolehan data, pembersihan data, integrasi data, padanan data, tapisan data, analisis deskriptif dan pengelasan data.

Penyediaan data memainkan peranan penting dalam kajian ini terutamanya untuk pembangunan model. Terdapat keperluan untuk menyemak penyediaan data semula selepas pembangunan model dan penilaian prestasi model.

3.4.1 Perolehan Data

Data yang digunakan dalam kajian ini ialah data penyata cukai sebenar yang telah dipohon dan dibekalkan secara rasmi oleh JKDM bagi tujuan penyelidikan ini sahaja dengan surat kelulusan di Lampiran A. Data mentah yang diperoleh adalah daripada borang GST-01 bagi maklumat pendaftaran sebanyak 688,747 baris data dan borang GST-03 bagi maklumat penyata GST dengan markah petunjuk risiko sebanyak 12,596,583 baris data daripada tempoh April 2015 sehingga Oktober 2018. Data mentah dipadankan dan dibuat pemprosesan menjadikan hanya 2,637,078 baris data dan 24 atribut sebagai set data penyata cukai bagi tujuan kajian ini iaitu sebanyak 21 peratus set data kajian daripada set data mentah yang diperoleh. Pengurangan data yang banyak adalah disebabkan duplikasi semasa perolehan data dan dibincangkan di para 3.4.3.

Data yang dibekalkan adalah data dalaman yang berstruktur yang diperoleh daripada modul penyata, modul pendaftaran dan modul pengurusan risiko dalam sistem MyGST. Data dijana melalui medan pertanyaan data yang menggunakan Penyataan Bahasa Pertanyaan Berstruktur (*Structured Qeury Languange*, SQL) seperti Rajah 3.2 dan contoh data mentah yang dijana adalah seperti Rajah 3.3.



Rajah 3.2 Penjanaan set data mentah bagi penyata GST-03 melalui sistem MyGST

Account ID	Return Filing	Account Filir	Standard Ra	Output Tax	Standard Ra	Input Tax	Amount Clai	Amount Pay	Local Supplie	Export Suppl	GST Relief St	Exempt Supp	Goods Impo	Suspended C	Capital Goo	Bad Debt Re	Ba
	30-Apr-15	Monthly	663157.59	39789.47	413934.99	24836.09	0	14953.38	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	51800	3108	70809.87	4304.78	1196.78	0	0	0	3299869.9	0	0	0	0	0	
	30-Apr-15	Monthly	3905825.4	234349.54	44486.35	2669.19	0	231680.35	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	3905825.4	234349.54	44486.35	2669.19	0	231680.35	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	3905825.4	234349.54	44486.35	2669.19	0	231680.35	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	484.91	29.09	0	0	0	29.09	50	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	484.91	29.09	0	0	0	29.09	50	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	178932.65	10735.96	136709.76	8202.59	0	2533.37	13020	1250455.4	0	0	299253.43	17955.21	0	0	
	30-Apr-15	Monthly	178932.65	10735.96	136709.76	8202.59	0	2533.37	13020	1250455.4	0	0	299253.43	17955.21	0	0	
	30-Apr-15	Three Montl	173732.22	9221.45	42622.91	2557.47	0	6663.98	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	42617.92	2557.08	8694.54	521.68	0	2035.4	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	42617.92	2557.08	8694.54	521.68	0	2035.4	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	1431746.6	85904.79	912268.43	54736.09	0	31168.7	0	125193.14	0	0	0	0	0	0	
	30-Apr-15	Monthly	45	2.7	35221.95	2107.82	2105.12	0	1004721.5	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	35713.7	2142.82	1077708.7	64662.56	62519.74	0	0	10808633	0	0	0	0	528709.72	0	
	30-Apr-15	Monthly	291596.32	17495.78	57176.65	3430.59	0	14065.19	1512.4	0	0	107.73	0	0	0	0	
	30-Apr-15	Monthly	380836.2	22850.17	460636.18	27638.17	4788	0	0	0	34171.2	0	0	0	0	0	
	30-Apr-15	Monthly	380836.2	22850.17	460636.18	27638.17	4788	0	0	0	34171.2	0	0	0	0	0	
	30-Apr-15	Monthly	663157.59	39789.47	413934.99	24836.09	0	14953.38	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	723060	43483.34	687893.26	41273.59	0	2209.75	0	0	35595.72	0	0	0	0	0	
	30-Apr-15	Monthly	0	0	50264.83	3015.92	3015.92	0	0	931907.56	0	0	0	0	550	0	
	30-Apr-15	Monthly	3166720	190024.37	328623.83	13655.19	0	176369.18	0	0	0	1678563	0	0	0	0	
	30-Apr-15	Monthly	51800	3108	70809.87	4304.78	1196.78	0	0	0	3299869.9	0	0	0	0	0	
	30-Apr-15	Monthly	3905825.4	234349.54	44486.35	2669.19	0	231680.35	0	0	0	0	0	0	0	0	
	30-Apr-15	Monthly	3905825.4	234349.54	44486.35	2669.19	0	231680.35	0	0	0	0	0	0	0	0	

Rajah 3.3 Contoh set data mentah daripada borang GST-03

Data yang diperoleh daripada modul penyata ialah maklumat penyata cukai yang dihantar secara bulanan dan suku tahunan. Bagi modul pendaftaran pula ia mengandungi maklumat kod industri pembayar cukai, jenis pendaftaran perniagaan dan jenis perniagaan. Data daripada modul pengurusan risiko adalah markah petunjuk risiko yang daripada sistem penarafan risiko dalam oleh sistem MyGST yang mengelaskan penyata cukai sebagai risiko rendah dan risiko tinggi. Senarai data mentah yang diperoleh dan keterangan atribut adalah seperti Jadual 3.3.

Jadual 3.3 Atribut dan diskripsi medan data

Bil	Perolehan Data	Medan (Atribut)	Diskripsi Medan
1	Borang Penyata Cukai GST-03	Account ID	Nombor GST
2		Return Filing Period	Tempoh penghantaran penyata cukai
3		Account Filing Frequency	Tempoh penyata cukai
4		Standard Rated Supply	Jumlah pembekalan bercukai
5		Output Tax	Jumlah cukai yang diperlu dibayar berdasarkan 6% daripada pembekalan bercukai
6		Standard Rated Acquisition	Jumlah perolehan bercukai
7		Input Tax	Jumlah cukai yang dikeluarkan berdasarkan 6% daripada perolehan bercukai
8		Amount Claimable	Jumlah tuntutan pulangbalik cukai
9		Amount Payable	Jumlah cukai yang perlu dibayar
10		Local Supplies	Jumlah pembekalan dalam negara
11		Export Supplies	Jumlah pembekalan luar negara, eksport
12		GST Relief Supplies	Jumlah pembekalan yang dilepaskan daripada cukai
13		Exempt Supplies	Jumlah pembekalan yang dikecualikan
14		Goods Imported	Jumlah perolehan yang diimport
15		Suspended GST	Jumlah pembekalan yang dilevikan
16		Capital Goods Acquired	Jumlah pembelian aset fizikal bagi tujuan perniagaan
17		Bad Debt Relief	Jumlah pelepasan hutang lapuk
18	Petunjuk Risiko	GST ID	Nombor GST
19		Lead Amount	Markah petunjuk risiko yang dijana oleh Modul Pengurusan Risiko MyGST
20	Borang Pendaftaran GST-01	GST ID	Nombor GST
21		Customer Type	Jenis pendaftaran perniagaan
22		Registration Type	Jenis pendaftaran GST
23		Section	Industri pembekalan
24		MSIC Code	Kod *Malaysia Standard Industrial Classification* (MSIC)
25		Total Turnover	Jumlah nilai ambang tahunan
26		State Office	Kod negeri mengawal

3.4.2 Integrasi Data

Pembersihan data dimulakan dengan integrasi data penyata cukai yang telah dimuat turun berdasarkan tempoh penyata cukai dalam format *.csv* daripada *Microsoft Excel* dan mengandungi setiap tempoh penyata cukai bermula daripada April 2015 sehingga Oktober 2018. Terdapat 43 buah fail *.csv* bagi 43 tempoh bercukai yang mengandungi penyata cukai dan markah petunjuk risiko dan satu *.csv* fail yang mempunyai maklumat pendaftaran pembayar cukai daripada borang pendaftaran GST-01.

Proses integrasi data dalam kajian ini melibatkan penggabungan kesemua fail tersebut dalam satu fail bagi memudahkan proses pembersihan dan penerokaan data dan dijalankan menggunakan pengaturcaraan Python seperti dalam Rajah 3.4 dibawah.

```
import glob
import pandas as pd

data = pd.concat(map(pd.read_csv, glob.glob('/content/*.csv')))
data.head()
```

	Account ID	Return Filing Period	Account Filing Frequency	Standard Rated Supply	Output Tax	Standard Rated Acquisition	Input Tax	Amount Claimable	Amount Payable	Local Supplies	Export Supplies	GST Relief Supplies	Exempt Supplies	Goods Imported	Suspended GST	Capital Goods Acquired	Bad Debt Relief	Bad Debt Recovered	Lead Amount
0	[redacted]	31-Dec-15	Three Monthly	61235.97	3674.18	21408.56	1284.46	0.00	2389.72	0.0	0.00	0.0	0.00	0.00	0.00	0.0	0.0	0.0	61.0
1		31-Dec-15	Monthly	293740.50	17624.43	500046.19	30002.77	12378.34	0.00	5340.0	2108942.34	0.0	11682.59	679237.38	40754.24	58500.0	0.0	0.0	43.0
2		31-Dec-15	Monthly	110729.01	6643.75	14813.73	888.82	0.00	5754.93	0.0	0.00	0.0	0.00	0.00	0.00	0.0	0.0	0.0	18.0
3		31-Dec-15	Monthly	110729.01	6643.75	14813.73	888.82	0.00	5754.93	0.0	0.00	0.0	0.00	0.00	0.00	0.0	0.0	0.0	48.0
4	1069056	31-Dec-15	Three Monthly	119747.18	7184.86	66693.38	4001.56	0.00	3183.30	0.0	0.00	0.0	0.00	0.00	0.00	0.0	0.0	0.0	113.0

Rajah 3.4 Proses integrasi data menggunakan pengaturcaraan python

3.4.3 Pembersihan Data

Setelah proses integrasi dijalankan, penelitian ke atas data mentah mendapati terdapat data duplikasi yang perlu dibersihkan. Walau bagaimanapun tiada data kosong (*missing values*) yang dikenal pasti. Pembersihan data dijalankan untuk membuang data duplikasi tersebut. Baki baris data adalah sebanyak 2,646,828 selepas pembersihan dibuat.

Data penyata cukai yang diperoleh daripada sistem MyGST telah dibuat padanan dengan atribut markah petunjuk risiko sebelum dimuat turun yang menyebabkan berlakunya duplikasi ke atas kebanyakan penyata cukai. Bagi tujuan ini semasa proses awal pembersihan data, nombor cukai dan tempoh penyata cukai diperlukan bagi tujuan pembersihan duplikasi seperti pengaturcaraan dalam Rajah 3.5.

Atas persetujuan dengan JKDM, maklumat ini tidak akan didedahkan di sepanjang kajian ini.

```
data = data.drop_duplicates(subset=['Account ID', 'Return Filing Period', 'Account Filing Frequency'])
```

Rajah 3.5 Proses pembersihan data menggunakan pengaturcaraan python

3.4.4 Padanan dan tapisan data

Bagi padanan data dengan data borang pendaftaran GST-01 yang mengandungi maklumat pendaftaran, padanan telah dibuat menggunakan *inner join* setelah proses membuang data duplikasi di para 3.4.3 selesai dijalankan. Terdapat empat (4) medan yang akan digabungkan ke dalam set data iaitu 'MSIC', '*Sector'*, '*Customer Type'* dan '*Registration Type'*.

Hasil daripada proses pemadanan dan tapisan data terdapat sebanyak 2,637,078 baris data yang tersedia untuk kajian ini daripada keseluruhan 12,596,583 baris data sebelum ditapis. Atribut '*Account ID*' juga digugurkan kerana tidak diperlukan untuk membangunkan model selain untuk menjaga kerahsiaan data. Rajah 3.6 menunjukkan kod untuk proses integrasi data.

```
new_data = pd.merge(data,reg_data, left_on='Account ID',right_on='GST ID')
```

```
new_data = new_data.drop(['Account ID'], axis = 'columns')
```

```
new_data.shape
```

```
(2637078, 27)
```

Rajah 3.6 Proses integrasi data menggunakan pengaturcaraan python

Pemerhatian ke atas set data penyata cukai menunjukkan medan atribut dalam bentuk kategori seperti dalam Jadual 3.4. Model pembelajaran mendalam perlu dikaji menggunakan set data dalam bentuk numerik.

Jadual 3.4 Jenis data dalam set data penyata cukai

Data Numerik	Data Nominal
Standard Rated Supply	*Account Filing Period*
Output Tax	*Account Filing Frequency*
Standard Rated Acquisition	*MSIC*
Input Tax	*Sector*
Amount Claimable	*Registration Type*
Amount Payable	*Customer Type*
Local Supplies	*Total Turnover*
Export Supplies	*State Office*
GST Relief Supplies	
Exempt Supplies	
Goods Imported	
Suspended GST	
Capital Goods Required	
Bad Debt Relief	
Bad Debt Recovered	

3.4.5 Pengelasan Data

Berdasarkan skor markah risiko yang ditetapkan, penyata cukai akan dihantar untuk audit meja jika skor markah melebihi nilai ambang atau penyata cukai boleh terus diproses untuk diterima. Bagi tujuan pengelasan atribut kelas, markah petunjuk risiko digunakan untuk mengelas penyata cukai yang risiko rendah kepada 0 dan penyata cukai risiko tinggi sebagai 1. Pengelasan ini dibuat berdasarkan ketetapan sedia ada JKDM mengikut jenis penyata dan skor markah risiko seperti dalam Rajah 3.7.

```
filters = [
    ((df['Account Filing Frequency'] ==1) & (df['Lead Amount'] >    )),
    ((df['Account Filing Frequency'] ==3) & (df['Lead Amount'] >    ))
]

values = ["1", "1"]
```

Rajah 3.7 Proses pelabelan data kepada 0 – risiko rendah, 1 – risiko tinggi

Selepas proses pengelasan dijalankan, atribut *'Lead Amount'* digugurkan. Medan atribut lain diberikan nama baru yang lebih generik untuk memudahkan untuk memudahkan proses penerokaan dan pemodelan data. Jadual 3.5 menunjukkan definisi medan atribut yang telah dinamakan.

Jadual 3.5 Medan bagi set data penyata cukai

Bil	Medan Baru	Atribut Asal
1	Period	*Return Filing Period*
2	V2	*Account Filing Frequency*
3	V3	*Standard Rated Supply*
4	V4	*Output Tax*
5	V5	*Standard Rated Acquisition*
6	V6	*Input Tax*
7	V7	*Amount Claimable*
8	V8	*Amount Payable*
9	V9	*Local Supplies*
10	V10	*Export Supplies*
11	V11	*GST Relief Supplies*
12	V12	*Exempt Supplies*
13	V13	*Goods Imported*
14	V14	*Suspended GST*
15	V15	*Capital Goods Acquired*
16	V16	*Bad Debt Relief*
17	V17	*Bad Debt Recovered*
18	V18	*MSIC Code*
19	V19	*Section*
20	V20	*Customer Type*
21	V21	*Registration Type*
22	V22	*Total Turnover*
23	V23	*State Office*
24	V24	*Class*

3.4.6 Analisis Deskriptif Bagi Set Data Penyata Cukai

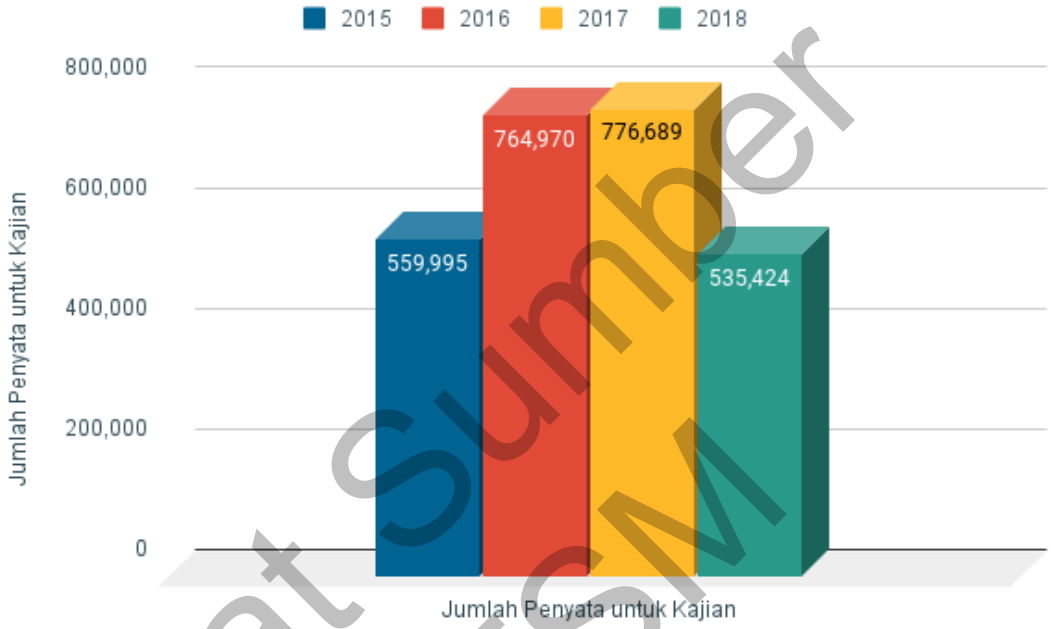
Analisis deskriptif menerangkan secara lebih terperinci mengenai set data penyata cukai daripada aspek statistik. Analisa deskriptif dapat memberikan gambaran data yang lengkap dalam bentuk jadual, numerik dan gambar rajah.

a. Set data penyata cukai

Set data penyata cukai yang lengkap setelah menjalani pra-pemprosesan mengandungi 2,637,078 baris data dan 24 atribut termasuk atribut kelas. Set data tersebut merangkumi data daripada borang penyata cukai GST-03, petunjuk risiko GST dan borang pendaftaran GST-01 daripada April 2015 sehingga Oktober 2018 untuk tempoh 43 bulan. Pecahan setiap tahun bercukai adalah seperti di Jadual 3.6 dan Rajah 3.8. Tahun 2015 dan 2018 mempunyai jumlah penyata yang kurang kerana GST hanya mula diperkenalkan pada April 2015 dan kemudiannya dimansuhkan pada Oktober 2018.

Jadual 3.6 Jumlah data MyGST mengikut tahun bercukai

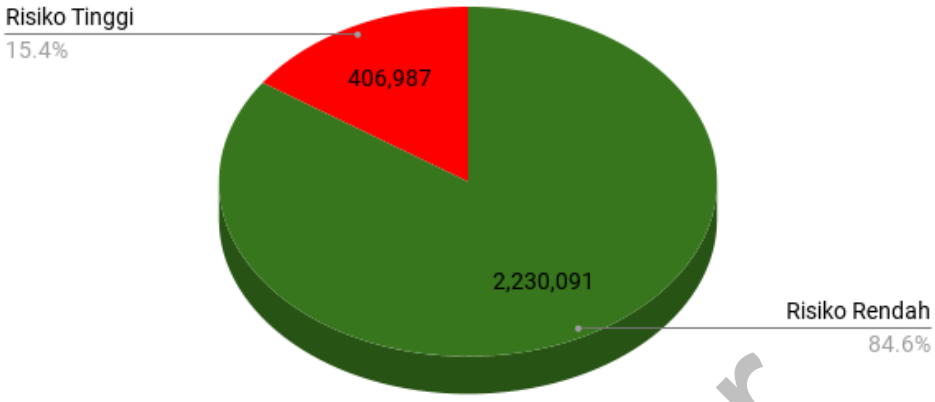
Bil	Tahun Bercukai	Jumlah Penyata untuk Kajian
1	2015	559,995
2	2016	764,970
3	2017	776,689
4	2018	535,424
	Jumlah	**2,637,078**



Rajah 3.8 Jumlah data MyGST mengikut tahun bercukai

b. Atribut '*Class*'

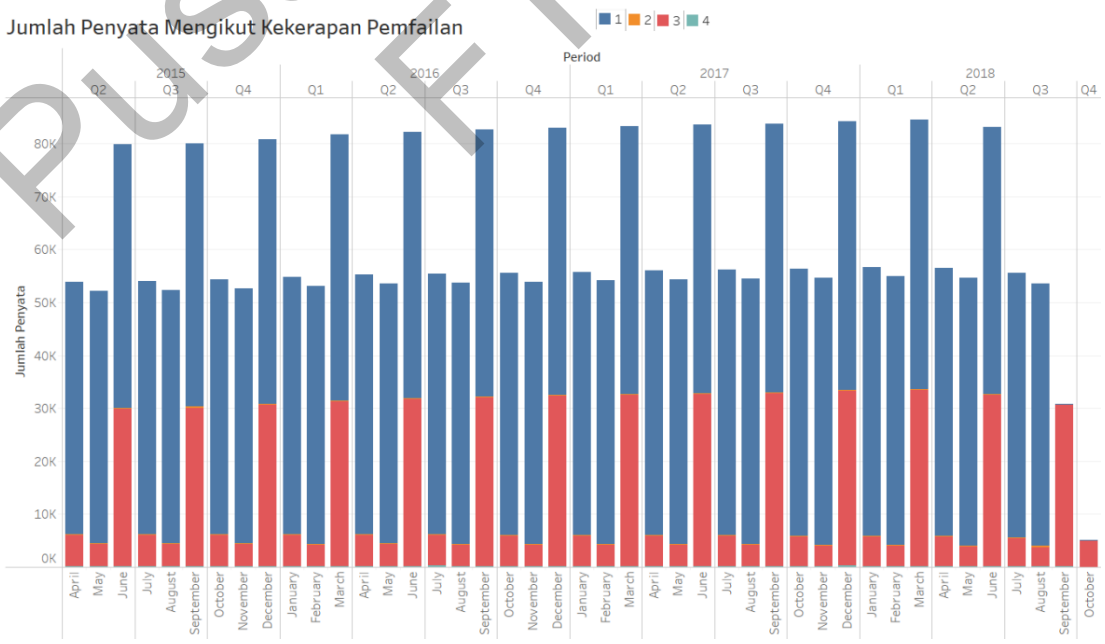
Seperti yang dinyatakan di para 3.4.5, atribut '*Class*' ialah label kelas bagi penyata cukai dengan label risiko rendah dan risiko tinggi melakukan pengelakan cukai. Rajah 3.9 menunjukkan pecahan atribut kelas mewakili 0, risiko rendah dan 1, risiko tinggi sebanyak 84.6% dan 15.4%.



Rajah 3.9 Pecahan atribut ‘Class’ berdasarkan label kelas 0,1

c. Atribut ‘Period’

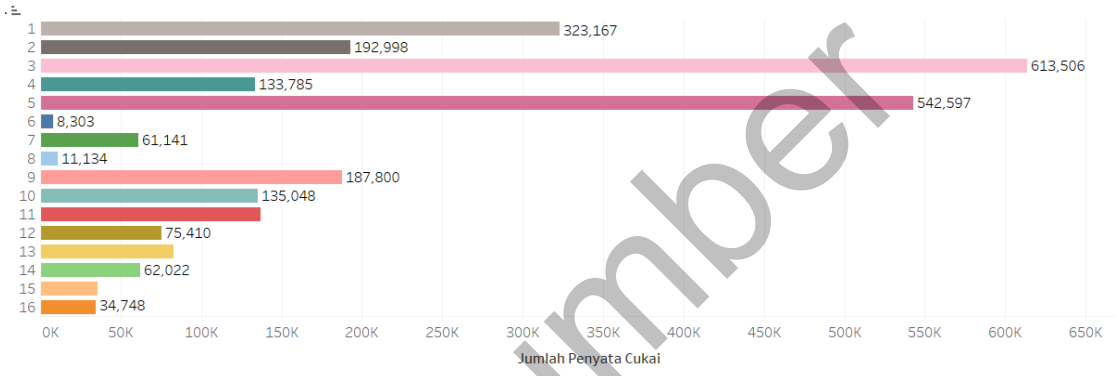
Rajah 3.10 menunjukkan taburan jumlah penyata GST berdasarkan tempoh bercukai iaitu bulanan, dwibulanan, suku tahun dan pelbagai. Dapat dilihat di rajah di bawah tempoh bercukai bulanan mewakili tempoh bercukai yang terbesar untuk penghantaran penyata GST dan terdapat penambahan jumlah penyata bagi setiap bulan ketiga bagi penyata suku tahun. Terdapat perbezaan pada jumlah penyata kerana kekerapan pemfailan adalah berbeza. Hal ini mempengaruhi prestasi model pembelajaran mendalam dan akan dibincangkan pada para 3.4.7.



Rajah 3.10 Taburan rekod penyata GST mengikut tempoh penyata cukai

d. Jumlah Penyata Cukai Mengikut Stesen Pengawal (Atribut 'V23')

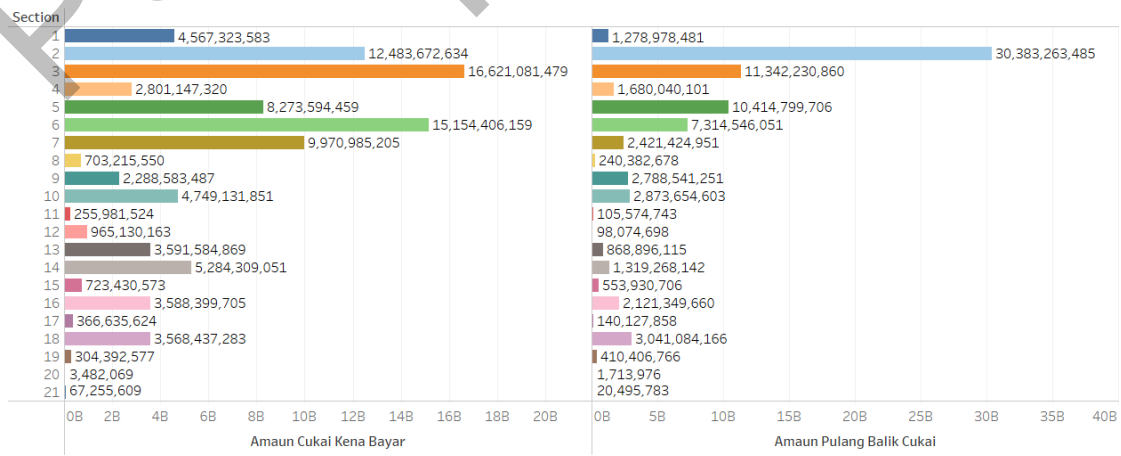
Merujuk kepada Jadual 3.5 di atas, atribut 'V23' mewakili stesen pengawal bagi pembayar cukai. Rajah 3.11 di bawah menunjukkan jumlah penyata cukai mengikut stesen pengawal. Stesen pengawal 3 dan 5 merupakan stesen yang paling banyak menerima dan memproses penyata cukai seperti semakan verifikasi dan audit cukai.



Rajah 3.11 Jumlah penyata cukai bagi set data penyata cukai mengikut stesen mengawal

e. Taburan kutipan cukai GST

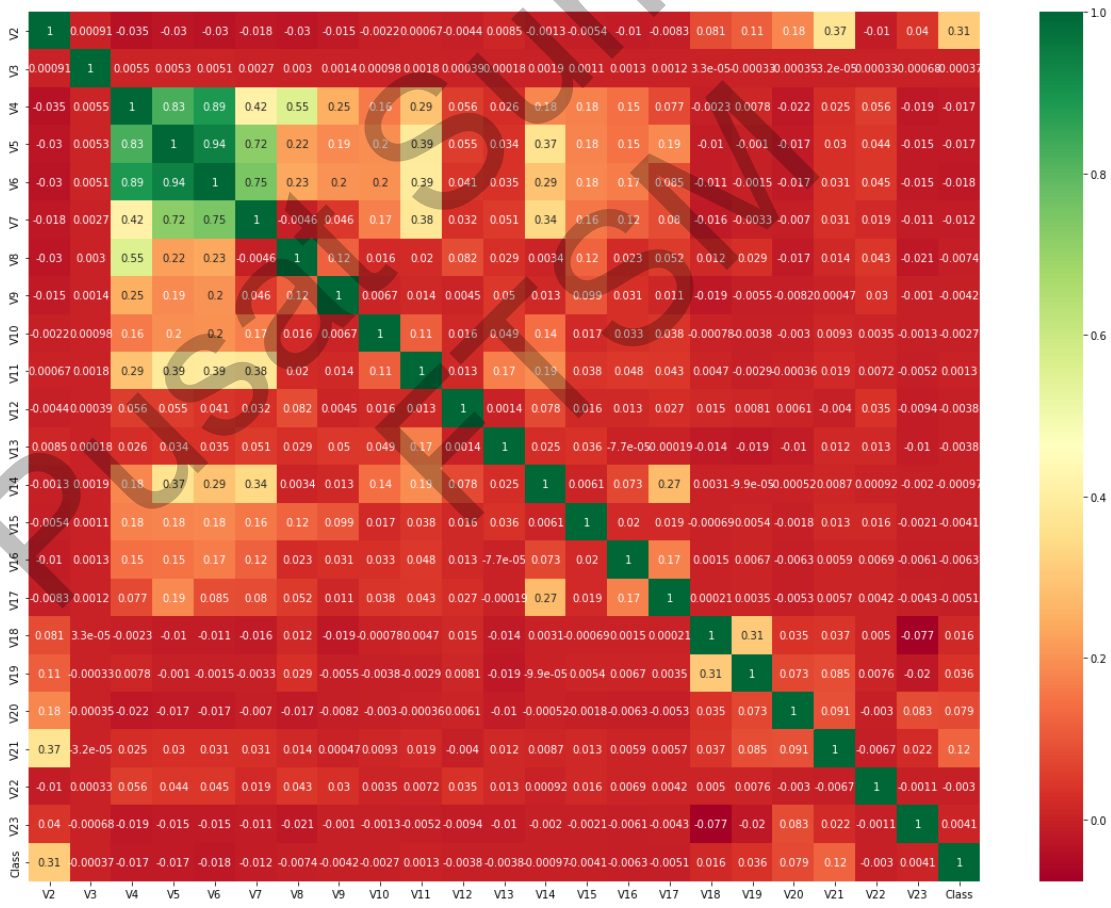
Rajah 3.12 menunjukkan jumlah amaun cukai kena bayar dan amaun pulang balik cukai berdasarkan atribut "V19" yang mewakili industri perniagaan. Industri 3 dan 6 adalah penyumbang terbesar bagi kutipan cukai manakala industri 2 merupakan industri yang mendapat pelepasan cukai.



Rajah 3.12 Taburan rekod penyata GST mengikut tempoh penyata cukai

f. Matriks korelasi set data penyata cukai

Rajah 3.13 menunjukkan matriks korelasi (*correlation matrix*) bagi set data penyata cukai. Berdasarkan matriks korelasi, terdapat atribut yang mempunyai yang mempunyai hubungan yang kuat seperti atribut V4, V5, V6 dan V7. Atribut ini merujuk kepada nilai cukai input dan cukai output yang menentukan amaun pembayaran cukai yang akan dikenakan kepada pembayar cukai dan berpotensi untuk dimanipulasikan bagi pengelakan cukai. Oleh itu, hubungan pemboleh ubahnya adalah tinggi. Selain itu, atribut V2 sebagai kategori pemfailan mempunyai hubungan sederhana dengan atribut V21 dan V22 yang mewakili jenis pendaftaran dan nilai ambang. Hubungan korelasi ini saling berkaitan kerana atribut V21 dan V22 menentukan jenis kategori pemfailan pembayar cukai.



Rajah 3.13 Matriks korelasi antara atribut dan label kelas bagi set data penyata cukai

3.4.7 Set Data Berkategori

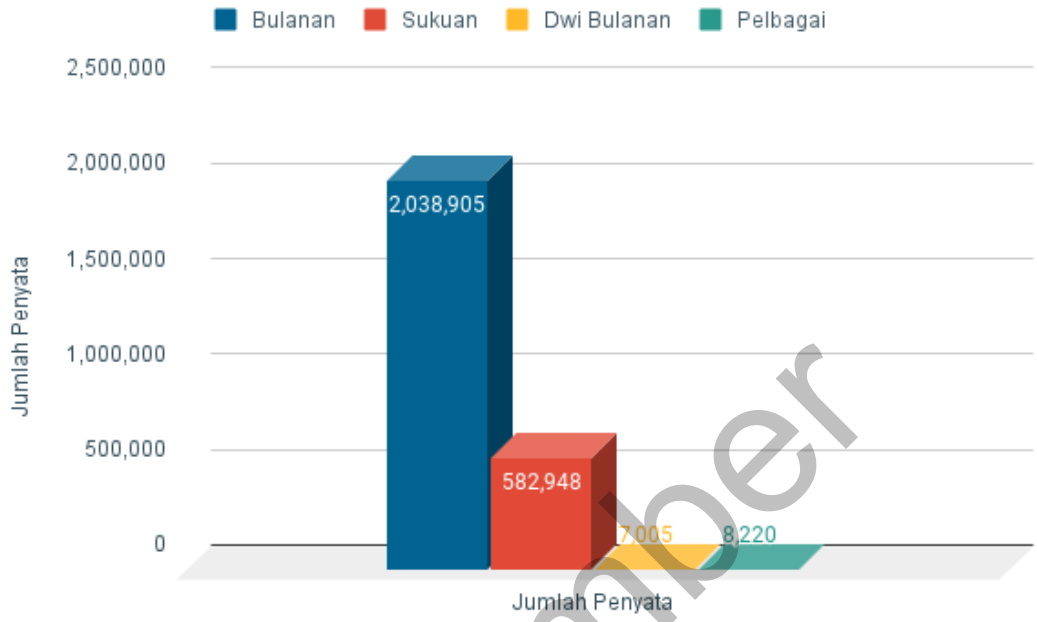
(Rahimikia et al. 2017) dalam kajiannya terhadap pengelakan cukai di Iran membahagikan set data kepada beberapa set data kecil berdasarkan kod industri untuk pensampelan set data kajian kerana masalah data yang tidak seimbang. Bagi kajian ini, set data dibahagikan berdasarkan empat (4) kategori kekerapan pemfailan penyata GST iaitu bulanan, suku tahun, dwibulanan dan pelbagai yang diwakili oleh atribut V2 dalam set data penyata cukai. Set data berkategori ini bertujuan untuk melihat prestasi model pembelajaran mendalam pada set data yang lebih kecil dan peratusan label kelas yang berbeza. Saiz sampel dan label kelas boleh mempengaruhi prestasi ramalan pengelasan pembelajaran mesin dan pembelajaran mendalam (Mary & Claret 2021).

Kategori pemfailan penyata GST ini ditentukan berdasarkan nilai ambang tahunan pembayar cukai dan juga berdasarkan penyata kewangan pembayar cukai. Bagi nilai ambang melebihi RM500,001, pemfailan penyata perlu dilakukan secara bulanan. Manakala pembayar cukai dengan nilai ambang RM500,000 dan ke bawah, pemfailan adalah pada suku tahun. Bagi kategori dwibulanan, ia ditetapkan berdasarkan tempoh penyata kewangan setiap dua (2) bulan dan kategori pelbagai adalah pembayar cukai dari luar negara yang mempunyai tahun kewangan yang berbeza seperti dua (2) minggu dan lain-lain.

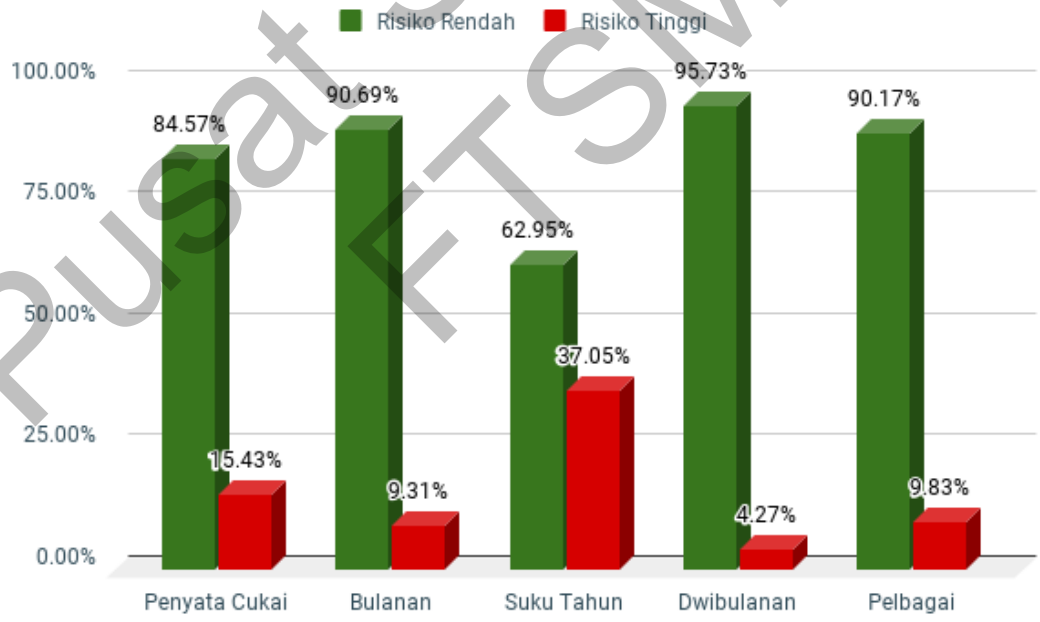
Selain daripada jumlah dan saiz data yang berbeza, set data berkategori ini juga mempunyai peratusan label kelas yang berbeza. Situasi ini boleh mempengaruhi prestasi model. Jumlah penyata bagi setiap kategori diperincikan dalam Jadual 3.7 dan Rajah 3.14 dan Rajah 3.15 menunjukkan peratusan label kelas bagi setiap kategori pemfailan penyata GST berbanding dengan jumlah keseluruhan set data penyata cukai.

Jadual 3.7 Set data berkategori mengikut kekerapan pemfailan GST

Set data berkategori	**Risiko Rendah, 0**		**Risiko Tinggi, 1**		**Jumlah**
Bulanan	1,849,040	90.68%	189,875	9.31%	2,038,905
Suku Tahun	366,943	62.95%	216,005	37.05%	582,948
Dwibulanan	6,706	95.73%	299	4.27%	7,005
Pelbagai	7,412	90.17%	808	9.83%	8,220
Set data penyata cukai	**2,230,091**	**84.57%**	**406,987**	**15.43%**	**2,637,078**



Rajah 3.14 Jumlah penyata cukai mengikut kategori pemfailan



Rajah 3.15 Peratusan label kelas berdasarkan set data penyata cukai dan set data berkategori

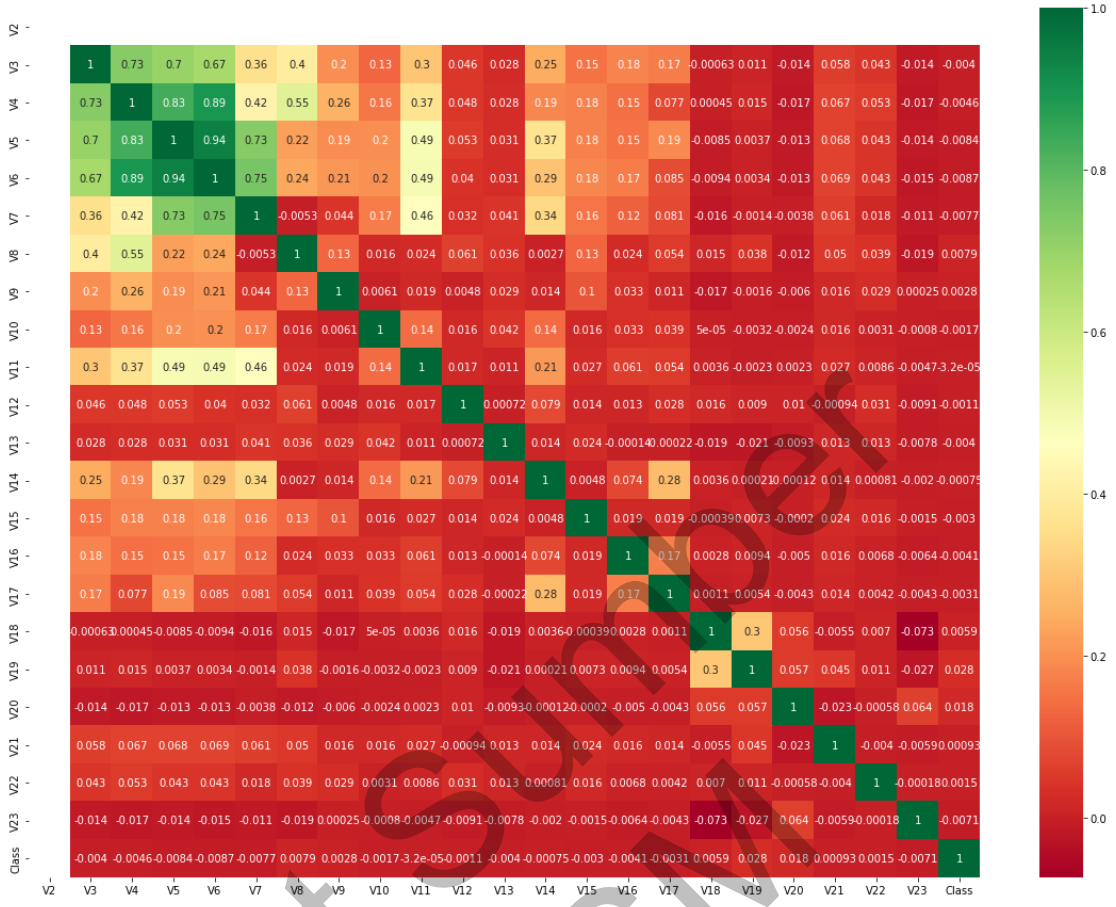
Set data berkategori dwibulanan dan pelbagai mempunyai jumlah data yang paling sedikit iaitu 7,005 dan 8,220 baris data. Manakala set data berkategori bulanan dan suku tahun mempunyai jumlah set data yang besar dan perbezaan peratusan label kelas risiko tinggi dan risiko rendah yang tidak ketara. Perbezaan bagi saiz set data dan

peratusan label kelas bagi set data penyata cukai dan set data berkategori boleh mempengaruhi prestasi model pembelajaran yang dibangunkan. Oleh itu, terdapat keperluan untuk menjalankan kajian terhadap set data penyata cukai dan set data berkategori untuk melihat prestasi model. Matriks korelasi digunakan untuk melihat hubungan linear antara pemboleh ubah berdasarkan set data yang telah dikategorikan.

a. Set Data Berkategori – Bulanan

Set data berkategori bulanan adalah daripada penyata GST-03 yang dihantar oleh pembayar cukai dalam kumpulan pengikraran bulanan. Set data berkategori bulanan mempunyai 2,039,905 baris data dengan peratusan label kelas sebanyak 90.68 peratus risiko rendah dan 9.31 peratus bagi label kelas risiko tinggi.

Matriks korelasi bagi set data berkategori bulanan adalah seperti dalam Rajah 3.16 di bawah. Dapat dilihat kelompokan pemboleh ubah yang mempunyai hubungan yang kuat adalah daripada atribut V3, V4, V5, V6 dan V7. Atribut-atribut tersebut adalah merujuk kepada nilai pengikraran cukai input, cukai output dan V7 sebagai nilai tuntutan pulang balik cukai yang menjadi salah satu faktor pengelakan cukai kerana pembayar cukai berpotensi memanipulasikan nilai-nilai tersebut untuk mengelak pembayaran cukai dan membuat tuntutan cukai yang tidak sepatutnya. Hubungan korelasi bagi atribut yang lain menunjukkan kekuatan hubungan yang lemah dan negatif seperti atribut V18, V19, V20, V21, V22 dan V23 yang merujuk kepada maklumat pendaftaran pembayar cukai.

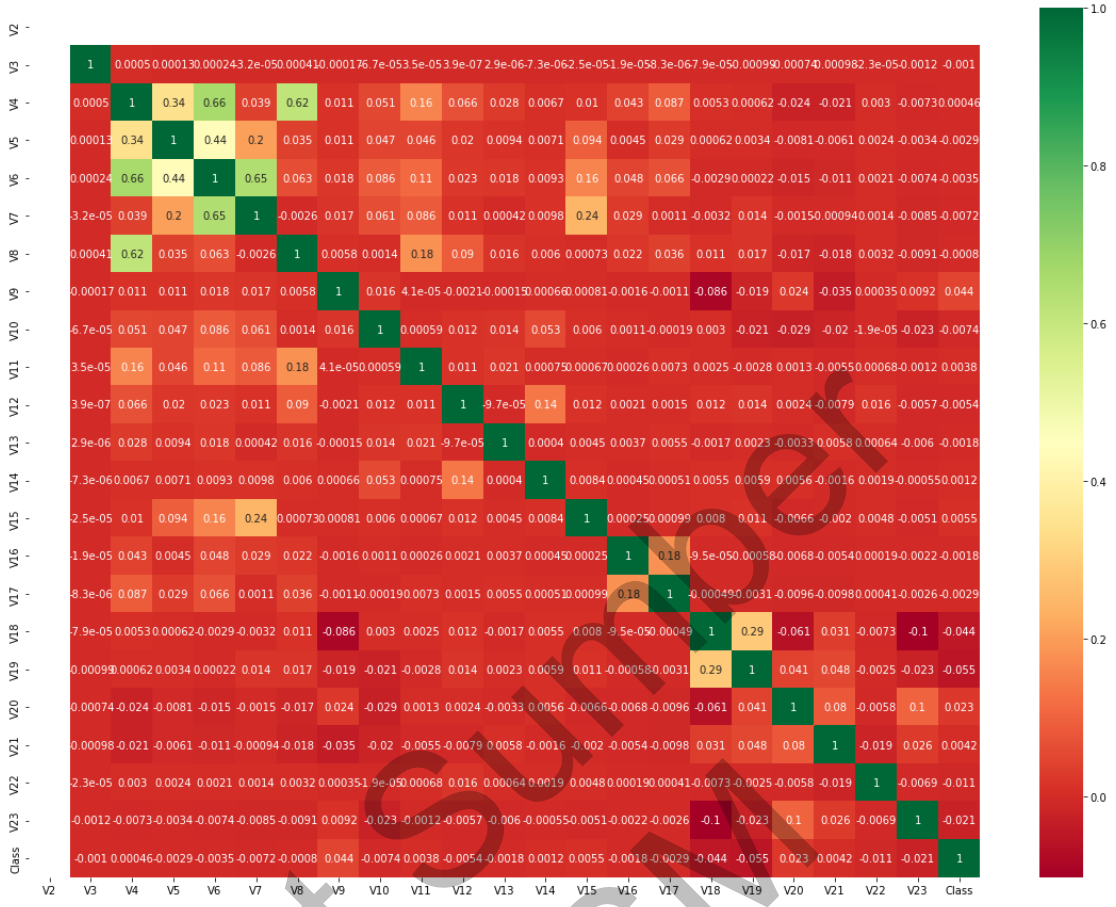


Rajah 3.16 Matriks korelasi antara atribut dan label kelas bagi set data bulanan

b. Set Data Berkategori – Suku Tahun

Set data berkategori suku tahun adalah daripada penyata GST-03 yang dihantar oleh pembayar cukai dalam kumpulan pengikraran suku tahun. Set data ini mempunyai 582,948 baris data dengan peratusan label kelas risiko rendah sebanyak 62.95% dan risiko tinggi sebanyak 37.05.

Rajah 3.17 menunjukkan matriks korelasi bagi set data berkategori suku tahun. Bagi set data berkategori suku tahun, kelompokan hubungan pemboleh ubah yang sederhana dan kuat adalah sedikit dan diwakili oleh atribut V5, V6 dan V7. Atribut yang selebihnya menunjukkan hubungan pemboleh ubah yang lemah dan faktor ini akan mempengaruhi ketepatan model pembelajaran mendalam nanti.



Rajah 3.17 Matriks korelasi antara atribut dan label kelas bagi set data suku tahun

c. Set Data Berkategori – Dwibulanan

Set data berkategori dwibulanan adalah senarai pembayar cukai dengan kategori pemfailan dwibulanan yang diambil daripada set data penyata cukai. Set data berkategori dwibulanan mewakili 7,005 baris data sahaja dengan peratusan label kelas 95.73% bagi risiko rendah dan 4.27% bagi risiko tinggi.

Rajah 3.18 menunjukkan matriks korelasi bagi set data berkategori dwibulanan yang menggambarkan hubungan antara atribut dan label kelas. Matriks korelasi menunjukkan banyak hubungan sederhana dan kuat berbeza dengan set data berkategori bulanan dan suku tahun yang mempunyai lebih banyak hubungan yang lemah dan sederhana. Bagi set data berkategori dwibulanan, atribut V3, V4, V5 dan V6 mempunyai hubungan yang kuat menunjukkan nilai cukai input dan cukai output sangat bergantung kepada label kelas.

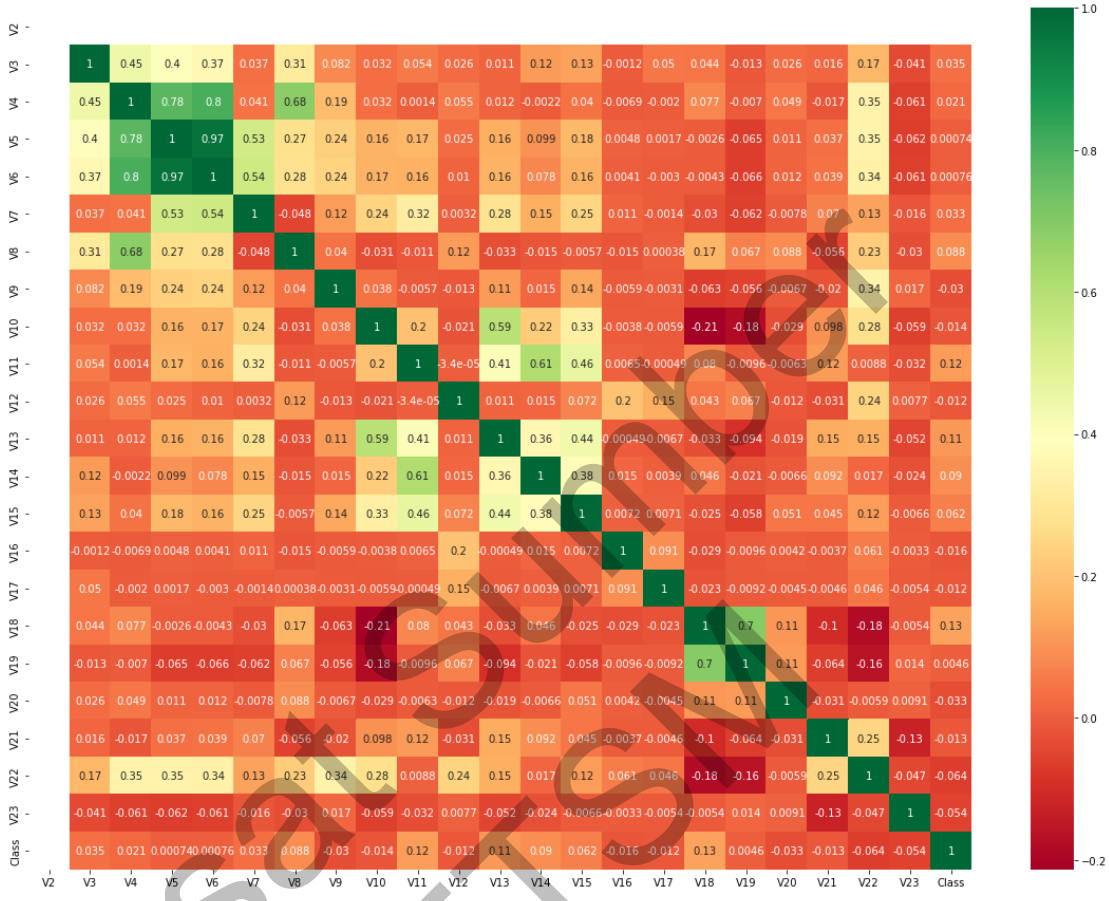
Rajah 3.18 Matriks korelasi antara atribut dan label kelas bagi set data dwibulanan

d. Set Data Berkategori – Pelbagai

Set data berkategori pelbagai adalah data pembayar cukai yang membuat pengikraran mengikut tempoh pelbagai iaitu tidak mempunyai tempoh yang tetap berbeza dengan set data berkategori yang lain. Set data berkategori pelbagai mempunyai 8,220 baris data dengan peratusan kelas label sebanyak 90.17% bagi kelas risiko rendah dan 9.83% bagi kelas label risiko tinggi.

Rajah 3.19 menunjukkan matriks korelasi untuk set data berkategori pelbagai. Matriks korelasi memvisualkan banyak hubungan pemboleh ubah yang sederhana dan kuat sama seperti set data berkategori dwibulanan. Walau bagaimanapun, hanya atribut V4, V5 dan V6 serta V19 yang mempunyai hubungan yang kuat. Atribut V18 dan V19 merujuk kepada sektor dan kod industri pembayar cukai. Bagi pembayar cukai dengan

kategori pelbagai, mereka merupakan pembayar cukai daripada luar negara dan juga pembayar cukai daripada industri khas yang tertakluk kepada kelulusan.



Rajah 3.19 Matriks korelasi antara atribut dan label kelas bagi set data pelbagai

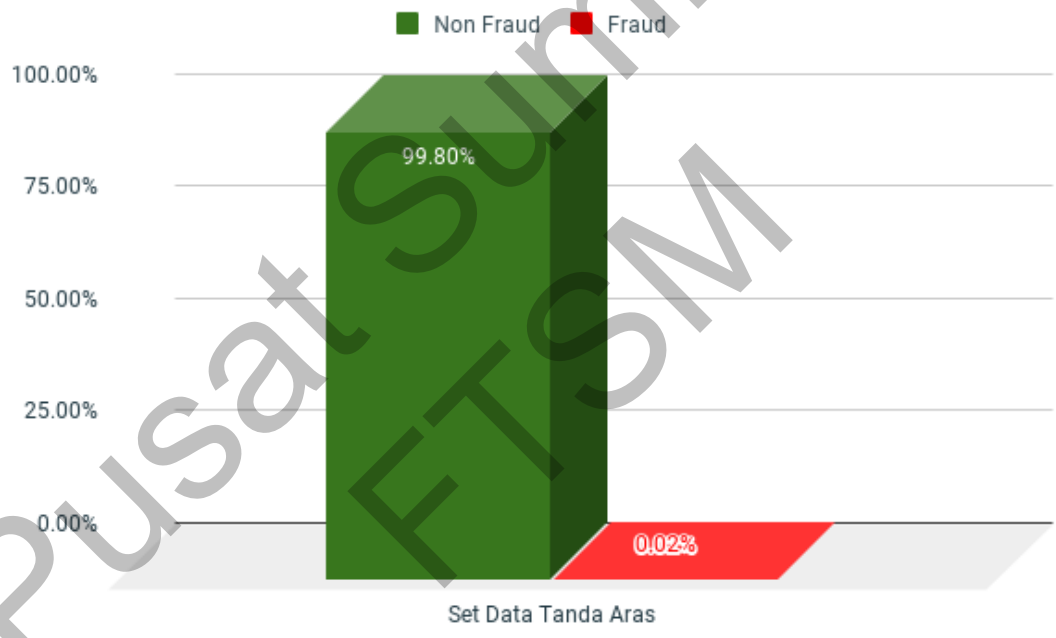
3.4.8 Data Tanda Aras (*Benchmark Data*)

Data tanda aras digunakan sebagai rujukan untuk melihat prestasi kaedah dan kajian yang bersesuaian dengan model pengesanan pengelakan cukai yang dibangunkan. Set data tanda aras dipilih berdasarkan set data yang mempunyai domain dan struktur yang sama dengan set data penyata cukai.

Data percukaian adalah sangat terhad dan selalunya sulit, oleh itu kajian ini menggunakan set data awam daripada domain fraud kewangan iaitu set data fraud kad kredit daripada *UCI Machine Learning* sebagai set data tanda aras. Set data ini mengandungi transaksi yang dibuat dengan kad kredit pada September 2013 oleh pemegang kad Eropah. Ia mempunyai 284,807 baris data dengan 31 atribut termasuk

label kelas fraud dan bukan fraud (*non* fraud) dengan ciri komponen utama yang diperoleh dengan PCA. Set data ini boleh dimuat turun melalui repositori *Kaggle*.

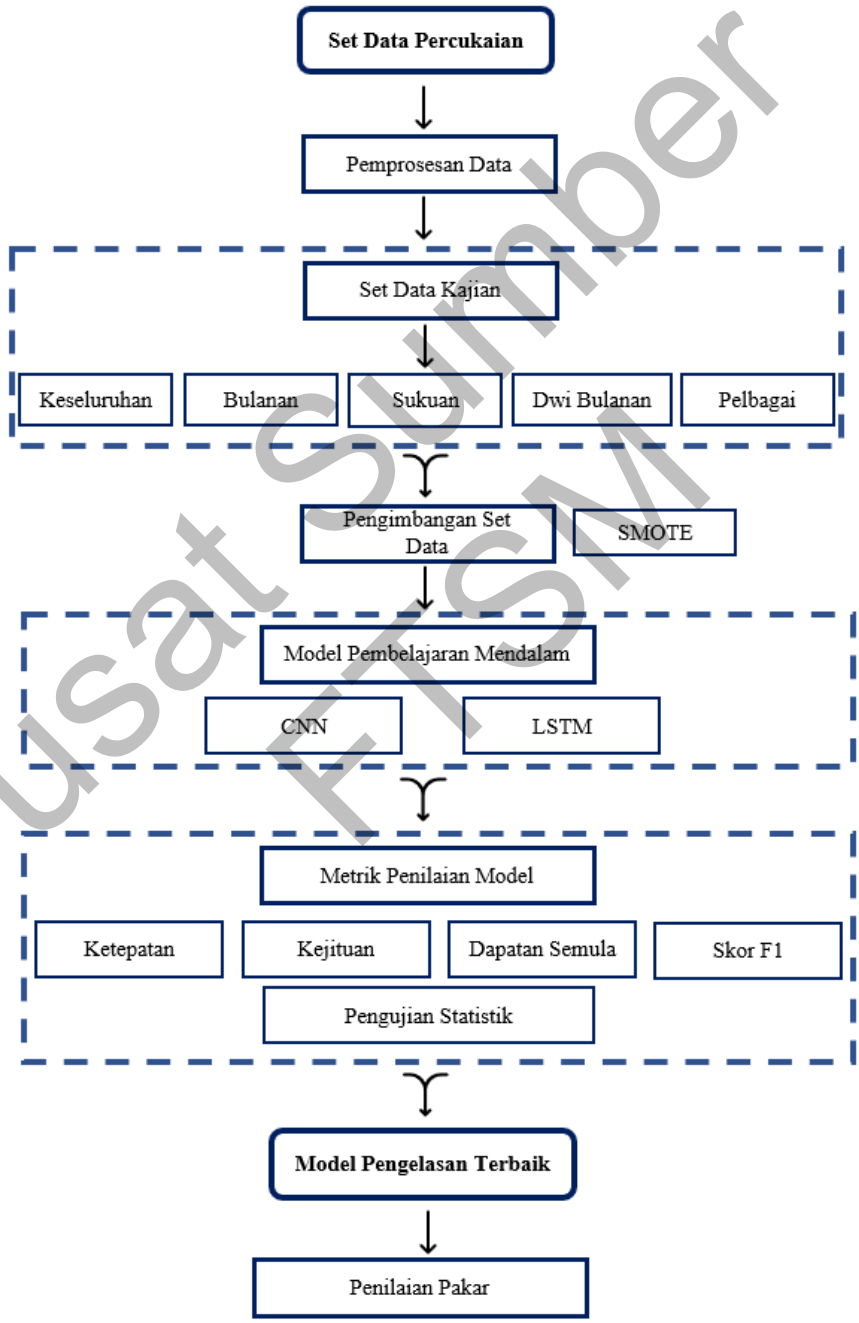
Set data ini mempunyai struktur ciri atribut dan label kelas yang sama untuk domain transaksi kewangan dan amat popular digunakan dalam kajian pengesanan fraud kewangan (Almuteer et al. 2021; Babu & Pratap 2020; Carrasco & Sicilia-Urbán 2020; Nancy et al. 2020; Nguyen et al. 2020; Rahimikia et al. 2017; Roy et al. 2018; Yu et al. 2020). Rajah 3.20 menunjukkan peratusan label kelas bagi set data tanda aras yang sangat tidak seimbang dengan 99.8% kelas bukan fraud (*non fraud*) dan 0.2% kelas fraud.



Rajah 3.20 Peratusan label kelas bagi set data tanda aras

3.5 PEMBANGUNAN MODEL

Fasa pembangunan model adalah saling berkait dengan fasa sebelumnya dan keputusan pada setiap fasa juga mempengaruhi teknik dan arkitektur model yang dibangunkan. Ia mestilah selaras dengan matlamat dan objektif kajian yang telah ditetapkan berdasarkan kes bisnes. Fasa pembangunan model dirangka seperti dalam Rajah 3.21.



Rajah 3.21 Aliran Proses Eksperiment Pemodelan Pembelajaran Mendalam

Model pengesanan pengelakan cukai jualan dan cukai perkhidmatan ini dibangunkan menggunakan pembelajaran mendalam melalui kaedah algoritma CNN dan LSTM daripada Rangkaian Neural Berulang (*Recurrent Neural Network*, RNN).

3.5.1 Set Latihan, Pengesahan dan Ujian

Set latihan, pengesahan dan ujian dibahagikan pada kedua-dua arkitektur CNN dan LSTM bagi setiap set data yang diuji. Penetapan adalah mengikut 70 peratus jumlah data untuk set latihan, 10 peratus untuk set pengesahan dan baki 20 peratus jumlah data untuk set ujian yang dibuat secara rawak. Set data berkategori ini dapat mengurangkan risiko kebocoran maklumat dalam set ujian semasa pembelajaran dan pengesahan dijalankan oleh model.

3.5.2 Pengendalian Data Tidak Seimbang

Data tidak seimbang akan mempengaruhi prestasi pembelajaran sesuatu algoritma kerana kebanyakan algoritma direka untuk mendapatkan ketepatan yang tinggi. Ini mempengaruhi model pembelajaran untuk membuat pengelasan pada label kelas majoriti yang tinggi pada set data yang tidak seimbang. Set data dalam kajian ini adalah tidak seimbang dengan peratusan label kelas risiko rendah yang tinggi dan label kelas risiko tinggi yang sedikit. Jadual 3.8 menunjukkan peratusan label kelas bagi setiap set data dalam kajian ini.

Jadual 3.8 Peratusan label kelas untuk set data

Set Data	Risiko Rendah	Risiko Tinggi	Jumlah
Tanda Aras	99.8%	0.176%	248,807
Penyata Cukai	84.57%	15.36%	2,637,078
Bulanan	90.68%	15.43%	2,038,905
Suku Tahun	62.95%	9.31%	582,948
Dwibulanan	95.73%	37.05%	7,005
Pelbagai	90.17%	4.27%	8,220

Masalah data tidak seimbang dapat di atasi melalui kaedah pengimbangan . Salah satu kaedah yang berkesan adalah kaedah *Synthetic Minority Oversampling Technique* (SMOTE). Kaedah SMOTE menjana kelas minoriti dan meningkatkan saiz kelas minoriti. Ia membuat interpolasi data daripada contoh kelas minoriti dengan

mencari jiran k-terhampir (*k-nearest*) dengan kelas minoriti untuk setiap contoh minoriti yang tersedia (Bauder et al. 2018).

Kaedah ini adalah bersesuai dengan set data kajian kerana set data kajian tidak dapat dikurangkan melalui kaedah pengimbangan data yang lain seperti kaedah pengurangan sampel (*undersampling*) atas risiko kehilangan terlalu banyak titik data semasa pengimbangan dijalankan disebabkan oleh perbezaan kelas yang sangat ketara. Selain itu, adaptasi kaedah SMOTE juga banyak digunakan dalam kajian pengesanan fraud sebagai teknik pengimbangan data yang berkesan (Benchaji et al. 2021; Youssef et al. 2020; Z. Zhang & Huang 2020).

3.5.3 Rangkaian Neural Pelingkaran (CNN)

CNN terdiri daripada beberapa lapisan yang memproses dan mengubah input untuk menghasilkan output dan ia sangat popular dalam pengelasan imej, pengesanan objek dan segmentasi dan pemprosesan imej. Dalam kajian masa kini, model CNN telah digunakan lebih banyak untuk masalah pengelasan terutamanya dalam kajian pengesanan fraud (Babu & Pratap 2020; Heryadi & Warnars 2018; Nguyen et al. 2020).

3.5.4 Memori Jangka Masa Panjang dan Pendek (LSTM)

Model LSTM adalah model yang membuat pembelajaran daripada data jujukan. LSTM mampu untuk mengenal pasti kebergantungan jarak jauh titik data dan mengenal pasti jujukan yang berbeza. LSTM mengekstrapolasi jumlah maklumat yang maksimum untuk membuat ramalan pengelasan dengan ketepatan tertinggi. Ia berfungsi seperti penghantar maklumat di mana data input dikawal oleh struktur terkawal yang dipanggil sebagai pagar. Model LSTM menunjukkan prestasi yang baik dalam pengelasan pengesanan fraud (Alghofaili et al. 2020; Almuteer et al. 2021; Benchaji et al. 2021; Jan 2021).

3.6 PENILAIAN MODEL

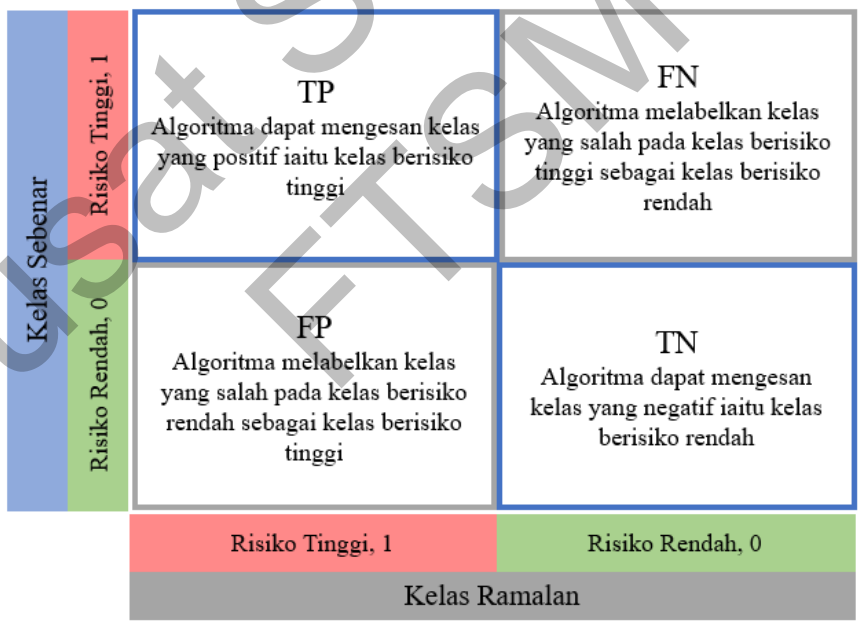
Hasil dapatan daripada kedua-dua model yang dibangunkan dinilai berdasarkan beberapa metrik prestasi yang telah dipilih. Hasil pengelasan daripada model terbaik yang dipilih akan dinilai dan disahkan oleh pakar.

3.6.1 Pemilihan Model Terbaik untuk Pengesanan Pengelakan Cukai

Pemilihan kaedah penilaian metrik yang sesuai hendaklah di jalan berdasarkan sifat set data penyata cukai dan objektif yang ingin dicapai. Pengujian model pengesanan fraud cukai yang berkesan adalah melalui matriks ralat (*confusion matrix*) dan kolerasi Matthews (Raghavan & Gayar 2019). Pengukuran dan pemilihan prestasi model menggunakan matriks ralat adalah bersesuaian memandangkan set data penyata cukai mempunyai kelas yang tidak seimbang.

a. Matriks Ralat

Matriks ralat mengukur prestasi model berdasarkan pengujian terhadap positif benar (*true positive*, TP), positif salah (*false positive*, FP), negatif salah (*false negative*, FN) dan positif salah (*true negative*, TN). Rajah 3.22 menunjukkan kedudukan pengelasan berdasarkan matriks ralat.



Rajah 3.22 Matriks ralat bagi pengelasan kelas positif (risiko tinggi) dan kelas negatif (risiko rendah)

Daripada matriks ralat ini, prestasi model dinilai secara lanjut dengan pengiraan ketepatan (*accuracy*), kejituan (*precision*), dapatan semula (*recall*) dan skor F1 (*F1 score*) seperti formula kiraan parameter di bawah:

$$Ketepatan = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Kejituan = \frac{TP}{TP + FP}$$

$$Dapatan\ Semula = \frac{TP}{TP + FN}$$

$$Skor\ F1 = \frac{2\ X\ Kejituan\ X}{FP + TN}$$

3.6.2 Penilaian Pakar

Kajian ini menggunakan data sebenar pembayar cukai yang memerlukan penilaian pakar terhadap keputusan dan pengelasan model. Penilaian pakar melibatkan penerangan kaedah yang digunakan dan output pengelasan yang dihasilkan oleh model pembelajaran mendalam kepada pakar yang berpengalaman dalam bidang percukaian. Seramai tiga orang pakar yang terlibat dalam kajian ini dari awal dipilih bagi memberi maklum balas terhadap data yang digunakan, pendekatan pra pemprosesan data dan hasil dapatan kajian. Pakar yang dipilih merupakan pegawai di JKDM dalam bidang tugas percukaian dan pengurusan risiko percukaian. Mereka juga adalah pegawai yang bertanggungjawab menguruskan laporan dan statistik bagi data percukaian GST dan SST di JKDM.

3.7 PENGATURAN MODEL

Objektif kajian adalah untuk menghasilkan model pembelajaran mendalam yang terbaik untuk pengesanan pengelakan cukai. Hasil dapat kajian digunakan untuk memilih model yang terbaik dan menyesuaikan aplikasi model dalam dunia sebenar. Dalam fasa ini, strategi untuk pengaturan model digariskan seperti kitaran perlombongan data dan aplikasi model pada masa hadapan. Bagi pengaturan model di JKDM, terdapat pengubahsuaian yang perlu diambil kira memandangkan pada masa ini SST dikuatkuasakan sebagai sumber percukaian tidak langsung. Model juga boleh disesuaikan kepada lain-lain percukaian yang dikawal oleh JKDM seperti TTX, STODS yang masih menggunakan konsep yang sama dengan GST. Penilaian pakar dijalankan

dengan menerangkan pengelasan yang dijana oleh model dan keberkesanannya dalam pengesanan pengelakan cukai.

3.8 KESIMPULAN

Bab ini menerangkan mengenai metodologi kajian yang merangkumi pendekatan kajian, penerokaan kes bisnes, penyediaan data, pembangunan model, penilaian model, pengaturan model dan pemantauan 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 perolehan data, integrasi data, pembersihan data, padanan dan tapisan data, pengelasan data dan analisis deskriptif. Setiap aktiviti ini diperlukan untuk memastikan kualiti data yang akan diuji dengan model pembelajaran mendalam adalah tersedia. Analisis deskriptif 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 pembayar cukai yang risiko tinggi melalukan pengelakan cukai berdasarkan model pembelajaran mendalam CNN dan LSTM yang dibangunkan. Hasil dapatan kajian adalah berdasarkan input daripada kajian literatur di Bab 2 dan metodologi yang dibincangkan di Bab 3. Dalam bab ini, penetapan eksperimen melalui penggunaan set data, pengendalian data tidak seimbang, dan arkitektur model diterangkan secara lebih terperinci. Hasil dapat kajian dan hasil analisis deskriptif dibincangkan secara analitikal.

4.2 TREND DAN HASIL DAPATAN DESKRIPTIF

Analisis deskriptif daripada set data dengan perincian kepada atribut dapat memberi gambaran yang lebih jelas dan pengetahuan yang lebih mendalam terhadap set data untuk membantu pembangunan model pembelajaran mendalam yang lebih berkesan. Analisis deskriptif menggunakan plot taburan dan plot ketumpatan untuk menunjukkan pengagihan atribut berdasarkan label kelas dan mengenal pasti taburan yang tidak seimbang. Bagi plot taburan dan plot ketumpatan, sebanyak 5,000 sampel data digunakan untuk menjadikan taburan label kelas sebagai normal pada set data yang tidak seimbang

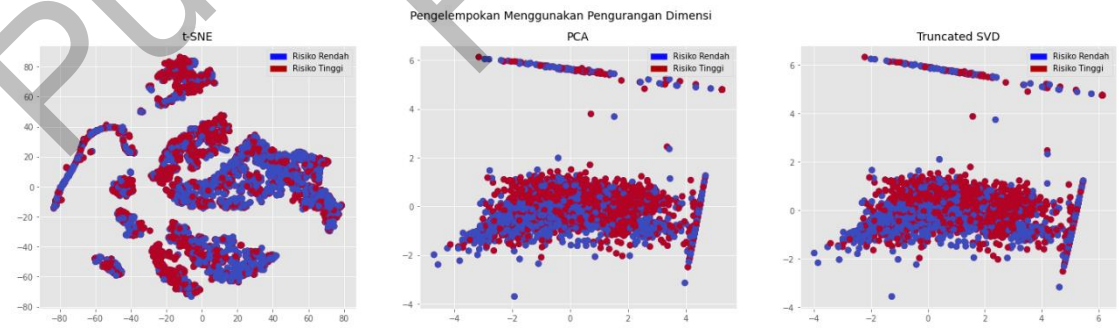
4.2.1 Visualisasi Data Melalui Pengurangan Dimensi

Kaedah pengurangan dimensi digunakan untuk meneroka pemahaman dan pandangan terhadap data dengan memvisualkan data kepada pemetaan dua (2) dimensi ada atau tiga (3) dimensi menerusi pengelompokan (Wang et al. 2021).

Pemetaan data ke dimensi yang lebih tingi juga dapat menunjukkan jarak antara dua (2) label kelas dalam tugasan pengelasan (Charitou et al. 2020). Bagi tujuan ini, pengurang dimensi dilakukan dengan menggunakan teknik *t-Distributed Stochastic Neighbour Embedding* (t-SNE), *Principal Componet Analysis* (PCA), dan *Truncated Singular-Value Decomposition* (*Truncated* SVD). t-SNE memaparkan struktur data berdimensi tinggi melalui dua (2) atau (3) dimensi.

t-SNE digunakan untuk eksplorasi data dan melihat bagaimana data disusun dalam ruang berdimensi tinggi tetapi masih mengekalkan struktur penting data (Anowar et al. 2021) dan mengukur persamaan antara dua titik data untuk setiap pasangan. Manakala PCA adalah salah satu kaedah pengurangan dimensi yang penting untuk menggambarkan data. PCA berfungsi dengan menukar dimensi *n* data kepada *k* dan mengekalkan seberapa banyak maklumat daripada set data asal. *Truncated* SVD memfaktorkan matriks data dengan bilangan lajur yang sama dengan pemangkasan.

Gambaran kelompok data menggunakan kaedah pengurangan dimensi dapat di lihat pada Rajah 4.1. t-SNE menunjukkan kelompok label kelas yang bercampur dengan beberapa kelompok kecil untuk titik data. Manakala PCA dan *Truncated* SVD menghasilkan gambaran kelompok yang hampir serupa dengan perbezaan fungsi pada matriks sampel.



Rajah 4.1 Pengelompokan menggunakan pengurangan dimensi

Berdasarkan pemerhatian ini, pengurangan dimensi tidak dapat memisahkan dua kelompok yang ketara dan memerlukan lebih banyak sampel untuk berfungsi dengan baik (F. Zhang et al. 2020). Ia juga memberi gambaran ramalan pembelajaran mesin tidak dapat mencapai ketepatan yang tinggi dengan model yang ringkas. Oleh

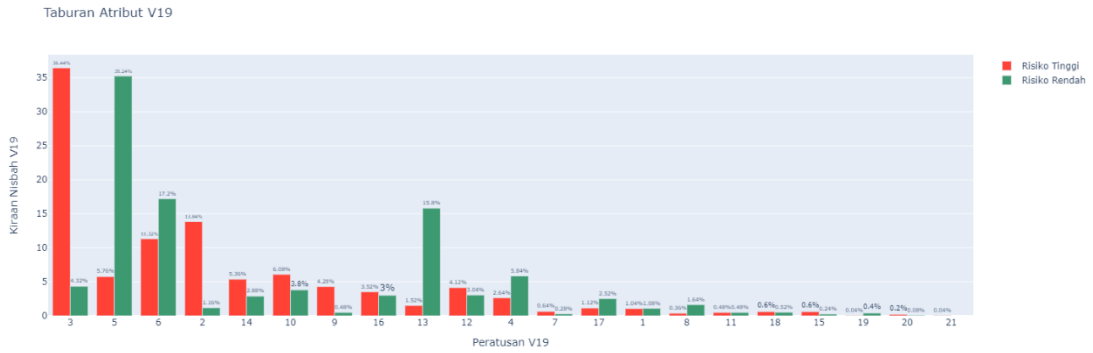
itu, model yang lebih kompleks seperti pembelajaran mendalam dapat membantu membuat ramalan pengelasan yang lebih berkesan.

4.2.2 Taburan Data Nominal

Plot taburan digunakan untuk analisis data nominal. Bagi Rajah 4.2, taburan atribut V18 menunjukkan taburan bagi kod MSIC yang mempunyai kiraan nisbah lebih daripada satu (1) yang mewakili 20 kod MSIC. Berdasarkan plot ini, taburan bagi setiap kod MSIC menunjukkan kod MSIC 1261, 46312 dan 47114 yang mempunyai kelas risiko tinggi yang besar dan kelas risiko rendah yang tinggi bagi kod MSIC 41001, 42009, 46594 dan 46639

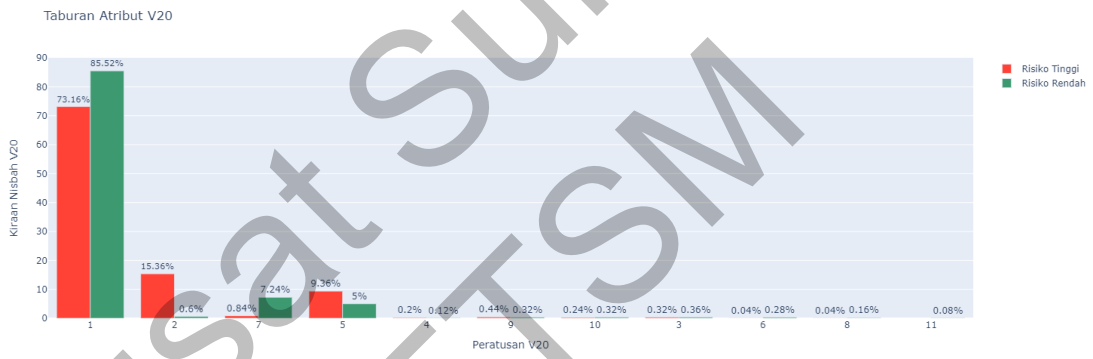
Rajah 4.2 Plot taburan untuk atribut “V18”

Bagi Rajah 4.3, taburan atribut V19 menunjukkan taburan 21 sektor industri dalam GST. Berdasarkan plot ini, taburan bagi sektor industri menunjukkan beberapa sektor yang mempunyai kelas risiko tinggi dengan kiraan nisbah yang besar iaitu sektor industri 3 dan 2. Manakala sektor industri 5 dan 13 mempunyai kiraan nisbah yang tinggi dalam kelas risiko rendah.



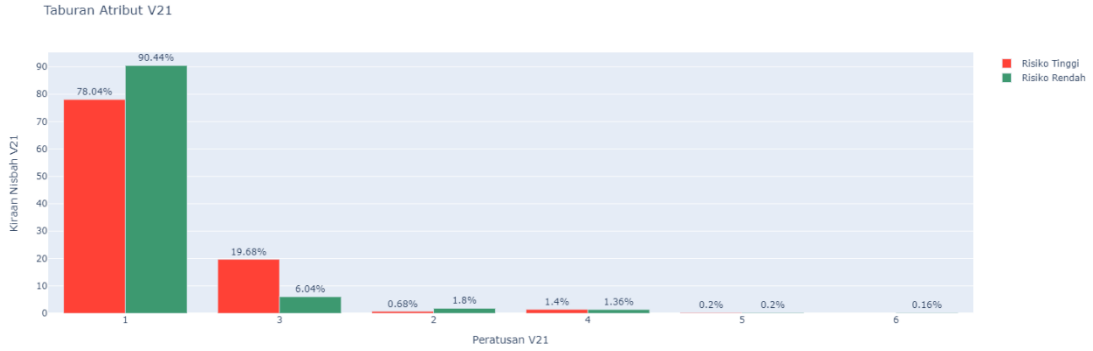
Rajah 4.3 Plot taburan untuk atribut "V19"

Bagi Rajah 4.4, taburan atribut V20 menunjukkan taburan 11 jenis pendaftaran perniagaan pembayar cukai GST. Berdasarkan plot ini, pendaftaran perniagaan 1, 2 dan 5 mempunyai taburan kelas risiko tinggi dalam kiraan nisbah yang besar.



Rajah 4.4 Plot taburan untuk atribut "V20"

Rajah 4.5 menunjukkan taburan atribut V21 yang mewakili jenis pendaftaran GST oleh pembayar cukai. Berdasarkan plot taburan, jenis pendaftaran 1 dan 2 mempunyai kiraan nisbah yang besar untuk kelas risiko tinggi.

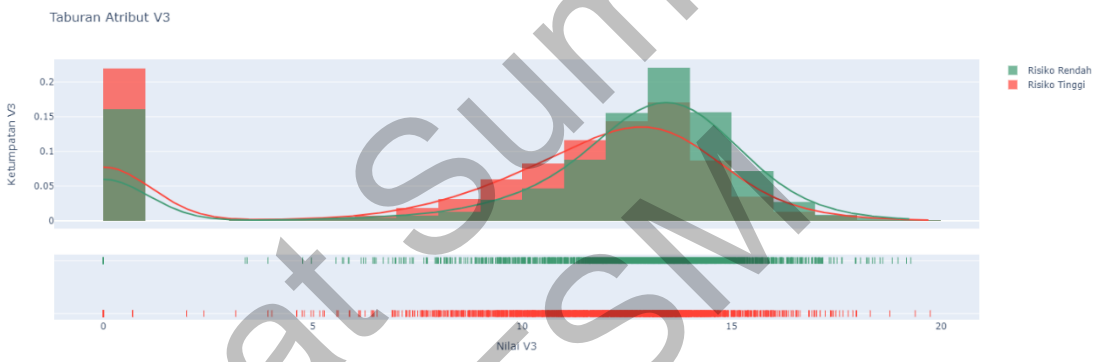


Rajah 4.5 Plot taburan atribut "V21"

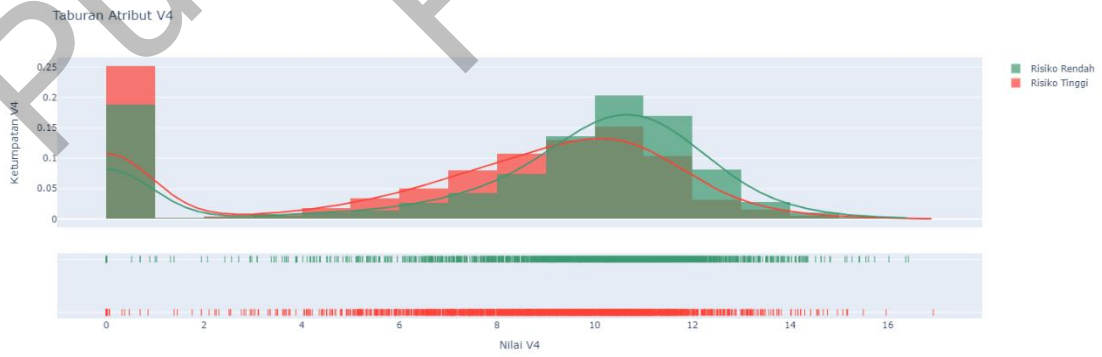
4.2.3 Taburan Data Numerik

Taburan data numerik menggunakan plot ketumpatan untuk menunjukkan taburan kelas bagi atribut utama yang dipilih iaitu atribut V3 sehingga V8 dan atribut V22.

Atribut “V3” adalah jumlah nilai pembekalan berkadar standard manakala atribut “V4” adalah jumlah cukai output berdasarkan 6% daripada perolehan berkadar standard. Rajah 4.6 dan Rajah 4.7 menunjukkan plot taburan bagi atribut “V3” dan “V4”. Taburan ketinggian bagi kelas risiko rendah dan risiko tinggi adalah mengikut corak lengkung loceng biasa dengan kepadatan tinggi pada nilai 0 kerana terdapat nilai 0 bagi penyata cukai yang tidak mempunyai perolehan bagi sesuatu tempoh bercukai.



Rajah 4.6 Plot taburan atribut “V3”

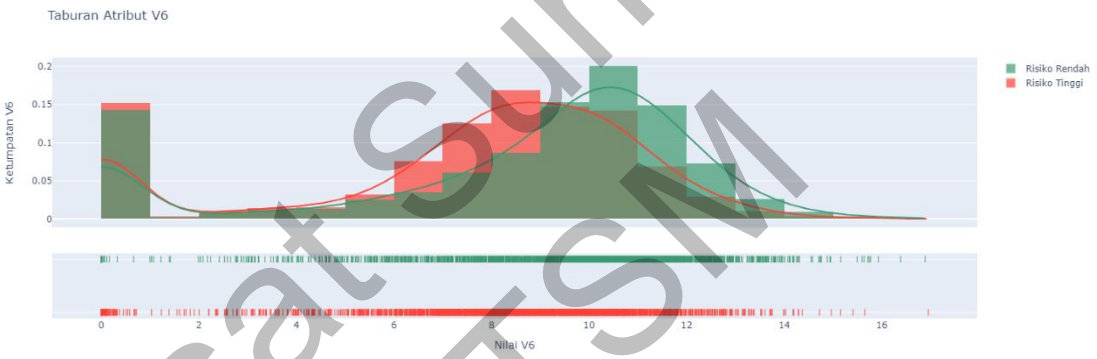


Rajah 4.7 Plot taburan atribut “V4”

Atribut V5 adalah nilai perolehan berkadar standard dan plot taburan dapat dilihat dalam Rajah 4.8. Manakala atribut V6 merupakan cukai input yang dikira daripada 6% nilai perolehan berkadar standard dan plot taburan dapat dilihat dalam

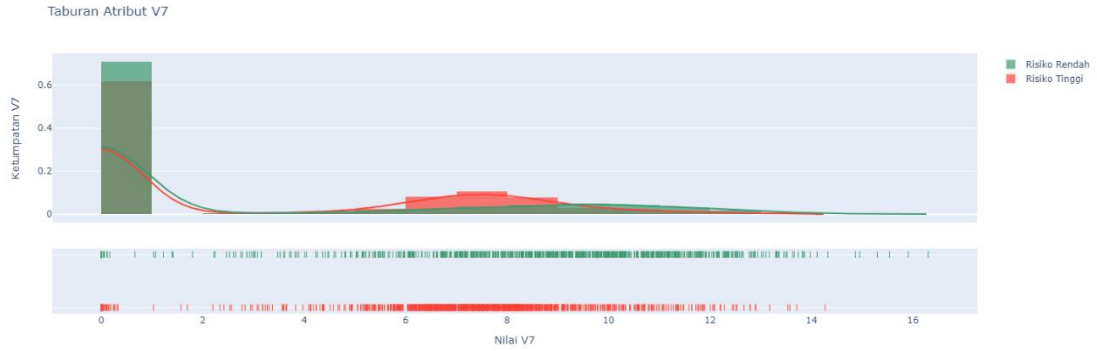
Rajah 4.9. Plot taburan bagi atribut V5 dan V6 menunjukkan taburan normal dalam lengkukan loceng bagi kelas risiko tinggi dan kelas risiko rendah.

Rajah 4.8 Plot taburan atribut "V5"



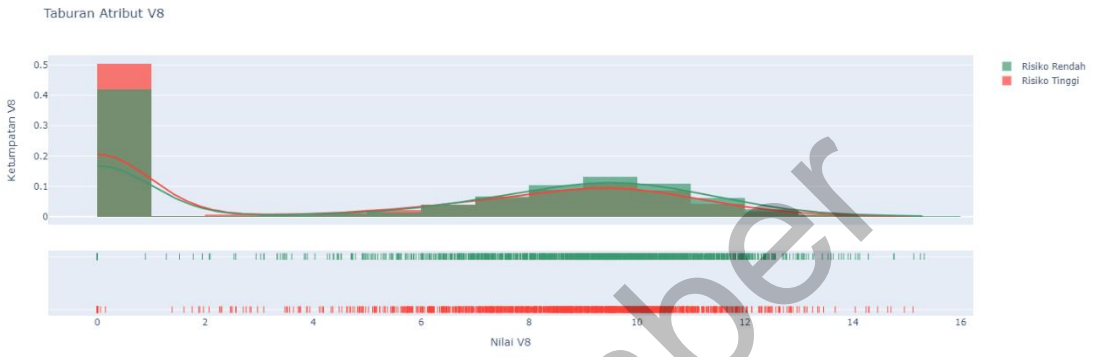
Rajah 4.9 Plot taburan atribut "V6"

Atribut V7 dalam Rajah 4.10 mewakili amaun cukai pulang balik yang dituntut oleh pembayar cukai. Plot taburan adalah sangat rendah bagi kelas risiko rendah berbanding kelas risiko tinggi.



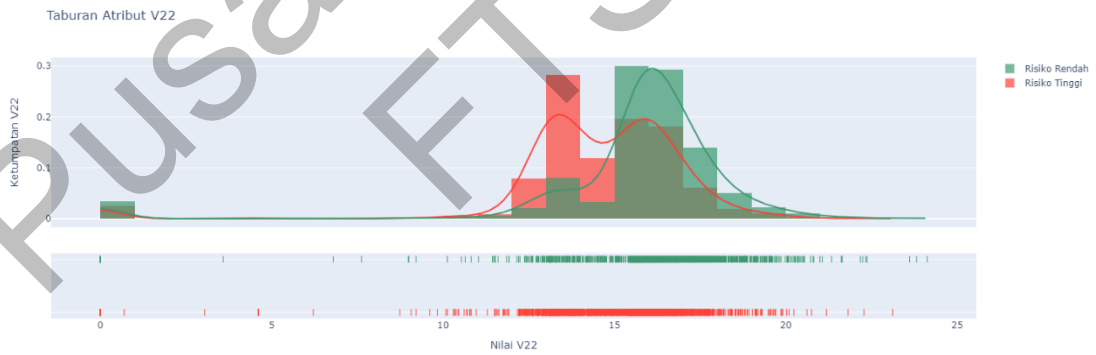
Rajah 4.10 Plot taburan atribut "V7"

Atribut V8 dalam Rajah 4.11 adalah amaun cukai perlu dibayar oleh pembayar cukai. Plot taburan menunjukkan taburan lengkukan loceng adalah normal dan selari bagi kedua-dua kelas risiko.



Rajah 4.11 Plot taburan atribut "V8"

Atribut V22 adalah jumlah nilai ambang tahunan pembayar cukai. Rajah 4.12 menunjukkan plot taburan bagi kelas risiko rendah dan risiko tinggi. Plot taburan kelas risiko tinggi menunjukkan dua taburan yang dikenali sebagai taburan pelbagai mod. Bentuk taburan ini menunjukkan data boleh dibahagikan kepada beberapa kumpulan.



Rajah 4.12 Plot taburan atribut "V22"

Nilai ambang tahunan digunakan sebagai asas kepada penentuan kekerapan pengikraran penyata cukai. Pembahagian data akan dijalankan untuk kajian ablasi dengan mengasingkan data kajian berdasarkan kumpulan kekerapan pengikraran penyata cukai.

4.3 HASIL DAPATAN PENGUJIAN MODEL PENGESANAN PENGELAKAN CUKAI

Pengujian dijalankan berdasarkan penetapan eksperimen dan hasil analisis deskriptif. Pengujian bagi kedua-dua model pembelajaran mendalam adalah menggunakan fasa pra pemprosesan yang sama, kaedah pengimbangan data melalui SMOTE dan bilangan epoc (*epoch*) yang sama iaitu pada nilai 100 untuk set data tanda aras dan 1,000 epoc bagi set data penyata cukai dan set data berkategori.

Pengujian dimulakan dengan set data tanda aras sebagai panduan kepada arkitektur yang digunakan oleh kajian lepas. Seterusnya pengujian dijalankan pada set data saiz kecil iaitu set data berkategori dwibulanan untuk mendapatkan penalaan parameter yang terbaik bagi set data penyata cukai. Hasil daripada penetapan arkitektur dan penalaan hiperparameter pada model CNN dan LSTM, pengujian diteruskan kepada set data penyata cukai dan set data berkategori bulanan, suku tahun dan pelbagai.

4.3.1 Model Pengelasan CNN

Model pengelasan CNN dibangunkan menggunakan modul *Keras* dengan memuat naik *Sequential*, lapisan *Flatten*, *Dense*, *Dropout*, *BatchNormalization* dan *Conv1D*. Perincian model CNN dan setiap lapisan diterangkan dalam Jadual 4.1 dan pengekodan model menggunakan *Python* adalah seperti di Lampiran B.

Jadual 4.1 Model dan lapisan CNN

Lapisan	Bentuk Output	Parameter
conv1d	(None, 21, 128)	384
batch_normalization	(None, 21, 128)	512
dropout	(None, 21, 128)	0
conv1d_1	(None, 20, 64)	16448
batch_normalization_1	(None, 20, 64)	256
dropout_1	(None, 20, 64)	0
flatten	(None, 1280)	0
dense	(None, 64)	89194
dropout_2	(None, 64)	0
dense_1	(None, 1)	65
Jumlah parameter	99,649	
Parameter yang boleh dilatih	99,265	
Parameter yang tidak boleh dilatih	388	