

IMPAK PENCEMARAN UDARA DAN  
METEOROLOGI TERHADAP PEMBENTUKAN  
OZON SEMASA PKP DI MALAYSIA  
MENGUNAKAN KAEDAH PEMBELAJARAN  
MESIN DAN NILAI SHAP

MUHAMAD HAZIQ BIN JUMALI

UNIVERSITI KEBANGSAAN MALAYSIA

IMPAK PENCEMARAN UDARA DAN METEOROLOGI TERHADAP  
PEMBENTUKAN OZON SEMASA PKP DI MALAYSIA MENGGUNAKAN  
KAEDAH PEMBELAJARAN MESIN DAN NILAI SHAP

MUHAMAD HAZIQ BIN JUMALI

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT  
UNIVERSITI KEBANGSAAN MALAYSIA  
BANGI

2021

**PENAKUAN**

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

14 October 2021

MUHAMAD HAZIQ BIN JUMALI  
P101449

PUSAT SUMBER  
FTSM

## PENGHARGAAN

Dengan nama Allah, yang Maha Pengasih lagi Maha Penyayang

Syukur ke hadrat Illahi dengan limpah kurniaNya dapat saya menyiapkan kajian projek akhir ini dengan jayanya. Setinggi penghargaan dan jutaan terima kasih kepada penyelia saya Prof. Madya. Dr. Zulaiha Ali Othman di atas ilmu, tunjuk ajar, bimbingan, nasihat, dan kesabaran sepanjang kajian ini dijalankan. Semoga ilmu yang diberikan memberi manfaat kepada saya dalam urusan kerjaya nanti. Terima kasih kepada Prof Talib Latif yang telah memberikan set data untuk kajian ini. Ucapan ribuan terima kasih juga kepada semua warga pendidik UKM dan staf UKM yang telah memberi bantuan kepada saya sepanjang pengajian saya di UKM.

Sekalung penghargaan saya berikan kepada kedua-dua ibu bapa saya yang sentiasa memberi kasih sayang dan doa yang tidak putus pada saya sepanjang pengajian saya di UKM. Terima kasih kepada rakan-rakan yang telah hadir sepanjang pengajian saya terutamanya *Family MCO* yang telah mencerikan hari-hari saya, memberi sokongan moral serta menjadi pendengar yang baik di UKM. Rakan-rakan sepejuangan Program Master Sains Data yang berkongsi ilmu dan pandangan, terima kasih di atas bantuan yang kalian berikan. Semoga jasa baik kalian di balas oleh Allah S.W.T.

PUSAT SUMBER  
FTSM

## ABSTRAK

Peningkatan tahap ozon adalah salah satu perkara yang perlu dikawal oleh manusia. Hal ini kerana kadar ozon yang tinggi boleh mengancam kesihatan manusia dan tumbuhan. Peningkatan tahap ozon berlaku disebabkan oleh asap yang dikeluarkan oleh kenderaan, kilang, pembakaran dan lain-lain. Semasa PKP dilaksanakan disebabkan oleh pandemik Covid-19, terdapat banyak kajian menyatakan bahawa tahap ozon telah menurun disebabkan oleh kurangnya pencemaran. Pada masa itu kebanyakan orang tinggal di rumah. Berdasarkan tinjauan literatur, terdapat beberapa model pembelajaran mesin yang digunakan untuk meramalkan tahap ozon di seluruh negara. Dalam kajian ini, data yang digunakan mengandungi parameter pencemaran udara setiap jam dan parameter meteorologi bermula dari 12 Feb 2020 hingga 21 April 2020 iaitu sebulan sebelum PKP dan sebulan semasa PKP. Set data ini direkodkan daripada stesen cuaca di 13 negeri yang mengandungi 8 parameter pencemaran udara dan 5 parameter meteorologi. Empat algoritma bergabung seperti Hutan Rawak, *XGboost*, *Catboost*, dan *Gradient Boosting* diuji untuk membina model ramalan untuk meramal tahap ozon. Nilai ketepatan, RMSE, dan MSE dibandingkan untuk menentukan model ramalan terbaik. Model yang terbaik digunakan untuk tafsiran faktor penting peningkatan tahap ozon dengan kepentingan Gini dan nilai SHAP. Hasil kajian menunjukkan bahawa model Random Forest adalah model ramalan terbaik dengan ketepatan di kawasan bandar, pinggir bandar, dan latar. Sementara itu, daripada penafsiran faktor penting, suhu dan kelembapan adalah faktor utama dalam meningkatkan tahap ozon di semua kawasan semasa PKP.

Kata kunci: Pembelajaran Mesin Bergabung, Nilai SHAP, Tahap Ozon, Hutan Rawak.

**THE IMPACT OF AIR POLLUTION AND METEOROLOGY ON OZONE  
FORMATION DURING MCO IN MALAYSIA USING MACHINE LEARNING  
METHODS AND SHAP VALUES**

**ABSTRACT**

The increment of ozone level is one of the things that humans need to control because high levels of ozone can threaten the health of humans and plants. The increase in ozone level occurs due to smoke released by vehicles, factories, combustion and others. When MCO was implemented due to the Covid-19 pandemic, many studies stated that the ozone level had decreased due to less pollution. During that time most people are staying at home. Based on the literature review, several machine learning models are used to predict ozone levels across the world. In this study, the data used contains hourly air pollution parameters and meteorological parameters starting from 12 February 2020 until 21 April 2020, one month before MCO and one month during MCO. This data set was collected from the weather station in 13 states with 8 air pollutant parameters and 5 meteorological parameters. Four ensemble algorithms such as Random Forest, XGBoost, CatBoost, and Gradient Boosting were tested to build predictive models to predict ozone levels. The accuracy, MSE, and RMSE for each model were compared in determining the best predictive model and can be used to find important factors of increasing ozone level with SHAP values. The results show that the Random Forest model is the best predictive level in urban, suburban, and background areas. Meanwhile, from the interpretation of important factors, temperature and humidity are the main factors in increasing the level of ozone in all areas during the MCO.

Keywords: Ensemble Machine Learning, SHAP value, Ozon level, Random Forest.

## KANDUNGAN

		<b>Halaman</b>
<b>PENGAKUAN</b>		<b>ii</b>
<b>PENGHARGAAN</b>		<b>iii</b>
<b>ABSTRAK</b>		<b>iv</b>
<b>ABSTRACT</b>		<b>v</b>
<b>KANDUNGAN</b>		<b>vi</b>
<b>SENARAI JADUAL</b>		<b>ix</b>
<b>SENARAI SINGKATAN</b>		<b>xiii</b>
<b>BAB I</b>	<b>Pengenalan</b>	
1.1	Pendahuluan	1
1.2	Latar Belakang	2
1.3	Penyataan Masalah	4
1.4	Persoalan Kajian	5
1.5	Objektif Kajian	5
1.6	Skop Kajian	6
1.7	Metodologi Kajian	6
1.8	Kepentingan Kajian	6
1.9	Organisasi Tesis	7
<b>BAB II</b>	<b>KAJIAN LITERATUR</b>	
2.1	Pengenalan	9
2.2	Ozon	10
	2.2.1 Impak Kualiti Udara Terhadap Fenomena PKP	11
	2.2.2 Faktor Mempengaruhi Tahap Ozon	12
2.3	Perlombongan data	13
	2.3.1 Teknik Bergabung Dalam Perlombongan Data	14
	2.3.2 Model Pembelajaran Mesin Bergabung	15
	2.3.3 Pbandingan Kajian Literatur Mengenai Ramalan Tahap Ozon Menggunakan Pembelajaran Mesin	17
2.4	Tafsiran dalam pembelajaran mesin	21
	2.4.1 Kepentingan Gini berdasarkan Impurity Based	23
	2.4.2 SHapley Additive Explanation (SHAP)	24

	2.4.3	Kajian Literatur Nilai SHAP Dalam Pembelajaran Mesin	25
2.5		Kesimpulan	28
<b>BAB III</b>		<b>METODOLOGI</b>	
3.1		Pengenalan	30
3.2		Perolehan Data	31
3.3		Pra-Pemrosesan Data	34
	3.3.1	Pembersihan Data	35
	3.3.2	Penggantian Nilai Hilang	35
	3.3.3	Data Integrasi dan Transformasi	35
	3.3.4	Pemilihan Atribut	39
3.4		Analisis Statistik	40
	3.4.1	Analisis Korelasi Data Sebelum PKP dan Semasa PKP	40
3.5		Pembangunan Model	41
	3.5.1	Pembangunan Model	41
3.6		Penilaian model	46
	3.6.1	Mean Square Error dan Root Mean Square Error	46
	3.6.2	Ketepatan ( <i>Accuracy</i> )	47
3.7		Pemilihan Model Terbaik Untuk Ramalan Tahap Ozon	47
3.8		Tafsiran Faktor Penting Menggunakan Gini dan Nilai SHAP	47
3.9		Kesimpulan	49
<b>BAB IV</b>		<b>DAPATAN KAJIAN</b>	
4.1		Pengenalan	50
4.2		Hasil Analisis Statistik	50
	4.2.1	Perbandingan Parameter Sebelum PKP dan Semasa PKP	52
4.3		Pemilihan Model Ramalan Tahap Ozon	55
	4.3.1	Perbandingan Prestasi Model di Kawasan Bandar	55
	4.3.2	Perbandingan Prestasi Model di Kawasan Pinggir Bandar	56
	4.3.3	Perbandingan Prestasi Model di Kawasan Latar	57
	4.3.4	Keputusan Prestasi Model Terbaik	58
4.4		Menafsir Faktor Penting dengan Gini Importance dan nilai SHAP	58



4.4.1	Plot Menggunakan Kepentingan Ciri Terbina di dalam Model Hutan Rawak (Gini Importance)	58
4.4.2	Tafsiran Faktor Penting Menggunakan Nilai SHAP	62
<b>BAB V RUMUSAN DAN CADANGAN</b>		
5.1	Pengenalan	74
5.2	Rumusan Hasil dan Pencapaian Objektif Kajian	74
5.1.1	Objektif 1: Mengenal pasti model terbaik ramalan ozon sebelum dan semasa PKP menggunakan teknik pembelajaran mesin	74
5.1.2	Objektif 2: Mengenal pasti faktor utama yang mempengaruhi pembentukan ozon berdasarkan model terbaik ramalan ozon sebelum dan semasa PKP menggunakan tafsiran pembelajaran mesin	75
5.3	Sumbangan Kajian	77
5.4	Cadangan Kajian Seterusnya	78
<b>RUJUKAN</b>		<b>79</b>
Lampiran A	86	

## SENARAI JADUAL

<b>No. Jadual</b>		<b>Halaman</b>
Jadual 2.1	Pembandingan kajian literatur terhadap ramalan ozon.	18
Jadual 3.1	Set data mentah kajian.	34
Jadual 3.2	Lokasi stesen udara yang telah dikategorikan di setiap kawasan.	36
Jadual 3.3	Nilai APU bagi ozon yang digunakan di Malaysia.	38
Jadual 3.4	Julat tahap ozon yang digunakan untuk mengkategorikan nilai ozon.	38
Jadual 3.5	Julat arah angin yang digunakan untuk mengkategorikan arah kompas.	38
Jadual 3.6	Analisis statistik kawasan bandar sebelum dan semasa PKP.	40
Jadual 3.7	Langkah-langkah uji kaji algoritma Hutan Rawak.	42
Jadual 3.8	Langkah-lankah uji kaji algoritma <i>XGboost</i> .	43
Jadual 3.9	Langkah-langkah uji kaji algoritma <i>Catboost</i> .	44
Jadual 3.10	Langkah-lankah uji kaji algoritma <i>GradientBoosting</i> .	45
Jadual 3.11	Langkah-langkah uji kaji menggunakan nilai SHAP dan <i>Gini</i>	48
Jadual 4.1	Analisis statistik kawasan pinggir bandar sebelum dan semasa PKP.	51
Jadual 4.2	Analisis statistik kawasan latar sebelum dan semasa PKP.	51
Jadual 4.3	Nilai ketepatan, MSE, dan RMSE model bagi kawasan bandar sebelum PKP.	55
Jadual 4.4	Nilai ketepatan, MSE, dan RMSE model bagi kawasan bandar semasa PKP.	56
Jadual 4.5	Nilai ketepatan, MSE dan RMSE model bagi kawasan pinggir bandar sebelum PKP.	56
Jadual 4.6	Nilai ketepatan, MSE dan RMSE model bagi kawasan pinggir bandar semasa PKP.	56
Jadual 4.7	Nilai ketepatan, MSE dan RMSE model bagi kawasan latar sebelum PKP.	57

Jadual 4.8 Nilai ketepatan, MSE dan RMSE model bagi kawasan latar semasa PKP.

57

PUSAT SUMBER  
FTSM

## SENARAI ILUSTRASI

<b>No. Rajah</b>		<b>Halaman</b>
Rajah 2.1	Pembentukan ozon troposfera terhasil daripada Nitrogen Dioksida, VOC, CH <sub>4</sub> , dan CO.	11
Rajah 2.2	Proses penemuan pengetahuan dalam pangkalan data.	14
Rajah 2.3	Nilai SHAP untuk setiap ciri ramalan model.	25
Rajah 2.4	Penjelasan ramalan pertama model pembelajaran mesin menggunakan nilai SHAP.	26
Rajah 2.5	SHAP <i>summary plot</i> untuk output pembelajaran mesin.	27
Rajah 2.6	Plot SHAP dependence bagi lima ciri persekitaran bandar.	28
Rajah 3.1	Metodologi CRISP-DM kajian.	31
Rajah 3.2	Lokasi Stesen Udara di Semenanjung Malaysia.	32
Rajah 3.3	Lokasi Stesen Udara di Sabah dan Sarawak.	33
Rajah 3.4	Illustrasi proses penggabungan set data dan gabungan set data “negeri” kepada kawasan bandar, pinggir bandar, dan latar .	37
Rajah 3.5	Set data yang telah dikategorikan kepada set data “sebelum PKP” dan “semasa PKP”.	39
Rajah 3.6	Contoh atribut dan data yang digunakan untuk pembangunan model ramalan.	39
Rajah 4.1	Perbezaan purata sebelum PKP dan semasa PKP di kawasan bandar.	52
Rajah 4.2	Perbezaan purata sebelum PKP dan semasa PKP di kawasan pinggir bandar.	53
Rajah 4.3	Perbezaan purata sebelum PKP dan semasa PKP di kawasan kawasan latar.	54
Rajah 4.4	Plot <i>Relative Importance</i> sebelum dan semasa PKP di kawasan bandar.	59
Rajah 4.5	Plot <i>Relative Importance</i> sebelum dan semasa PKP di kawasan pinggir bandar.	60
Rajah 4.6	Plot <i>Relative Importance</i> sebelum dan semasa PKP di kawasan latar.	61

Rajah 4.7	SHAP <i>summary plot</i> sebelum dan semasa PKP di kawasan bandar.	64
Rajah 4.8	SHAP feature important plot di kawasan bandar sebelum PKP.	65
Rajah 4.9	SHAP feature important plot di kawasan bandar semasa PKP.	66
Rajah 4.10	SHAP plot dependences bagi 3 faktor utama peningkatan tahap ozon untuk kelas 0 di kawasan bandar.	66
Rajah 4.11	SHAP <i>summary plot</i> sebelum dan semasa PKP di kawasan pinggir bandar.	68
Rajah 4.12	SHAP feature important plot di kawasan pinggir bandar sebelum PKP.	69
Rajah 4.13	SHAP feature important plot di kawasan pinggir bandar semasa PKP.	69
Rajah 4.14	SHAP plot dependences bagi 3 faktor utama peningkatan tahap ozon untuk kelas 0 di kawasan pinggir bandar.	70
Rajah 4.15	SHAP <i>summary plot</i> sebelum dan semasa PKP di kawasan latar.	71
Rajah 4.16	SHAP feature important plot di kawasan latar sebelum PKP.	72
Rajah 4.17	SHAP feature important plot di kawasan latar semasa PKP.	72
Rajah 4.18	Plot SHAP dependence bagi 3 faktor utama peningkatan tahap ozon untuk kelas 0 di kawasan latar.	73

**SENARAI SINGKATAN**

m/s	Meter per saat ( <i>meter per second</i> )
µg/m <sup>3</sup>	Mikrogram setiap meter padu ( <i>micrograms per cubic meter</i> )
ppm	Bahagian per sejuta ( <i>parts per million</i> )
ppbv	Bahagian per billion ( <i>parts per billion</i> )
w/m <sup>2</sup>	Watt per meter persegi ( <i>watt per square meter</i> )
°C	Darjah Celsius (Degree Celcius) m/s
PKP	Perintah Kawalan Pergerakan
O <sub>3</sub>	Ozon
PM <sub>2.5</sub>	Zarah Terampai ( PM <sub>2.5</sub> )
PM <sub>10</sub>	Zarah Terampai ( PM <sub>10</sub> )
SO <sub>2</sub>	Sulfur Dioksida
NO <sub>x</sub>	Nitrogen Oksida
NO	Nitrium Oksida
NO <sub>2</sub>	Nitrogen Dioksida
CO	Carbon Oksida
XGboost	eXtreme Gradient Boosting
SHAP	Shapley Additive exPlanations

## **BAB I**

### **PENGENALAN**

#### **1.1 PENDAHULUAN**

Kualiti udara yang bersih adalah keperluan asas bagi semua mahluk yang hidup di bumi ini. Pemantauan kualiti udara yang baik adalah sangat penting kerana kualiti udara berbeza-beza di setiap lokasi. Sebagai contoh, pencemaran udara boleh berlaku di kawasan yang sangat terpencil, sehingga ke kawasan yang sangat sibuk seperti kawasan bandar, industri mahupun di sekitar jalan raya yang dipenuhi kenderaan (Ghorani Azam et al. 2016).

Kualiti udara yang tidak bersih boleh mengancam kesihatan manusia mahupun tumbuh-tumbuhan. Oleh itu, sistem pemantauan udara yang baik dan efisien diperlukan agar ia dapat membantu dalam menilai tahap pencemaran selari dengan piawai kualiti udara ambien. Baru-baru ini, wabak coronavirus telah tersebar di seluruh dunia. Wabak ini telah menyebabkan kerajaan untuk mengambil langkah yang berjaga-jaga dengan mengarahkan rakyat untuk duduk dirumah dan hanya keluar rumah ketika mempunyai urusan yang penting (Cucinotta dan Vanelli 2020). Kesan daripada ini, terdapat pengurangan terhadap penggunaan pengangkutan di jalan raya, di lautan, mahupun di udara. Selain itu, banyak juga perniagaan dan kilang ditutup lantas telah memberikan impak positif terhadap kualiti udara (Singh et al. 2020).

Pembangunan model ramalan untuk meramalkan kualiti udara adalah sangat berguna kerana dengan adanya model tersebut, ia dapat memberikan amaran atau peringatan awal kepada manusia apabila tahap pencemaran udara meningkat ke tahap yang bahaya. Selain itu, pembangunan model ini juga boleh digunakan untuk

membantu dalam menentukan faktor atau punca mengapa pencemaran udara ini berlaku apabila satu set dataset diaplikasikan kepada set data sebelum dan semasa PKP. Untuk membangunkan model ramalan ini, pelbagai kajian literatur yang telah dijumpai menggunakan teknik perlombongan data sebagai kaedah ramalan pencemaran udara. Selain itu, kaedah statistik juga telah digunakan untuk mencari korelasi diantara kualiti udara dengan parameter pencemaran udara dan parameter meteorologi.

Berdasarkan kajian literatur yang dijumpai, teknik pembelajaran mesin bergabung seperti Hutan Rawak dan *XGboost*, telah menunjukkan prestasi yang baik dalam meramal tahap ozon. Teknik pembelajaran mesin bergabung ini juga telah diaplikasikan di beberapa bidang seperti perubatan, ekonomi, perbankan dan mereka mendapati bahawa prestasi teknik ini lebih baik daripada algoritma pembelajaran mesin yang lain (Maryam Al Janabi et al. 2020).

Oleh kerana masalah kerumitan dan masalah hubungan antara parameter yang tidak linear, maka kesukaran untuk mengenal pasti polar dan faktor utama mempengaruhi penghasilan ozon. Oleh itu, tafsiran menggunakan pembelajaran mesin digunakan bagi mengatasi masalah ini. Tafsiran dalam pembelajaran mesin digunakan bagi membantu menilai model yang dipelajari (WJ Murdoch et al. 2019). Terdapat pelbagai cara dalam penafsiran faktor penting dan *Shapley Additive exPlanations* (SHAP) antara kaedah yang telah digunakan secara meluas oleh penyelidik. Mereka mendapati bahawa teknik pembelajaran mesin bergabung dapat mengolah faktor penting lebih baik dengan menggunakan nilai SHAP (R. Rodríguez-Pérez dan Bajorath 2020). Kaedah ini telah dikembangkan kepada pelbagai jenis seperti visualisasi yang menunjukkan polar yang lebih terperinci yang berguna untuk perolehan pengetahuan dalam bentuk pelbagai jenis graf.

## 1.2 LATAR BELAKANG

Wabak pertama Coronavirus Novel (COVID-19) dilaporkan pada bulan Disember 2019 di Wuhan, Wilayah Hubei, China (Heikal Ismail et al. 2020). Wabak ini telah memaksa semua kerajaan di dunia termasuk kerajaan Malaysia untuk melaksanakan Perintah Kawalan Pergerakan (PKP) untuk mengurangkan bilangan yang dijangkiti, bermula pada 18 Mac 2020 (Ash'aari et al. 2020). Kesan daripada arahan PKP dan



arahan untuk duduk dirumah menyebabkan aktiviti di jalan raya, lautan dan udara berkurang. Justeru, hal ini telah mengurangkan pencemaran udara di seluruh negara (Abdullah et al. 2020).

Beberapa kajian telah dijalankan, dan mereka mendapati bahawa terdapat pengurangan pencemaran udara semasa PKP berbanding dengan sebelum PKP (Mokhtari et al. 2019). Sebagai contoh, Berman dan Ebisu (2020) telah membandingkan kualiti udara  $PM_{2.5}$  dan  $NO_2$  di Amerika Syarikat pada tahun 2017 hingga 2020. Hasil kajian ini mendapati bahawa terdapat penurunan sebanyak 11% kepekatan  $PM_{2.5}$  dan 26% kepekatan  $NO_2$ .

Namun, terdapat beberapa kajian juga yang menyatakan bahawa terdapat peningkatan tahap ozon semasa PKP di jalankan. Kajian Li et al. (2010) menyatakan bahawa faktor peningkatan tahap ozon adalah kerana jangka masa cahaya matahari yang lama, contohnya pada musim panas cahaya matahari akan lebih lama berbanding musim sejuk yang bermaksud suhu yang tinggi pada musim panas berbanding musim sejuk. Selain itu, kajian ini juga mendapati kelajuan angin juga menjadi faktor peningkatan tahap ozon. Kajian Sabah Abdul et al. (2003) pula mendapati bahawa parameter meteorologi seperti suhu, tenaga suria menyumbang kepada peningkatan ozon pada waktu siang (Abdul Wahab et al. 2005) dan Kajian KL So et al. (2003) mendapati peningkatan parameter pencemaran udara seperti  $SO_2$  dan  $NO_x$  juga mempengaruhi peningkatan tahap ozon.

Pencarian faktor penting yang menyebabkan peningkatan tahap ozon telah dikaji menggunakan kaedah statistik dan kaedah pembelajaran mesin. Kajian E Kovac et al. (2009) telah menggunakan "*Fourier Analysis*", "*Principal Component Analysis*", dan "*Regression Analysis*" untuk mengenal pasti kolerasi diantara parameter meteorologi dan tahap ozon. Kajian ini mendapati bahawa faktor meteorologi sangat memberi impak terhadap tahap ozon dari pukul 10 pagi sehingga 3 petang. Manakala Rui Feng et al. (2019) telah menggunakan teknik pembelajaran mesin seperti "*Extreme Learning Machine*", "*Multi-layer Perceptron*", hutan rawak dan "*Recurrent Neural Network*" untuk meramal dan mencari faktor penting menggunakan parameter pencemaran udara dan meteorologi yang menghasilkan ozon di Hangzhou China.

Di Malaysia, model ramalan ozon telah dibangunkan untuk mengenal pasti faktor yang mempengaruhi tahap ozon di Klang dengan mengambil kira 5 parameter pencemaran udara dan 3 parameter meteorologi. Kajian telah membangunkan model “*Multiple Linear Regression*” (MLR) dan menggunakan analisis statistik untuk mencari faktor penting dalam pembentukan ozon dan mendapati tiga faktor pembentukan ozon adalah NO, Kelembapan, dan NO<sub>2</sub> (Abdullah et al. 2019).

Pada tahun 2020, Pal Vegard et al. (2020) menyatakan pembelajaran mesin bergabung mempunyai kemampuan untuk mengolah faktor penting lebih baik dengan nilai SHAP. Menurut kajian ini, tafsiran model pembelajaran mesin melalui SHAP digunakan dalam bidang perubatan membolehkan kajian menjelaskan ramalan secara berasingan. Oleh itu, teknik pembelajaran mesin bergabung akan digunakan bagi menghasilkan model ramalan tahap ozon.

### 1.3 PENYATAAN MASALAH

Peningkatan tahap ozon di Malaysia telah menjadi salah satu pencemaran udara semenjak tahun 2004 (Hashim dan Noor 2017) dan fluktuasi tahap ozon di sesuatu lokasi bergantung kepada pencemaran udara yang terdapat di sesebuah kawasan. Antara faktor utama yang mempengaruhi tahap ozon adalah seperti SO<sub>2</sub>, NO<sub>x</sub>, NO<sub>2</sub>, dan CO. Semasa PKP dilaksanakan terdapat penurunan terhadap pencemaran udara direkodkan di seluruh dunia. Fenomena PKP selama sebulan telah memberi impak penurunan tahap pencemaran udara yang dikatakan penyebab utama penurunan ozon. Kajian ini bertujuan mengenal pasti sejauh mana impak penurunan pencemaran udara mempengaruhi pembentukan ozon.

Berdasarkan kajian lepas, pembangunan model ramalan untuk meramal kualiti udara amatlah penting agar dapat memberi amaran awal kepada manusia. Model ramalan gabungan seperti hutan rawak, *XGboost*, telah menunjukkan prestasi yang baik dalam meramal tahap ozon (Marvin et al. 2021). Namun, terdapat juga algoritma bergabung seperti *Catboost* dan *Gradient Boosting* yang juga menunjukkan prestasi yang baik. Justeru, terdapat keperluan untuk menggunakan semula algoritma yang telah digunakan oleh kajian lepas seperti hutan rawak, *XGboost*, *Catboost*, dan *Gradient Boosting* dan membandingkan prestasi mereka.

Selain itu, faktor yang mempengaruhi penurunan tahap ozon juga berbeza di setiap negara. Sebagai contoh, kajian Li et al. (2010) menyatakan bahawa faktor peningkatan tahap ozon adalah suhu dan sinaran radiasi kerana jangka masa matahari yang lama seperti di musim panas. Manakala berdasarkan kajian yang dijalankan di Malaysia mendapati faktor peningkatan tahap ozon adalah NO, NO<sub>2</sub>, dan kelembapan (Abdullah et al. 2019). Oleh itu, tafsiran faktor penting ozon diperlukan untuk menambah pengetahuan baru terhadap peningkatan ozon di Malaysia.

#### **1.4 PERSOALAN KAJIAN**

Terdapat tiga (3) persoalan bagi kajian ini seperti disenaraikan di bawah.

objektif bagi kajian ini untuk mengkaji kepekatan ozon seperti yang disenaraikan berikut:

- i. Apakah pembelajaran mesin terbaik untuk pembangunan ramalan tahap ozon menggunakan parameter pencemaran udara dan meteorologi ?
- ii. Bagaimana nilai SHAP digunakan oleh model ramalan terbaik bagi mengenal pasti faktor utama pembentukan ozon dan sejauh mana nilai SHAP dapat mengeksplotasikannya bagi menjelaskan fenomena ozon ?

#### **1.5 OBJEKTIF KAJIAN**

Terdapat dua (2) objektif bagi kajian ini untuk mengkaji tahap ozon seperti yang disenaraikan berikut:

- i. Mengetahui model terbaik ramalan ozon sebelum dan semasa PKP menggunakan teknik pembelajaran mesin.
- ii. Mengetahui faktor utama yang mempengaruhi pembentukan ozon berdasarkan model terbaik ramalan ozon sebelum dan semasa PKP menggunakan tafsiran pembelajaran mesin.

## 1.6 SKOP KAJIAN

Skop kajian ini adalah untuk mengenal pasti faktor penting yang menyumbang kepada tahap ozon di Malaysia. Data parameter pencemaran udara dan meteorologi seperti PM10, PM2.5, O<sub>3</sub>, CO, NO<sub>2</sub>, SO<sub>2</sub>, NO, dan NO<sub>x</sub> dikumpulkan dari 12 Februari 2020 sehingga 21 April 2020 daripada 46 stesen seluruh Malaysia, satu bulan sebelum dan selepas perintah kawalan pergerakan (PKP). Data ini dikategorikan kepada 3 jenis kawasan iaitu kawasan bandar, pinggir bandar, dan kawasan latar di Malaysia. Kajian ini menggunakan analisis deskriptif statistik dan pembelajaran mesin untuk mengenal pasti faktor penting peningkatan tahap ozon.

## 1.7 METODOLOGI KAJIAN

Kajian ini merupakan kajian berbentuk eksperimental. Kajian ini menggunakan metodologi CRISP-DM dengan menggunakan teknik perlombongan data yang terdiri daripada pengkelasan, ramalan, dan visualisasi. Secara terperinci, metodologi kajian ini adalah seperti pemerolehan data, tapisan data, pembersihan data, analisa diskriptif, penerokaan data, pembangunan dan pemilihan model terbaik dengan menilai model tersebut serta penggunaan model terbaik untuk tafsiran pembelajaran mesin.

Empat algoritma pembelajaran mesin bergabung seperti hutan rawak, *XGboost*, *Catboost*, dan *Gradient Boosting* diuji untuk meramal tahap ozon. Model yang terbaik akan digunakan untuk menafsir faktor penting. Dua kaedah penilaian peringkat yang akan digunakan, pertama adalah dengan menggunakan penilaian melalui kepentingan Gini dan penilaian menggunakan nilai SHAP. Kaedah ini akan dibentangkan sebagai hasil dapatan kajian ini.

## 1.8 KEPENTINGAN KAJIAN

Penyelidikan ini berharap dapat menyumbangkan pengetahuan baru berkenaan tahap ozon supaya dapat memberi manfaat kepada kerajaan Malaysia, ahli penyelidik, dan Jabatan Meteorologi Malaysia. Dalam jangka masa panjang, masyarakat sekeliling juga akan mendapat manfaat daripada penyelidikan ini.

Pra-kajian dilakukan untuk mengenal pasti masalah dan jurang dalam memahami faktor penting tahap ozon. Secara menyeluruh, pengetahuan yang meluas tentang pencemaran udara seperti ozon adalah sangat penting kerana tahap ozon yang tinggi dapat memberi kesan yang negatif terhadap manusia dan tumbuh-tumbuhan.

## 1.9 ORGANISASI TESIS

Penyelidikan ini terdiri daripada 5 bab utama yang menjelaskan kajian ini secara terperinci berkenaan kerja-kerja yang dilakukan bagi menjayakan kajian ini. Tesis telah disusun dalam 5 bab dan ringkasan bagi setiap bab yang akan menerangkan keseluruhan perjalanan tesis adalah seperti yang dijelaskan.

BAB II menerangkan kajian literatur yang lepas tentang isu berkaitan dengan ramalan tahap ozon. Bab ini menerangkan maksud ozon, kesan ozon terhadap alam sekeliling, kesan kualiti udara semasa Pandemi Covid-19, faktor penting peningkatan tahap ozon, kaedah untuk mencari faktor penting, perbandingan model-model ramalan pembelajaran mesin terhadap kajian lepas, tafsiran faktor penting menggunakan penilaian peringkat seperti kepentingan Gini dan nilai SHAP untuk menafsir faktor penting.

BAB III menjelaskan metodologi kajian dengan menerangkan proses penyelidikan ini dengan lebih terperinci bagi mencapai objektif. Terdapat 7 fasa yang perlu dilakukan untuk mendapatkan hasil tafsiran tahap ozon. Antara langkah tersebut adalah seperti proses penyediaan data bersih yang melalui proses analisis dan pra-pemrosesan seperti penggantian nilai hilang, data integrasi, dan transformasi. Teknik pembangunan model ramalan tahap ozon daripada empat algoritma dan seterusnya penilaian model dilakukan untuk mencari model yang terbaik. Fasa yang terakhir adalah menafsir faktor penting daripada model menggunakan penilaian peringkat kepentingan Gini dan nilai SHAP.

BAB IV menjelaskan tentang hasil dapatan kajian yang dijalankan. Bab ini membentangkan hasil dapatan analisa statistik sebelum dan semasa PKP, penilaian terhadap model-model ramalan tahap ozon, dan plot tafsiran faktor penting daripada

kepentingan Gini dan nilai SHAP di kawasan bandar, pinggir bandar, dan latar. Kebanyakan hasil dapatan kajian dilampirkan dalam bentuk graf dan jadual.

BAB V merupakan pencapaian keseluruhan untuk kajian ini berpandukan objektif kajian. Selain itu, sumbangan kajian akan dijelaskan daripada hasil dapatan kajian dan cadangan kajian akan diutarakan untuk tujuan kajian pada masa akan datang.

PUSAT SUMBER  
FTSM

## **BAB II**

### **KAJIAN LITERATUR**

#### **2.1 PENGENALAN**

Bab ini membincangkan secara terperinci tentang tinjauan kajian berkaitan dengan kajian terkini domain. Bab ini dibahagikan kepada tiga bahagian utama: memahami ozon dan kaitannya dengan fenomena covid, mengenal pasti faktor utama menggunakan teknik perlombongan data, dan pencarian faktor penting menggunakan tafsiran pembelajaran mesin seperti nilai SHAP serta teknik visualisasinya.

Bahagian pertama membincangkan kajian literatur mengenai apa itu ozon, bagaimana ozon itu terhasil, secara semula jadi atau sebaliknya dan bagaimana ozon memberi kesan terhadap manusia dan mahluk sekeliling. Seterusnya kajian lepas membincangkan kesan pandemik COVID-19 telah mempengaruhi kualiti udara termasuk ozon di seluruh negara termasuk Malaysia. Bahagian ini juga menganalisa faktor yang mempengaruhi penghasilan ozon di atas muka bumi dan menjelaskan kaedah yang digunakan untuk mencari faktor penting tahap ozon.

Bahagian kedua membincangkan teknik perlombongan data dengan menggunakan teknik pembelajaran mesin bergabung. Dalam bahagian ini, akan dijelaskan literatur kajian lepas yang menggunakan pembelajaran mesin untuk membina model ramalan tahap ozon. Bahagian ketiga membincangkan tentang kaedah tafsiran dalam pembelajaran mesin dan kaedah yang sesuai digunakan bagi pembelajaran mesin bergabung seperti kepentingan Gini, nilai SHAP.

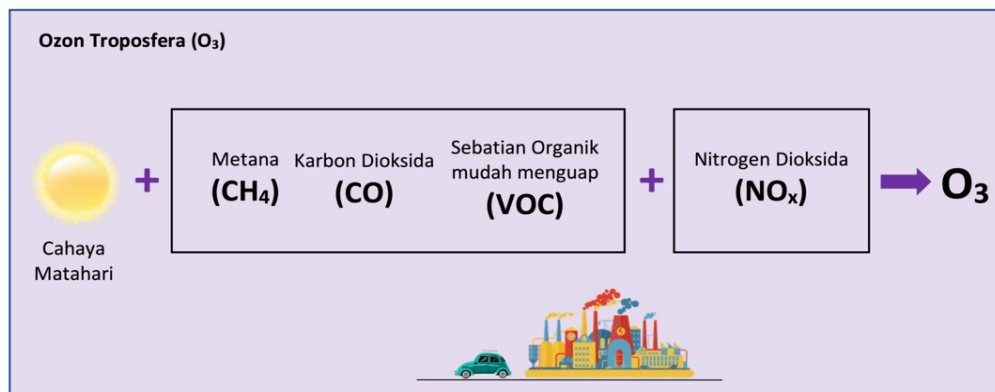
## 2.2 OZON

Ozon ( $O_3$ ) adalah gas reaktif yang terdapat di dua lapisan atmosfera dimana di lapisan atas adalah Ozon Stratosfera dan lapisan bawah adalah Ozon Troposfera. Ozon adalah gas yang terdiri daripada tiga atom oksigen ( $O_3$ ). Pada atmosfera bawah,  $O_3$  dihasilkan oleh tindak balas kimia di antara sebatian kimia seperti sebatian organik yang mudah meruap (VOC) dan nitrogen oksida ( $NO_x$ ) sewaktu kehadiran cahaya matahari (Zhang et al. 2019).

Menurut “*National Ambient Air Quality Standards*”, pembentukan  $O_3$  berasal daripada bahan pencemaran udara seperti asap kenderaan, pelepasan gas daripada industri, serta sumber semula jadi seperti  $NO_x$  dan VOC (Suh et al. 2000). Oleh kerana pembentukan ozon memerlukan sinaran cahaya matahari, ozon hanya mudah terbentuk pada waktu siang dan boleh mencapai tahap tertinggi pada waktu panas (Warmiński dan Bęś 2018).

Tahap ozon yang tinggi dapat mengancam hidupan-hidupan di bumi. Menurut Li et al. (2015), untuk melindungi diri ketika tahap ozon tinggi, manusia disarankan untuk tidak kerap melakukan aktiviti di luar rumah sewaktu musim panas. Hal ini kerana, apabila seseorang itu terlalu banyak terdedah kepada pencemaran udara seperti ozon, ia boleh mendatangkan pelbagai masalah kesihatan. Antara kesan ozon terhadap kesihatan termasuk sakit dada, batuk, kerengsaan tekak, dan sesak nafas (Amann 2008). Berdasarkan kajian yang di jalankan oleh Semple dan Moore (2020) menyatakan bahawa terlalu terdedah kepada  $O_3$  akan menyebabkan kerosakan kapsid dan mengganggu kitaran pembiakan. Rajah 2.1 menunjukkan ilustrasi pembentukan ozon. Berdasarkan ilustrasi pada rajah tersebut, ozon troposfera terhasil daripada partikel  $CH_4$ ,  $CO$ , NMVOC, dan  $NO_x$  dengan kehadiran cahaya matahari.





Rajah 2.1 Pembentukan ozon troposfera terhasil daripada Nitrogen Dioksida, VOC, CH<sub>4</sub>, dan CO.

### 2.2.1 Impak Kualiti Udara Terhadap Fenomena PKP

Pada Disember 2019, wabak pertama Coronavirus Novel (COVID-19) telah dilaporkan di Wuhan, Wilayah Hubei, China (Surveillances 2020). Pada 11 Mac 2020, Pertubuhan Kesihatan Sedunia (WHO) telah membuat pengumuman bahawa wabak novel coronavirus (COVID-19) sebagai wabak global. Ketua Pengarah WHO, Dr. Tedros Adhanom Ghebreyesus menjelaskan bahawa jumlah kes di luar negara China telah meningkat sebanyak 13 kali ganda dan jumlah negara dengan kes tersebut telah meningkat 3 kali ganda. Hal ini kerana virus ini telah tersebar dengan meluas di negara-negara seperti Eropah, Amerika Utara, Asia, dan Timur Tengah, dengan kes pertama di rekodkan di negara Afrika dan Amerika Latin (Hua dan Shaw 2020).

Berdasarkan kajian yang dilakukan di JAMA, penyelidik melaporkan bahawa SARS-CoV-2, nama virus COVID-19, dapat merebak dengan mudah dan cepat. Berdasarkan uji kaji yang dijalankan terhadap pesakit, virus ini sering dijumpai dalam sistem pernafasan manusia. Manusia yang menghidap virus ini biasanya akan mengalami penyakit seperti selesema biasa (Rahimi et al. 2020). Selain itu, pesakit yang menghidap virus ini juga dapat dicirikan dengan gejala sakit seperti demam, lesu, batuk kering, dan limfopenia. Menurut Abdullah et al. (2019), warga emas yang berusia lebih daripada 80 tahun berisiko dengan kadar kematian yang lebih tinggi dengan kadar sebanyak 21.9% setelah dijangkiti virus ini. Namun, sehingga hari ini masih belum ada rawatan atau vaksin yang berkesan yang dapat digunakan untuk merawat pesakit yang dijangkiti coronavirus.

Oleh itu, pada 18 Mac 2020, kerajaan Malaysia telah mengumumkan Pelaksanaan Perintah Kawalan Pergerakan (PKP) dan Arina et al. (2020) menyatakan sewaktu PKP dilaksanakan, hanya sektor-sektor penting yang dibenarkan beroperasi. Kegiatan di luar rumah dilarang kecuali untuk membeli barang keperluan rumah dan mendapat rawatan perubatan. Selain itu juga, sebahagian besar syarikat-syarikat di Malaysia membenarkan pekerjaannya bekerja dari rumah.

Selain itu, PKP juga menyebabkan aktiviti komersial dan perindustrian dikurangkan lantas mengurangkan pergerakan lalu lintas kereta, bas, trak dan kapal terbang. Kesan daripada ini menyebabkan pengurangan pencemaran udara secara drastik dan memberi impak yang besar terhadap perubahan iklim. Menurut kajian Mohd Shahrul et al. (2020), sumber utama pencemaran udara di Malaysia adalah dari asap kenderaan, kilang, tapak pelupusan sampah, dan loji rawatan air sisa. Kehadiran tahap pencemaran udara yang tinggi pada masa kini perlu dikawal sebaiknya kerana ia dapat mengancam kesihatan manusia dan tumbuh-tumbuhan.

Beberapa kajian telah dilakukan, dan mereka mendapati bahawa terdapat pengurangan pencemaran udara semasa PKP dilaksanakan berbanding sebelumnya. Nur Faseeha et al. (2020) telah mengkaji kualiti udara di beberapa kawasan di Malaysia sebelum dan semasa PKP dilaksanakan. 4 parameter meteorologi dan 6 parameter pencemaran udara digunakan. Berdasarkan hasil kajian, mereka mendapati bahawa parameter seperti O<sub>3</sub>, PM<sub>10</sub>, PM<sub>5</sub>, SO<sub>2</sub>, dan CO mengalami penurunan diantara 5% hingga 50% pada minggu pertama PKP dilaksanakan.

### **2.2.2 Faktor Mempengaruhi Tahap Ozon**

Kajian literatur menerangkan apakah faktor yang menyebabkan peningkatan tahap ozon. Sabah Abdul et al. (2002) mengkaji faktor yang mempengaruhi tahap ozon di bandar Kuawit yang kering sewaktu musim panas dan sejuk dan kering semasa musim sejuk. Kajian mereka mendapati bahawa parameter seperti NO, SO<sub>2</sub>, dan kelembapan adalah tiga teratas yang memberi kesan paling besar terhadap penghasilan ozon. Selain itu, parameter seperti suhu juga menyumbang kepada peningkatan ozon pada waktu siang.

Pada tahun 2003, So et al. (2003), menjalankan kajian di beberapa wilayah Hong Kong terhadap tahap ozon. Mereka menggunakan data daripada tahun 1999 sehingga 2000 yang mengandungi parameter 5 pencemaran udara. Hasil analisa mendapati bahawa peningkatan tahap ozon selari dengan peningkatan SO<sub>2</sub> dan NO<sub>2</sub> dengan keadaan cahaya matahari yang kuat dan kelajuan angin yang rendah. Pada tahun 2012, Hone-Jay Chu et al. (2012) telah mengkaji data siri masa yang mengandungi tahap ozon dan parameter meteorologi untuk mengenal pasti kenaikan tahap ozon di Taiwan. Berdasarkan analisis mereka, mereka mendapati parameter seperti suhu, kelajuan angin, VOC, dan NO<sub>x</sub> adalah punca utama yang menyebabkan naik turunnya tahap ozon.

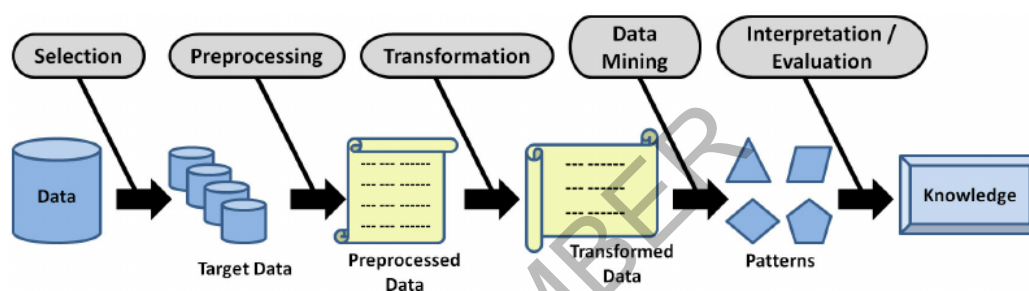
Lee et al. (2014) telah mengkaji korelasi NO<sub>2</sub>, NO<sub>x</sub>, CO dan VOC terhadap tahap ozon di Hong Kong dan China Selatan berdasarkan data daripada tahun 1990 hingga ke tahun 2010. Menurut kajian mereka, mereka mendapati bahawa penurunan NO<sub>2</sub>, NO<sub>x</sub>, CO tidak banyak mempengaruhi tahap kenaikan O<sub>3</sub>. Walau bagaimanapun, kenaikan sebatian organik yang tidak menentu (VOC) adalah penyebab utama kenaikan O<sub>3</sub>.

Jabatan Alam Sekitar, Makanan dan Hal Ehwal Luar Bandar (Defra) melakukan penyelidikan terhadap kadar peningkatan ozon dan apakah faktor ia berlaku di United Kingdom. Kajian ini dijalankan dengan menggunakan data daripada tahun 1992 sehingga pertengahan tahun 2019. Hasil daripada kajian mereka, mereka mendapati bahawa tahap ozon masih meningkat walaupun kadar NO<sub>x</sub> turun. Mereka juga mendapati bahawa tahap ozon tertinggi direkodkan iaitu pada akhir musim bunga dan semasa musim panas, dimana ketika itu sinaran matahari berterusan.

### **2.3 PERLOMBONGAN DATA**

Perlombongan data secara umumnya adalah aplikasi algoritma kepada set data untuk mencari bentuk dan menilai bentuk pengetahuan yang tersebut. Perlombongan data juga dikenali sebagai proses penemuan pengetahuan, pengekstrakan pengetahuan atau analisis data (Bharati dan Ramageri 2010).

Proses pengekstrakan maklumat dalam pangkalan data merangkumi beberapa langkah daripada pengumpulan data mentah sehingga ke beberapa data yang dapat digunakan untuk proses pemodelan. Langkah tersebut terdiri daripada pembersihan data, data integrasi, pemilihan data, transformasi data, perlombongan data, penilaian corak, dan persembahan pengetahuan (Chamatkar dan Butey 2014). Rajah 2.2 menunjukkan proses bagaimana proses pengekstrakan sehingga maklumat daripada data mentah diperoleh.



Rajah 2.2 Proses penemuan pengetahuan dalam pangkalan data.

Sumber: F Gullo (2015).

Selain itu, teknik utama dalam penerokaan pengetahuan adalah seperti *Association Rules*, *Sequence Mining*, *Deviation Detection*, *Classification*, *Regression*, dan *Clustering* (Ajay Kumar dan Indranath 2016).

### 2.3.1 Teknik Bergabung Dalam Perlombongan Data

Pembelajaran mesin bergabung digunakan di dalam kajian ini untuk meramal tahap ozon. Pembelajaran mesin bergabung adalah untuk melatih beberapa *base learner* sebagai satu ahli kemudian menggabungkan ramalan mereka kepada satu output yang mempunyai prestasi yang lebih baik daripada ahli bergabung lain dengan ralat yang tidak berkaitan pada set data sasaran. Untuk tugas klasifikasi, banyak penyelidik telah menguji bahawa pembelajaran mesin bergabung menunjukkan prestasi yang terbaik (Dietterich et al. 2002).

Secara umumnya, gabungan yang baik mempunyai *base learner* yang tepat, dan meluas. Terdapat pelbagai proses untuk menganggar ketepatan sesuatu model seperti pengesahan bersilang, ujian *hold-out*, dan sebagainya. Pengambilan proses

penjanaan *base learner* yang berbeza atau kombinasi yang berbeza membawa kepada kaedah bergabung yang berbeza. Terdapat pelbagai kaedah bergabung yang berkesan. Tiga contoh kaedah bergabung adalah seperti *Boosting*, *Bagging*, dan *Stacking* (Zhi-Hua Zhao 2009).

### 2.3.2 Model Pembelajaran Mesin Bergabung

Untuk kajian ini, teknik pembelajaran mesin bergabung digunakan untuk meramal tahap ozon di Malaysia. Setelah meneliti kajian-kajian literatur lepas, algoritma pembelajaran mesin bergabung mampu menunjukkan prestasi yang baik untuk meramal tahap ozon. Antara algoritma yang akan digunakan untuk kajian ini adalah seperti hutan rawak, *XGboost*, *Catboost* dan *Gradient Boosting*.

#### a. Hutan Rawak

Algoritma Hutan Rawak diperkenalkan oleh Breiman (2001). Algoritma ini merupakan salah satu teknik bergabung yang melatih beberapa pokok keputusan secara selari dengan *bootstrap*. Pokok keputusan yang dilatih secara selari oleh *bootstrap* menggunakan subset dari ciri yang tersedia kerana *bootstrap* akan memastikan setiap output hutan rawak adalah unik. Algoritma hutan rawak berjaya memberikan anggaran yang lebih tepat dan baik berbanding pokok keputusan, dengan mengurangkan korelasi antara pokok yang dibina.

Bagi mengeluarkan output akhir, pengklasifikasi hutan rawak akan menggabungkan keputusan setiap pokok agar output digeneralisasi dengan baik. Kelebihan hutan rawak adalah ia dapat mengatasi kebanyakan kaedah klasifikasi tanpa ada masalah berlebihan. Hutan rawak juga mempunyai kelebihan terhadap pemilihan sampel latihan dan kebisingan dalam set data latihan (Siddharth Mirsa dan Hao Li 2020).

**b. XGboost**

Algoritma *XGboost (Extreme Gradient Boosting)* adalah model yang diusulkan dalam pertandingan . Algoritma *XGboost* adalah salah satu pembelajaran mesin bergabung pokok keputusan yang berdasarkan model pokok *Boosting*. Selain itu, ia juga menggunakan hubungan antara *boosting* dan *Gradient Boosting Machine (GBM)*. *XGBoost* merupakan GBM yang lebih baik dan efisien. Algoritma ini juga merupakan algoritma yang terdiri atas kumpulan klasifikasi dan pokok regresi (Tianqi Chen dan Carlos 2011).

**c. Catboost**

Algoritma *Catboost* merupakan algoritma pembelajaran mesin yang dicipta dari Yandex yang digabungkan daripada *Gradient Boosting Decision Tree (GBDT)* dan ciri kategori (M Kinnander 2020). Mekanisme ini juga menangani masalah kecenderungan dan masalah ramalan sehingga meningkatkan kemampuan generalisasi dan ketahanan algoritma. Semasa algoritma ini memproses ciri kategori, ia memasukan semua set data sampel ke dalam algoritma untuk proses latihan. Kemudian, ia mengatur semua set data ini diatur secara rawak dan menyaring sampel dengan kategori yang sama ciri. *CatBoost* juga menggunakan prinsip mengarah untuk menyelesaikan masalah GBDT dengan memodifikasi algoritma GBDT dengan ramalan, dan juga menjadi algoritma baru untuk memproses ciri yang dinamakan dengan *ordered target statistic* (Prokhorenkova et al 2018).

**d. Gradient Boosting**

Algoritma *Gradient Boosting* menggabungkan kumpulan yang lemah, yang bermaksud kumpulan yang sedikit lebih baik daripada kumpulan yang dipilih secara rawak. Dengan melatih kumpulan tersebut menjadi kumpulan yang kuat dengan cara berulang supaya dapat meminimumkan fungsi kerugian. Fungsi kerugian ini diukur dengan idea bahawa setiap model baru dapat meminimumkan fungsi kerugian dan diukur dengan kaedah keturunan kecerunan. Dengan adanya fungsi kerugian, setiap model baru akan lebih menjadi lebih tepat, dan nilai ketepatan dapat ditingkatkan.

Walaupun bagaimanapun, *boosting* perlu diperhatikan, agar model tidak akan cenderung berlebihan (Candice Bentejac et al. 2019) (Rahman et al. 2020).

### **2.3.3 Perbandingan Kajian Literatur Mengenai Ramalan Tahap Ozon Menggunakan Pembelajaran Mesin**

Dalam kajian ini, kaedah pembelajaran bergabung digunakan untuk meramal tahap ozon di Malaysia. Berdasarkan definisi, pembelajaran mesin bergabung adalah langkah untuk melatih pelbagai algoritma dan menggabungkan hasil mereka. Pembelajaran mesin bergabung juga adalah algoritma yang membina sekumpulan pengelasan dan mengklasifikasikan titik data baru dengan mengambil kira output ramalan model tersebut (J Kittler dan F Roli 2003). Jadual 2.1 menunjukkan ringkasan perbandingan kajian literatur lepas terhadap ramalan tahap ozon dengan menggunakan pembelajaran mesin.

PUSAT SUMBER  
FTSM

Jadual 2.1 Pembedingan kajian literatur terhadap ramalan ozon.

Penulis	Tajuk	Set Data	Eksperimen Algoritma	Penilaian	Rumusan
Asha B. Chelani, 2009	<i>Prediction of daily maximum ground ozone concentration using support vector machine</i>	Set Masa Multi – Pembolehkan di Bahadurshah Zahar Marg, Delhi.	Multiple Regression, SVM, MLP	MAPE, NRMSE, RMSE, dan Index of agreement.	
Nahun Loya et al. 2012	<i>Forecast of Air Quality Based on Ozone by Decision Trees and Neural Network</i>	Set data Multi – Variate di Bandar Mexico	Random Forest, C4.5 dan MLP	Accuracy	Ozon di bahagikan kepada 5 kelas. Daripada kelas “Good” kepada “Highly bad”.
Ningbo Jiang and Matthew L. Riley, 2015	<i>Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY</i>	Set data Multi – Variate di Sydney.	Decision Tree, Random Forest	Accuracy , Bias, False alarm rate, Critical success index, Probability, dan Skill	Algoritma pokok keputusan di majukan daripada CART algoritma.
Eman S. Al Abri et al. 2015	<i>Modelling Ground-Level Ozone Concentration using Ensemble Learning Algorithms</i>	Set data Multi- Variate, di “Sohar Highway”, Oman	ANN, SVM, Random Forest, dan Bagging Classifier	Correlation Coefficient, dan MAE	Membuat perbandingan diantara Ensemble Classifier Bagging dan Standard Single Classifier.
					bersambung...



---

...sambungan					
Sanjiban Sekhar Roy et al. 2017	<i>Predicting Ozone Layer Concentration Using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree</i>	Set data UCI Multi – Variate di Itali	Random Forest, Multivariate Adaptive Regression Splines and	RMSE, MSE, GCV, MAD, MRAD, SSY, SSE, R <sup>2</sup> , R <sup>2</sup> Norm, GCV R-Sr	Menggunakan MARS dan Random Forest untuk mencari faktor penting peningkatan tahap ozon.
Zhiying Meng, 2019	<i>Ground Ozone Level Prediction Using Machine Learning</i>	Set data UCI (Ozone level Detection) di Bostn	Logistic Regression, Decision Tree, Random Forest, AdaBoost, dan SVM	“Principal Compenent Analysis” (PCA) dan “Logistic Regression”	Mengelaskan hari yang ada ozon sebagai kelas “1” dan “0” tiada ozon.
Mohan, S. dan Saranya,P 2019	<i>A novel bagging ensemble approach for predicting summertime ground-level ozone concentration</i>	Set data Multi – Variate di Gummidipoondi, India.	Random Tree, REP Tree, Random Forest, dan Base Classifier	Willmott’s index of agreement (IoAd) , R <sup>2</sup> , dan PEP	Menggunakan Willmotts ‘s index of agreement (IoAd) untuk penilaian model.
Rui Feng et al. 2019	<i>Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison:A case study in hangzhou, China</i>	Set data Multi – Variate di Westbrook National Park	Random Forest, ELM, MLP, RNN	R, RMSE, NMB(%), NME(%), MFB(%), dan MPE(%)	Menggunakan hutan rawak untuk menafsir faktor penting.
					bersambung...

---

---

...sambungan					
Roberta Valentina Gagliardi dan Claudio Andenna, 2020	<i>A Machine Learning Approach to Investigate the Surface Ozone Behaviour</i>	Set data Multi-Variate di Wilayah Basilicata, Itali	Advance ML, dan Boosted Regression Tree	Kaedah Statistik, R <sup>2</sup> , RMSE, MAE. Dan MBE	Menggunakan model BRT untuk mencari faktor penting ozon.
Weeberb J. Requia et al. 2020	<i>An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States</i>	Set data Multi – Variate di Amerika	Neural Network, Random Forest, dan Gradient Boosting	R <sup>2</sup> , RMSE, nilai cerun,	Menuggunakan pengaturcaraan R dengan menggunakan paket H2O.
Maryam Aljanabi et al. 2020	<i>Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan</i>	Siri Masa Multi-Variate, di Jordan	MLP, SVR, DTR, dan XGboost	R <sup>2</sup> , RMSE, dan MSE	Melakukan pemilihan yang intensif untuk mengurangkan jumlah ciri dan mengurangkan masa yang diperlukan untuk ramalan.
Yu Wang et al. 2021	<i>Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on Himawari-8 using a self-adaptive geospatially local model</i>	Siri Masa Multi – Variate di China	SGLboost, Catboost, Light GBM, XGboost, RF, dan ERT	R <sup>2</sup> , RMSE, dan RPE	Menggunakan model yang dapat menyesuaikan keadaan secara geospasial.

---

Jadual 2.1 menunjukkan literatur lepas dari tahun 2009 sehingga 2021 yang telah menggunakan teknik pembelajaran mesin untuk meramal tahap ozon. Berdasarkan analisa, parameter multi-variate yang direkodkan di negara sendiri lebih banyak digunakan berbanding data yang di ekstrak daripada laman sesawang UCI. Set data yang digunakan mengandungi parameter pencemaran udara dan parameter meteorologi. Pada bahagian pemprosesan data, setiap kajian literatur menggunakan teknik yang berbeza untuk mengatasi masalah seperti nilai hilang, pemilihan atribut dan lain-lain bergantung kepada data masing-masing.

Untuk meramal tahap ozon, beberapa algoritma telah digunakan seperti hutan rawak, pokok keputusan, MLP, ANN dan lain-lain. Berdasarkan hasil kajian, mereka mendapati bahawa algoritma bergabung seperti hutan rawak telah menunjukkan prestasi ramalan yang terbaik berbanding algoritma yang lain. Menurut Eman Abri (2015), pembelajaran mesin bergabung adalah pendekatan yang terbaik untuk meramal tahap ozon kerana ia menggunakan teknik klasifikasi yang berbeza untuk membina satu model dan output dari model yang berbeza disatukan ke dalam satu model ramalan.

Oleh itu, kajian ini akan menggunakan semula algoritma bergabung yang telah digunakan pada kajian lepas seperti algoritma hutan rawak, *XGboost*, *Catboost*, dan *Gradient Boosting* untuk meramal tahap ozon. Dalam penilaian prestasi model, prestasi algoritma ini akan di nilai menggunakan kaedah yang sama seperti kajian lepas dengan mengambil kira nilai ketepatan, RMSE, dan MSE.

#### **2.4 TAFSIRAN DALAM PEMBELAJARAN MESIN**

Tafsiran dalam pembelajaran mesin adalah sangat penting untuk pengekstrakan pengetahuan yang relevan mengenai hubungan yang terkandung dalam data atau yang dipelajari oleh model (James Murdoch et al. 2019). Ia memberi model pembelajaran mesin kemampuan untuk menerangkan istilah yang dapat difahami oleh manusia. Dari perspektif pembangun dan penyelidik sistem pembelajaran mesin, tafsiran yang diberikan dapat membantu mereka memahami masalah dan data dengan lebih baik (Liu et al. 2017).

Menurut Gregor Stiglic et al. (2020) tafsiran dalam pembelajaran mesin dapat di kategorikan kepada dua kategori: kebolehtafsiran intrinsik dan kebolehtafsiran pasca-hoc. Kebolehtafsiran intrinsik dicapai dengan model yang mempunyai kemampuan untuk menerangkan sendiri dengan menggabungkan kebolehtafsiran secara lansung ke strukturnya. Contoh bagi kategori ini merangkumi pokok keputusan, model *rule-based*, model linear, dan lain-lain. Manakala bagi pasca-hoc, ia merujuk kepada penerapan kaedah pentafsiran selepas latihan model dan contohnya adalah seperti kepentingan ciri permutasi. Perbezaan diantara kedua kategori ini terletak terhadap pertukaran antara ketepatan model dan penjelasan.

Carvalho et al. (2019) menyatakan terdapat pendekatan umum untuk mengklasifikasikan kaedah keterangan dalam pembelajaran mesin adalah dengan Model-Generality. Dalam taksonomi ini, kaedah tersebut adalah seperti model-spesifik atau model-khusus (*Agnostik*). Teknik penafsiran model-khusus adalah teknik yang baik untuk menafsir output model pembelajaran mesin yang sangat kompleks. Teknik ini adalah untuk mengeluarkan penjelasan pasca-hoc dengan menggunakan model asal sebagai “*Black Boxes*” (Cynthia Rudin et al. 2021)

Bagi pendekatan model-spesifik, ia menggunakan pemberat fitur untuk mengenal pasti penerangan yang menentukan ramalan model pembelajaran mesin. Menurut Sunghyeon Choi dan Jin Hur (2020), model pembelajaran mesin bergabung yang berasaskan pokok, seperti *Gradient Boosting*, hutan rawak, dan *XGboost* adalah dikategorikan di dalam kategori pasca-hoc yang menggunakan kaedah keterangan model-khusus dengan mengira ketepatan apabila ciri digunakan pada pokok.

Menze et al. (2009) menyatakan penggunaan kaedah pasca-hoc digunakan pada pembelajaran mesin bergabung seperti hutan rawak dengan melatih model tersebut menggunakan pendekatan peringkat yang berbeza untuk mengenal pasti faktor penting di set data. Dua penilaian peringkat tersebut adalah berasaskan permutasi (*Permutation based*) dan *Impurity based* seperti kepentingan *Gini*. Kajian Matthew Valazquez dan Yugyung Lee (2021), menggunakan kaedah pasca-hoc pada pembelajaran mesin bergabung seperti hutan rawak, *XGboost*, dan *Gradient Boosting*. Mereka menggunakan penilaian peringkat berasaskan *Impurity based* untuk menafsir

faktor penting bagi penyakit Alzheimer dan Refluks Esofagus. Andreas Messalas et al. (2019) menyatakan penerangan model-khusus akan menjadi lebih dengan penggunaan penilaian seperti nilai SHAP kerana ia memberikan tafsiran penuh. Pal Vegard et al. (2020) menyatakan bahawa pembelajaran mesin bergabung mempunyai keupayaan untuk memproses faktor penting dan ia dapat ditingkatkan dengan menggunakan nilai SHAP.

#### 2.4.1 Kepentingan Gini berdasarkan Impurity Based

Kepentingan Gini (*Gini-importance*) adalah ukuran bagi setiap kali perpecahan nod dibuat pada pemboleh ubah, dan kekotoran Gini bagi dua nod adalah kurang daripada nod yang asal. Dengan menambah penurunan Gini bagi setiap pemboleh ubah ke atas semua pokok di hutan memberikan kepentingan pemboleh ubah yang sangat konsisten dengan ukuran kepentingan permutasi. Terdapat beberapa kaedah penilaian yang berbeza dalam penilaian faktor penting sebagai contoh penggunaan purata penurunan ketepatan digunakan untuk model hutan rawak. Bagi pemboleh ubah kepentingan berdasarkan indeks kekotoran Gini, penurunan purata Gini digunakan untuk pengiraan pembahagian semasa latihan.

Setiap masa pembahagian nod dibuat semasa latihan dijalankan pada pemboleh ubah, kriteria bagi kekotoran Gini adalah untuk kedua-dua nod kurang daripada nod asal. Pengiraan penurunan Gini,  $G$  untuk pemboleh ubah individu terhadap jumlah pokok di model memberikan kepentingan ciri yang cepat.

$$G = \sum_{i=1}^{n_c} P_i(1 - P_i) \quad \dots(2.1)$$

Di mana  $n_c$  adalah bilangan kelas didalam sasaran pemboleh ubah dan  $P_i$  adalah nisbah kelas ini. Oleh itu, bagi kes klasifikasi binari  $G$  dimaksimumkan untuk sampel dengan jumlah contoh yang sama bagi setiap kelas dan diminimumkan untuk set homogen :

$$\text{Importance(variable)} = G_{\text{parent}} - G_{\text{split1}} - G_{\text{split2}} \quad \dots(2.2)$$

Kepentingan berubah adalah purata pemboleh ubah yang digunakan dan purata penurunan Gini akan menjadi min bagi kepentingan ini (Uma dan Valarmathi, 2019).

#### 2.4.2 SHapley Additive Explanation (SHAP)

Konsep nilai Shapley (SHAP) pada awalnya dibangunkan untuk menganggarkan faktor penting pemain individu di dalam pasukan kolaborasi. Ia diperkenalkan oleh Lloyd Shapley pada tahun 1953. Konsep ini bertujuan untuk mendistribusikan jumlah keuntungan atau pembayaran terhadap pemain, bergantung kepada kepentingan faktor sumbangan mereka terhadap hasil akhir permainan. Nilai Shapley memberikan penyelesaian kepada pemberian ganjaran kepada setiap pemain dan mewakili hasil unik yang dicirikan oleh sifat semula jadi atau aksioma (Kelly 2003).

Menurut Futagami et al. (2021), kesan setiap ciri pada ramalan yang dipelajari oleh model dikira melalui nilai SHAP. Dalam SHAP, di beri input  $x = [x_1, \dots, x_p]$  dan model  $f$  terlatih, SHAP menganggar model  $f$  dengan model  $g$  yang dapat memudahkan penjelasan sumbangan setiap ciri. Formula bagi mendapatkan model  $g$  dapat dirumuskan seperti berikut.

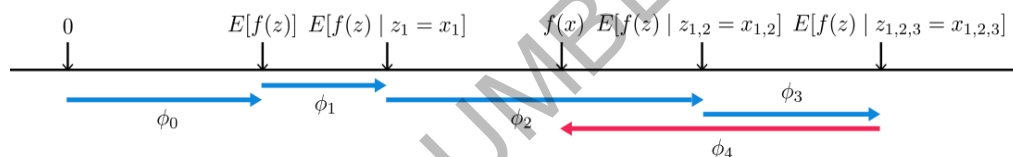
$$g(z) = \phi_0 + \sum_{i=1}^p \phi_i z_i. \quad \dots(2.3)$$

Dimana  $p$  adalah bilangan ciri,  $z = [z_1, \dots, z_p]^T$  adalah ringkasan bagi input  $x$ , di mana nilai  $z$  yang bersesuaian dengan ciri yang digunakan dalam ramalan data adalah 1 dan nilai  $z$  yang sesuai dengan ciri yang tidak digunakan adalah 0. Selain itu,  $\phi_0$  mewakili sumbangan setiap ciri kepada model. Seterusnya, formula bagi mendapatkan nilai  $\phi_0$  adalah seperti yang berikut.

$$\phi_i(f, x) = \sum_{z \subseteq x} \frac{|z|!(p - |z| - 1)!}{p!} [f(z) - f(z \setminus i)]. \quad \dots(2.4)$$

Nilai yang dikeluarkan oleh formula  $\Phi_i$  dikenali sebagai nilai SHAP, dan ia sama dengan nilai SHAP di dalam teori permainan. Nilai SHAP adalah nilai yang mewakili sumbangan setiap pemain ketika pemain bekerjasama dalam permainan dengan beberapa pemain. Dengan kata lain, SHAP menghitung nilai SHAP setiap ciri sebagai pemain dalam model yang dipelajari.

Penafsiran model pembelajaran mesin melalui nilai SHAP akan membolehkan kita menerangkan ramalan secara berasingan dan nilai SHAP dapat memenuhi tiga ciri seperti *Local Accuracy*, *Missingness*, dan *Consistency*. Rajah 2.3 di bawah menunjukkan SHAP menerangkan faktor penting terhadap sesuatu atribut.



Rajah 2.3 Nilai SHAP untuk setiap ciri ramalan model.

Sumber: Beford (2020)

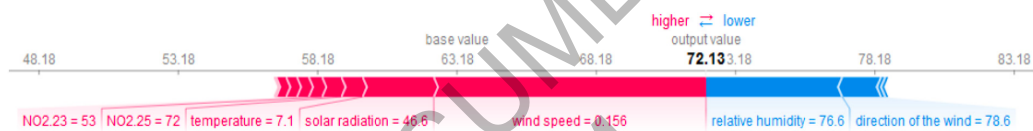
Kaedah mencari kepentingan *Additive feature* dikenali sebagai *SHAP framework*. Ia menunjukkan cara penyelesaian yang unik dalam kelas ini yang memenuhi ciri yang diperlukan (Lundberg dan Lee 2017). Dalam pengiraan nilai SHAP, Lundberg menyusun algoritma pokok SHAP yang digunakan untuk menyediakan kaedah yang cepat dan tepat dengan menggunakan pohon keputusan. Selain itu, pokok SHAP memberikan keterangan khusus maklumat kotak hitam mengenai pembelajaran mesin berasaskan pokok. Keupayaan menghasilkan penjelasan tempatan dengan mudah dan tepat menggunakan nilai SHAP pada set data dapat mengembangkan pengetahuan baru untuk memahami model algoritma dan mengenal pasti kejadian anomali yang mencurigakan (Donghyun Kim et al. 2021).

### 2.4.3 Kajian Literatur Nilai SHAP Dalam Pembelajaran Mesin

Maria Vaga Garcia and Jose Aznarte (2020) telah menggunakan SHAP untuk mencari faktor penting NO<sub>2</sub> di Sepanyol. Menurut literatur tersebut, model pembelajaran

mesin mempunyai masalah dalam menafsirkan ramalan mereka kerana seni bina “kotak hitam” yang kompleks. Masalah ini sering berlaku walaupun mereka memperoleh model ramalan yang baik, lantas ia membebankan mereka untuk mencari faktor penting sesuatu ramalan. Namun begitu, mereka berhasil menafsir faktor penting dengan menggunakan SHAP.

Mereka menggunakan nilai SHAP untuk menafsir output daripada model “*Deep Neural Network*” yang digunakan untuk meramal tahap NO<sub>2</sub>. Beberapa rajah telah ditunjukkan untuk menjelaskan bagaimana nilai SHAP menafsir output model mereka. Rajah 2.4 menunjukkan contoh pertama dari set data pengujian, menerangkan bagaimana setiap atribut menyumbang terhadap ramalan purata model berbanding set data latihan.

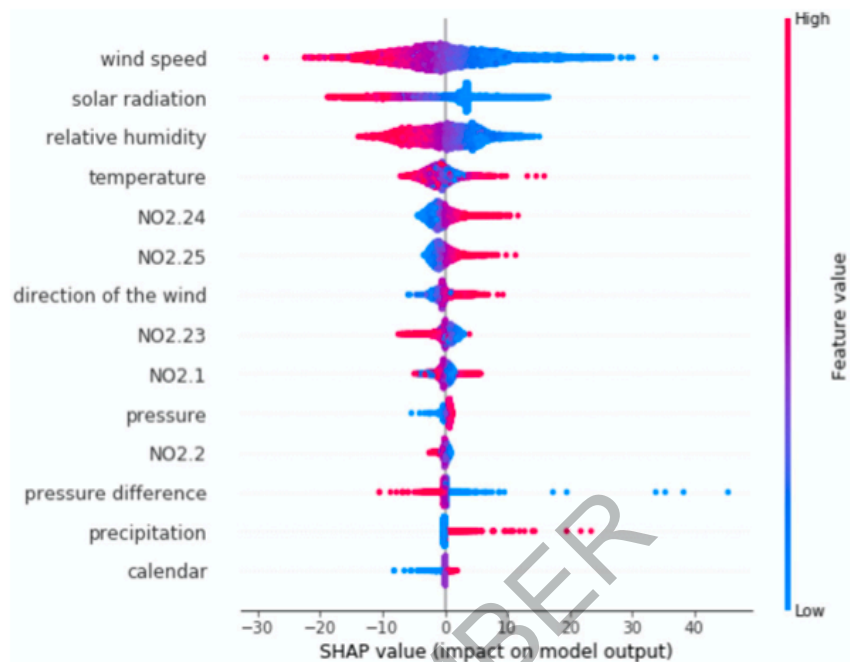


Rajah 2.4 Penjelasan ramalan pertama model pembelajaran mesin menggunakan nilai SHAP.

Sumber: Maria da Jose (2020)

Kemudian, Rajah 2.5 adalah analisis yang menunjukkan bagaimana nilai-nilai SHAP dikaitkan dengan atribut di peringkat global. Analisis ini dinamakan plot *SHAP summary*. Dalam kes ini, garis menegak menunjukkan semua atribut berterusan yang disusun mengikut kesan mereka terhadap ramalan. Menurut Tseng et al. (2020), plot *SHAP summary* digunakan untuk menggambarkan kesan positif atau kesan negative daripada atribut penting di dalam set data.

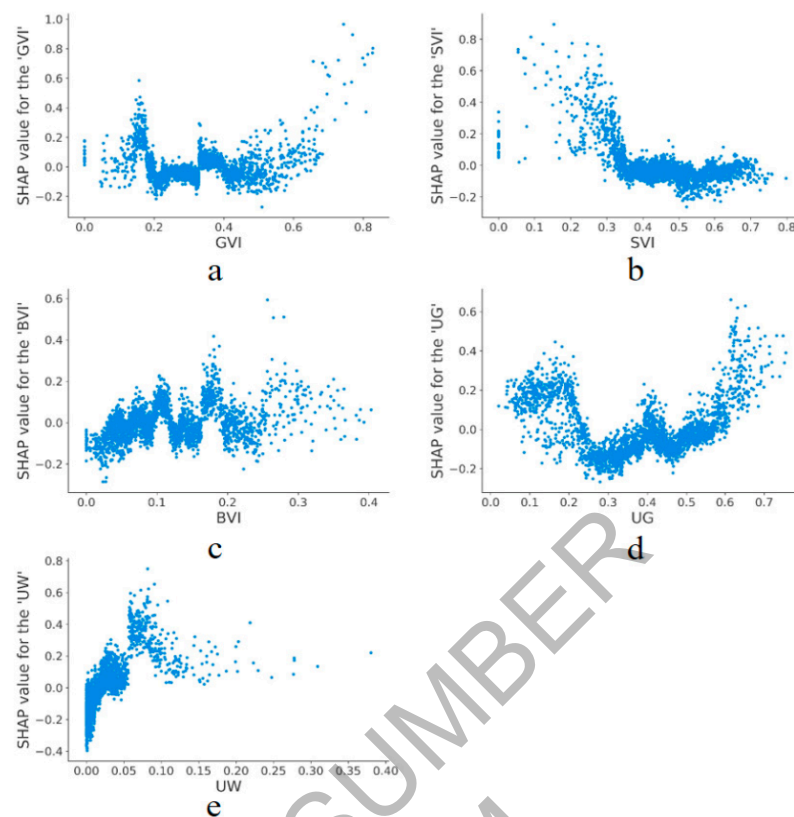




Rajah 2.5 SHAP *summary plot* untuk output pembelajaran mesin.

Sumber: Po-Yu Tseng et al. (2020)

Selain daripada itu, plot *SHAP feature dependence* juga perlu ditunjukkan dalam sesebuah kajian. Plot ini adalah plot penyebaran yang menunjukkan kesan sesuatu ciri pada ramalan yang diramal oleh algoritma pembelajaran mesin. Liujia Chen et al. 2020 telah menggunakan plot *SHAP feature dependence* terhadap lima ciri persekitaran bandar untuk menerangkan kesan ciri tersebut terhadap harga rumah. Menurut literatur tersebut, plot ini digunakan kerana *SHAP summary plot* tidak dapat menafsir sepenuhnya hubungan kompleks dan tidak linear antara ciri persekitaran bandar dan harga perumahan. Namun, Rajah 2.6 menunjukkan bagaimana plot *SHAP feature dependence* digunakan bagi mencari hubungan diantara ciri persekitaran bandar dan harga rumah dengan menunjukkan nilai yang lebih terperinci.



Rajah 2.6 Plot SHAP dependence bagi lima ciri persekitaran bandar.

Sumber: Liujia Chen et al. (2020)

## 2.5 KESIMPULAN

Di dalam bab ini telah menerangkan mengenai kajian-kajian literatur terdahulu yang menjadi panduan dan rujukan untuk kajian ini. Bab ini telah membincangkan bagaimana ozon ini terhasil daripada pencemaran udara dan faktor meteorologi. Pencemaran-pencemaran ini berlaku diseluruh negara yang majoritinya disebabkan oleh asap kenderaan dan kilang yang beroperasi setiap hari. Selain itu, peningkatan tahap ozon juga berlaku apabila berlakunya perubahan cuaca terutamanya di negara yang mempunyai 4 musim.

Hasil daripada tinjauan literatur ini, boleh disimpulkan bahawa algoritma bergabung seperti hutan rawak banyak digunakan dalam meramal tahap ozon dan ia menunjukkan prestasi ramalan yang baik. Algoritma seperti hutan rawak, *XGboost*,