MODEL RAMALAN RISIKO PESAKIT COVID-19 MENGGUNAKAN PENDEKATAN PEMBELAJARAN MESIN

TENG WEI ZHUN
AZURALIZA ABU BAKAR

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

SARS-CoV-2 virus berasal dari Wuhan, China telah mencetuskan pandemik Covid-19 dan kini merebak ke negara-negara di seluruh dunia. Pembangunan bidang kecerdasan buatan amat penting bagi proses diagnosis dan ramalan untuk mengawal penyebaran Covid-19. Menurut worldometer, terdapat 240 juta orang di dunia yang menjangkiti Covid-19 dan 4 juta kes kematian tercatat sehingga bulan Oktober 2021. Oleh itu, kaedah pembelajaran mesin perlu diguna untuk menjalankan diagnosis Covid-19 dengan segera. Kajian ini menggunakan enam algoritma pembelajaran mesin untuk pembangunan model dan perbandingan prestasi dilakukan untuk mendapatkan model yang terbaik dalam ramalan risiko pesakit Covid-19 di Malaysia. Enam model ramalan yang berbeza akan digunakan iaitu regresi logistik, mesin vektor sokongan, *Gaussian Naïve Bayes*, hutan rawak, K-jiran terdekat (KNN), dan pokok keputusan. Set data projek ini diperoleh dari laman sesawang Kementerian Kesihatan Malaysia yang terdiri daripada tarikh, jenis vaksin, lokasi, jantina, umur, dan kewarganegaraan. Bilangan data yang digunakan dalam projek ini adalah sebanyak 405947 yang diekstrak daripada pesakit COVID-19 di seluruh Malaysia. Hasil ramalan akan dinilai menggunakan metrik penilaian iaitu *accuracy*, *precision*, Skor F1, *recall* dan nilai AUC. Skor penilaian model yang tinggi menentukan algoritma model tersebut sesuai digunakan dalam ramalan risiko pesakit Covid-19. Projek ini dijalankan dengan kerjasama Pejabat Kesihatan Daerah Hulu Langat. Selain dari mengunakan data seluruh negara, projek ini juga melaporkan model ramalan bagi data kes di Hulu Langat. Hasil kajian ini dapat membantu dalam pembuatan keputusan klinikal dalam pandemik Covid-19 dan pengagihan kapasiti tempat tidur ICU.

1 PENGENALAN

Penyakit Coronavirus (COVID-19) merupakan jenis virus Corona yang boleh mengakibatkan jangkitan kepada sistem saluran pernafasan. World Health Organization (WHO) mengumumkan penyakit Coronavirus adalah coronavirus yang baru ditemui pada tahun 2019 dan penyakit ini dinamakan COVID-19 pada 12 Januari 2020. Berdasarkan rekod data worldometer, jumlah sebanyak 240 juta kes positif COVID-19 di seluruh dunia dan 4 juta kes kematian disebabkan penyakit COVID-19 tercatat setakat bulan Oktober 2021. COVID-19 merupakan pandemik penyakit di seluruh dunia kerana berlaku di kawasan geografi yang luas dan mempengaruhi populasi yang sangat tinggi. Usaha global amat diperlukan untuk menghalang transmisi COVID-19 supaya tidak mencapai tahap yang amat kritikal.

Kes positif COVID-19 masih meningkat dan mencapai kes seharian yang paling tinggi di Malaysia adalah 24599 jumlah kes seharian pada 26 Ogos 2021 (Sinar Harian, 2021).

Program Vaksinasi COVID-19 dilancarkan dalam tempoh tersebut untuk memastikan risiko jangkitan di tempat-tempat yang tumpuan orang ramai dapat dikurangkan (Bernama, 2021). Terdapat 3 jenis vaksin COVID-19 dibekalkan untuk rakyat Malaysia iaitu *AstraZeneca*, *Sinovac* dan *Pfizer*. Kementerian Malaysia juga memperketatkan SOP untuk memastikan risiko penyebaran COVID-19 di Malaysia ke tahap yang paling rendah.

Pembelajaran mesin merupakan salah satu cabang dalam bidang kecerdasan buatan yang menggunakan algoritma dalam kapasiti pembelajaran untuk meningkatkan prestasi dan ketepatan ramalan. Kewujudan kaedah pembelajaran mesin bukan sahaja mengenal pasti risiko kematian pesakit dengan tepat, tetapi juga boleh dijadikan alat diagnosis serta ramalan pandemik penyakit-penyakit kritikal. Pembelajaran mesin juga merupakan mesin pintar yang canggih dan boleh digunakan untuk pengesanan, pencegahan dan ramalan untuk memerangi wabak COVID-19.

Kajian ini akan menggunakan algoritma pembelajaran mesin yang berbeza untuk memproses data pesakit COVID-19 agar dapat mengeluarkan hasil ramalan yang tepat. Data pesakit COVID-19 adalah diperoleh dari laman web Kementerian Kesihatan Malaysia yang mengandungi atribut jantina, umur, gejala jangkitan, tarikh jangkitan, dan jenis vaksin. Pemprosesan data akan dijalankan sebelum hantar ke model pembelajaran mesin. Model pembelajaran mesin akan dinilai dan seterusnya membuat perbandingan antara satu sama lain untuk mengenal pasti model yang terbaik bagi ramalan risiko penyakit COVID-19. Hasil kajian ini dapat membantu kakitangan hospital untuk membuat keputusan klinikal dan pengagihan kapasiti katil unit rawatan rapi (ICU) di hospital Malaysia.

2 PENYATAAN MASALAH

Penyebaran pandemik COVID-19 di Malaysia masih dalam keadaan kritikal walaupun pelbagai jenis analitik statistik dijalankan untuk menganalisis arah aliran kes jangkitan COVID-19. Teknik pengiraan nilai R-Naught yang merujuk kepada kebolehjangkitan COVID-19 telah diperkenalkan oleh KKM (A. Suhael, 2021). Nilai R-Naught ini mengira dan menjana hasil nilai untuk hari tertentu sahaja. Pengiraan nilai R-Naught sememangnya dapat menggambarkan tahap penularan jangkitan COVID-19, tetapi tidak boleh digunakan sebagai angka nilai untuk membuat ramalan terhadap risiko wabak penyakit COVID-19 pada masa depan. Kebanyakan orang yang tidak mematuhi SOP yang telah ditetapkan

menyebabkan penularan penyakit COVID-19 menjadi serius. Hal ini mengakibatkan peningkatan kadar penggunaan katil ICU bagi rawatan pesakit COVID-19 (A. Lee, A. Razak, 2021). Terdapat beberapa hospital kerajaan mengalami masalah kekurangan katil ICU dan kakitangan hospital terpaksa membuat pertimbangan bagi pengagihan katil ICU untuk pesakit COVID-19 (R. Noraina, 2022). Situasi yang kritikal ini merupakan faktor kenaikan kes kematian COVID-19 di Malaysia. Menurut laporan CPRC KKM, sebanyak jumlah 55 daripada 60 katil ICU iaitu lebih kurang 92 peratus katil ICU sudah digunakan untuk rawatan pesakit COVID-19 (R. Noraina, 2022). Tambahan pula, kekurangan bekalan oksigen juga menyebabkan kadar kes kematian meningkat secara mendadak terutamanya pesakit COVID-19 varian Delta. Hal ini menyebabkan kerajaan Malaysia perlu menyediakan jumlah perbelanjaan yang tinggi untuk membeli fasiliti serta bekalan perubatan dan kelengkapan perlindungan diri untuk menangani wabak COVID-19 (H. Muhamad, 2021). Oleh itu, ciri-ciri klinikal dan data demografi pesakit COVID-19 amat diperlukan dalam kajian saintifik untuk membuat ramalan mengenai tahap penularan wabak COVID-19. Antara contoh ciri-ciri data pesakit COVID-19 termasuk umur, jantina, kewarganegaraan, negeri, tarikh jangkitan, simptom dan jenis vaksin diterima. Setakat ini masih kekurangan kajian dan penyelidikan saintifik yang menggunakan maklumat data pesakit dalam model pembelajaran mesin kerana kuantiti data yang banyak amat diperlukan dalam pembinaan algoritma (D. Jemilah, 2021). Kewujudan algoritma pembelajaran mesin amat penting dalam usaha membendung penularan wabak COVID-19 dengan pemodelan ramalan risiko kesihatan pesakit COVID-19. Hasil keputusan ramalan dapat membantu perancangan kapasiti hospital dan peruntukan bagi sumber kelengkapan rawatan COVID-19.

3 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah:

1. Melaksanakan proses pra-pemprosesan dan penyediaan data pesakit COVID-19 di Malaysia untuk meningkatkan kualiti data sebelum diintegrasikan dalam model pembelajaran mesin.

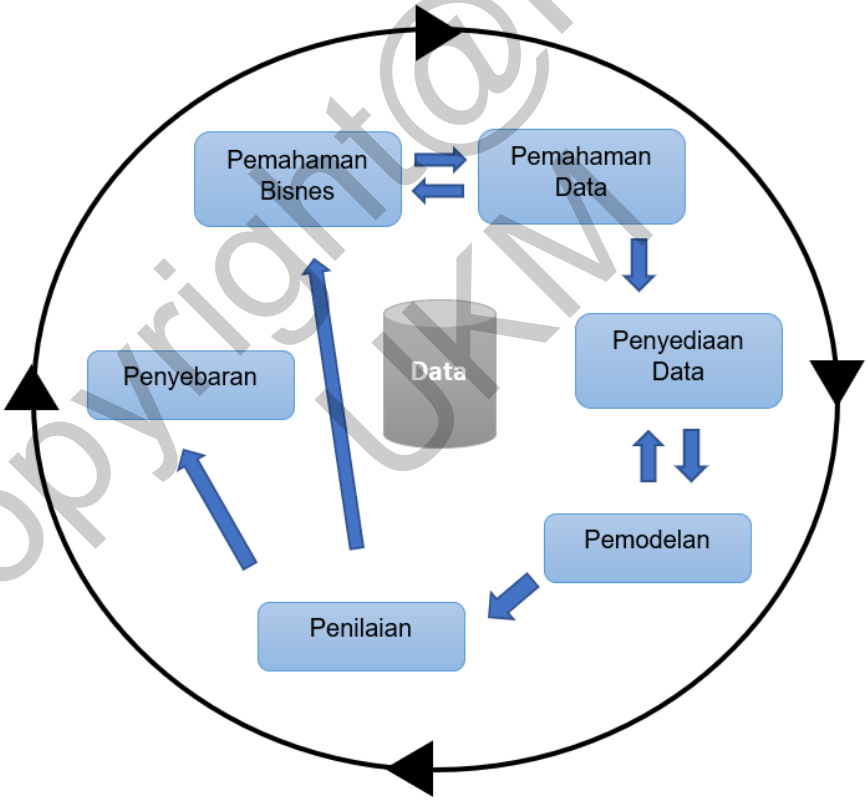
2. Membangunkan model pembelajaran mesin dengan menggunakan data yang sudah melalui proses-prapemprosesan untuk ramalan risiko kesihatan pesakit

COVID-19 dan perbandingan prestasi akan diuji atas enam jenis algoritma pembelajaran mesin.

3. Menilai prestasi perbandingan dan hasil penilaian model pembelajaran mesin dengan grafik profesional menerusi alat visualisasi *Tableau*.

4 METOD KAJIAN

Metodologi yang akan digunakan dalam projek ini adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*). CRISP-DM merupakan panduan untuk perancangan dan pengaturan dalam projek yang melibatkan pembelajaran mesin. Rajah 4.1 menunjukkan urutan fasa metodologi CRISP-DM.



Rajah 4.1 Fasa-fasa Metodologi CRISP-DM

4.1 Fasa Pemahaman Bisnes (*Business Understanding*)

Fasa pemahaman bisnes merujuk kepada fasa yang memberi tumpuan kepada pemahaman objektif dan kriteria projek. Objektif utama dalam projek ini adalah untuk menghasilkan keputusan ramalan terhadap risiko kesihatan pesakit COVID-19 dengan pendekatan

pembelajaran mesin. Pencarian maklumat bahan rujukan dan laporan kajian mengenai pandemik COVID-19 perlu dilakukan kerana mempunyai pemahaman dan pengetahuan yang berkaitan dengan domain amat penting. Hal ini disebabkan situasi wabak COVID-19 yang kritikal berlaku di negara Malaysia dan memerlukan laporan analisis yang tepat untuk membantu pengagihan kelengkapan perubatan di hospital.

4.2 Fasa Pemahaman Data (*Data Understanding*)

Fasa pemahaman data fokus kepada pengumpulan data dan menerokai data untuk mengenal pasti masalah-masalah yang terlibat dalam data yang diperoleh. Sumber set data yang digunakan dalam kajian ini diperoleh daripada laman web repositori *github* Kementerian Kesihatan Malaysia. Sebanyak 405947 rekod data pesakit COVID-19 dan 15 atribut rekod pesakit COVID-19 dalam tempoh April 2020 hingga April 2021 telah diekstrak untuk projek kajian ini. Laporan kualiti data yang merangkumi kriteria stastistik iaitu purata, maksimum, minimum, sisihan piawai, varians, kecondongan dan sebagainya perlu dijana untuk meningkatkan pemahaman terhadap data.

4.3 Fasa Penyediaan Data (*Data Preparation*)

Fasa penyediaan data melibatkan pemprosesan data untuk menghasilkan set data yang bersih dan sesuai untuk diintegrasikan ke dalam model pembelajaran mesin. Langkah-langkah pemprosesan data termasuklah penggabungan data, penerokaan data, pembersihan data, pendiskretan data, dan pemilihan ciri-ciri penting. Langkah pertama adalah menggabungkan atribut yang diperoleh daripada set data yang berlainan. Penerokaan data dijalankan untuk memastikan kita mempunyai kefahaman terhadap bentuk set data. Pembersihan data menghasilkan set data yang bebas daripada nilai kosong dan tidak lengkap. Data-data yang berbentuk berterusan akan menjalankan transformasi dan menukar kepada bentuk kategori melalui proses pendiskretan data. *ExtraTreesClassifier* digunakan dalam projek ini untuk menjalankan pengekstrakan ciri-ciri yang penting dalam set data. Fasa ini akan mengeluarkan set data yang siap sedia untuk fasa pembangunan model supaya dapat mencapai ketepatan penilaian model ramalan yang tinggi.

4.4 Fasa Pemodelan (*Modelling*)

Fasa Pemodelan mengintegrasikan set data yang lengkap dan bersih dengan algoritma pembelajaran mesin dalam model ramalan. Enam jenis algoritma pembelajaran mesin digunakan untuk ramalan risiko kesihatan pesakit COVID-19 iaitu regresi logistik, mesin

vektor sokongan, *Gaussian Naïve Bayes*, hutan rawak, KNN dan pokok keputusan. Algoritma-algoritma tersebut dapat menyelesaikan masalah pengelasan dan dikenali sebagai pendekatan pembelajaran terkawal yang sesuai untuk model ramalan. Sebelum pembinaan model ramalan, set data terbahagi kepada 75 peratus bagi set latihan dan 25 peratus bagi set ujian.

4.5 Fasa Penilaian (*Evaluation*)

Fasa Penilaian melakukan penilaian terhadap model pembelajaran mesin agar dapat mencapai objektif bisnes. Prestasi model akan dinilai untuk memastikan keputusan yang diperoleh adalah memuaskan. Jika memperoleh keputusan yang tidak memuaskan, proses pemodelan akan mengulangi semula dan mengkaji balik punca-punca yang menghasilkan keputusan yang tidak tinggi. Dalam projek kajian ini, penilaian metrik yang digunakan untuk menilai peramalan risiko kesihatan pesakit COVID-19 termasuklah *accuracy*, *precision*, *recall*, dan Skor F1 dan nilai AUC. *T-test* digunakan sebagai ujian hipotesis untuk membandingkan prestasi model ramalan dan membuktikan model yang terpilih mempunyai prestasi yang tinggi dalam ramalan risiko kesihatan pesakit COVID-19.

4.6 Fasa Penyebaran (*Deployment*)

Fasa penyebaran merupakan fasa terakhir proses pembangunan model ramalan risiko kesihatan pesakit COVID-19. Hasil keputusan kajian yang dikaji bersedia untuk disebarkan kepada orang ramai. Hasil laporan kajian ramalan risiko kesihatan pesakit COVID-19 akan dipaparkan melalui alat visualisasi *Tableau* secara atas talian supaya pengguna dapat memahami hasil keputusan analisis dengan cepat dan mudah. Pemantauan dan penyelenggaraan perlu dilaksanakan dari semasa ke semasa untuk memastikan tiada kesilapan berlaku atas hasil kajian dan tidak menyebarkan maklumat yang salah kepada orang ramai.

4.7 Spesifikasi Keperluan

Perkakasan dan perisian perlu ditentukan untuk memastikan pembangunan model pembelajaran mesin dengan lancar. Jadual 4.1 menunjukkan spesifikasi keperluan perkakasan manakala jadual 4.2 menunjukkan spesifikasi keperluan perisian.

Jadual 4.1 Spesifikasi perkakasan

Processor	**Ruang Cakera Keras (SSD)**	**Ingatan Capaian Rawak (RAM)**	**Sistem Operasi**	**Kad Grafik**
Intel Core i5-9300H	256GB	12GB	*Window 10 Home 64-bit*	*NVIDIA GeForce GTX 1650*

Jadual 4.2 Spesifikasi perisian

Perisian	**Perincian**
Microsoft Word	Menulis tesis projek akhir tahun
Microsoft Excel	Tempat penyimpanan sumber data
Google Colab	Sebagai IDE untuk menulis algoritma *Python* bagi proses pemodelan
Tableau	Alat visualisasi keputusan ramalan risiko kesihatan pesakit COVID-19

5 HASIL KAJIAN

Dalam projek ini, set data mengandungi 7 atribut dalam jenis *int* dan 1 atribut dalam jenis *float* selepas pemprosesan data dijalankan. Set data pesakit COVID-19 terbahagi kepada set latihan dan set ujian dalam nisbah 75 : 25. Set latihan diperlukan untuk pembelajaran algoritma manakala set ujian digunakan untuk mengesahkan proses pembangunan set latihan dan melaraskan untuk mendapatkan hasil yang terbaik.

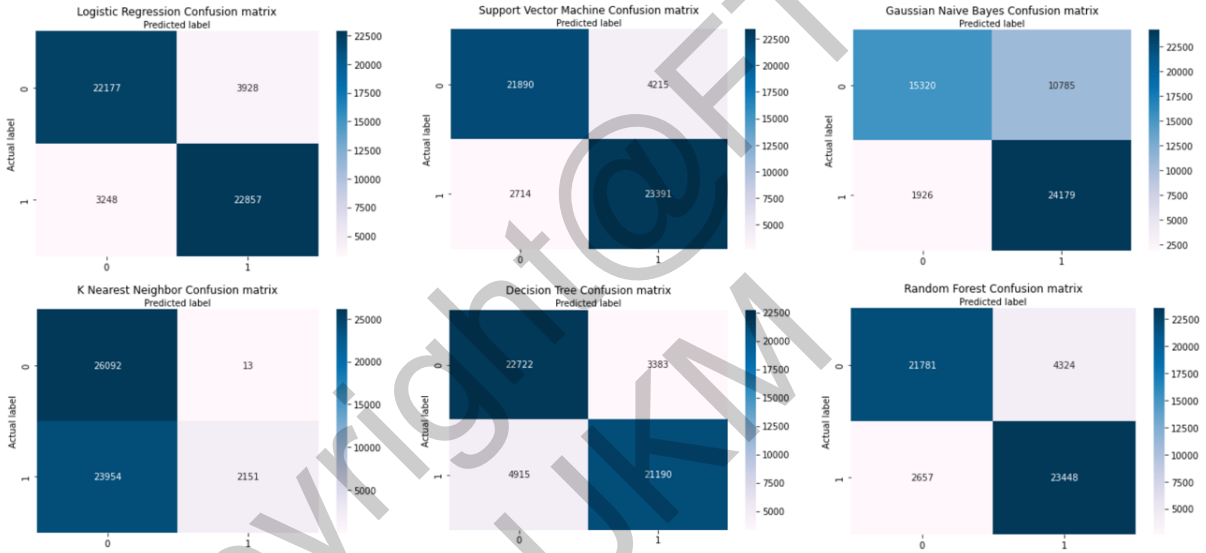
Dalam set data ini mendapati bahawa masalah ketidakseimbangan kelas label akan menjejaskan ketepatan pemodelan. Cara untuk mengatasi masalah ini adalah mengambil contoh yang berlebihan dalam kelas minoriti dengan kaedah duplikasi data. Justeru, *Synthetic Minority OverSampling Technique* (SMOTE) digunakan untuk menjalankan pengimbangan data.

5.1 Keputusan Prestasi Pemodelan Negeri Selangor

Penilaian metrik dijalankan terhadap algortima pembelajaran mesin yang telah dibangunkan untuk menentukan algoritma yang paling sesuai untuk mengklasifikasikan risiko kesihatan

pesakit COVID-19. Metrik penilaian yang digunakan untuk menilai prestasi model ramalan adalah *accuracy*, *precision*, *recall*, *score F1*, *confusion matrix* dan *ROC curve*.

Berdasarkan *confusion matrix* regresi logistik, terdapat 22177 kelas data positif yang telah diklasifikasikan dengan betul oleh regresi logistik. Kelas negatif yang diklasifikasikan dengan betul adalah sebanyak 22857. Kelas negatif yang salah diklasifikasikan sebagai kelas positif adalah sebanyak 3928. Terdapat 3248 bilangan kelas data positif yang salah diklasifikasikan sebagai kelas negatif oleh model regresi logistik. Rajah 5.1 menunjukkan *confusion matrix* yang dihasilkan oleh enam model pembelajaran mesin menggunakan set data negeri Selangor.



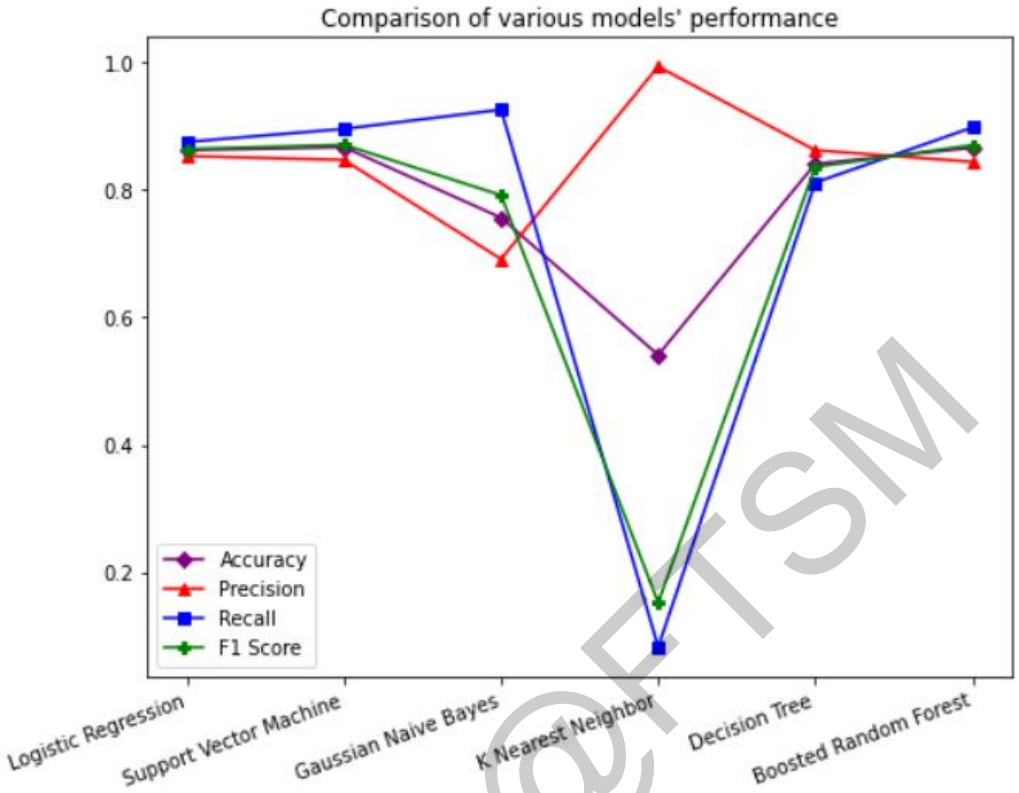
Rajah 5.1 *Confusion Matrix* bagi set data Selangor

Penilaian algoritma klasifikasi akan dijalankan perbandingan untuk menentukan model ramalan risiko kesihatan pesakit COVID-19 yang dapat membuat ramalan dengan prestasi yang tinggi di negeri Selangor. Jadual 5.1 meringkaskan semua penilaian metrik model ramalan negeri Selangor dalam satu jadual.

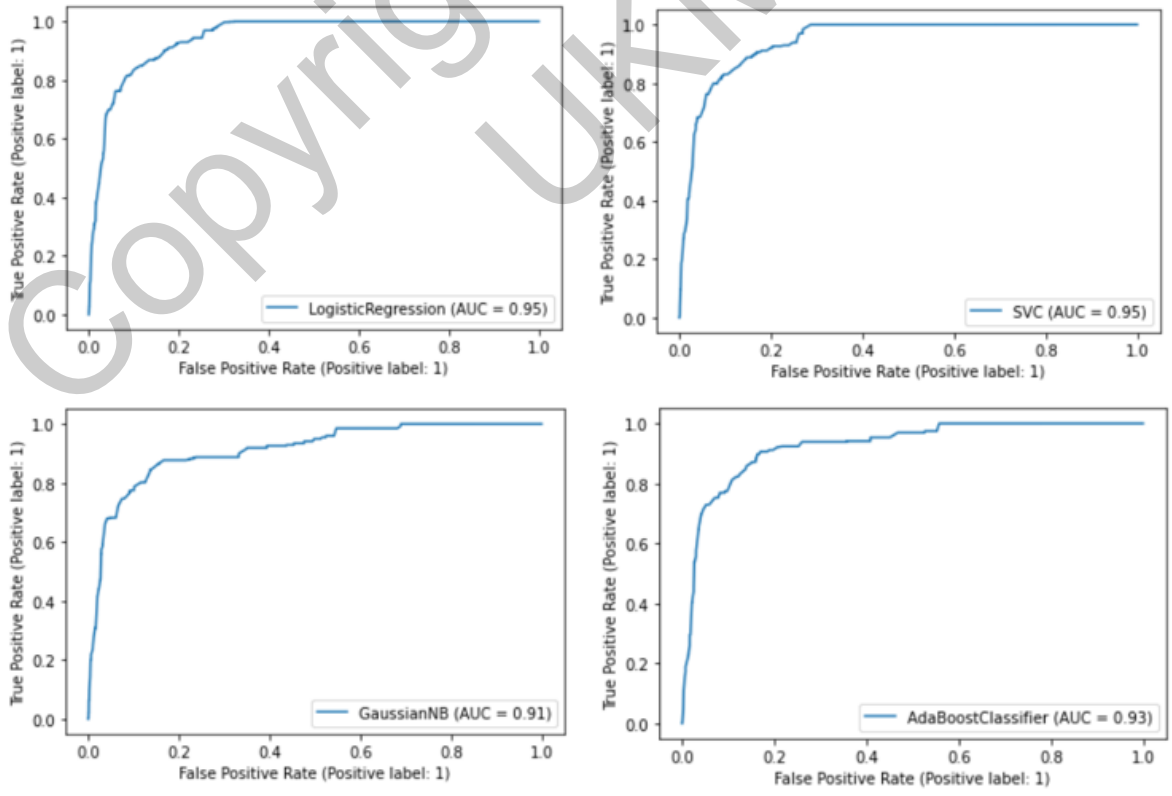
Jadual 5.1 Penilaian metrik model ramalan (Selangor)

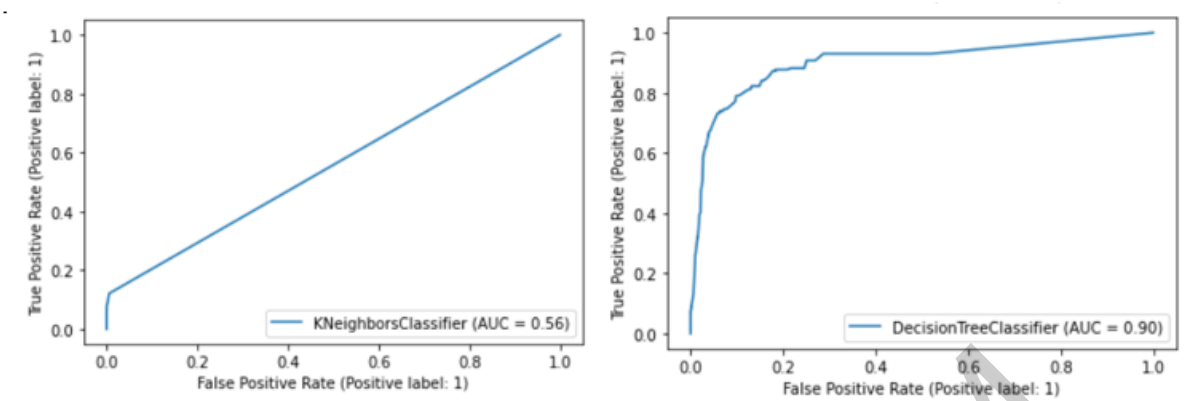
Metrik Penilaian	Accuracy	Precision	Recall	Skor F1	Nilai AUC
Regresi Logistik	0.86	0.85	0.88	0.86	0.95
Mesin Vektor Sokongan	0.87	0.85	0.90	0.87	0.95
Gaussian Naïve Bayes	0.76	0.69	0.93	0.79	0.91
K Nearest Neighbor	0.54	0.99	0.08	0.15	0.56
Pokok Keputusan	0.84	0.86	0.81	0.84	0.90
Hutan Rawak dengan AdaBoost	0.87	0.84	0.90	0.87	0.93

Jadual 5.1 mempamerkan metrik penilaian yang digunakan untuk mengukur prestasi model ramalan risiko kesihatan pesakit COVID-19. Dalam jadual 5.1, terdapat 2 model yang mencapai *accuracy* yang tinggi iaitu sebanyak 0.87 dalam membuat ramalan iaitu mesin vektor sokongan dan hutan rawak yang gabung *AdaBoost*. *Precision* yang dicapai oleh KNN adalah paling tinggi antara model ramalan iaitu sebanyak 0.99. Nilai *recall* yang paling tinggi adalah 0.93 dicapai oleh *Gaussian Naïve Bayes*. Skor F1 bagi gabungan hutan rawak dengan *AdaBoost* dan mesin vektor sokongan adalah sama tinggi dengan nilai sebanyak 0.87. Regresi Logistik mencapai skor F1 dengan nilai sebanyak 0.86. Nilai AUC yang diperoleh oleh regresi logistik dan mesin vektor sokongan adalah sama tinggi iaitu nilai sebanyak 0.95. Selepas perbandingan dijalankan di antara model ramalan, model regresi logistik, mesin vektor sokongan dan gabungan hutan rawak dengan *Adaboost* memperoleh penilaian metrik yang tinggi dan tiada perbezaan yang nyata. Oleh itu, *T-test* digunakan untuk membuktikan bahawa mana model yang mempunyai metrik penilaian paling tinggi dalam ramalan risiko kesihatan pesakit COVID-19. Rajah 5.2 menunjukkan graf perbandingan prestasi model ramalan di negeri Selangor. Rajah 5.3 menggambarkan *ROC curve* yang dihasilkan oleh model ramalan.



Rajah 5.2 Graf perbandingan prestasi model ramalan (Selangor)





Rajah 5.3 ROC curve (Selangor)

5.2 *T-test*

T-test merupakan ujian statistik yang digunakan untuk membuat perbandingan purata (min) antara dua kumpulan yang berbeza dan menentukan sama ada perbezaan min adalah signifikan. Nilai t merupakan nilai piawai yang dihasilkan daripada data sampel semasa ujian hipotesis.

Dalam kajian projek ini, *t-test* digunakan sebagai ujian hipotesis untuk membandingkan prestasi model ramalan dan membuktikan model yang terpilih mempunyai prestasi yang tinggi dalam ramalan risiko kesihatan pesakit COVID-19. Untuk negeri Selangor, terdapat dua kes akan dijalankan termasuklah kes antara mesin vektor sokongan dengan regresi logistik dan kes antara mesin vektor sokongan dengan gabungan hutan rawak dengan *AdaBoost*. Aras signifikan yang digunakan adalah *alpha* dengan nilai 0.05 (95% sela keyakinan). Jadual 5.2 menunjukkan t-test bagi model mesin vektor sokongan dan regresi logistik.

Jadual 5.2 *T-test (SVM* dan *LR*)

T-test	Perincian
H0	Prestasi Mesin Vektor Sokongan kurang tinggi daripada Regresi Logistik dan tiada perbezaan min yang ketara antara satu sama lain. SVM <= LR
H1	Prestasi Mesin Vektor Sokongan adalah lebih tinggi daripada Regresi Logistik dan terdapat perbezaan min yang ketara antara satu sama lain. SVM > LR
Statistik ujian T	-2.138090
Nilai-p	0.049650
Kesimpulan	Oleh kerana nilai-p(=0.049650) < *alpha*(=0.05), hipotesis H0 ditolak. SVM >LR

Bagi kes pertama, hipotesis sifar menyatakan prestasi mesin vektor sokongan tidak tinggi daripada regresi logistik manakala hipotesis alternatif menyatakan prestasi mesin vektor sokongan lebih tinggi daripada regresi logistik. Nilai T-statistik adalah -2.138 dan nilai-p adalah 0.0497. Keputusan menunjukkan nilai-p adalah lebih rendah daripada aras signifikan. Oleh itu, kita dapat membuktikan bahawa prestasi mesin vektor sokongan adalah lebih tinggi daripada regresi logistik pada 95% sela keyakinan. Jadual 5.3 menunjukkan t-test bagi model mesin vektor sokongan dan hutan rawak dengan *Adaboost*.

T-test	**Perincian**
H0	Prestasi Mesin Vektor Sokongan kurang tinggi daripada Hutan Rawak dan tiada perbezaan min yang ketara antara satu sama lain. SVM <= RF
H1	Prestasi Mesin Vektor Sokongan adalah lebih tinggi daripada Hutan Rawak dan terdapat perbezaan min yang ketara antara satu sama lain. SVM > RF
Statistik ujian T	1.500000
Nilai-p	0.104000
Kesimpulan	Oleh kerana nilai-p(=0.104000) > *alpha*(=0.05), hipotesis H0 diterima. SVM <= RF

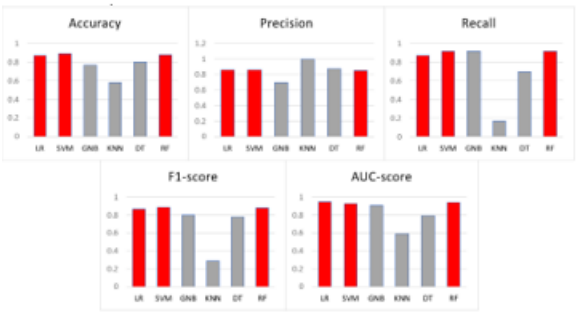
Jadual 5.3 *T-test* (*SVM* dan *RF*)

Bagi kes kedua, hipotesis sifar menyatakan prestasi mesin vektor sokongan tidak tinggi daripada hutan rawak manakala hipotesis alternatif menyatakan prestasi mesin vektor sokongan lebih tinggi daripada hutan rawak. Nilai T-statistik adalah -1.5 dan nilai-p adalah 0.104. Keputusan menunjukkan nilai-p adalah lebih tinggi daripada aras signifikan. Oleh itu, kita dapat mengesahkan bahawa prestasi mesin vektor sokongan adalah lebih rendah daripada gabungan hutan rawak dengan *AdaBoost* pada 95% sela keyakinan.

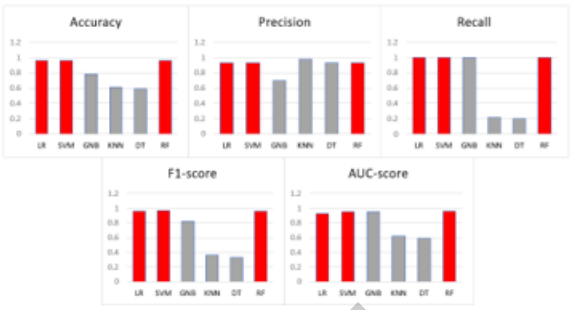
5.3 Keputusan Prestasi Model Ramalan Mengikut Negeri

Bahagian ini menunjukkan hasil keputusan penilaian metrik model ramalan mengikut negeri-negeri di Malaysia. Keputusan penilaian metrik model ramalan dipaparkan dalam bentuk carta bar seperti rajah 5.4 berikut.

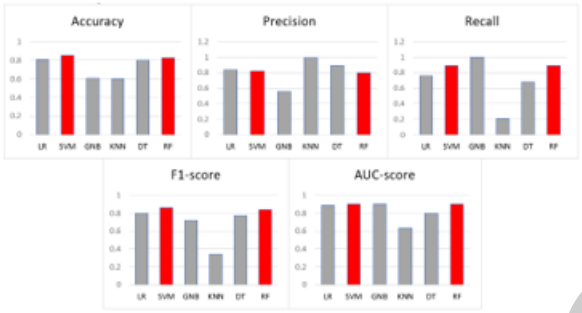
Johor



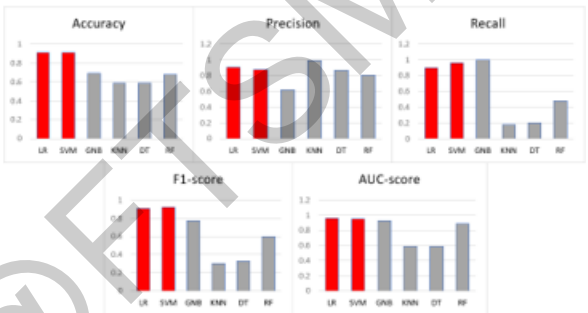
Kedah



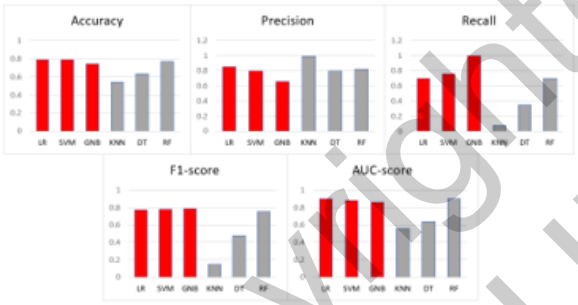
Kelantan



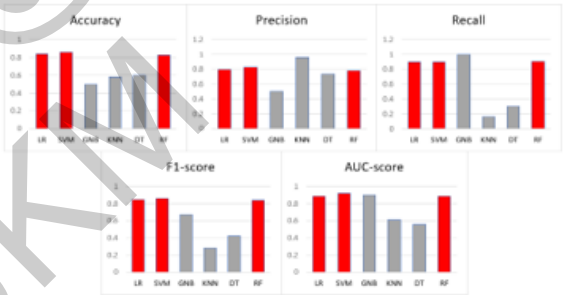
Melaka



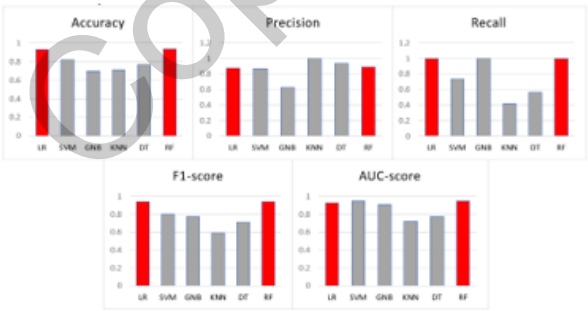
Negeri Sembilan



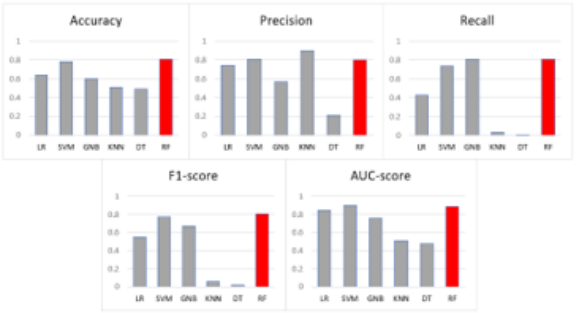
Pahang



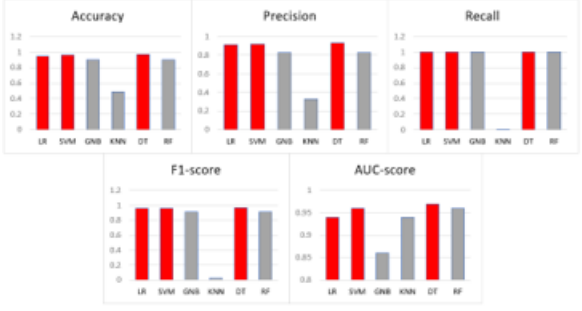
Pulau Pinang



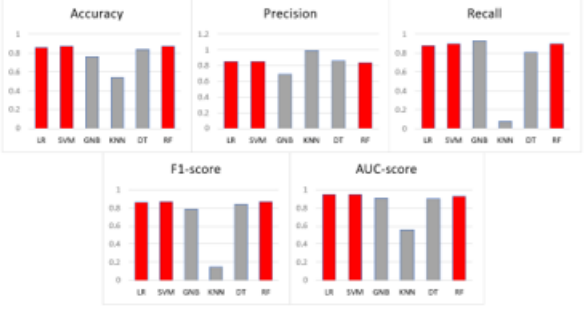
Perak



Perlis



Selangor



Rajah 5.4 Keputusan Prestasi Model Ramalan Mengikut Negeri

Kebanyakan negeri di Malaysia mengesahkan bahawa model regresi logistik, mesin vektor sokongan dan hutan rawak memperoleh prestasi model ramalan yang tinggi dari keseluruhan penilaian metrik. Selepas ujian hipotesis *T-test* dijalankan, model terbaik bagi setiap negeri telah ditentukan dan dipamerkan dalam jadual 5.4 di bawah.

Jadual 5.4 Model terbaik bagi setiap negeri

Negeri	**Model Terbaik**	**Negeri**	**Model Terbaik**
Johor	Hutan Rawak	Perak	Hutan Rawak
Kedah	Hutan Rawak	Perlis	Mesin Vektor Sokongan
Kelantan	Mesin Vektor Sokongan	Selangor	Hutan Rawak
Melaka	Mesin Vektor Sokongan	Terengganu	Regresi Logistik
Negeri Sembilan	Regresi Logistik	Sabah	Regresi Logistik
Pahang	Mesin Vektor Sokongan	Sarawak	Hutan Rawak
Pulau Pinang	Hutan Rawak		

Model gabungan hutan rawak dengan *AdaBoost* merupakan model majoriti yang diberi pengesahan sebagai model terbaik di setiap negeri. Oleh itu, model gabungan hutan rawak dengan *AdaBoost* adalah model yang paling sesuai digunakan untuk melakukan ramalan COVID-19 di Malaysia.

5.4 Perbandingan keputusan model dengan kajian lepas

Jadual 5.5 menunjukkan prestasi dan penilaian model ramalan kajian ini bersama dengan hasil keputusan penilaian model ramalan yang telah dihasilkan dalam kajian lepas.

Jadual 5.5 Perbandingan prestasi model ramalan secara umum dengan kajian lepas

	Kajian					**Kajian Buvana**				
Metrik Penilaian	**Accuracy**	**Precision**	**Recall**	**Skor F1**	**Nilai AUC**	**Accuracy**	**Precision**	**Recall**	**Skor F1**	**Nilai AUC**
Regresi Logistik	0.849 *	0.845 *	0.845 *	0.843 *	0.908 *	0.809	0.798	0.831	0.814	0.925
Mesin Vektor Sokongan	0.851 *	0.837 *	0.869 *	0.851 *	0.911 *	0.850	0.811	0.916	0.860	0.905
Gaussian Naïve Bayes	0.682	0.632	0.956	0.756	0.880	0.799 *	0.786 *	0.828 *	0.806 *	0.894 *
K Nearest Neighbor	0.579	0.925	0.165	0.264	0.631	0.923 *	0.937 *	0.909 *	0.923 *	0.966 *
Pokok Keputusan	0.703	0.806	0.487	0.576	0.716	0.945 *	0.939 *	0.952 *	0.946 *	0.977 *
Hutan Rawak dengan AdaBoost	0.831 *	0.825 *	0.842 *	0.826 *	0.909 *	X	X	X	X	X

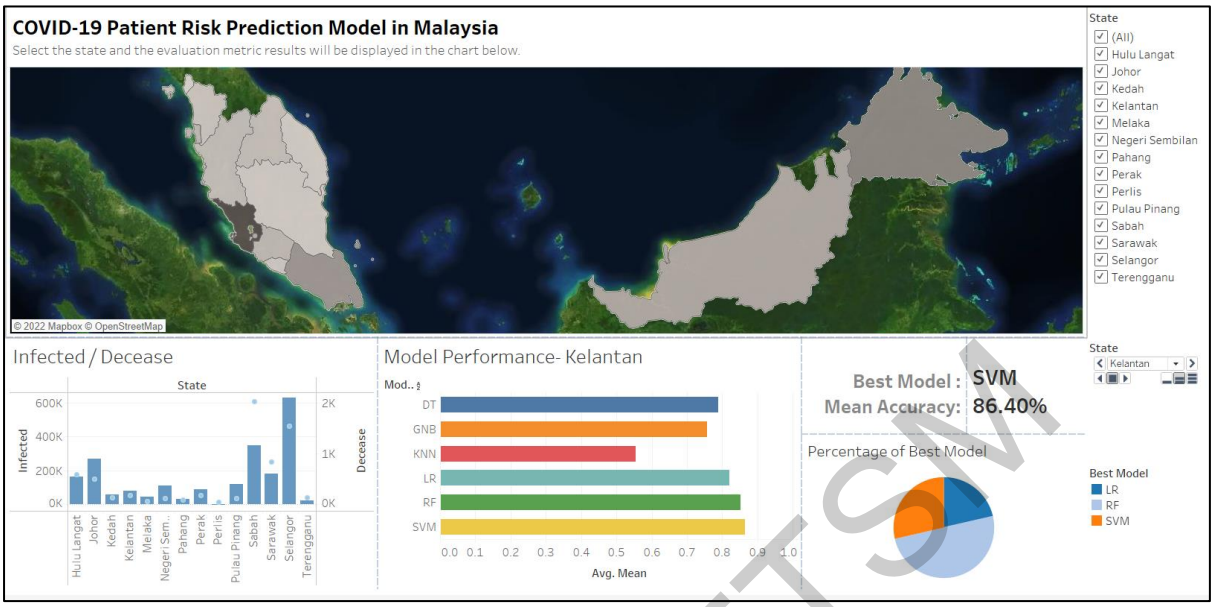
Berdasarkan jadual 5.5, kajian Buvana tidak mengaplikasikan model hutan rawak dengan *AdaBoost* dalam kajiannya. Justeru, tiada keputusan mengenai model hutan rawak dengan *AdaBoost* dipaparkan dalam jadual 5.5 dan digantikan dengan symbol ‘X’. Kajian Buvana mendapati bahawa pokok keputusan merupakan antara model ramalan yang mencapai prestasi yang paling tinggi dalam ramalan COVID-19. Metrik penilaian model pokok keputusan dalam kajian Buvana mencapai *accuracy* sebanyak 95%, *precision* sebanyak 94%, *recall* sebanyak 95%, skor F1 sebanyak 95% dan *AUC-score* sebanyak 98%.

Pokok keputusan dalam kajian ini hanya mencapai keputusan yang sederhana dalam ramalan risiko COVID-19.

Secara umumnya, prestasi model kajian ini adalah tinggi untuk model regresi logistik dan mesin vektor sokongan berbanding dengan kajian Buvana. Manakala, model *Gaussian Naïve Bayes*, *K-Nearest Neighbor*, dan pokok keputusan menunjukkan prestasi yang rendah jika berbanding dengan kajian Buvana. Hal ini disebabkan model ramalan tersebut tidak dapat meramal keputusan dengan tepat dengan set data yang besar. Kajian ini mendapati bahawa model gabungan hutan rawak dengan *AdaBoost* berprestasi tinggi selepas penggunaan ujian hipotesis *T-test* manakala kajian buvana melaporkan bahawa model pokok keputusan berprestasi tinggi dalam ramalan COVID-19. Hal ini kerana kajian ini mengaplikasikan set data di Malaysia sahaja dan kajian Buvana menggunakan set data yang terdiri daripada negara-negara lain. Secara umumnya, faktor hasil keputusan yang berbeza dengan kajian lepas terpengaruh oleh bilangan set data dan atribut data yang digunakan dalam kajian ini. Bilangan set data yang digunakan oleh kajian Buvana adalah kecil iaitu sebanyak jumlah 2500 baris. Oleh itu, masalah kolinearitи akan berlaku kerana kolerasi antara atribut adalah tinggi untuk set data yang kecil. Pokok keputusan dapat mengendalikan masalah kolinearitи dengan cekap berbanding dengan model ramalan yang lain. Justeru, pokok keputusan diiktiraf sebagai model berprestasi tinggi dalam kajian Buvana. Kajian ini menggunakan jumlah set data yang besar iaitu sebanyak 405947 data. Model KNN, Gaussian Naïve Bayes, dan pokok keputusan tidak dapat meramal keputusan dengan tepat dengan set data yang besar kerana memerlukan kos pengiraan yang tinggi. Hutan Rawak dipilih sebagai model yang berprestasi tinggi kerana hutan rawak merupakan model ensemble yang menggabungkan pokok keputusan untuk menghasilkan model yang cekap dan mengendali masalah *overfitting* dengan baik berbanding dengan model ramalan yang lain.

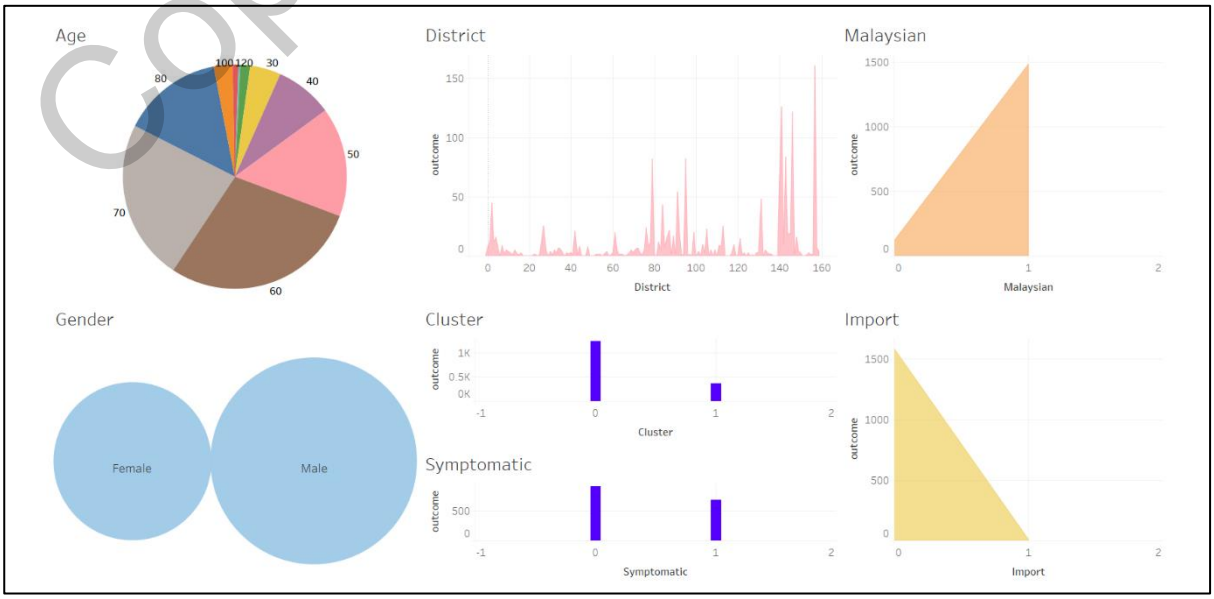
5.5 Papan Pemuka

Rajah 5.5 mempamerkan papan pemuka *Tableau* yang mengandungi analisis keputusan model ramalan yang terdiri daripada 13 negeri di Malaysia. Pengguna boleh memilih jenis negeri dan analisis keputusan ramalan model akan berubah mengikut negeri yang terpilih.



Rajah 5.5 Papan Pemuka yang dihasilkan dengan *Tableau*

Atribut data yang penting untuk ramalan risiko pesakit Covid-19 adalah termasuk umur, jantina, daerah, kluster, gejala, warganegara dan kes import. Keputusan menunjukkan individu yang berumur 60 tahun ke atas merupakan golongan yang berisiko akibat jangkitan COVID-19. Risiko pesakit COVID-19 untuk lelaki adalah tinggi berbanding dengan perempuan. Kebanyakan kes risiko tinggi adalah berasal daripada negeri Selangor. Warganegara Malaysia lebih terdedah kepada risiko jangkitan COVID-19. Kebanyakan kes jangkitan COVID-19 yang berisiko tinggi bukan dari kluster, tiada gejala, dan bukan kes import. Statistik deskriptif ditunjukkan dalam rajah 5.6 untuk memudahkan kefahaman penggnuna.



Rajah 5.6 Papan Pemuka *Tableau* (Statistik deskriptif)

6 KESIMPULAN

Secara keseluruhannya, projek ini melibatkan algoritma pengelasan dalam menjalankan proses pembelajaran mesin dalam ramalan risiko kesihatan pesakit COVID-19. Set data pesakit COVID-19 menjalankan pemprosesan pada peringkat awal untuk memastikan kualiti data terjamin dan tiada ralat yang boleh menjejaskan keputusan akhir ramalan. Keputusan model ramalan risiko pesakit COVID-19 dihasilkan melalui perbandingan prestasi antara model ramalan dengan menggunakan metrik penilaian. Hasil kajian yang didapati dalam kajian ini adalah model gabungan hutan rawak dengan *AdaBoost* mencapai prestasi yang paling tinggi berbanding dengan model ramalan lain seperti regresi logistik, mesin vektor sokongan, *Gaussian Naïve Bayes*, *K-Nearest Neigbor*, dan pokok keputusan. Keseluruhan keputusan ramalan dipaparkan dalam papan pemuka *Tableau.* Keputusan ramalan yang diperoleh diharapkan dapat membantu pihak hospital dalam pengagihan kelengkapan perubatan dan memberi maklumat yang bermakna kepada orang ramai.

7 RUJUKAN

Ahmad Suhael. 2021. Kebolehjangkitan COVID-19 di 11 negeri melebihi purata kebangsaan. https://www.bharian.com.my/berita/nasional/2021/05/817933/kebolehjangkitan-covid-19-di-11-negeri-melebihi-purata-kebangsaan [18 Mei 2021].

Annabelle Lee, Aidila Razak. 2021. Katil wad ICU kian berkurangan ketika kes parah Covid-19 makin bertambah. https://www.malaysiakini.com/news/559241 [16 Januari 2021].

AJ Myles, RN Feudale. 2004. An introduction to decision tree modelling. *Journal of Chemometrics.*18(6) pg 275-285.

A. Paul, DP. Mukherjee. 2018. Improved Random Forest for Classification. *Journal of IEEE Xplore.* 27(8) pg 4012-4024.

AS. Ritonga. 2018. Penerapan metode support vector machine dalam klasifikasi kualitas pengelasan smaw. *Journal of Edutic Pendidikan dan Informatika* 5(1).

Avinash Naviani. 2018. AdaBoost Classifier in Python. https://www.datacamp.com/community/tutorials/adaboost-classifier-python [8 Oktober 2021].

Ayush Pant. 2019. Introduction to Logistic Regression. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 [6 Oktober 2021].

Bernama. 2021. COVID-19: Lebih 22.9 juta golongan dewasa kini lengkap divaksin. https://www.astroawani.com/berita-malaysia/covid19-lebih-229-juta-golongan-dewasa-kini-lengkap-divaksin-362857 [11 November 2021].

Buvana, M, Muthumayil, K. 2021. Prediction of COVID-19 Patient using Supervised Machine Learning Algorithm. *Journal of Science Malaysia* 50(8) :2479-2497.

Celestine Iwendi , Ali Kashif Bashir , Atharva Peshkar. 2020. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Journal of Frontiers Public Health* 357(8) : 1-9.

Chris Nicholson. 2020. Evaluation Metrics for Machine Learning - Accuracy, Precision, Recall, and F1 Defined. https://wiki.pathmind.com/accuracy-precision-recall-f1 [1 Disember 2021].

Dr. Jemilah. 2021. Covid-19 mungkin jadi endemik, perlukan pendekatan berbeza. https://www.hmetro.com.my/mutakhir/2021/06/721727/covid-19-mungkin-jadi-endemik-perlukan-pendekatan-berbeza-dr-jemilah [23 Jun 2021].

ET. Pamungkas. 2017. Metode Regresi Logistik Biner pada Faktor yang Mempengaruhi Kesembuhan Pasien Penderita Demam Berdarah Dengue di RSUD dr. Iskak Kabupaten Tulungagung. *Journal of School of Electronics and Computer Science.*

Fauziyah Dewi. 2019. K-Nearest Neighbor (KNN) in R. https://medium.com/@fauziyahdewi16/k-nearest-neighbor-knn-in-r-87a3064e6e31 [6 November 2021].

Fernanda Sumika Hojo De Souza , Natália Satchiko Hojo-Souza. 2021. Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study With Brazilian Data. *Journal of Frontiers* 4 :1-13.

Hasimi Muhamad. 2020. KKM belanja RM1.17 bilion tangani COVID-19. Berita Malaysia. https://www.astroawani.com/berita-malaysia/kkm-belanja-rm117-bilion-tangani-covid19-271393 [3 Disember 2020].

Hidayah Hairom. 2021. Larangan rentas negeri kekal, SOP Aidilfitri masih diperhalusi https://www.sinarharian.com.my/article/135746/KHAS/Covid-19/Larangan-rentas-negeri-kekal-SOP-Aidilfitri-masih-diperhalusi. [27 April 2021].

Hilal Azmi. 2019. Memahami teknologi kecerdasan buatan (AI). Berita Malaysia. https://www.astroawani.com/berita-malaysia/memahami-teknologi-kecerdasan-buatan-ai-213109 [20 Julai 2019].

ICHI.PRO. 2020. Metrik Regresi dan Klasifikasi dalam Pembelajaran mesin dengan Python. https://ichi.pro/id/metrik-regresi-dan-klasifikasi-dalam-pembelajaran-mesin-dengan-python-120533115696042 [14 November 2021].

Jason Brownlee. 2020. SMOTE for Imbalanced Classification with Python. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ [1 Januari 2022].

Jason Brownlee. 2020. Tour of Evaluation Metrics for Imbalanced Classification. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/ [3 Januari 2021].

Joseph Kaos JR. 2020. Health DG: Centres needed because 15% did not comply with home quarantine. https://www.thestar.com.my/news/nation/2020/03/31 [4 Jun 2021].

Kirsten Barkved. 2022. The Difference Between Training Data vs. Test Data in Machine Learning. https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machinelearning [11 Februari 2022].

KKM. 2021. Situasi Terkini COVID-19 di Malaysia. https://covid-19.moh.gov.my/ [7 Julai 2021].

Lilly Chen. 2019. Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)-Step by Step Explained. https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725 [5 Disember 2021].

Mario A. Quiroz-Juárez, Armando Torres-Go´mez2, Irma Hoyo-Ulloa2, Roberto de J. Leo´n-Montiel3, Alfred B. U'Ren. 2021. Identification of high-risk COVID-19 patients using machine learning. *Journal of Institute Plos One* 16(9) : 1-21.

Matthew Martin. 2021. What is Non-Functional Requirement in Software Engineering? Types and Examples. https://www.guru99.com/non-functional-requirement-type-example.html [18 Oktober 2021].

ML Zhang, ZH Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Journal of Pattern Recognition Society* 40(7) pg 2038-2048.

N. Sakke. 1999. Bab 5 Analisis kajian, hlm 144-145. http://studentsrepo.um.edu.my/817/6/BAB5.pdf [25 Februari 2022].

Prateek Majumder. 2021. Gaussian Naive Bayes. https://iq.opengenus.org/gaussian-naive-bayes/ [15 November 2021].

Purva Huilgol. 2020. Precision vs. Recall – An Intuitive Guide for Every Machine Learning Person. https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/ [3 Mac 2022].

Raja. Noraina. 2022. Tujuh negeri guna katil ICU lebih 50 peratus. https://www.bharian.com.my/berita/nasional/2022/02/927236/tujuh-negeri-guna-katil-icu-lebih-50-peratus [26 Februari 2022].

ReQtest. 2012. Why is the difference between functional and Non-functional requirements important? https://reqtest.com/requirements-blog/functional-vs-non-functional-requirements/ [22 Februari 2022].

Rohith Gandhi. 2018. Support Vector Machine — Introduction to Machine Learning Algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [12 November 2021].

Saishruthi Swaminathan. 2018. Logistic Regression — Detailed Overview. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc [11 November 2021].

Simon James Fong, Nilanjan Dey and Jyotismita Chaki. 2020. An Introduction to COVID-19. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7307707/ [5 Julai 2021].

Sinar Harian. 2021. 26 Ogos.

Stacey Ronaghan. 2018. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 [13 Jun 2021].

Sujoy Kar1, Rajesh Chawla, Sai Praveen Haranath, Suresh Ramasubban, Nagarajan Ramakrishnan, RajuVaishya, Anupam Sibal & Sangita Reddy. 2021. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). Journal of Nature 11: 1-11.

Sunil Ray. 2017. Understanding Support Vector Machine(SVM) algorithm from examples. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [23 November 2021].

Thpanorama, M . 2022. Penjelasan teorem Bayes, aplikasi, latihan. https://ms.thpanorama.com/articles/matemticas/teorema-de-bayes-explicacin-aplicaciones-ejercicios.html [22 Mac 2022].

Tuga Mauritsius dan Faisal Binsar. 2020. Cross-Industry Standard Process For Data Mining (Crisp-Dm) https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/ [5 Jun 2021].

Venu Gopal. 2021. Evaluation Metrics for Classification Problems with Implementation in Python https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-problems-with-implementation-in-python-a20193b4f2c3 [27 Jun 2021].

Will Koehrsen. 2018. Hyperparameter Tuning the Random Forest in Python. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 [25 Disember 2021].

World Health Organization. 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-COVID-19---11-march-2020 [18 Jun 2021].

Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q. 2020. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. https://pubmed.ncbi.nlm.nih.gov/32173574/ [3 Jun 2021].

Teng Wei Zhun (A176160)
Azuraliza Abu Bakar
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia