PEMBANGUNAN MODEL DESKRIPTIF DAN DIAGNOSTIK KES COVID-19 HULU LANGAT

Nurin Nabilah Binti Bahrin, Suhaila Binti Zainudin

[1]*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia*

Abstrak

Laporan teknik ini bertujuan untuk menganalisis perkembangan COVID-19 di Daerah Hulu Langat menggunakan pemodelan data pembelajaran mesin dan menunjukkan hasilnya melalui paparan interaktif di Tableau. Dalam fasa pemahaman data, data diperoleh dalam empat fail yang berbeza dan dibersihkan untuk memastikan kebolehpercayaan data. Proses pembersihan data ini melibatkan penyingkiran rekod yang tidak berguna serta menangani kesalahan ejaan. Fasa penyediaan data melibatkan pembahagian data menggunakan kaedah pecahan peratusan untuk pengelasan. Empat jenis model klasifikasi digunakan, iaitu Support Vector Machine (SVM), Logistic Regression (LR), Rangkaian Neural Buatan (ANN), dan Peningkatan Kecerunan Ekstrem (XGBoost). Hasil penilaian model menunjukkan prestasi model klasifikasi yang terbaik berdasarkan metrik laporan klasifikasi seperti ketepatan, skor f1, ingat semula (Recall), skor kuasa dua, min kuasa dua ralat (MSE), min ralat mutlak (MAE), ketepatan klasifikasi, dan ralat jadual. Kesimpulannya, kajian ini membuktikan bahawa analisis dan model klasifikasi berdasarkan pembelajaran mesin dan paparan interaktif juga membantu memvisualisasikan data dengan lebih baik. Hasil kajian ini diharapkan memberi sumbangan penting kepada kaedah menangani pandemik COVID-19.

Kata kunci: COVID-19, Daerah Hulu Langat, pemodelan data klasifikasi, paparan interaktif, visualisasi data COVID-19

Pengenalan

Pembelajaran mesin merupakan salah satu cabang kecerdasan buatan yang menerapkan algoritma yang mempelajari data untuk menentukan prestasi dan ketepatan sesuatu ramalan. Untuk mendapat predikasi model yang tepat dan berkesan, kajian ini memerlukan data sebenar yang mengandungi kriteria perubatan dan demografik pesakit Data sebenar ini diperoleh dari PKD Hulu Langat yang mengandungi data pesakit sebenar COVID-19 di daerah Hulu Langat. Setiap data ini berkait rapat dalam membezakan tahap penyebaran virus kepada setiap individu. Penggunaan kaedah ini adalah untuk mengenal pasti risiko kematian daripada jangkitan virus. Dengan bantuan algoritma dan model analisis, COVID-19 dapat dihindarkan, dihalang dan diramal untuk penggunaan masa akan datang.

Melalui pembelajaran mesin, pemodelan data melibatkan penggunaan pelbagai algoritma pembelajaran mesin untuk memproses data dan menghasilkan model yang boleh meramal atau mengelaskan dengan ketepatan yang tinggi. Proses ini memberi tumpuan kepada mengenal pasti corak dan hubungan dalam data yang boleh digunakan untuk membuat keputusan atau melaksanakan tugas tertentu secara automatik. Dengan model ini, diharapkan ia dapat memberikan pemahaman yang lebih baik tentang corak penyebaran penyakit, menyokong pembuatan keputusan untuk pihak berkuasa dan kakitangan perubatan, dan membantu dalam usaha mencegah dan menangani pandemik COVID-19.

Dalam konteks pemantauan COVID-19, terdapat dua jenis paparan interaktif yang penting, iaitu WHO Coronavirus (COVID-19) Dashboard yang menunjukkan data global mengenai seksyen kematian, kes aktif, dan vaksinasi, dan juga paparan interaktif COVIDNOW di Malaysia yang

menampilkan data mengenai seksyen kematian, kes aktif, vaksinasi, dan kemasukan ke hospital mengikut negeri di Malaysia.

Pernyataan Masalah

COVID-19 merupakan penyakit berjangkit skala global yang menyerang negara. Penyebaran COVID-19 ini masih berlaku di Malaysia dan berada dalam tahap berbahaya. Kesesakan di hospital, pusat penilaian COVID-19 yang menyebabkan kuarantin kendiri di rumah antara punca virus merebak dengan lebih pantas. Keadaan ini meningkatkan keperluan untuk melakukan proses ramalan tahap jangkitan & penyebaran virus. Dalam kajian ini, terdapat beberapa permasalahan yang perlu diselesaikan berkaitan dengan situasi COVID-19 di kawasan Hulu Langat.

Permasalahan pertama yang dihadapi adalah kekurangan kajian analisis yang menyeluruh bagi data pesakit COVID-19 di kawasan Hulu Langat dalam mengenal pasti jangkitan dan status pesakit secara tepat. Data yang ada memerlukan analisis yang lebih mendalam untuk mengidentifikasi faktor-faktor yang mempengaruhi kebarangkalian seseorang pesakit dijangkiti virus COVID-19, yang pada gilirannya mempengaruhi langkah-langkah proaktif untuk menguruskan penyebaran virus dengan berkesan.

Permasalahan kedua adalah ketiadaan pemodelan klasifikasi yang menggunakan algoritma pembelajaran mesin yang khusus dalam menguji prestasi pesakit COVID-19 di Hulu Langat. Kekurangan pemodelan yang tepat dapat menghalang kemampuan untuk memahami pola jangkitan dan penyebaran virus dengan lebih mendalam, serta mengidentifikasi potensi tindakan pencegahan dan penanganan yang lebih baik.

Permasalahan ketiga adalah ketiadaan papan pemuka interaktif yang menyajikan deskriptif dan diagnostik kes pesakit COVID-19 secara lengkap dan informatif bagi Daerah Hulu Langat. Papan

pemuka yang sedia ada mungkin tidak memberikan gambaran yang cukup untuk merangka strategi pencegahan yang efektif dan mengambil langkah-langkah keselamatan yang sesuai dalam mengurangkan penyebaran virus di kawasan tersebut. Oleh itu, penyelidikan lebih mendalam dan penambahbaikan dalam analisis, pemodelan klasifikasi, dan papan pemuka interaktif perlu dilakukan untuk mengatasi permasalahan dalam situasi COVID-19 di Hulu Langat.

Cadangan Penyelesaian

Cadangan penyelesaian kajian ini terbahagi kepada dua iaitu dari segi analisis menggunakan pembelajaran mesin dan analisis dalam paparan interaktif. Cadangan pertama untuk kajian ini adalah dengan melakukan analisis deskriptif dengan menggunakan berbagai model pembelajaran mesin untuk mempelajari data pesakit COVID-19 pada tahun 2021 di daerah Hulu Langat. Pendekatan ini akan memberikan gambaran yang lebih lengkap tentang corak dan pola data pesakit, dengan setiap model memberikan pandangan yang berbeza yang membantu memahami perbezaan dan persamaan dalam data. Cadangan penyelesaian kedua ialah menghasilkan paparan interaktif yang memaparkan gambaran situasi COVID-19, jangka masa, dan punca jangkitan di daerah Hulu Langat. Paparan interaktif ini akan memudahkan pengguna untuk meneroka dan memahami data dengan lebih mendalam melalui peta interaktif, graf trend jangka masa kesembuhan dan kematian, serta visualisasi yang membantu mengenal pasti punca jangkitan yang paling dominan.

Objektif Kajian

Objektif utama projek ini adalah:

1. Menganalisis data pesakit COVID-19 untuk mengenal pasti jangkitan dan status pesakit.
2. Membangunkan pemodelan klasifikasi dengan menggunakan algoritma pembelajaran mesin untuk menguji prestasi data pesakit COVID-19.

3. Membangunkan papan pemuka interaktif bagi menunjukkan deskriptif dan diagnostik kes pesakit COVID-19.

Jadual 1 : Sorotan Susastera

Pengarang	Set Data	Atribut Data	Kaedah	Keputusan
Sumana H V, Ashok Neelammanavar, Sudeep K R, Veena R dan Dr Rajashree Shetter (2021)	Set Data COVID-19 Pusat Sains dan Kejuruteraan Sistem Johns Hopkins (CSSE)	1 fail data dan 10 atribut	1)"*Long short-term memory (LSTM)"* 2)"*Polynomial regression (PR)"* 3)"*Support Vector Machine (SVM)"*	Keputusan dalam RSME 1)47.15 2)26463.63 3)26857.69.
Adwitiya Sinha Dan Megha Rathi (2021)	Set Data untuk COVID-19 (DS4C) di Korea Selatan dari Disember 2019 hingga Mac 2020	1 fail data dan 9 atribut	1)"*Linear Regression (LR)*" 2)"*Support Vector Machine*" 3)"*Artificial Neural Network (ANN)*"	Keputusan dalam Accuracy 1)91% 2)97% 3)99%
Siti Nurhidayah Sharin, Mohamad Khairil Radzali, dan Muhamad Shirwan Abdullah Sani (2021)	Kementerian Kesihatan Malaysia (KKM) bermula dari Julai 2020 (S3 2020) hingga Jun 2021 (S2 2021).	2 fail data, 24 atribut dan 10 atribut	1)"*Support Vector Regression (SVR)*"	Keputusan dalam RSME 1)0.413
G. Vetrugno, P. Laurenti, F. Franceschi, F. Foti; F. D'Ambrosio (2021)	Data pesakit "*warded*" COVID-19 di Fondazione Policlinico A. Gemelli IRCCS, hospital universiti penjagaan tertiari di Rom, Itali, bermula daripada 5 Mac 2020 hingga 5 Jun 2020.	1 fail data, dan 4 atribut	1)"*Decision Tree Analysis*"	Keputusan dalam ROC 1)0.9633 2)0.9124 3)0.9188 4)0.9768
Arjun Dutta, Aman Gupta, dan Farhan Hai Khan (2021)	Set Data COVID-19 Pusat Sains dan Kejuruteraan Sistem Johns Hopkins (CSSE)	1 fail data, dan 10 atribut	1)"*Long short-term memory*" (LSTM) 2)"*Gated recurrent unit*" (GRU) 3)"*Bidirectional long-short term memory*" (BI-LSTM).	Keputusan dalam RSME 1.1)LSTM bagi kes yang disahkan (2.2428) 1.2)LSTM bagi kes kematian (0.0103) 2.1)GRU bagi kes yang disahkan (3.3158) 2.2)GRU bagi kes kematian (0.0402) 2.3)GRU bagi kes sembuh (8.4009 3.1) BI-LSTM bagi kes kematian (0.0077)

Jadual 1 memaparkan sorotan susastera berkenaan dengan kajian-kajian lepas yang telah dijalankan oleh para penyelidik berkenaan dengan tajuk kajian.

Metodologi Kajian

Kajian ini akan menggunakan pendekatan analisis berstruktur *CRISP-DM* yang terbukti berkesan dalam membangunkan sistem berbasis perlombongan data. Pendekatan ini menggabungkan analisis dengan pendekatan *agile* dan melibatkan empat fasa penting, yaitu Fasa Pengumpulan dan Pemahaman Data, Fasa Penyediaan Data, Fasa Pembangunan Model Klasifikasi, dan Fasa Penilaian Model.

1. Fasa Pemahaman Data

Antara perkara yang dilaksanakan dalam peringkat ini adalah:

i) Pernyataan Masalah

ii) Sorotan susastera

Fasa ini merupakan fasa awal dalam analisis berstruktur yang bertujuan untuk memahami data yang ada dan isu-isu yang ingin diselesaikan melalui kajian. Dalam fasa ini, perlu dilakukan pernyataan masalah dengan jelas untuk menggariskan objektif kajian. Selain itu, sorotan susastera juga dilakukan untuk mengumpulkan maklumat berkaitan dengan topik kajian dari sumber-sumber berbeza. Proses ini membantu dalam mengenal pasti punca data, kualiti data, dan memahami konteks kajian secara menyeluruh.

2. Fasa Pengumpulan Data

Fasa ini mempunyai dua bahagian iaitu pengumpulan data dan pra-pemprosesan data. Berdasarkan pemahaman data yang dilakukan di fasa pertama, fasa ini diteruskan dengan menentukan jenis dan bilangan atribut yang dipilih untuk digunakan dalam kajian. Berdasarkan empat fail set data yang diberikan oleh pihak Pejabat Kesihatan Daerah Hulu Langat, setiap atribut dalam setiap fail dikaji dan ditapis. Dengan memastikan setiap atribut mempunyai kesinambungan

untuk memenuhi objektif kajian, kaedah pemilihan fitur dilakukan dengan mengenalpasti jenis data pada setiap kolum.

Setelah pemilihan atribut dilakukan, proses pembersihan data dilakukan dengan menyemak dari segi format data, kuantiti rekod, serta sifat data. Penerokaan set data pesakit COVID-19 sangat penting dalam untuk memahami data dan pilih set data yang sesuai dengan kriteria projek. Statistik analisis yang deskriptif akan membantu memahami distribusi data dan mengenal pasti jenis dan nilai data seperti bunyi bising, data yang hilang, atau data lebihan. Data berkualiti rendah akan menghasilkan perlombongan keputusan data yang tidak tepat. Mengisi nombor yang hilang, menala data bunyi, mengenal pasti atau memadam bahagian lebihan dan menghapuskan masalah data yang tidak konsisten adalah sebahagian daripada proses pembersihan data.

Setelah itu, proses pembahagian data kepada set data latihan dan ujian yang berbeza adalah penting untuk memastikan algoritma pembelajaran mesin memahami dan mempelajari data secara berperingkat. Proses pembahagian ini dilakukan menggunakan kaedah pecahan peratusan dimana set data dibahagikan kepada dua bahagian, iaitu 80% sebagai set data latihan dan 20% sebagai set data ujian. Set data latihan akan digunakan untuk melatih algoritma dan membangunkan model klasifikasi, manakala set data ujian akan digunakan untuk menguji prestasi dan kebolehpercayaan model tersebut. Pembahagian peratusan ini membantu memastikan model dapat digeneralisasi dengan baik untuk data yang belum pernah dilihat sebelumnya.

3. Fasa Pembangunan Model Klasifikasi

Fasa Pembangunan ini mempunyai dua bahagian iaitu melatih dan menguji model klasifikasi dan analisis prestasi model klasifikasi. Pembangunan ini akan menggunakan empat jenis model klasifikasi iaitu Mesin Sokongan Vektor (SVM), Logistic Regression (LR), Rangkaian Neural Buatan (ANN) dan Peningkatan Kecerunan Ekstrem (XGBoost) menggunakan set data latihan bertujuan untuk mengenalpasti ketepatan dan kejituan prestasi model.

Setelah kesemua model menjalani proses klasifikasi, prestasi dan keputusan hasil daripada model akan dianalisis menggunakan empat kaedah iaitu metrik laporan klasifikasi, termasuk ketepatan, skor f1, ingat semula(Recall), skor kuasa dua, min kuasa dua ralat (MSE), min ralat mutlak (MAE), ketepatan klasifikasi dan ralat jadual. Sebagai contoh, kaedah ralat jadual digunakan untuk hasil model yang mempunyai lebih daripada dua jenis kelas. Kaedah ini mempunyai empat pecahan hasil gabungan diantara nilai yang diramalkan dan nilai sebenar. Pecahan gabungan ini dipanggil sebagai positif sebenar, positif palsu, negatif sebenar dan negatif palsu. Berdasarkan hasil kaedah ini, pengiraan akan dilakukan menggunakan teknik ketepatan, skor f1, ingat semula(Recall) seperti rajah dibawah.

4. Fasa Penilaian dan Visualisasi Model

Fasa Penilaian dan visualisasi ini mempunyai dua bahagian iaitu penilaian model prestasi terbaik dan visualisasi model menggunakan paparan interaktif. Penilaian ini adalah berdasarkan daripada prestasi model yang digunakan. Fasa Penilaian dan visualisasi ini mempunyai dua bahagian iaitu penilaian model prestasi terbaik dan visualisasi model menggunakan paparan interaktif. Dalam penilaian model prestasi, data yang telah dibersihkan di fasa pengumpulan data ditukar kepada bentuk binari melalui proses penggubahan data. Penggubahan data ini bertujuan untuk mengubah

data menjadi format yang sesuai untuk analisis klasifikasi. Selain itu, sampel data dalam penilaian model menggunakan kaedah sampel hibrid yang menggabungkan beberapa teknik sampel, seperti *oversampling* dan *undersampling*, untuk mengatasi ketidakseimbangan dalam kelas data yang dihadapi.

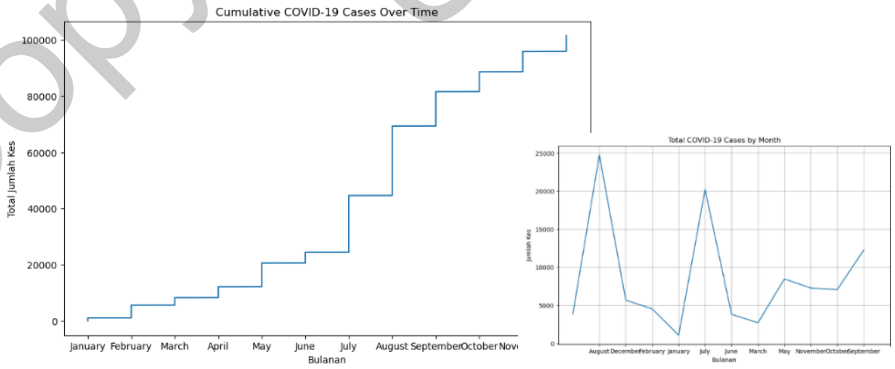
Prestasi model dinilai menggunakan berbagai metrik laporan klasifikasi, termasuk ketepatan (accuracy), skor f1, ingat semula (recall), skor kuasa dua, min kuasa dua ralat (MSE), min ralat mutlak (MAE), ketepatan klasifikasi, dan matriks ingat semula. Metrik-metrik ini digunakan untuk menilai kemampuan model dalam menangani masalah klasifikasi dengan kelas yang berbeza dan memberikan gambaran sejauh mana model dapat memprediksi kelas data dengan benar. Hasil penilaian ini akan membantu memilih model dengan prestasi terbaik untuk digunakan dalam analisis dan pemodelan data lebih lanjut.

Keputusan dan Perbincangan

Bahagian ini terbahagi kepada tiga iaitu visualisasi data, keputusan penilaian model dan papan pemuka data. Ketiga-tiga bahagian ini menjawab objektif utama kajian ini.

1. Visualisasi Data Bersih

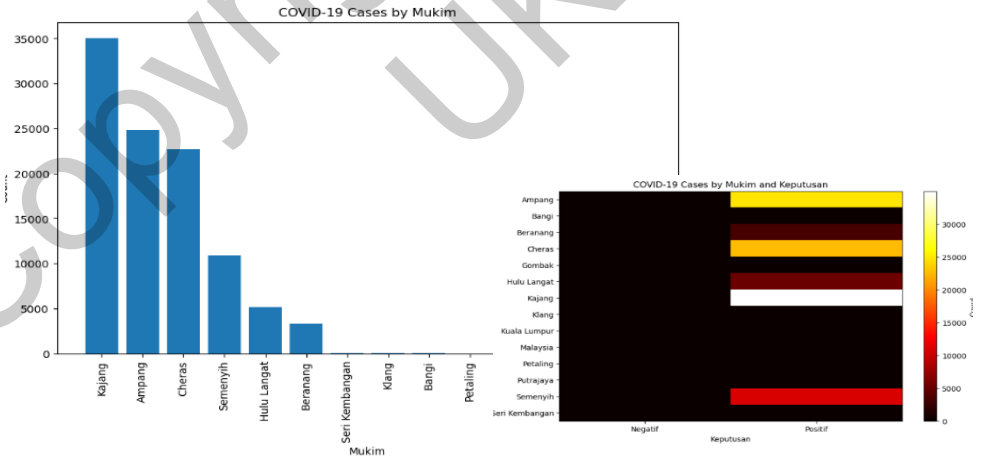
Visualisasi pertama adalah jumlah kes kumulatif COVID-19 melawan masa. Disini jumlah kes bermula daripada kes kosong hingga puluhan ribu. Berdasarkan paparan graf pertama, data menunjukkan peningkatan daripada bulan Januari 2021 hingga Disember 2021. Terdapat sedikit peningkatan yang mendadak pada bulan Julai iaitu sebanyak 40000 kes pesakit COVID-19. Setelah peningkatan yang tinggi di bulan Julai, kes terus meningkat hingga mencecah 100000 ribu kes pada bulan Disember. Bagi visualisasi kedua yang mana Jumlah kes mengikut bulanan, terdapat peningkatan yang agak tinggi pada dua bulan iaitu bulan Ogos dengan jumlah kes sebanyak 25000 ribu manakala bulan kedua tertinggi adalah bulan Julai bersamaan dengan 20000 kes.



Rajah 1: Visualisasi Jumlah Kes dan Kumulatif Kes menentang Bulanan

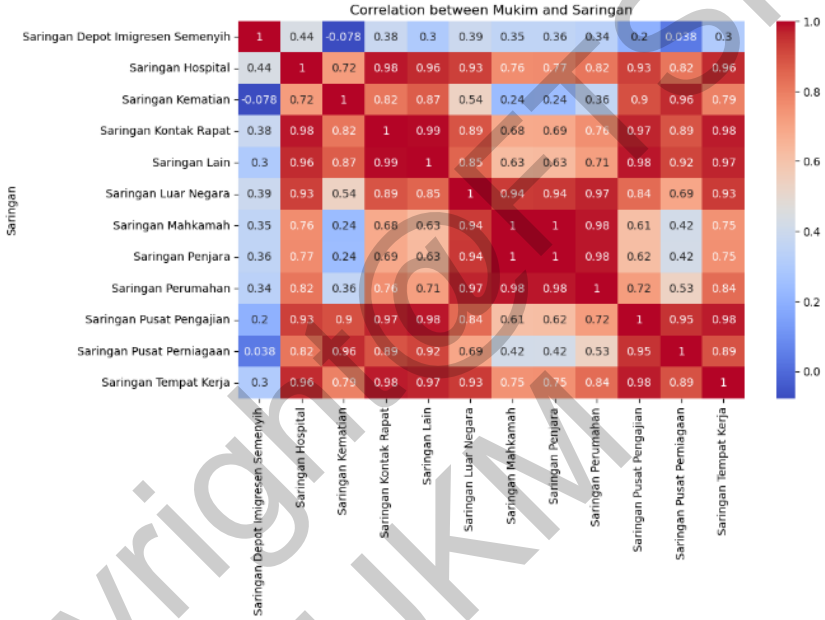
Bagi visualisasi yang ketiga, visualisasi ini menumpukan pada kolum mukim dimana kes COVID-19 mengikut muukim telah dikaji. Berdasarkan rajah paparan grafik bar dibawah, terdapat 14 mukim daripada mukim kajang hingga putrajaya. Graf ini menunjukkan peningkatan kes COVID-19 di 6 mukim yang dianggap mempunyai kes aktif pada tahun 2021. Kajang merupakan mukim yang tertinggi mempunyai kes COVID-19 iaitu sebanyak 35000 ribu manakala Ampang merupakan mukim kedua tertinggi dengan jumlah kes sebanyak 25000 ribu. Bagi mukim terendah iaitu Putrajaya dengan kes sebanyak 1 kes.

Seterusnya, visualisasi berkaitan dengan mukim dimana paparan grafik ini berbentuk Peta Panas dimana kawasan yang berwarna coklat gelap merangkumi mukim yang mempunyai data kurang dari 5000 ribu data. Bagi kawasan peta yang berwarna kemerahan merangkumi data bermula daripada 10000 ribu hingga 20000 ribu manakala selain daripada warna itu sudah merangkumi data yang mempunyai kes COVID-19 sebanyak 25000 ribu keatas. Peta panas menunjukkan kawasan yang panas dan kes tertinggi adalah mukim kajang.



Rajah 2 : Visualisasi Jumlah Kes menentang Mukim

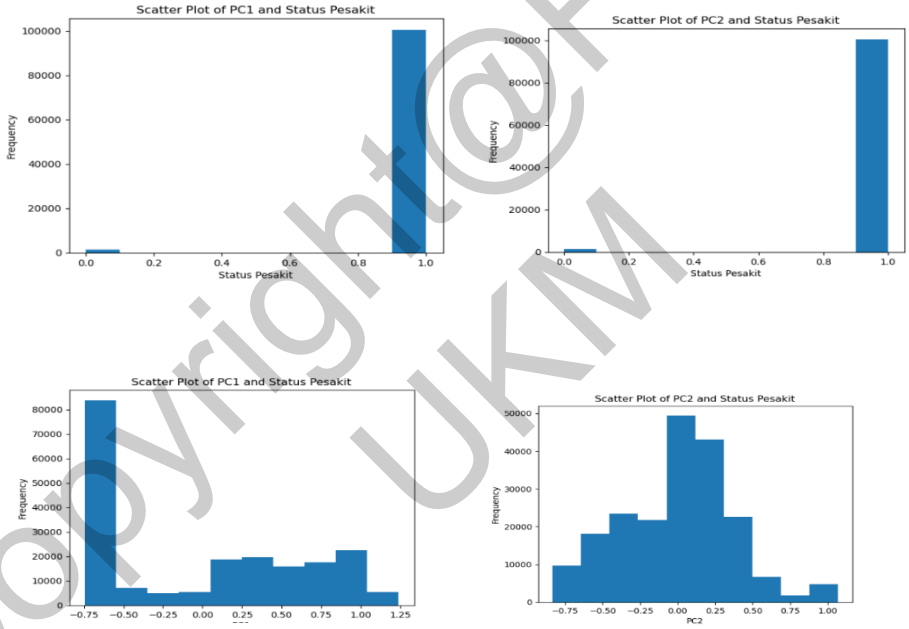
Visualisasi yang terakhir adalah visualisasi peta panas menunjukkan Jadual Kejadian Bersyarat bagi Jumlah Mukim dan Saringan. Menggunakan jumlah daripada kedua-dua nilai itu, peta panas kolerasi diantara Mukim dan Saringan COVID-19. Peta panas ini menggambarkan taburan suhu purata merentas bandar dalam rantau ini. Warna pada peta ini dikaitkan dengan tahap suhu iaitu warna biru mewakili suhu yang lebih rendah dan merah mewakili suhu yang lebih tinggi. Lebih cerah warna, lebih tinggi suhu di rantau ini.



Rajah 3 : Visualisasi Peta Panas Kolerasi bagi Mukim dan Saringan

2. Visualisasi Data PCA

Berdasarkan teknik PCA yang dilakukan untuk mengurangkan dimensi data kepada dua dimensi, paparan grafik dibawah yang berbentuk plot histogram menunjukkan plot sebelum dan selepas dimensi data dikurangkan. Berdasarkan plot yang pertama, kedua-dua data iaitu *PC1* dan *PC2* berada sangat jauh dari satu sama lain dan mempunyai jurang jumlah data yang sangat jauh. Data sembuh berada di frekuensi 100000 manakala data meninggal berada dekat dengan kosong. Namun, setelah proses PCA dilakukan, plot menjadi lebih tersusun dan graf tersebar dengan sangat baik. Tiada jurang jauh diantara data dan tidak hanya mengukur data berdasarkan nilai 0.0 dan 1.0 sahaja.



Rajah 4 : Visualisasi sebelum dan selepas proses pengurangan dimensi menggunakan PCA

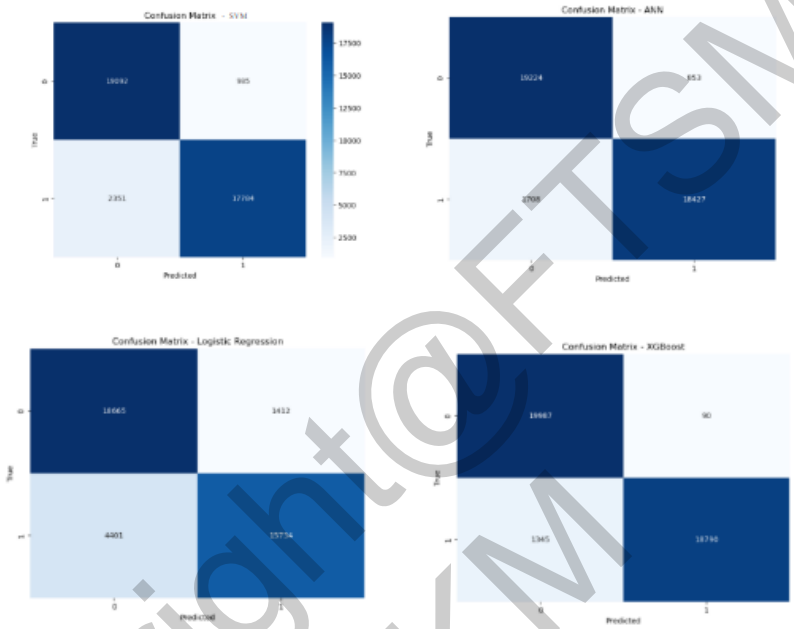
3. Keputusan Penilaian Model Ralat Jadual

Penilaian model menggunakan metrik ralat jadual adalah penting untuk menilai keupayaan model dalam mengklasifikasikan kelas yang berbeza. Ralat jadual menggambarkan jumlah kelas yang salah diklasifikasikan oleh model dalam bentuk matriks 2x2. Jadual ini mengandungi empat kemungkinan keputusan klasifikasi iaitu iaitu positif palsu, negatif palsu, positif benar dan negatif benar bagi setiap kelas yang berada dalam masalah klasifikasi. Dalam kajian ini, kita mempunyai empat model yang dinilai, iaitu SVM, ANN, LR, dan XGBoost.

Bagi metrik ralat jadual bagi model SVM, nilai ramalan bagi positif benar ialah sebanyak 17784 sampel manakala sebanyak 985 sampel salah dalam positif palsu. Bagi ramalan negatif pula, model dengan betul meramalkan sebanyak 19092 contoh yang termasuk dalam negatif benar manakala model salah dalam meramalkan 2351 contoh yang negatif palsu. Bagi model ANN pula, sebanyak 18427 sampel, manakala sebanyak 853 sampel salah dalam positif palsu. Bagi ramalan negatif, model dengan betul meramalkan sebanyak 19224 contoh yang termasuk dalam negatif benar, manakala model salah dalam meramalkan 1708 contoh sebagai negatif palsu.

Bagi metrik ralat jadual yang seterusnya iaitu model LR, nilai positif benar adalah sebanyak 15734 sampel, manakala sebanyak 4401 sampel salah dalam positif palsu. Bagi ramalan negatif, model dengan betul meramalkan sebanyak 18665 contoh yang termasuk dalam negatif benar, manakala model salah dalam meramalkan 1412 contoh sebagai negatif palsu. Bagi model XGBoost, terdapat 19987 sampel yang betul diramalkan sebagai negatif dan 18790 sampel yang betul diramalkan sebagai positif. Namun, terdapat 90 sampel yang salah diramalkan sebagai negatif palsu dan 1345 sampel yang salah diramalkan sebagai positif palsu.

Berdasarkan nilai ralat jadual di atas, model terbaik untuk projek ini adalah XGBoost. Model XGBoost menunjukkan nilai ralat jadual yang paling baik dibandingkan dengan model lain. Ia mempunyai jumlah kesalahan jenis 1 dan jenis 2 yang paling rendah, menunjukkan keupayaan model dalam mengklasifikasikan dengan lebih tepat.



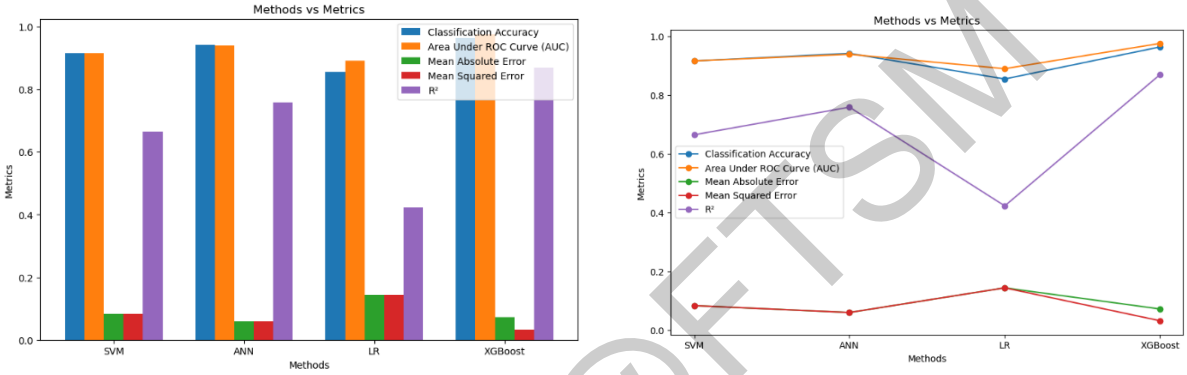
Rajah 5 : Metriks jadual ralat untuk setiap model

4. Keputusan Penilaian Model Keseluruhan

Setiap model telah dinilai menggunakan beberapa metrik prestasi berbeza seperti Ketepatan Klasifikasi, Kawasan dibawah lengkung *AUC*, Ralat Min Mutlak (MAE), Ralat Kuasa Dua (MSE) dan Skor R^2.

Berdasarkan dua graf dibawah, untuk Ketepatan Pengelasan, model *XGBoost* menunjukkan hasil yang memberangsangkan dengan nilai sekitar 0.964, diikuti oleh model *ANN* dan *SVM*, dan model *LR* mempunyai ketepatan yang paling rendah. Bagi nilai *AUC*, model *XGBoost* juga menunjukkan hasil terbaik dengan nilai sekitar 0.976, diikuti oleh model *ANN*, *SVM* dan *LR*.

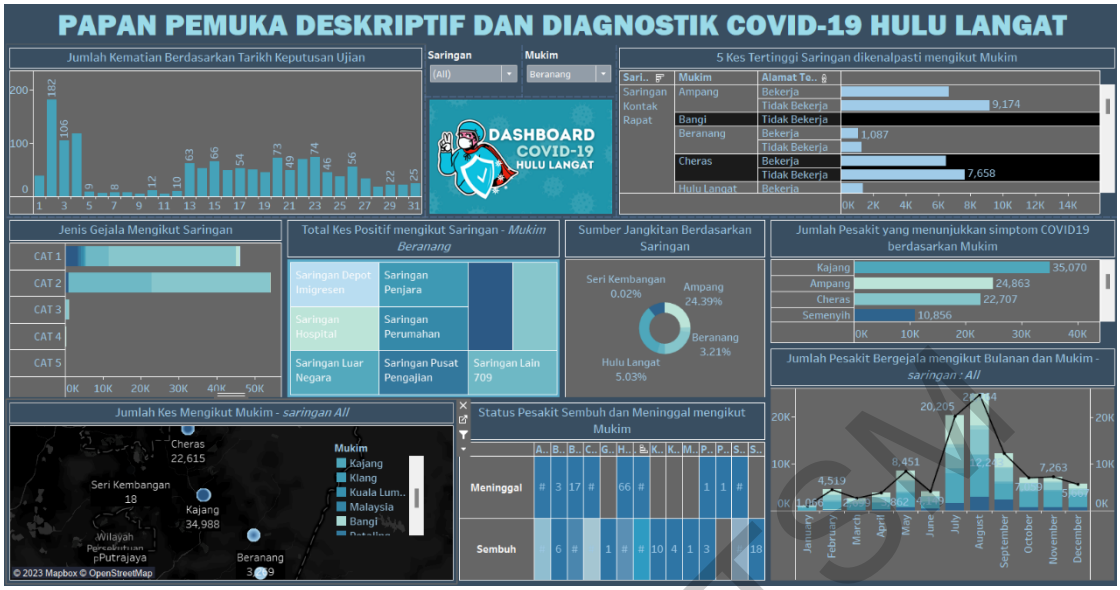
Seterusnya, bagi *MAE* dan *MSE*, model *ANN* menunjukkan hasil terbaik dengan nilai sekitar 0.060, manakala model *LR* mempunyai nilai *MAE* dan *MSE* tertinggi. Bagi nilai Skor R^2, model *XGBoost* menunjukkan hasil terbaik dengan nilai sekitar 0.870, diikuti oleh model *ANN*, *SVM* dan *LR* yang mempunyai Skor R^2 terendah.



Rajah 6 : Plot penilaian model yang digunakan

5. Papan pemuka data

Menerusi pemprosesan data yang bersih, pembangunan papan pemuka dapat dilaksanakan. Papan pemuka ini menggunakan perisian Tableau untuk menvisualisasikan data daripada set data kepada carta dan plot. Para pengguna boleh melihat paparan interaktif dalam bentuk carta agar mudah difahami dan dipelajari. Terdapat pelbagai jenis carta digunakan seperti carta bar, carta jadual, carta bar melintang, carta peta panas, carta pie, carta peta dan carta garisan. Setiap carta mempamerkan hubungan data-data yang berbeza. Selain itu, terdapat dua menu lungsur turun di papan pemuka dimana pengguna boleh memilih jenis saringan atau jenis mukim yang diinginkan. Fungsi terakhir adalah fungsi sorotan nama mukim dimana pengguna boleh menekan nama-nama mukim di kawasan peta dan carta yang berkaitan dengan mukim itu akan disorotkan untuk penglihatan pengguna.



Rajah 7 : Papan Pemuka Deskriptif dan Diagnostik COVID-19 Hulu Langat

Secara keseluruhan, papan pemuka deskriptif dan diagnostik COVID-19 Hulu Langat ini memberikan sumbangan yang besar kepada pengguna, penduduk Hulu Langat, dan Pejabat Kesihatan Daerah Hulu Langat dengan menyediakan visualisasi data yang interaktif, informatif, dan mudah digunakan. Hal ini memungkinkan semua pihak terlibat untuk membuat keputusan yang lebih tepat dan strategik dalam menangani pandemik COVID-19 dan menyumbang kepada upaya keselamatan dan kesejahteraan komuniti di daerah ini.

Dalam menjayakan kajian ini dari segi pemodelan data dan pembangunan papan pemuka interaktif, terdapat beberapa kelemahan yang tidak dapat dielakkan. Kelemahan yang berlaku membuatkan proses kajian yang dilakukan lebih panjang dan sukar. Bagi kelemahan projek ini, data mentah mengandungi banyak kesalahan dan kesalahan ejaan yang memerlukan usaha yang besar untuk pembersihan. Beberapa fail data terlalu besar dengan jumlah data yang sangat banyak, menyebabkan proses pembersihan dan analisis menjadi lebih rumit dan masa yang diambil lebih lama. Data kurang dalam segi kelas data, terutamanya kolum Status Pesakit yang menunjukkan jurang besar antara pesakit sembuh dan meninggal.

Namun begitu, terdapat beberapa kekuatan yang membantu dalam menjayakan kajian ini diantaranya adalah model-model yang digunakan memberikan keputusan yang baik dan informatif tentang perkembangan COVID-19 di Daerah Hulu Langat. Penggunaan perlombongan data berjaya membantu menyatukan dan membersihkan data mentah menjadi satu fail yang berguna. Proses pembersihan data yang mendalam menjaga keautentikan data dan meningkatkan kebolehpercayaan analisis.

Untuk meningkatkan kebolehpercayaan data, usaha harus diberikan untuk memperoleh data yang lebih lengkap dan terkini. Penggunaan teknik pembelajaran mesin mungkin boleh dipertimbangkan untuk membantu proses pembersihan data dan analisis. Selain itu, dengan menambahkan data populasi dan faktor sosioekonomi Daerah Hulu Langat dapat memberikan konteks lebih luas dalam analisis COVID-19 di kawasan tersebut.

Kesimpulannya, kajian ini berjaya memberikan analisis dan deskriptif yang informatif tentang situasi COVID-19 di Daerah Hulu Langat. Walau bagaimanapun, ia memerlukan usaha yang besar dalam pembersihan dan analisis data. Penambahbaikan dan penggunaan teknik terkini di masa hadapan dapat meningkatkan lagi keberkesanan projek ini dalam menyediakan maklumat yang berguna untuk pemantauan dan pengurusan COVID-19 di kawasan tersebut.

Penghargaan

Dengan nama Allah, Yang Maha Mengasihani lagi Maha Penyayang. Segala puji dan terima kasih kepada Allah S.W.T. kerana memberkati saya dengan peluang dan kekuatan untuk menyiapkan laporan projek tahun akhir dengan jayanya. Pertama sekali saya ingin mengucapkan terima kasih kepada penyelia saya, Assoc. Prof. Dr. Suhaila Zainudin, atas arahan, cadangan, insentif, maklumat, dan nasihat beliau. Tanpa pengawasan dan bantuannya, saya tidak akan berada di mana saya sekarang. Laporan projek tahun akhir ini tidak akan dapat dilaksanakan tanpa bantuan beliau. Di samping itu, sepanjang sesi konsultasi, beliau cukup membantu memberi penerangan yang jelas dalam menyiapkan projek tahun akhir ini. Tidak boleh dilupakan, saya ingin merakamkan ucapan terima kasih kepada semua tenaga pengajar UKM atas bantuan dan kerjasama sepanjang pengajian saya. Ucapan tidak terhingga kepada kepada ahli keluarga yang amat disayangi, Bahrin bin Selamat(Ayah), Norazah binti Abd Rahman(ibu) dan adik-beradik saya yang lain yang tidak putus-putus mendoakan kejayaan saya.

Tambahan pula, terima kasih juga kepada rakan-rakan sekelas saya iaitu Aisyah Hanis, Nur Hazrina Asyikin, Afiq Aiman, Nuruddin Naim, Izqalan, Muhammad Abu Bakar As Siddiq atas sokongan dan galakan anda untuk sama sama berjuang siang dan malam demi melengkapkan projek tahun akhir ini. Akhir sekali, ribuan terima kasih saya ucapkan kepada rakan yang membimbing saya dari awal kajian ini iaitu Khairul Idham yang sentiasa memberi idea yang bernas, kritikan yang membina serta semangat yang tidak putus sejak awal pembelajaran di Universiti Kebangsaan Malaysia hingga kini. Terima kasih kerana memberi saya kekuatan untuk maju jaya dalam perjuangan ini.

RUJUKAN

Sinha, A. & Rathi, M. 2021b. COVID-19 prediction using AI analytics for South Korea. *Applied Intelligence* 51(12): 8579–8597. https://doi.org/10.1007/s10489-021-02352-z

World Health Organization.2019 'Naming the coronavirus disease (COVID-19) and the virus that causes it.' https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

Üstebay, S., Sarmis, A., Kaya, G.K. & Sujan, M.-A. 2022. A comparison of machine learning algorithms in predicting COVID-19 prognostics. Internal and Emergency Medicine 18(1): 229–239.

DataScience RoadMap. 2023. Create an Amazing Interactive Tableau Dashboard in 40 minutes | Airbnb NYC. Video,

Hv, S. 2021. Covid-19 outbreak prediction using Machine Learning and Deep Learning. https://www.jetir.org/view?paper=JETIR2106427.

Dutta, A., Gupta, A. & Khan, F.H. 2020. COVID-19 : Detailed Analytics & Predictive Modelling using Deep Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology: 95–104. https://doi.org/10.32628/ijsrst207517

Niralipatel. 2023. BAGGING technique on Covid-19 dataset. *Kaggle*. https://www.kaggle.com/code/niralipatel20/bagging-technique-on-covid-19-dataset

WHO Coronavirus (COVID-19) dashboard. (n.d.). . https://covid19.who.int/.

COVIDNOW in Malaysia - COVIDNOW. (n.d.). . https://covidnow.moh.gov.my/.

Abhigyan. 2021. Calculating accuracy of an ML model. - Analytics Vidhya - Medium. *Medium*. https://medium.com/analytics-vidhya/calculating-accuracy-of-an-ml-model-8ae7894802e.

Introduction to balanced and imbalanced datasets in Machine Learning. (n.d.). . https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/.

Canuma, P. 2023. How to deal with imbalanced classification and regression data. *neptune.ai*. https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data

Choudhury, K. 2021. Accuracy visualisation: supervised machine learning classification algorithms. *Medium*. https://towardsdatascience.com/accuracy-visualisation-supervised-machine-learning-classification-algorithms-af13d18fcc6c

Python Pandas dataframe cross-referencing and new column generation. (n.d.). . https://stackoverflow.com/questions/59060523/python-pandas-dataframe-cross-referencing-and-new-column-generation.

Wagh, H. 2022. Handling missing value | Data cleaning | Analytics Vidhya. *Medium*. https://medium.com/analytics-vidhya/the-why-and-how-to-handle-missing-values-46ab8c1b9034

SOP PERINTAH KAWALAN PERGERAKAN (PKP). 2020. https://covid-19.moh.gov.my/faqsop/sop-perintah-kawalan-pergerakan-pkp.

Yusof, A. 2021. Timeline: How the COVID-19 pandemic has unfolded in Malaysia since January 2020. *CNA Lifestyle*. https://cnalifestyle.channelnewsasia.com/asia/timeline-how-covid-19-pandemic-has-unfolded-malaysia-january-2020-289286

Team, D. 2021. Artificial Neural Networks for Machine Learning – Every aspect you need to know about. *DataFlair*. https://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/

Kumar, A. 2021. Deep Learning : Artificial Neural Networks(ANN) - DataDrivenInvestor. *Medium*. https://medium.datadriveninvestor.com/deep-learning-artificial-neural-networks-ann-9eb56cf95be8

Linear Regression (LR) Model | Kaggle. (n.d.). . https://www.kaggle.com/discussions/general/278185.

Prashant. 2020. SVM Classifier Tutorial. *Kaggle*. www.kaggle.com/code/prashant111/svm-classifier-tutorial.

Korstanje, J. 2022. SMOTE | Towards Data Science. *Medium*. https://towardsdatascience.com/smote-fdce2f605729

Eswarchandt. 2020. COVID-19 Forecasting XGBOOST. *Kaggle*. https://www.kaggle.com/code/eswarchandt/covid-19-forecasting-xgboost

What is Logistic regression? | IBM. (n.d.). https://www.ibm.com/topics/logistic-regression.

Universiti Kebangsaan Malaysia