

**MODEL PEMBELAJARAN MESIN MULTIVARIATE UNTUK
RAMALAN KES COVID-19 MENGGUNAKAN INTEGRASI DATA
METEOROLOGI DAN KUALITI UDARA DI DAERAH HULU**

LANGAT

Peggy Yeo^{1*} & Azuraliza Abu Bakar²

^{1,2}*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM
Bangi,, Selangor Darul Ehsan, Malaysia*

ABSTRAK

Pada Disember 2019, kes pertama wabak koronavirus 2019 (COVID-19) telah dilaporkan di Wuhan, China dengan kadar penularan yang sangat pantas. Sehingga 28 Oktober 2022, wabak itu telah menjangkiti lebih enam ratus juta orang di seluruh dunia, dengan enam juta kematian berdasarkan penyiasatan *World Health Organization* (WHO). Malaysia juga telah menjadi salah satu negara yang mengalami pandemik COVID-19 yang teruk dan memberi impak kepada ekonomi dan masyarakat. Gejala biasa COVID-19 termasuk demam, batuk kering, cirit-birit, sakit tekak, dan hilang rasa/bau deria. Kajian sedia ada menunjukkan pemboleh-ubah seperti meteorologi dan kualiti udara merupakan faktor- faktor kritikal bagi wabak seperti influenza. Oleh itu, dalam kajian ini model pembelajaran mendalam akan dibina menggunakan eni bina rangkaian neural berasaskan *Long-Short-Term Memory* (LSTM) dan juga *Convolutional Neural Network* (CNN). Model ini dibina dengan menggunakan data meteorologi dan data kualiti udara yang dikumpul daripada Jabatan Alam Sekitar serta kes COVID-19 yang dikumpul daripada Pejabat Kesihatan Hulu Langat. Data meteorologi dan

kualiti udara yang telah dikumpul akan diintegrasikan dengan kes COVID-19 berdasarkan mukim. Keputusan ramalan akan ditunjukkan dalam papan pemuka tableau supaya dapat membantu pihak berkenaan untuk menganalisis dan meramal gelombang COVID-19 yang akan berlaku pada masa depan.

Kata kunci: COVID-19, LSTM, CNN, Data Meteorologi, Kualiti Udara

PENGENALAN

Penyakit Koronavirus (COVID-19) ialah sesuatu penyakit jangkitan yang disebabkan oleh sejenis virus bernama Severe Acute Respiratory Syndrome Coronavirus 2 (SARS- CoV-2). Virus tersebut akan menyebabkan jangkitan berlaku pada saluran pernafasan. Kes pertama yang dikesan oleh dunia ialah kes COVID-19 yang berlaku di Wuhan, Hubei, China pada bulan Disember 2019. Seterusnya kes-kes telah muncul di Thailand, Nepal, Amerika Syarikat, Korea Selatan, Singapura, Perancis, Vietnam, Malaysia, Australia, Kanada, Kemboja, Sri Lanka dan Jerman. Hal ini kerana penduduk China yang sangat padat telah menjadi salah satu pemangkin kepada pembentukan pandemik COVID-19 yang diiktiraf oleh Pertubuhan Kesihatan Sedunia (WHO). Situasi ini berlaku oleh sebab kekurangan kajian yang mengenai virus tersebut dari setiap aspek dan seterusnya menyebabkan kekurangan pemahaman dan menghasilkan sikap yang mengambil mudah terhadap penyakit COVID-19 pada fasa awal wabak. Sehingga 28 Oktober 2022, wabak tersebut telah menjangkiti lebih enam ratus juta orang di seluruh dunia, dengan enam juta kematian berdasarkan penyiasatan Pertubuhan Kesihatan Sedunia (WHO 2022). Malaysia juga tidak dapat dikecualikan daripada wabak COVID- 19 dengan kemunculan kes pertama COVID-19 yang iaitu seorang wanita berusia 66 tahun serta dua cucunya yang datang dari Singapura telah disahkan positif koronavirus pada 25 Januari 2020 (Harits Asyraf Hasnan 2020).

Kerajaan telah mengambil inisiatif untuk mengekang kemasukan pengunjung luar negara dari China, Korea, Itali dan Iran ke Malaysia mulai 13 Februari 2020. Selain daripada itu, Perintah Kawalan Pergerakan (PKP) yang diisytiharkan oleh kerajaan Malaysia pada 18 Mac 2020 hingga 3 Mei 2020 juga telah menyumbangkan tenaga dalam mengurangkan pergerakan dan aktiviti seluruh rakyat Malaysia bagi mengawal penyebaran coronavirus ("Portal Rasmi Kementerian Kesihatan Malaysia" 2022). Perintah Kawalan Pergerakan (PKP) yang hanya membenarkan bidang-bidang yang tertentu sahaja beroperasi dengan mempunyai syarat waktu operasi premis bagi mengawalkan penularan coronavirus dalam kalangan rakyat mereka telah menghasilkan juga impak yang negatif daripada perintah tersebut. Impak-impak tersebut dapat dilihat melalui pelbagai sektor seperti kemelesetan dalam sektor ekonomi negara dan juga peningkatan tekanan emosi dan perasaan murung dalam sektor sosial masyarakat. Sejalan dengan itu, aplikasi MySejahtera juga diwajibkan dimuat turun oleh setiap warganegara supaya memudahkan proses penjejakan dan pemantauan oleh pihak berkuasa untuk tujuan menangani wabak COVID-19. Pihak berkuasa akan mendapatkan data-data kes positif daripada aplikasi tersebut dengan melalui ujian-ujian COVID-19 seperti Antigen Rapid Test Kit (RTK-Ag) dan Reverse Transcription Polymerase Chain Reaction (RT-PCR) yang dilakukan oleh sama ada hospital, klinik atau pesakit sendiri.

Keaktifan Koronavirus diyakini akan dipengaruhi oleh faktor-faktor persekitaran kerana virus tersebut disebar melalui udara. Terdapat juga kajian mendapati bahawa faktor-faktor persekitaran seperti suhu cuaca dan kelembapan dapat mempengaruhi penyebaran Koronavirus. (VALSAMATZI-PANAGIOTOU & PENCHOVSKY 2022). Kajian tersebut telah membuktikan bahawa coronavirus dapat hidup dalam masa yang lebih panjang dalam suhu yang lebih rendah dan kelembapan yang lebih tinggi. Selain itu, terdapat juga kajian yang menyatakan bahawa terdapat korelasi positif antara penyebaran COVID-19 dengan pencemaran udara (KHAN ET AL. 2021). Semakin panjang masa terdedah kepada pencemaran udara semakin

tinggi risiko dijangkit COVID-19 (Ali & Islam 2020). Kebanyakan kajian menyatakan bahawa penyebaran COVID-19 akan lebih luas terutamanya di lokasi-lokasi yang mempunyai iklim sejuk, kelembapan tinggi dan juga pencemaran udara yang tinggi.

Salah satu kajian yang melibatkan COVID-19 adalah untuk meramal COVID-19 berdasarkan pemboleh ubah PM_{2.5}, suhu dan kelembapan udara yang berbeza (Ogunjo, Fuwape & Rabiou 2022). Kajian ini diterbitkan pada tahun 2022 di Nigeria dan bertujuan untuk meramal kes COVID-19 dengan menggunakan parameter PM_{2.5}, suhu iklim dan juga kelembapan udara. Kajian ini telah mendapatkan data daripada National Centre for Disease Control manakala data atmosfera (suhu dan relatif kelembapan) dan bahan zarah (PM_{2.5}, PM_{1.0}, PM_{10.0}) diperolehi daripada kempen berterusan Pusat Penyelidikan Atmosfera, Agensi Penyelidikan dan Pembangunan Angkasa Negara menggunakan penderia Purple Air dari 1 November 2020 hingga 31 Mac 2021. Kajian ini dijalankan dengan menggunakan algoritma Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) dan juga K-Nearest Neighbour (KNN). Set data dibahagikan kepada set latihan (80%) dan set ujian (20%) sebelum menjalankan proses pemodelan. Skor RMSE yang digunakan untuk menilai prestasi pemodelan dalam kajian ini dan nilai RMSE yang diperhatikan ialah 67, 87, 86, dan 65 bagi model DT, KNN, SVM, RF masing-masing. Hasil dapatan kajian ialah suhu iklim merupakan parameter yang paling sesuai digunakan untuk meramal kes COVID-19 seterusnya diikuti dengan atribut kelembapan udara.

Kajian yang seterusnya ialah kajian yang berkaitan dengan kesan faktor meteorologi dan PM_{2.5} terhadap penularan COVID-19 di United States (Zhou 2022) dan tempoh kajian adalah dari 12 April 2020 hingga 13 Oktober 2020. Kajian ini bertujuan untuk menganalisis hubungan antara faktor meteorologi dan bilangan pesakit COVID-19. Data kes COVID-19 berjumlah tujuh ratus tujuh puluh lapan ribu seratus tujuh puluh empat dikumpulkan daripada

Coronavirus Resource Center of the JOHNS HOPKINS UNIVERSITY dan data meteorologi seperti tarikh, kelajuan angin purata harian (AWND), purata hujan harian (PRCP), suhu purata harian (TAVG), PM2.5 dan indeks kualiti udara harian (AQI) telah dikumpulkan daripada National Centers for Environmental Information. Kajian ini menunjukkan hujan tidak mempengaruhi bilangan pesakit COVID-19 secara terus tetapi boleh mempengaruhi bilangan pesakit melalui kelembapan udara. Suhu purata harian dan kelajuan angin purata harian telah dibuktikan berkait rapat dengan bilangan pesakit COVID-19 dalam kebanyakan situasi. Hal ini demikian kerana suhu dapat menjejaskan kemandirian virus dan kelajuan angin dapat menjejaskan penyebaran virus. Kajian ini menunjukkan bahawa PM2.5 dan indeks kualiti udara harian berkemungkinan mempunyai kesan terhadap penyebaran COVID-19 oleh sebab terdapat peningkatan dalam kadar kemasukan ke hospital dengan kepekatan PM2.5 yang lebih tinggi.

Salah satu kajian lepas yang melibatkan ramalan COVID-19 melalui data cuaca dan kualiti udara dengan menggunakan algoritma LSTM telah diterbitkan pada tahun 2022 (Hernandez-resendiz 2022). Kajian ini bertujuan untuk mengenal pasti hubungan antara meteorologi dan kualiti udara dengan kes COVID-19 di negara Mexico dan untuk menghasilkan model ramalan kes COVID-19. Data kes COVID-19 yang digunakan dalam kajian tersebut adalah data yang dikumpulkan pada masa gelombang ketiga COVID-19 di Victoria, Mexico yang telah disahkan oleh kerajaan Mexico manakala data meteorologi dan kualiti udara diukur daripada parameter seperti Particulate Matter PM 2.5, PM 10, suhu iklim, kelembapan udara, tekanan udara, kelajuan angin dan arah angin telah dikumpulkan dari 27 Mei hingga 13 Oktober 2021. Kajian ini menggunakan algoritma clustering berdasarkan K-means untuk mendapati parameter dengan membentuk kelompok yang mempunyai kesepaduan yang tinggi seterusnya menggunakan algoritma LSTM untuk membina model

ramalan yang mempunyai ketepatan yang tinggi pada tahap peringkat latihan model. Model ramalan yang telah dibina telah dinilai melalui ralat punca min kuasa dua (RMSE), ralat mutlak min (MAE) dan ralat peratusan mutlak min (MAPE). Prestasi tertinggi diperolehi dengan RMSE 0.0892, MAE 0.0592 dan MAPE 0.2061 dalam peringkat ujian manakala dalam peringkat pengesahan pula, model ramalan memperoleh nilai antara 0.4152 dan 3.9084 bagi RMSE dan kurang daripada 4.1% bagi MAPE dengan menggunakan tiga set data berbeza. Hasil dapatan kajian ini ialah pemboleh ubah kelembapan udara dan suhu iklim merupakan pemboleh ubah yang lebih berpengaruh dalam meramal kadar kematian COVID-19.

Di samping itu, terdapat juga kajian lepas yang berkaitan dengan ramalan kesan alam sekitar terhadap gelombang kerap COVID-19 pada tahun 2021 (Yu et al. 2021). Tujuan kajian ini adalah untuk mengkaji kesan pencemaran udara terhadap kes COVID-19 di Itali. Data kualiti udara didapatkan daripada “World’s Air Pollution: Real-time Air Quality Index” dan data kes COVID-19 didapat daripada laman portal Our World in Data. Atribut yang akan digunakan termasuk PM2.5, ozon, O₃, NO₂, dan kelembapan udara. Artificial neural networks (ANNs) digunakan untuk menganggarkan maksimum dan minimum nilai pemboleh ubah. Keputusan kajian telah melaporkan bahawa pendedahan jangka panjang kepada pencemaran udara telah membuktikan akan meningkatkan kebarangkalian kematian kepada penyakit COVID-19. Hal ini demikian kerana kepekatan pencemaran udara yang lebih tinggi akan meningkatkan kadar kematian pesakit COVID-19 oleh sebab pencemaran udara boleh mempengaruhi penyebaran virus.

Kajian yang terakhir ialah mengkaji kesan lag ozon, PM2.5 dan faktor meteorologi pada kes COVID-19 di New York yang diterbitkan pada tahun 2021 (ADHIKARI & YIN 2021). Objektif utama kajian penyelidikan ini adalah untuk mengkaji kesan lag ozon, PM2.5 dan data

meteorologi seperti kelajuan angin, suhu iklim, kelembapan relatif, kelembapan mutlak dan peratusan awan terhadap kes COVID-19. Set data yang digunakan dalam kajian ini adalah diperolehi daripada laman portal USA FACTS yang merangkumi kes COVID-19 yang telah disahkan daripada 1 Mac sehingga 11 April 2020 manakala data meteorologi dikumpulkan daripada United States Environmental Protection Agency (AQS). Kajian ini telah menunjukkan bahawa suhu yang lebih rendah menyumbang kepada kemandirian virus dalam jangka masa yang lebih panjang dan kesan ozon, PM2.5 dan lima faktor meteorologi telah menunjukkan kesan yang ketara pada kes baharu COVID-19 dengan lag9-lag13 hari.

Ramalan yang tepat dan cepat dapat dilakukan melalui teknologi kecerdasan buatan (AI) pada era globalisasi yang canggih ini melalui pembelajaran mesin dan pembelajaran mendalam (deep learning). Teknik ini dapat digunakan dalam kajian sektor kesihatan terutamanya pada epidemiologi seperti COVID-19 untuk memberi manfaat kepada masyarakat. Tujuan kajian ini adalah untuk menggunakan set data meteorologi yang sedia ada dengan algoritma berkenaan seperti pemodelan memori jangka pendek (LSTM) dan Convolutional Neural Network (CNN) untuk menghasilkan model yang dapat menunjukkan keputusan ramalan yang mempunyai ketepatan yang tinggi supaya dapat menganggarkan dan menganalisis trend COVID-19 di Malaysia. Akhir sekali, kajian ini adalah untuk mengaji hubungan antara faktor persekitaran seperti suhu cuaca, kelembapan udara dan kualiti udara dengan kes COVID-19 di Hulu Langat Malaysia supaya dapat menghasilkan model ramalan yang tepat dan berguna.

PENYATAAN MASALAH

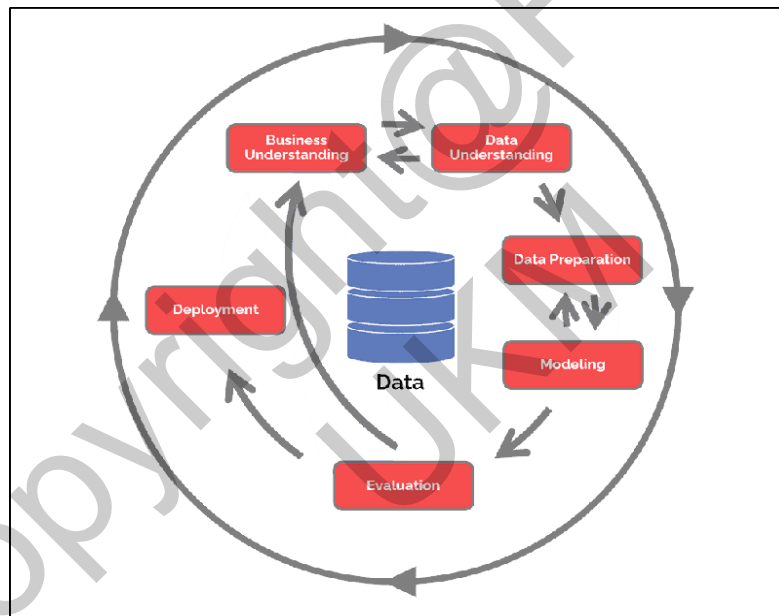
Kekurangan pemahaman berkaitan dengan COVID-19 telah menyebabkan penularan meluas virus COVID-19 yang berlaku di seluruh dunia pada gelombang pertama COVID-19. Pertanyaan tentang hubungan antara faktor persekitaran dan pencemaran udara dengan penyebaran COVID-19 telah muncul. Oleh sebab penyelidikan yang berkaitan dengan COVID-19 hanya mula sejak dua tahun kebelakangan ini, kajian yang berkaitan masih terhad di Malaysia, jadi kekurangan pemahaman tentang keaktifan Koronavirus dengan faktor persekitaran dan pencemaran udara telah berlaku. Situasi ini menyebabkan keputusan analisis data yang dilakukan tidak begitu bersepadu dan seterusnya meningkatkan kesukaran untuk menghasilkan model ramalan yang tepat. Kesukaran dalam ramalan wabak gelombang yang seterusnya akan menyebabkan pihak berkuasa tidak sempat mengambil tindakan yang diperlukan sebelum masa tersebut untuk mengelakkan keadaan wabak yang kritikal pada masa depan. Terdapat beberapa kajian yang telah membuktikan terdapat korelatif dengan ketara antara meteorologi, kualiti udara dengan kes positif COVID-19 (HERNANDEZ- RESENDIZ 2022), (ZHOU 2022), (YU ET AL. 2021). Selain daripada itu, kajian di Malaysia hanya mengaitkan hubungan antara parameter meteorologi dan data COVID- 19 yang umum daripada laman portal sahaja (Sheng 2021). Perkembangan teknologi penyimpanan data raya yang meningkat di Malaysia menyebabkan sejumlah data-data telah disimpan. Oleh itu, untuk memanfaatkan data-data di Malaysia dengan teknologi pembelajaran mendalam, kajian ini akan bekerjasama dengan Pejabat Kesihatan Hulu Langat dengan menggunakan data profil pesakit COVID-19 yang besar dan data meteorologi dan kualiti udara yang terdapat banyak parameter di Selangor untuk mengkaji hubungan antara meteorologi, kualiti udara dengan COVID-19 di Selangor.

OBJEKTIF KAJIAN

Objektif kajian ini adalah untuk:

1. Membangunkan model pembelajaran mendalam dengan menggunakan data kes COVID-19 di Malaysia yang telah diintegrasikan dengan data meteorologi dan data kualiti udara.
2. Menilai prestasi model pembelajaran mesin ke atas kesan cuaca dan kualiti udara kepada corak kes COVID-19 di Malaysia.

METOD KAJIAN



Rajah 4.1 Fasa-Fasa Metodologi CRISP-DM

4.1 Fasa Pemahaman Bisnes

Tahap pertama dalam CRISP-DM adalah untuk memastikan mempunyai pemahaman yang utuh tentang objektif dan keperluan kajian. Tahap ini adalah sangat penting supaya dapat mengenal pasti permasalahan yang dihadapi dan seterusnya menghasilkan penyelesaian

masalah yang sesuai dengan situasi semasa. Objektif utama dalam projek ini adalah untuk membangunkan satu model ramalan pembelajaran mesin untuk menggunakan bahasa pengaturcaraan *Python* untuk meramalkan kes COVID-19 dengan menggunakan data meteorologi dan data kualiti udara dengan menggunakan algoritma LSTM dan algoritma CNN. Pencarian maklumat mengenai COVID-19 amat diperlukan untuk memastikan model yang dihasilkan dapat memberi ramalan yang tepat supaya dapat dimanfaatkan pihak berkenaan.

4.2 Fasa Pemahaman Data

Pengumpulan data merupakan langkah pertama dalam fasa ini, data yang dikumpulkan perlu disemak dan dianalisis supaya dapat memahami hubungan antara atribut-atribut data. Dengan berbuat demikian, pemahaman yang cukup tentang set data dapat membantu dalam mengenal pasti kesesuaian data tersebut dalam projek ini. Data-data yang digunakan dalam penyelidikan ini adalah data kes COVID-19 harian, data meteorologi dan data kualiti udara.

4.3 Fasa Penyediaan Data

Fasa ini juga dikenali sebagai proses pra-pemprosesan data untuk menghasilkan set data akhir yang bersih dan paling sesuai untuk diguna dalam pemodelan ramalan pesakit COVID-19. Pembersihan, transformasi dan integrasi data juga diliputi dalam fasa tersebut supaya dapat menghasilkan data yang sesuai dan berkualiti dalam proses pembelajaran mesin dalam projek ini dan juga dapat menjamin ketepatan model ramalanyang akan dibangunkan nanti.

4.4 Fasa Pemodelan

Algoritma LSTM dan CNN telah dipilih untuk membangunkan model pembelajaran mesin yang dapat meramal kes COVID-19 dengan data meteorologi dan kualiti udara. Algoritma LSTM dapat menjalankan ramalansiri masa, dan juga dapat meramalkan data pada masa depan

dengan menggunakan data yang lama manakala CNN pula dapat menghasilkan keputusan dalam jangka masa yang amat pendek. Oleh itu, dua algoritma inidipilih untuk meramal kes COVID-19.

4.5 Fasa Penilaian

Fasa ini akan menilai ketepatan ramalan yang dilakukan oleh model yang dihasilkan dan seterusnya menilai model dengan menggunakan Ralat Relatif(RE). Perbandingan ketepatan antara algoritma sama ada LSTM atau CNN akan dilakukan supaya dapat memilih algoritma yang dapat dilaksanakan dengan lebih baik dan seterusnya dapat mencapai objektif projek ini iaitu meramal pesakit COVID-19 dengan meteorologi dan kualiti udara dengan lancar.

4.6 Fasa Penyebaran

Tahap ini merupakan langkah yang terakhir yang akan menentukan algoritma dan juga membangunkan model ramalan yang telah dilengkapi. Pengetahuan yang diperolehi melalui hasil kajian yang dilakukan akandisebarkan kepada awam untuk kegunaan awam. Algoritma juga akandiperlengkapkan dari segi sokongan teknikal iaitu pemantauan danpenyelenggaraan yang dilakukan pada semasa ke semasa supaya mencapai prestasi yang memuaskan dan juga mengelakkan kemunculan ralat sistem.

4.7 Spesifikasi Keperluan

Sistem ini akan diuji sama ada dapat mencapai objektif atau tidak. Sistem ini akan diuji oleh pensyarah dan pelajar di FTSM untuk menentukan keberkesannya. Maklum balas tinjauan juga akan dijalankan untuk menentukan tahap mesra pengguna sistem yang dibangunkan.

Jadual 4.1: Spesifikasi Perkakasan

Kriteria	Spesifikasi
Jenama	<i>ACER</i>
Model	<i>Swift SF314-55</i>
Pemproses	<i>Intel Core i5-8265U</i>
Pemacu keadaan pepejal	512GB
Memori Capaian Rawak (RAM)	8GB
Sistem Operasi	<i>Windows 10</i>
Unit Pemprosesan Grafik (GPU)	<i>Intel UHD Graphics 620</i>

Jadual 4.2: Spesifikasi Keperluan Perisian

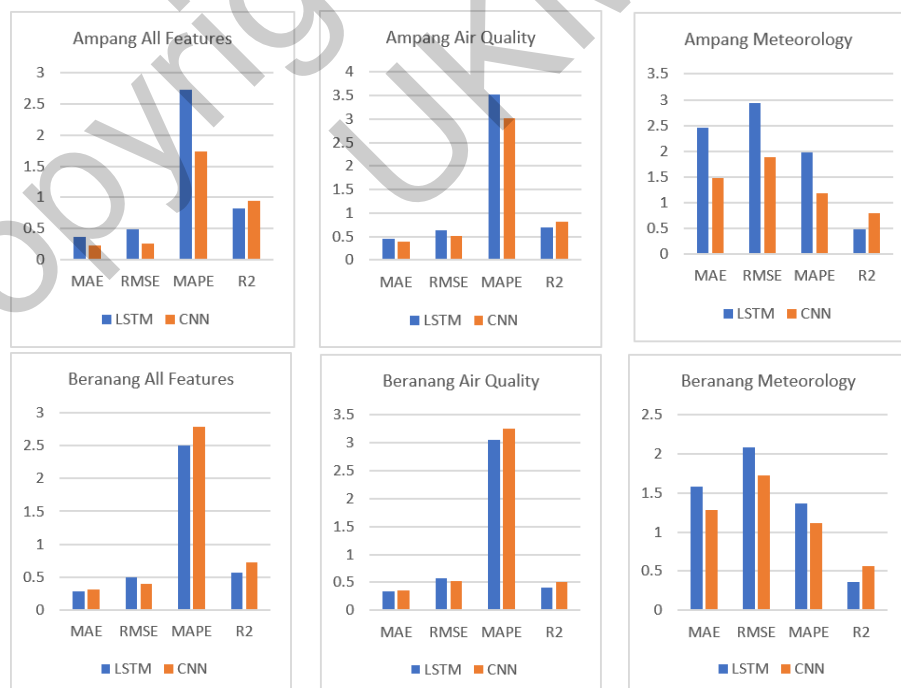
Perisian	Perincian
<i>Microsoft Word</i>	Menulis tesis projek akhir tahun
<i>Microsoft Excel</i>	Tempat simpanan maklumat dan data yang diperoleh
<i>Google Colab</i>	IDE untuk menulis algoritma <i>Python</i> bagi proses pemodelan
<i>Tableau Public</i>	Alat visualisasi keputusan ramalan kes COVID-19 melalui data meteorologi

Windows 10 Digunakan untuk membangunkan dan menyokong model ramalan

HASIL KAJIAN

Pemodelan LSTM Multivariate dan CNN Multivariate telah dibangunkan untuk meramalkan kes aktif COVID-19 di setiap mukim di Hulu Langat, Selangor. Model akan dilatih secara 3 bahagian iaitu model yang menggunakan semua ciri kualiti udara dan meteorologi, model yang menggunakan kualiti udara dan model yang menggunakan meteorologi sahaja bagi tujuan ramalan.

5.1 Keputusan Ramalan LSTM dan CNN





Rajah 5.1 Graf keputusan ramalan model LSTM dan CNN

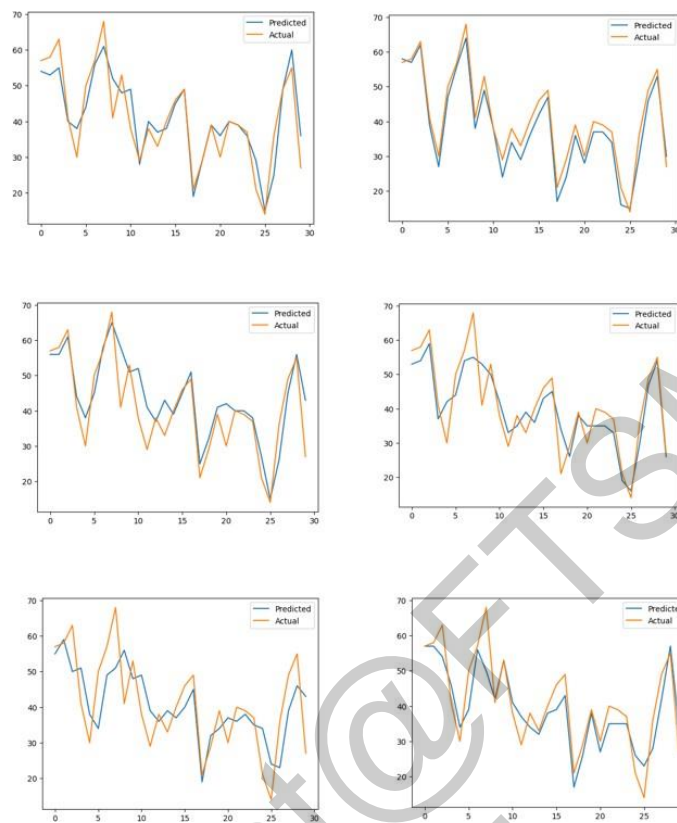
Rajah 5.1 telah menunjukkan perbandingan penilai matriks iaitu MAE, RMSE, MAPE dan R2 antara dua algoritma iaitu LSTM dan CNN berdasarkan ciri yang berlainan diguna dan mukim masing-masing. Setiap mukim akan mempunyai 6 model dibina iaitu model 3 LSTM dan 3 model CNN. Setiap algoritma akan dilatih berdasarkan 3 model iaitu berdasarkan semua ciri, berdasarkan ciri kualiti udara sahaja dan berdasarkan ciri meteorologi sahaja. Daripada Rajah 5.1, dapat mengenalpasti bahawa CNN mempunyai prestasi yang umumnya lebih baik daripada model LSTM dalam model yang dibina dengan mencapai R2 yang lebih tinggi serta keralatan MAE, RMSE dan MAPE yang lebih rendah. Di samping itu, konklusi lain yang dapat dibuat daripada keputusan ramalan ialah model yang dilatih menggunakan semua atribut data meteorologi dan kualiti udara dapat menghasilkan keputusan ramalan yang lebih tepat sekiranya berbanding dengan model yang dibina menggunakan data kualiti dan data meteorologi sahaja. Seterusnya, model yang dibina dengan menggunakan data kualiti udara berprestasi baik sedikit daripada model data meteorologi bermaksud kualiti udara memainkan peranan yang lebih tinggi sekiranya berbanding dengan meteorologi terhadap kes COVID-19.

Jadual 4.3: Keputusan Pemilihan Ciri-Ciri Penting set data

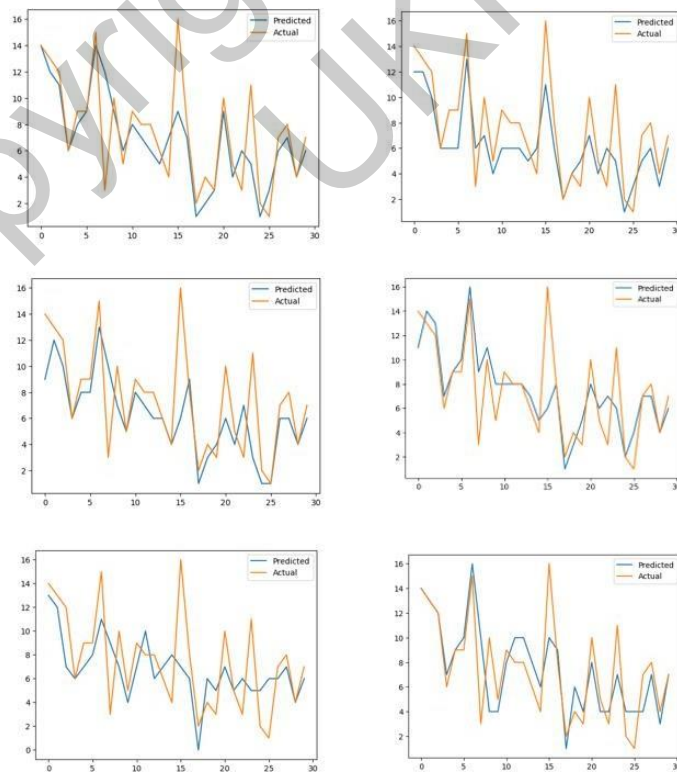
	Ciri	Kepentingan
0	CO	0.116
1	SO2	0.108
2	NO2	0.104
3	PM10	0.101
4	O3	0.100

5	<i>Relative Humidity Avg</i>	0.098
6	<i>Wind Direction Avg</i>	0.096
7	<i>Ambient Temperature Avg</i>	0.095
8	PM2.5	0.092
9	<i>Wind Speed Avg</i>	0.090

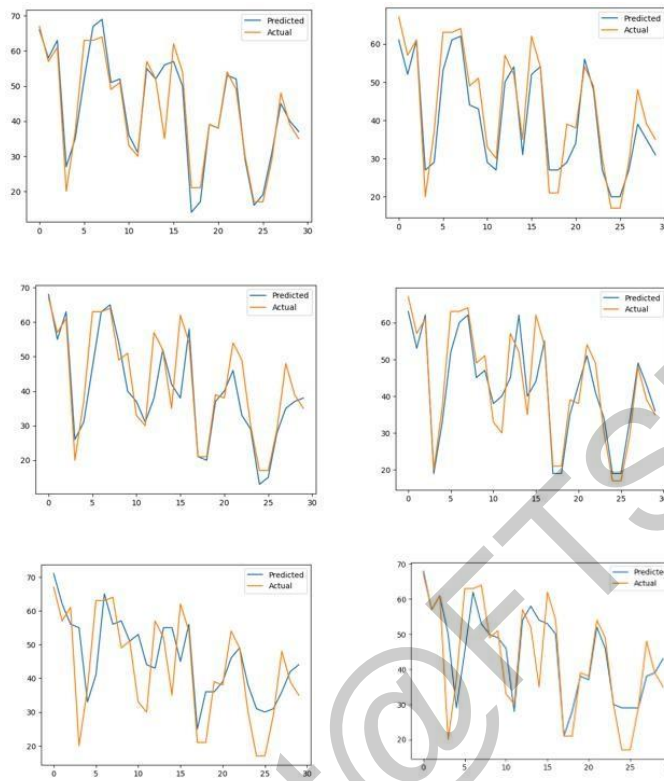
Jadual 4.3 menunjukkan keputusan pengestrakan ciri dalam set data tersebut. Pengestrakan ciri ialah langkah penting dalam pembelajaran mesin dan analisis data yang melibatkan pemilihan ciri yang memberi pengaruh yang lebih penting kepada setdata. Teknik pemilih ciri-ciri penting yang akan digunakan ialah *ExtraTreesClassifier*. Daripada Jadual 4.3 boleh mendapati nampaknya ciri yang paling penting ialah Karbon Monoksida (CO) dengan nilai kepentingan 0.116. Ini menunjukkan bahawa CO mempunyai kepentingan tertinggi dalam konteks yang sedang dipertimbangkan. Mengikut rapat ialah Sulfur Dioksida (SO₂) dengan kepentingan 0.108 dan Nitrogen Dioksida (NO₂) dengan kepentingan 0.104. Ciri penting lain termasuk PM₁₀, Ozon (O₃), Kelembapan, Arah Angin, Suhu, PM_{2.5}, dan Kelajuan Angin. Ciri-ciri ini menyumbang kepada pemahaman dan analisis keseluruhan, walaupun dengan nilai kepentingan yang rendah sedikit antara 0.100 hingga 0.090. walaupun dengan nilai kepentingan yang rendah sedikit antara 0.100 hingga 0.090. Nilai kepentingan menunjukkan kepentingan relatif setiap ciri dalam konteks yang diberikan. Oleh itu, dapat mendapatkan kesimpulan bahawa kualiti udara memainkan peranan lebih penting daripada meteorologi.



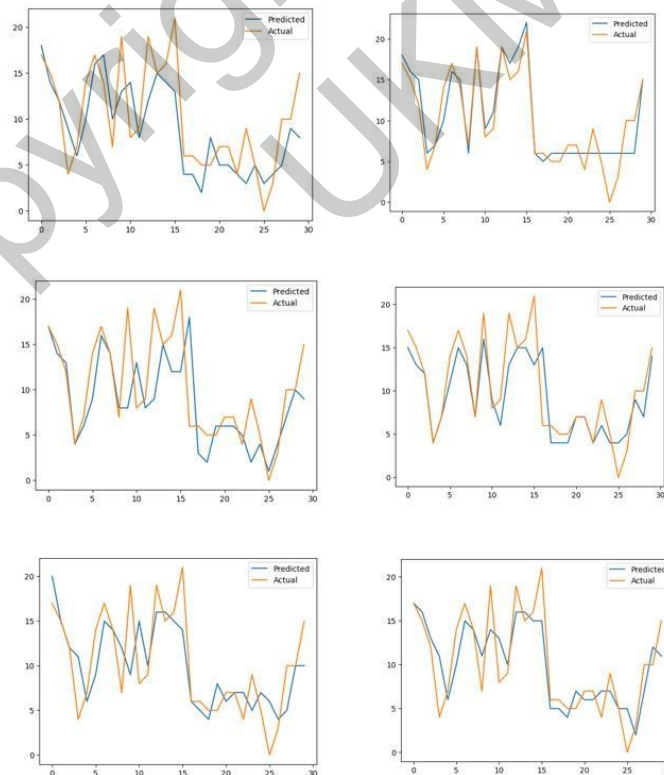
Rajah 5.2 Graf keputusan ramalan model LSTM dan CNN mukim Ampang



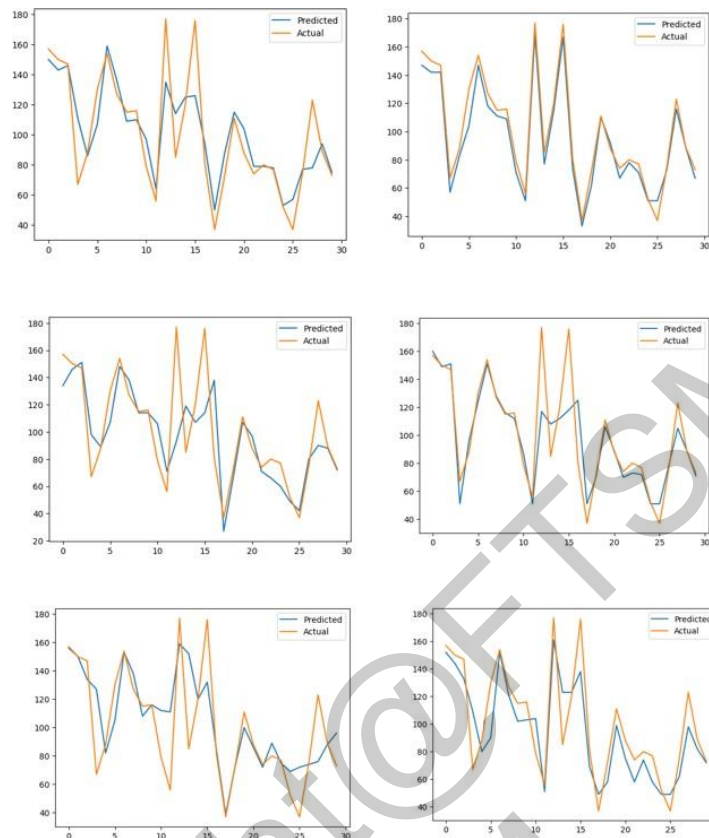
Rajah 5.3 Graf keputusan ramalan model LSTM dan CNN mukim Beranang



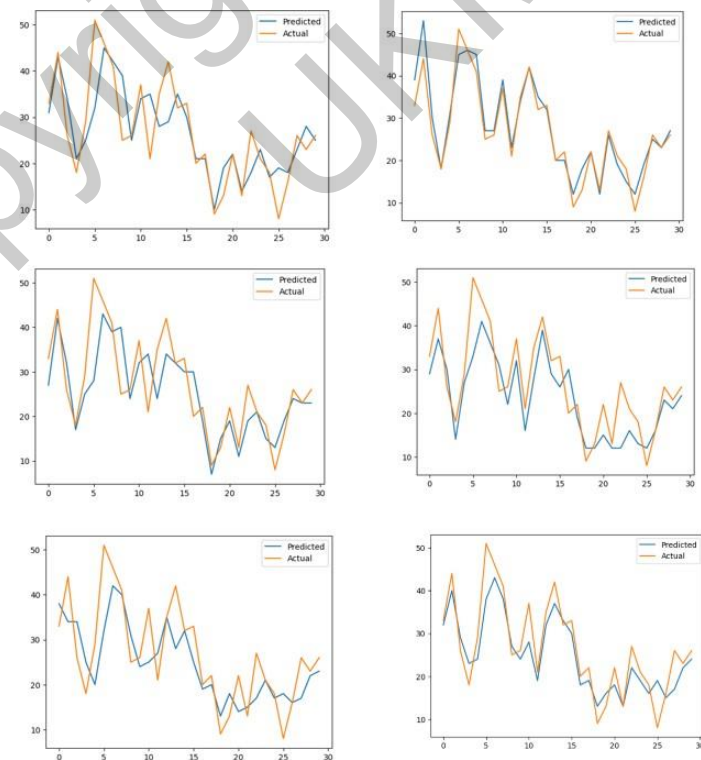
Rajah 5.4 Graf keputusan ramalan model LSTM dan CNN mukim Cheras



Rajah 5.5 Graf keputusan ramalan model LSTM dan CNN mukim Hulu Langat



Rajah 5.6 Graf keputusan ramalan model LSTM dan CNN mukim Kajang

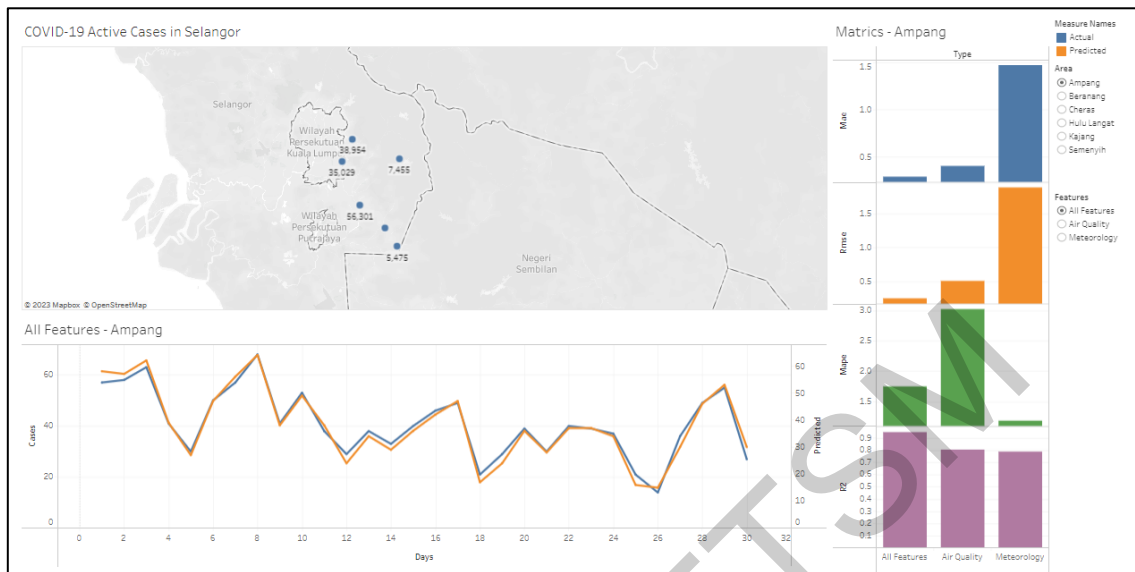


Rajah 5.7 Graf keputusan ramalan model LSTM dan CNN mukim Semenyih

Daripada Rajah 5.2 hingga Rajah 5.7 telah menunjukkan graf keputusan ramalan LSTM dan CNN bagi setiap mukim, rajah-rajah telah menunjukkan bahawa model CNN yang dilatih menggunakan semua ciri menghasilkan keputusan yang sangat baik di mukim Ampang Kajang dan Semenyih. Seterusnya diikuti dengan model CNN semua ciri di Cheras dengan keputusan yang memuaskan. Model yang menghasilkan keputusan yang kurang memuaskan ialah model di Beranang dan Hulu Langat oleh sebab data kes COVID-19 yang lebih rendah berbanding dengan mukim yang lain.

PAPAN PEMUKA

Rajah 5.8 telah mempamerkan antara muka *Tableau Public* yang mengandungi analisis keputusan model ramalan yang terdiri daripada enam mukim di Hulu Langat. Pandangan yang menyeluruh dalam satu antara muka dapat memudahkan pengguna selain mengurangkan masa yang diperlukan untuk memuatkan data dan seterusnya dapat melancarkan proses pemerhatian data. Penapis yang berada di sebelah kanan antara muka dapat digunakan oleh pengguna untuk memudahkan proses pemahaman dan menganalisis hasil kajian yang dipaparkan. Selain itu, pengguna boleh memilih mukim untuk memerhatikan bilangan data COVID-19. Di samping itu, keputusan ramalan juga akan dipaparkan dalam berdasarkan pilihan pengguna dan data meteorologi di Malaysia.



Rajah 5.8 Antara Muka Ramalan Kes COVID-19 di Hulu Langat

KESIMPULAN

Projek kajian ini adalah menggunakan algoritma-algoritma pembelajaran mesin untuk meramalkan kes COVID-19 melalui data meteorologi dan data kualiti udara di daerah Selangor. Matlamat bisnes yang telah dinyatakan pada perancangan awal projek ini telah dicapai melalui pembangunan model ramalan kes COVID-19. Set data yang digunakan dalam projek ini merupakan set data kes COVID-19 daripada PejabatKesihatan Hulu Langat dan set data meteorologi dan kualiti udara daripada Jabatan Alam Sekitar. Set data tersebut telah menjalankan pemprosesan data pada fasa penyediaan data untuk menjamin kualiti data supaya tidak mempengaruhi keputusan ramalan. Keputusan ramalan kes COVID-19 melalui data meteorologi dan kualiti udaratelah dihasilkan melalui perbandingan pencapaian antara modal ramalan LSTM dan CNN dengan menggunakan penilaian metrik yang sama. Projek ini telah membangunkan model ramalan menggunakan beberapa algoritma untuk meramalkan kes COVID-19 berdasarkan data meteorologi dan data kualiti udara. Keputusan ramalan yang

dihasilkan dan dipaparkan di Tableau diharapkan dapat memberi fungsi kepada pihak tertentu dalam meramalkan kes COVID-19 pada masa depan dan memberi informasi yang berguna kepada masyarakat umum.

PENGHARGAAN

Terlebih dahulu, saya ingin mengambil peluang ini untuk mengucapkan jutaan terima kasih kepada penyelia saya, Prof. Dr. Azuraliza binti Abu Bakar yang telah sabar membimbing dan memberi tunjuk ajar serta ilmu yang berkenaan kepada saya sepanjang kajian ilmiah dijalankan. Jutaan terima kasih juga saya hendak ucapkan kepada para pensyarah dari FTSM yang telah menaburkan ilmu pengetahuan kepada saya sepanjang pengajian saya di Universiti Kebangsaan Malaysia (UKM).

Selain itu, saya juga ingin mengucapkan penghargaan terima kasih yang tidak terhingga kepada ibu ayah saya yang menjadi sumber inspirasi, memberi semangat dan sentiasa berdoa terhadap kejayaan saya dalam menyiapkan kajian ini. Tidak lupa juga kepada rakan-rakan yang banyak membantu saya dalam memberikan pandangan yang bernas serta sokongan kepada saya semasa diperlukan.

Akhirnya, saya ingin mengucapkan terima kasih kepada semua pihak yang terlibat secara langsung atau tidak langsung dalam pembikinan kajian tahun akhir ini.

Sekian, terima kasih.

RUJUKAN

- Adhikari, A. & Yin, J. 2021. Lag effects of ozone, PM_{2.5}, and meteorological factors on COVID-19 new cases at the disease epicenter in queens, New York. *Atmosphere* 12(3).
- Ali, N. & Islam, F. 2020. The Effects of Air Pollution on COVID-19 Infection and Mortality— A Review on Recent Evidence. *Frontiers in Public Health* 8(2).
- Comprehensive guide to LSTM & RNNs. 2022. . <https://www.turing.com/kb/comprehensive-guide-to-lstm-rnn> [2 Desember 2022].
- Sheng, N.W. 2021. MODEL PEMBELAJARAN MENDALAM UNTUK MERAMALKES COVID-19 DI MALAYSIA MENGGUNAKAN INTEGRASI DATA CUACA.
- Valsamatzi-Panagiotou, A. & Penchovsky, R. 2022. Environmental factors influencing the transmission of the coronavirus 2019: a review. *Environmental Chemistry Letters* 20(3): 1603–1610.
- Vandeput, N. 2019. Forecast KPI: RMSE, MAE, MAPE & Bias | Towards Data Science. <https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d> [19 Jun 2023].
- Why do we need LSTM. An in depth analysis of the vanishing... | by Ampatishan Sivalingam | Towards Data Science. (n.d.). . <https://towardsdatascience.com/why-do-we-need-lstm-a343836ec4bc> [2 Desember 2022].
- Yu, Z., Abdel-Salam, A.S.G., Sohail, A. & Alam, F. 2021. Forecasting the impact of environmental stresses on the frequent waves of COVID19. *Nonlinear Dynamics* 106(2): 1509–1523.
- Zhou, N. 2022. The impact of meteorological factors and PM_{2.5} on COVID-19 transmission.

Peggy Yeo (A186179)

Prof. Dr. Azuraliza Abu Bakar

Fakulti Teknologi & Sains Maklumat,

Universiti Kebangsaan Malaysia

Copyright@FTSM
UKM