IMPAK PENCEMARAN UDARA DAN METEOROLOGI TERHADAP PEMBENTUKAN OZON SEMASA PKP DI MALAYSIA MENGGUNAKAN 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

PENGAKUAN

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

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 menceriakan 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.

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

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

m/s	Meter per saat (*meter per second*)
µg/m3	Mikrogram setiap meter padu (*micrograms per cubic meter*)
ppm	Bahagian per sejuta (*parts per million*)
ppbv	Bahagian per billion (*parts per billion*)
w/m2	Watt per meter persegi (*watt per square meter*)
°C	Darjah Celsius (Degree Celcius) m/s
PKP	Perintah Kawalan Pergerakan
O3	Ozon
PM2.5	Zarah Terampai ($PM_{2.5}$)
PM10	Zarah Terampai (PM_{10})
SO2	Sulfur Dioksida
NOx	Nitrogen Oksida
NO	Nitrium Oksida
NO2	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 terdapat 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 mengunakan teknik perlombongan data sebagai kaedah ramalan pencemaran udara. Selain itu, kaedah statistik juga telah digunakan untuk mencari korelasi diantara kualiti udara dengan parameter penceramaran 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 penghasilkan 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 mengunakan 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 mengunakan analisis statistik untuk mencari faktor penting dalam pembentukan ozon dan mendapati tiga faktor pembentukan ozon adalah NO, Kelembapan, dan NO2 (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_2, NOx, NO_2, 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 *Grandient 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, NO2, 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. Mengenal pasti model terbaik ramalan ozon sebelum dan semasa PKP menggunakan teknik pembelajaran mesin.

ii. Mengenal pasti 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_3, CO, NO_2, SO_2, NO, dan NO_x 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 terperincinya, 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 seperi 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 Pendamik Covid-19, faktor penting peningkatan tahap ozon, kaedah untuk mencari faktor peting, pembandingan model-model ramalan pembelajaran mesin terhadap kajian lepas, tafsiran faktor penting menggunakan penilaian peringkat seperi 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-pemprosesan 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 kepetingan 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.

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 femonena 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 pendamik COVID-19 telah mempengaruhi kualiti uadara termasuk ozon di seluruh negara termasuk Malaysia. Bahagian ini juga menganalisa faktor yang mepengaruhi penghasilkan ozon di atas muka bumi dan menjelaskan kaedah yang digunakan untuk mencari faktor penting tahap ozon.

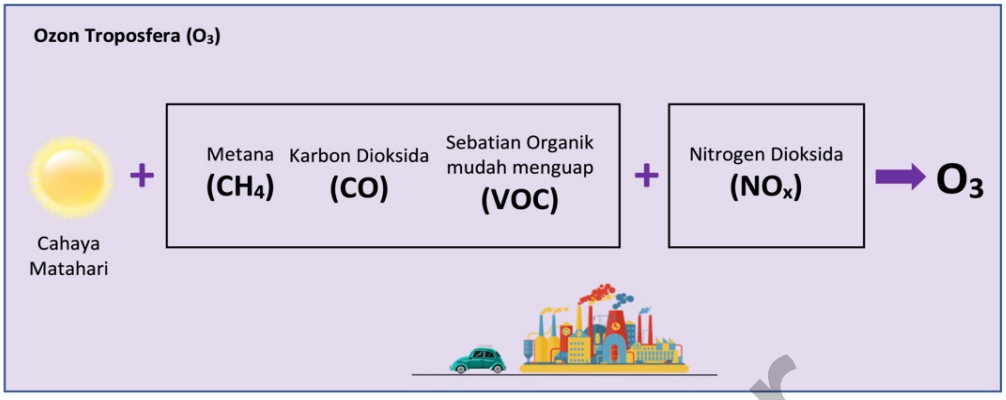
Bahagian kedua membincangkan teknik perlombongan data dengan menggunakan teknik pembelajaran mesin bergabung. Dalam bahagian ini, akan dijelaskan literatur kajian lepas yang menggunakan pembelajaran mesin untuk membina model ramalan tahap ozon. Bahagian ketiga membincangkan tentang kaedah tafsiran dalam 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 kahadiran 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 illustrasi pembentukan ozon. Berdasarkan illustrasi 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_4, 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. Berdasakan 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 pekerjanya 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 pencamaran 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 O3, PM10, PM5, SO2, 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, SO2, 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 SO2 dan NO2 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 NOx adalah punca utama yang menyebabkan naik turunnya tahap ozon.

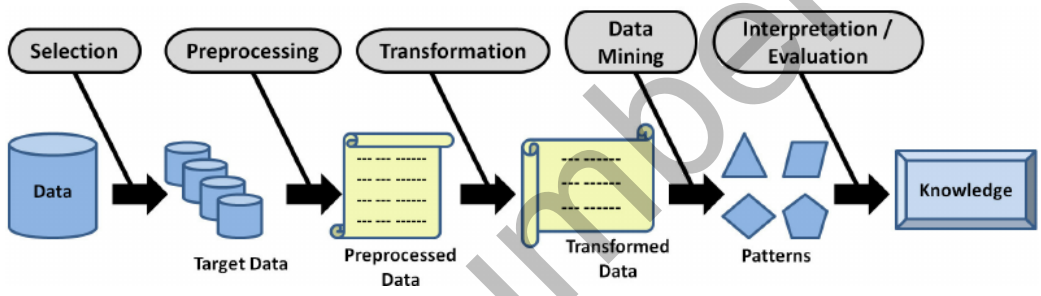
Lee et al. (2014) telah mengkaji korelasi NO2, NOx, CO dan VOC terhadap tahap ozon di Hong Kong dan China Selatan berdasarkan data daripada tahun 1990 hingga ke tahun 2010. Menurut kajian mereka, mereka mendapati bahawa penurunan NO2, NOx, CO tidak banyak mempengaruhi tahap kenaikan O3. Walau bagaimanapun, kenaikan sebatian organik yang tidak menentu (VOC) adalah penyebab utama kenaikan O3.

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 mengunakan data daripada tahun 1992 sehingga pertengahan tahun 2019. Hasil daripada kajian mereka, mereka mendapati bahawa tahap ozon masih meningkat walaupun kadar NOx 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 *boostrap* 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.

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

2.3.3 Pembandingan 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.

Jadual 2.1 Pembandingan kajian literatur terhadap ramalan ozon.

Penulis	Tajuk	Set Data	Eksperimen Algoritma	Penilaian	Rumusan
Asha B. Chelani, 2009	*Prediction of daily maximum ground ozone concentration using support vector machine*	Set Masa Multi – Pembolehubah di Bahadurshah Zahar Marg, Delhi.	Multiple Regression, SVM, MLP	MAPE, NRMSE, RMSE, dan Index of agreement.	
Nahun Loya et al. 2012	*Forecast of Air Quality Based on Ozone by Decision Trees and Neural Network*	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	*Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY*	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	*Modelling Ground-Level Ozone Concentration using Ensemble Leaning Algorithms*	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	*Predicting Ozone Layer Concentration Using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree*	Set data UCI Multi – Variate di Itali	Random Forest, Multivariate Adaptive Regression Splines and	RMSE, MSE, GCV, MAD, MRAD, SSY, SSE, R^2, R^2 Norm, GCV R-Sr	Menggunakan MARS dan Random Forest untuk mencari faktor penting peningkatan tahap ozon.
Zhiying Meng, 2019	*Ground Ozone Level Prediction Using Machine Learning*	Set data UCI (Ozone level Detection) di Bostn	Logistic Regression, Decision Tree, Random Forest, AdaBoost, dan SVM	"Principal Compenent Analysis" (PCA) dan "Logistic Regression"	Mengkelaskan hari yang ada ozon sebagai kelas "1" dan "0" tiada ozon.
Mohan, S. dan Saranya,P 2019	*A novel bagging ensemble approach for predicting summertime ground-level ozone concentration*	Set data Multi – Variate di Gummidipoondi, India.	Random Tree, REP Tree, Random Forest, dan Base Classifier	Willmott's index of agreement (IoAd) , R^2, dan PEP	Menggunakan Willmotts 's index of agreement (IoAd) untuk penilaian model.
Rui Feng et al. 2019	*Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison:A case study in hangzhou, China*	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	*A Machine Learning Approach to Investigate the Surface Ozone Behaviour*	Set data Multi-Variate di Wilayah Basilicata, Itali	Advance ML, dan Boosted Regression Tree	Kaedah Statistik, R^2, RMSE, MAE. Dan MBE	Menggunakan model BRT untuk mencari faktor penting ozon.
Weeberb J. Requia et al. 2020	*An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States*	Set data Multi – Variate di Amerika	Neural Network, Random Forest, dan Gradient Boosting	R^2, RMSE, nilai cerun,	Menuggunakan pengaturcaraan R dengan menggunakan paket H2O.
Maryam Aljanabi et al. 2020	*Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan*	Siri Masa Multi-Variate, di Jordan	MLP, SVR, DTR, dan XGboost	R^2, RMSE, dan MSE	Melakukan pemilihan yang intensif untuk mengurangkan jumlah ciri dan mengurangkan masa yang diperlukan untuk ramalan.
Yu Wang et al. 2021	*Spatiotemporal estimation of hourly 2-km ground-level ozone over China based on Himawari-8 using a self-adaptive geospatially local model*	Siri Masa Multi – Variate di China	SGLboost, Catboost, Light GBM, XGboost, RF, dan ERT	R^2, 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 umtuk 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 at. (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.

$$\mathrm{G} = \sum_{i=1}^{n_c} P_i(1 - P_i) \qquad ...(2.1)$$

Di mana n_c adalah bilangan kelas didalam sasaran pemboleh ubah dan *Pi* 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 :

$$\textbf{Importance(variable)} = \mathbf{G_{parent} - G_{split1} - G_{split2}} \qquad ...(2.2)$$

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 Explaination (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.

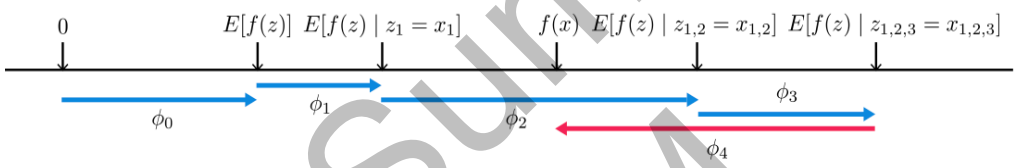
$$g(z) = \phi_0 + \sum_{i=1}^{p} \phi_i z_i. \qquad ...(2.3)$$

Dimana *p* adalah bilangan ciri, $z = [z_1, ..., z_p]^T$ adalah ringkasan bagi input *x*, di mana nilai *z* yang bersesuaian dengan ciri yang digunakan dalam ramalan data adalah 1 dam nilai *z* yang sesuai dengan ciri yang tidak digunakan adalah 0. Selain itu, Φ_0 mewakili sumbangan setiap ciri kepada model. Seterusnya, formula bagi mendapatkan nilai Φ_0 adalah seperti yang berikut.

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

Nilai yang dikeluarkan oleh formula Φ_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 permain. 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 fakor penting terhadap sesuatu atribut.



Rajah 2.3 Nilai SHAP untuk setiap ciri ramalan model.

Sumber: Beford (2020)

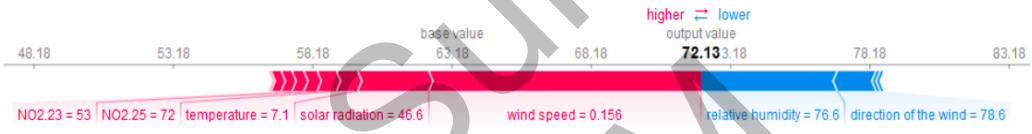
Kaedah mencari kepentingan *Additive feature* dikenali sebagai *SHAP framework*. Ia menunjukkan cara penyelesaian yang unik dalam kelas ini yang memenuhi ciri yang diperlukan (Lunberg dan Lee 2017). Dalam pengiraan nilai SHAP, Lundberg menyusulkan algortima pokok SHAP yang digunakan untuk menyediakan 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 NO2 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.

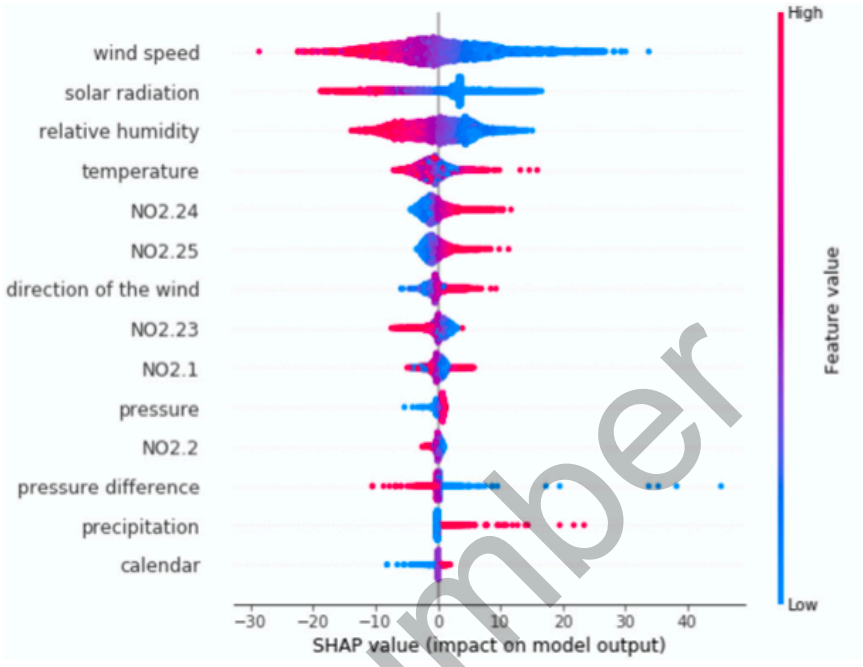
Mereka menggunakan nilai SHAP untuk menafsir output daripada model "*Deep Neural Network*" yang digunakan untuk meramal tahap NO_2. Beberapa rajah telah ditunjukan untuk menjelaskan bagaimana nilai SHAP menafsir output model mereka. Rajah 2.4 menunjukkan contoh pertama dari set data pengujian, menerangkan 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)

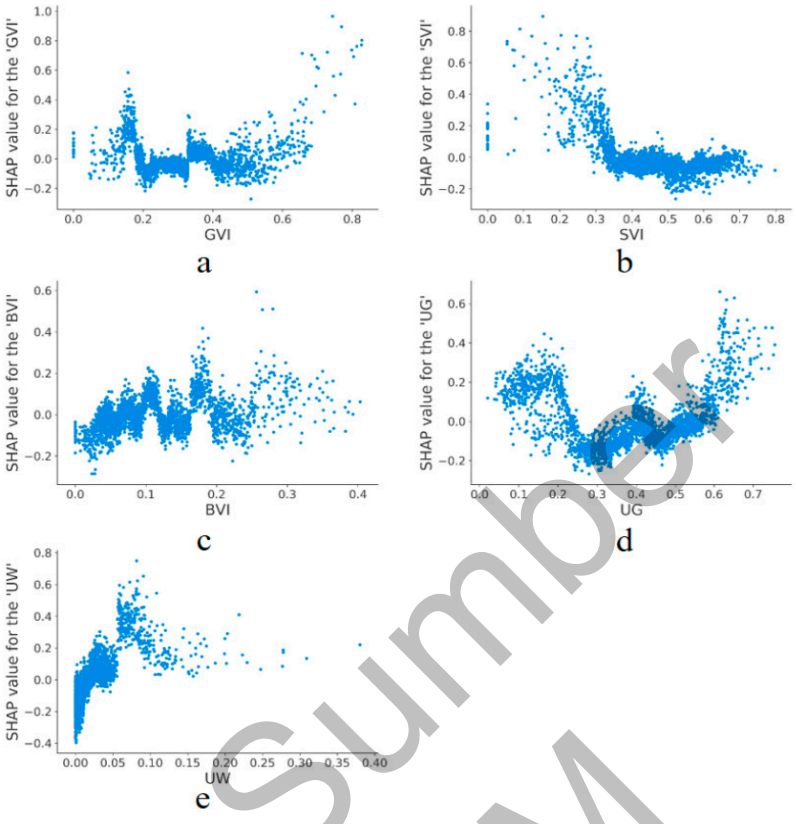
Kemudian, Rajah 2.5 adalah analisis yang menunjukkan bagaimana nilai-nilai SHAP dikaitkan dengan atribut di peringkat global. Analisis ini dinamakan plot *SHAP summary*. Dalam kes ini, garis menegak menunjukkan semua atribut berterusan yang disusun mengikut kesan mereka terhadap ramalan. Menurut Tseng 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*,

Catboost, dan *Gradient Boosting* yang bakal digunakan pada kajian ini juga antara algoritma bergabung yang telah digunakan dalam meramal tahap ozon.

Selain itu, pembelajaran mesin bukan sahaja digunakan untuk meramal, malah ia juga digunakan untuk memperolehi pengetahuan seperti faktor penting. Tafsiran menggunakan pembelajaran mesin telah digunakan oleh penyelidik untuk mencari faktor penting sesuatu ramalan. Berdasarkan literatur lepas, boleh disimpulkan bahawa terdapat dua penilaian peringkat yang sering diguna untuk tafsiran menggunakan pembelajaran mesin bergabung iaitu dengan menggunakan kepentingan Gini berdasarkan *Impurity Based* dan nilai SHAP. Kepentingan Gini merupakan penilaian yang terbina di dalam model pembelajaran mesin dan nilai SHAP merupakan teknik penafsiran baru yang dikatakan lebih baik daripada penilaian menggunakan kepentingan Gini. Oleh itu, kajian-kajian lepas ini akan menjadi rujukan dalam eksperimen yang akan dibincangkan dalam bab seterusnya.

BAB III

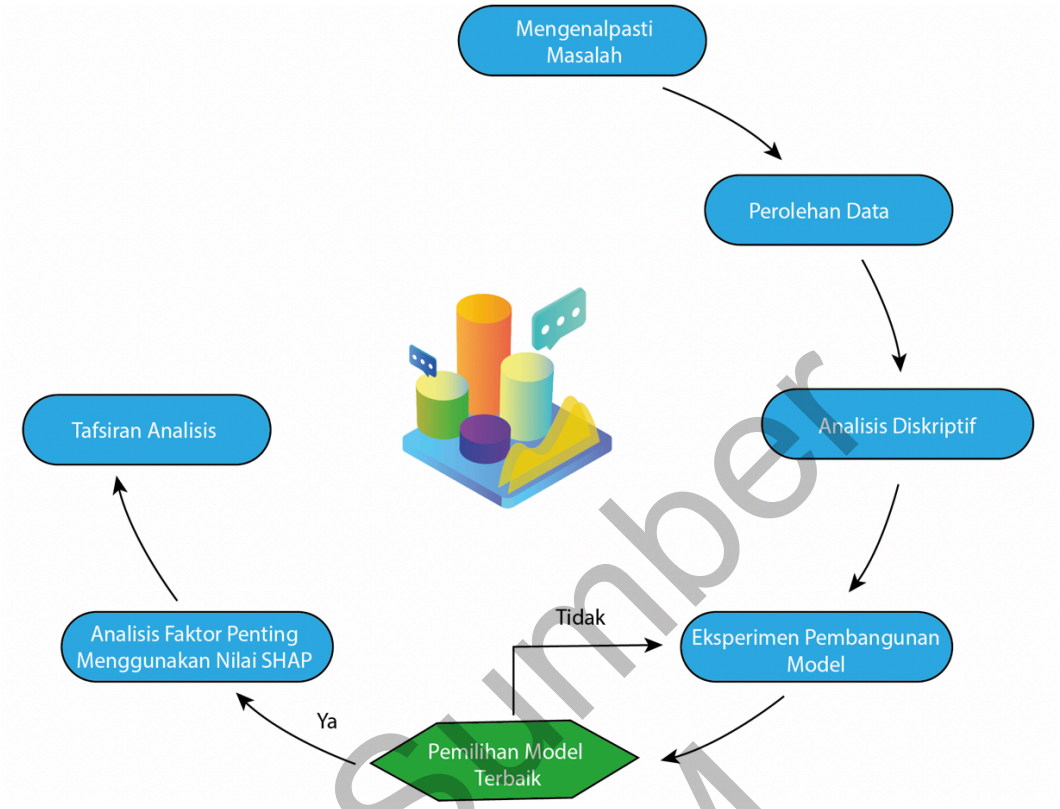
METODOLOGI

3.1 PENGENALAN

Metodologi kajian ini mengadaptasi metodologi CRISP-DM untuk mencapai 2 objektif kajian seperti yang digambarkan di dalam Rajah 3.2. Fasa pertama adalah untuk menghasilkan model terbaik bagi yang mempunyai ketepatan ramalan yang tertinggi. Manakala, fasa kedua adalah untuk menafsir faktor penting yang mempengaruhi peningkatan tahap ozon di Malaysia.

Pada fasa yang pertama terdapat langkah dan teknik digunakan untuk mengenal pasti masalah, dan mengoptimumkan penggunaan data. Langkah tersebut adalah seperti perolehan data, pemprosesan data, analisis data, penggantian nilai hilang, data integrasi, data transformasi, pemilihan atribut dan seterusnya pembangunan model serta pengujian prestasi model tersebut. Daripada Fasa pertama, model algoritma bergabung yang menunjukkan prestasi yang terbaik dipilih untuk meramal tahap ozon.

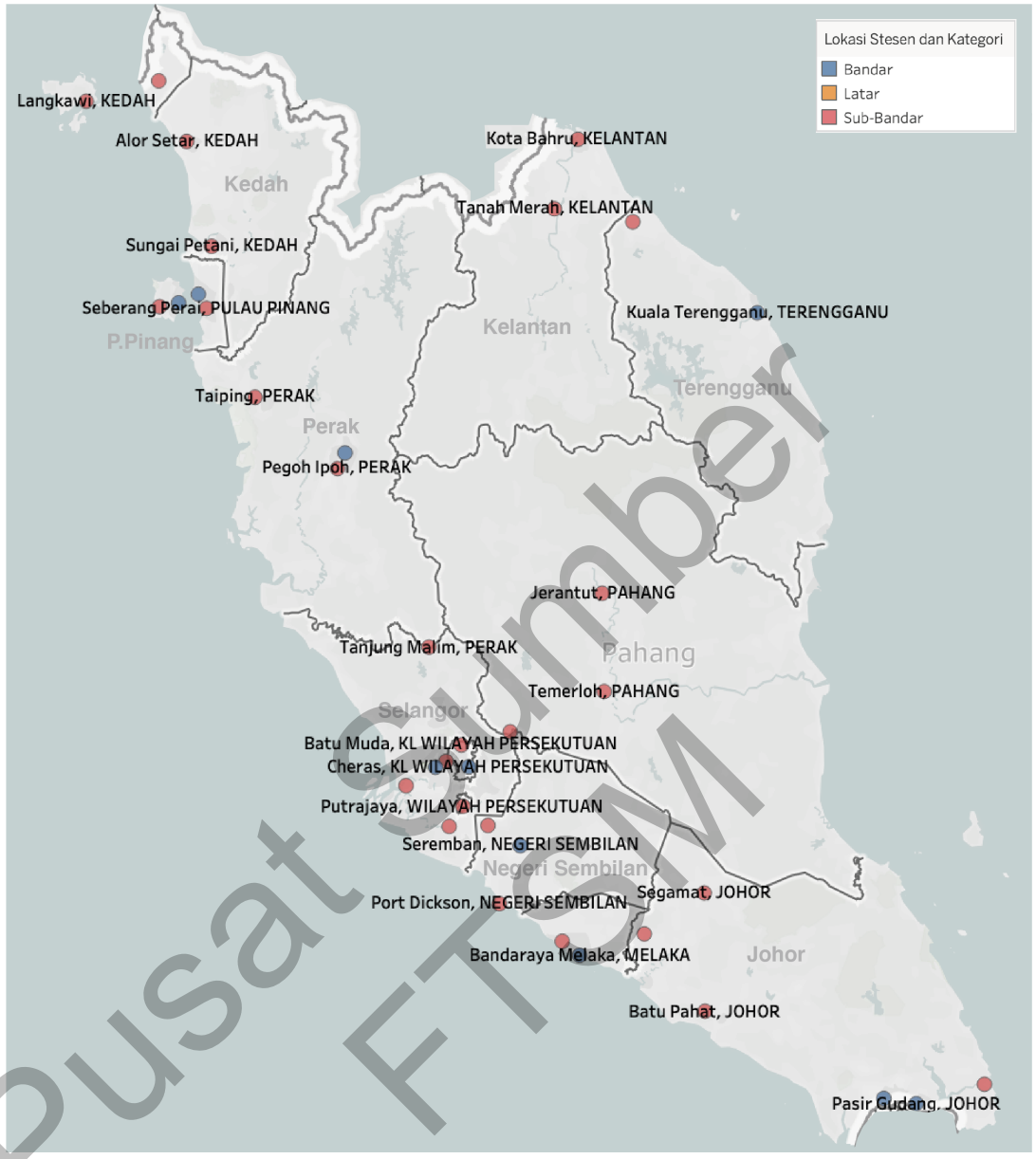
Model ramalan untuk meramal tahap ozon yang bakal dibangunkan menggunakan 4 kaedah pembelajaran mesin seperti hutan rawak, *XGboost*, *GradientBoosting*, dan *Catboost*. Kemudian, prestasi model tersebut akan dibandingkan. Pada Fasa kedua, model yang terbaik akan digunakan untuk mencari faktor penting tahap ozon dengan menggunakan kepentingan Gini dan nilai SHAP. Rajah 3.2 dibawah menunjukkan teknik CRISP-DM.



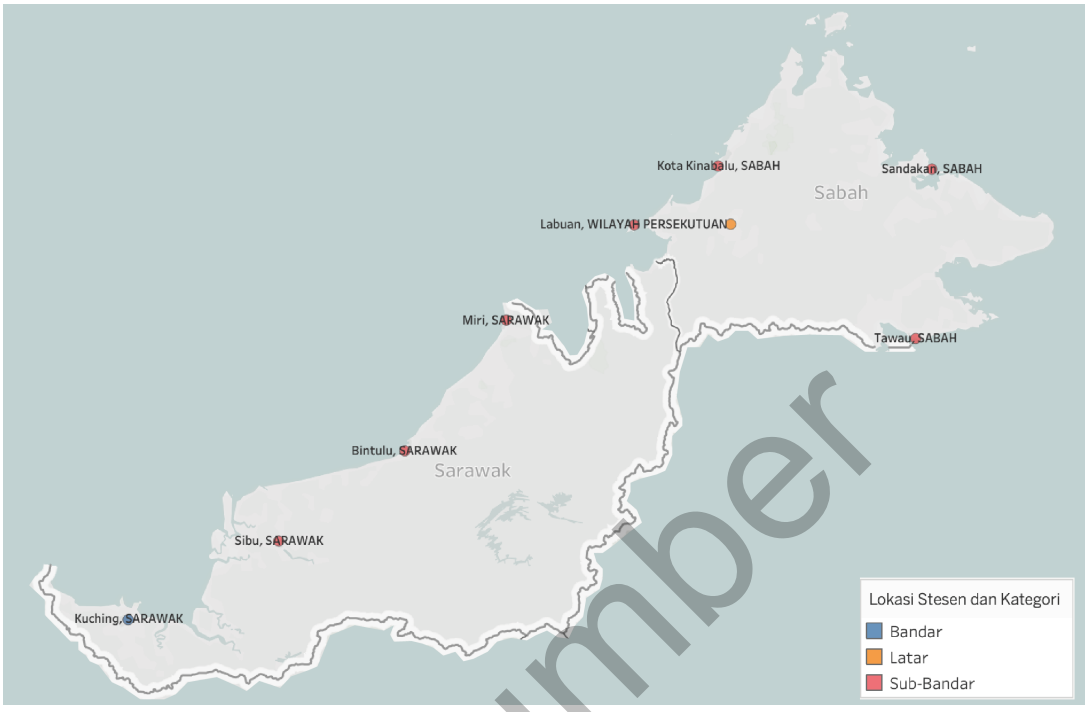
Rajah 3.1 Metodologi CRISP-DM kajian.

3.2 PEROLEHAN DATA

Data yang digunakan untuk kajian ini adalah data yang diperoleh oleh Jabatan Sains Bumi dan Alam Sekitar, Fakulti Sains dan Teknologi, UKM yang mengandungi 66 stesen udara yang ditempatkan di kawasan bandar, luar bandar, latar, perindustrian, dan pendalaman. Namun, oleh kerana kawasan perindustrian dan pendalaman mengandungi purata data hilang yang tinggi, set data yang digunakan bagi kajian ini hanya 46 stesen udara yang ditempatkan di kawasan bandar, pinggir bandar dan latar di 13 negeri di seluruh Malaysia seperti yang ditunjukan pada Rajah 3.2 dan Rajah 3.3.



Rajah 3.2 Lokasi Stesen Udara di Semenanjung Malaysia.



Rajah 3.3 Lokasi Stesen Udara di Sabah dan Sarawak.

Senarai penuh kesemua 46 stesen udara yang ditempatkan di suluruh Semenanjung Malayasia, Sabah dan Sarawak adalah seperti di **Lampiran A**. Set data ini mengandungi rekod data sebanyak 1,236,480 bagi tempoh masa yang bermula dari 12 Februari 2020 sehingga 21 April 2020 dan terdapat 16 atribut berkenaan maklumat cuaca yang di kumpul seperti ID Stesen, Tarikh, parameter pencemaran udara, dan parameter meteorologi.

Jadual 3.1 di bawah menunjukkan senarai atribut yang diperoleh daripada Jabatan Meteorologi Malaysia. Lajur pertama mewakili parameter pencemaran udara dan meteorologi. Lajur kedua mewakili diskripsi atribut data mentah dan lajur seterusnya menunjukkan unit ukuran untuk setiap atribut. Lajur terakhir mewakili bentuk data tersebut.

Jadual 3.1 Set data mentah kajian.

Bil.	Atribut	Diskripsi	Unit Ukuran	Bentuk Data
1	Temp	Suhu	^oC	Numerik
2	Humid	Kelembapan	%	Numerik
3	Radiation	Sinaran Radiasi	W/m^2	Numerik
4	WSp	Kelajuan Angin	m/s	Numerik
5	WDr	Arah Angin	^{o}	Numerik
6	PM_{10}	Zarah Terampai (PM_{10})	µg/m3	Numerik
7	$PM_{2.5}$	Zarah Terampai ($PM_{2.5}$)	µg/m3	Numerik
8	O_3	Ozon	ppm	Numerik
9	CO	Karbon Dioksida	ppm	Numerik
10	NO_2	Nitrogen Dioksida	ppm	Numerik
11	SO_2	Sulfur Dioksida	ppm	Numerik
12	NO	Nitrium Oksida	ppm	Numerik
13	NO_x	Nitrogen Oksida	ppm	Numerik
14	Station ID	Nombor Rujukan Stesen		Nominal
15	Location	Lokasi Stesen		Nominal
16	Date Time	Masa dan Tarikh		Nominal

Daripada 46 stesen udara, terdapat 1,236,480 rekod data yang diperolehi, purata sebanyak 2.67% adalah data yang hilang (*missing values*). Oleh itu, proses pre-pemprosesan adalah sangat penting untuk menghasilkan set data yang lengkap dan berkualiti.

3.3 PRA-PEMPROSESAN DATA

Setelah data diterima, set data mentah melalui fasa pemprosesan data. Fasa pemprosesan data merupakan fasa yang penting dalam pembelajaran mesin kerana ia mengubah set data mentah kepada set data yang lebih difahami dan berguna. Terdapat beberapa fasa dalam pemprosesan data dan antaranya adalah pembersihan data, data integrasi, data transformasi, pengurangan data, dan penafsiran data. Dalam proses penyediaan data kajian ini, perisian Microsoft Excel 2016, dan pengaturcaraan Python digunakan.

3.3.1 Pembersihan Data

Pembersihan data melibatkan pengesanan data yang tidak konsisten dalam kumpulan data kerana kemasukan yang tidak tepat. Data yang tidak lengkap, tidak relevan atau tidak tepat akan diubah suai atau dihapuskan supaya data yang akan dibersihkan mempunyai kualiti data yang baik.

Untuk kajian ini, rekod data yang tidak lengkap bermaksud rekod data yang tidak mempunyai nilai bagi parameter pencemaran udara dan meteorologi paling tinggi yang melebihi 4 atau 5 jam di setiap stesen udara. Bagi menyelesaikan masalah ini, stesen udara yang mengandungi nilai kehilangan yang tinggi tidak akan digunakan untuk proses seterusnya. Antara stesen udara yang tidak digunakan untuk penggantian nilai hilang adalah seperti stesen udara Nilai, stesen udara Seremban, stesen udara Port Dickson, dan stesen udara Labuan.

3.3.2 Penggantian Nilai Hilang

Set data mentah yang mempunyai nilai hilang, maklumat yang tidak relevan seperti data daripada stesen lain tidak akan digunakan. Data dalam ukuran unit yang berbeza akan digantikan mengikut format yang sesuai. Semua masalah ini dapat diatasi dengan kod yang berbeza yang dilaksanakan dalam pengaturcaraan Python. Untuk penggantian nilai yang hilang, kaedah *KNN-Imputation* daripada *Sckit-learn* akan digunakan. Pada peringkat pertama, semua data hilang perlu digantikan dengan nilai (*NaN*) di pengaturcaraan Python. Pada peringkat kedua, nilai hilang yang telah digantikan dengan nilai (*NaN*) akan digantikan dengan nilai k=5.

3.3.3 Data Integrasi dan Transformasi

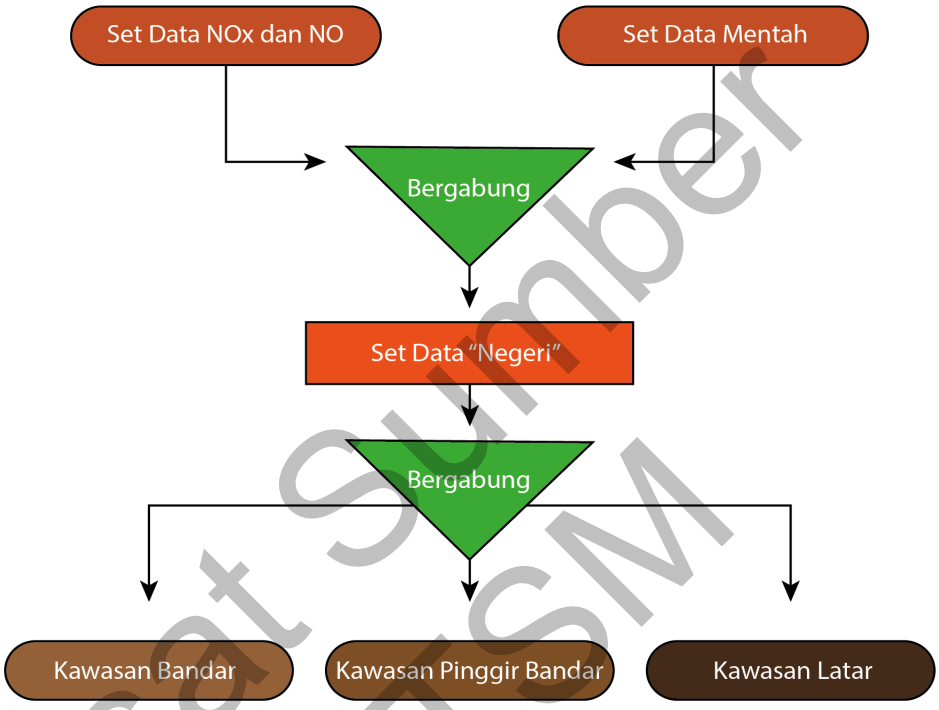
Data integrasi secara umumnya merujuk kepada syarat untuk menggabungkan data daripada beberapa buah set data kepada satu gabungan set data baru yang mengandungi kesemua data tersebut. Sebagai contoh, data parameter NO dan NOx berada di fail Excel yang berbeza. Selain itu, data setiap stesen udara juga berada di fail Excel yang berbeza dan diasingkan mengikut negeri. Setelah itu, kesemua data ini digabungkan menjadi tiga set data yang baru mengikut kategori kawasan. Kawasan

tersebut adalah seperti kawasan bandar, kawasan pinggir bandar, dan kawasan latar seperti yang ditunjukan di Jadual 3.2.

Jadual 3.2 Lokasi stesen udara yang telah dikategorikan di setiap kawasan.

Kawasan	Stesen Udara	Tempat
Bandar	Seberang Jaya, PULAU PINANG	Sek.Keb. Seberang Jaya II
	Minden, PULAU PINANG	Universiti Sains Malaysia
	Tasek Ipoh, PERAK	Sek. Men. Keb. Jalan Tasek
	Cheras, KL WILAYAH PERSEKUTUAN	Sek. Men. Keb. Seri Permaisuri
	Shah Alam, SELANGOR	Sek. Keb. TTDI Jaya
	Bandaraya Melaka, MELAKA	Sekolah Tinggi Melaka
	Larkin, JOHOR	Institut Perguruan Temenggong Ibrahim
	Pasir Gudang, JOHOR	Sek. Men. Pasir Gudang 2
	Kuala Terengganu, TERENGGANU	Sek. Keb. Chabang Tiga
	Kuching, SARAWAK	Depot Ubat Kementerian Kesihatan Malaysia
Sub-Bandar	Kangar, PERLIS	Institut Latihan Perindustrian (Kangar)
	Langkawi, KEDAH	Kompleks Sukan Langkawi
	Alor Setar, KEDAH	Sek. Men. Agama Kedah
	Sungai Petani, KEDAH	Sek.Men. Keb. Tunku Ismail
	Seberang Perai, PULAU PINANG	Kolej Vokasional Seberang Perai
	Balik Pulau, PULAU PINANG	Kolej Vakasional Balik Pulau
	Taiping, PERAK	Sek. Keb. Ayer Puteh
	Pegoh Ipoh, PERAK	Sek. Keb. Jalan Pegoh
	Tanjung Malim, PERAK	Universiti Perguruan Sultan Idris
	Batu Muda, KL WILAYAH PERSEKUTUAN	Sek. Keb. Batu Muda
	Putrajaya, WILAYAH PERSEKUTUAN	Sek. Keb. Presint 18
	Petaling Jaya, SELANGOR	Sek. Keb. Bandar Utama
	Klang, SELANGOR	Klinik Kesihatan Pandamaran
	Banting, SELANGOR	Kolej Mara Banting
	Bukit Rambai, MELAKA	Sek. Keb Tanjung Minyak 2
	Segamat, JOHOR	Klinik Kesihatan Bandar IOI
	Batu Pahat, JOHOR	MRSM Batu Pahat
	Kota Tinggi, JOHOR	Sek. Men. Agama Bandar Penawar
	Tangkak, JOHOR	Tapak Semaian Majlis Daerah Tangkak
	Temerloh, PAHANG	Pejabat Meteorologi Temerloh
	Jerantut, PAHANG	SMK. Jerantut
	Indera Mahkota, Kuantan, PAHANG	Sek. Keb. Indera Mahkota
	Besut, TERENGGANU	Sek. Keb. Nyiur Tujuh
	Tanah Merah, KELANTAN	Sek. Men. Tanah Merah
	Kota Bahru, KELANTAN	Sek. Men. Keb. Tanjong Chat
	Tawau, SABAH	Pejabat JKR Tawau
	Sandakan, SABAH	Pejabat JKR Sandakan
	Kota Kinabalu, SABAH	Sek. Men. Keb. Tansau, Kota Kinabalu
	Miri, SARAWAK	Sek. Men. Dato Permaisuri
	Bintulu, SARAWAK	Balai Polis Pusat Bintulu
	Sibu, SARAWAK	Ibu Pejabat Polis Sibu
Latar	Keningau, SABAH	Sek. Men. Keb. Gunsanad

Rajah 3.4 menunjukkan ilustrasi proses data integrasi untuk kajian ini. Proses yang pertama adalah menggabungkan parameter NO dan NO_x kepada set data mentah dan dilabelkan sebagai set data "negeri", "negeri" merujuk kepada negeri set data tersebut. Seterusnya menggabungkan set data tersebut kepada kawasan bandar, pinggir bandar, dan latar.



Rajah 3.4 Illustrasi proses penggabungan set data dan gabungan set data "negeri" kepada kawasan bandar, pinggir bandar, dan latar .

Tranformasi data adalah proses menukar format, struktur, atau nilai sesebuah data. Antara faktor mengapa perlunya tranformasi data adalah supaya data yang di ubah lebih mudah difahami dan dapat memberi maklumat kepada sesebuah organisasi. Nilai indeks pencemaran udara bagi ozon yang digunakan oleh Malaysia adalah seperti yang dilampirkan pada Rajah 3.3 (Ibrahim 2000). Namun, oleh kerana purata nilai indeks pencemaran ozon bagi set data berada di dalam kategori yang baik, maka kelas bagi atribut ozon (O_3) lebih sesuai dikategorikan kepada empat kelas yang menggunakan kaedah persentil. Rajah 3.4 menunjukkan julat tahap ozon yang digunakan bagi kajian ini.

Jadual 3.3 Nilai APU bagi ozon yang digunakan di Malaysia.

Atribut	Nilai APU	Kelas
O_3	< 50	Baik
	≥ 51 dan < 100	Sederhana
	≥ 101 dan < 200	Tidak Sihat
	≥ 201 dan < 300	Sangat Tidak Sihat
	≥ 201	Berbahaya

Jadual 3.4 Julat tahap ozon yang digunakan untuk mengkategorikan nilai ozon.

Atribut	Nilai O_3 (ppm)	Kelas
O_3	< 0.0325	0
	≥ 0.0325 dan < 0.065	1
	≥ 0.065 dan < 0.0975	2
	≥ 0.0975	3

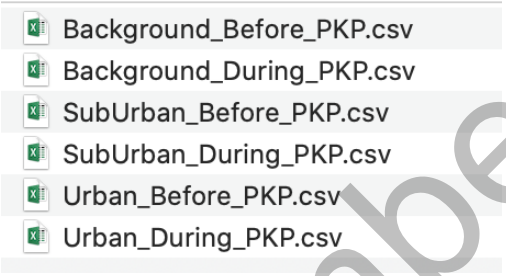
Setelah itu, atribut arah angin juga ditukar kepada kompas. Kemudian arah kompas tersebut dikategorikan dengan nilai digit. Jadual 3.5 menunjukkan set atribut arah angin yang telah ditukar kepada nilai digit.

Jadual 3.5 Julat arah angin yang digunakan untuk mengkategorikan arah kompas.

Atribut	Arah angin (°)	Arah Kompas	Nilai
Arah Angin (WDr)	≥ 335.6 dan < 22.5	Utara	1
	≥ 22.5 dan < 67.5	Timur Laut	2
	≥ 67.5 dan < 112.5	Timur	3
	≥ 112.5 dan < 157.5	Tenggara	4
	≥ 157.5 dan < 202.5	Selatan	5
	≥ 202.5 dan < 247.5	Barat Daya	6
	≥ 247.5 dan < 295.5	Barat	7
	≥ 295.5 dan < 335.5	Barat Laut	8

Seterusnya, sebelum set data melalui proses seterusnya di pengaturcaraan Python. Semua set data haruslah ditukarkan format daripada ".xlsx" kepada ".csv". Bagi mencapai objektif kajian iaitu untuk mencari faktor tahap penting ozon sebelum dan semasa PKP, set data kawasan akan dipecahkan mengikut dua kategori iaitu "sebelum PKP" dan "semasa PKP". Set data "sebelum PKP" mengandungi data

daripada tarikh 12 Feb 2020 sehingga 17 Mar 2020, manakala set data "semasa PKP" mengandungi data daripada tarikh 18 Mar 2020 sehingga 21 April 2020. Rajah 3.5 menunjukkan set data Excel yang telah di ubah format kepada ".csv" dan di bahagikan kepada dua fail yang berbeza dan data yang telah di dikategorikan kepada kategori "sebelum PKP" dan "semasa PKP".



Rajah 3.5 Set data yang telah dikategorikan kepada set data "sebelum PKP" dan "semasa PKP".

3.3.4 Pemilihan Atribut

Sebelum set data diteruskan dengan proses latihan, pemilihan atribut dilakukan untuk memilih atribut yang berguna untuk kajian ini. Oleh kerana objektif kajian ini adalah untuk mengetahui faktor penting peningkatan tahap ozon, atribut seperti tarikh, ID stesen, lokasi akan dikeluarkan. Pengeluaran atribut yang tidak digunakan akan dilakukan di pengaturcaraan Python dengan menggunakan *df.drop*. Rajah 3.6 menunjukkan atribut yang dipilih untuk model ramalan.

PM10	PM2.5	SO2	NO2	CO	WSp	WDr	Humid	Solar	Temp	NO	NOX
15.614	8.521	0.5030	7.142	0.8150	0.811	8	89.983	0.000	26.662	0.00100	0.00820
22.863	15.296	0.5140	9.190	0.9400	0.799	7	91.054	0.000	26.067	0.00300	0.01220
21.946	14.568	0.4960	8.682	0.8040	0.308	8	92.467	0.000	25.697	0.00270	0.01140
18.003	10.199	0.0540	7.130	0.7050	0.596	7	91.667	0.000	25.878	0.00130	0.00850
18.034	9.567	0.1050	4.707	0.6010	0.995	7	92.326	0.000	25.588	0.00060	0.00530

Rajah 3.6 Contoh atribut dan data yang digunakan untuk pembangunan model ramalan.

3.4 ANALISIS STATISTIK

Analisis statistik digunakan dalam kajian ini bagi menganalisis data kualiti udara. Ia membantu dalam menggambarkan, menunjukkan atau meringkaskan data dengan cara yang bermakna. Melalui analisis ini, ukuran kecenderungan utama termasuk min, max, dan purata dapat diketahui. Selain itu, perbezaan purata, sisihan piawai, dan t-test juga dinilai. Pengaturcaraan Python dan excel digunakan dengan bagi mendapatkan bacaan statistik. Jadual 3.6 menunjukkan analisis statistik bagi kawasan bandar dan jadual bagi kawasan pinggir bandar dan latar akan dilampirkan di bab 4.

	Sebelum PKP				**Semasa PKP**				**Perbezaan**		
Atribut	**Min**	**Max**	**SP**	**Purata**	**Min**	**Max**	**SP**	**Purata**	**ΔPurata**	**ΔSP**	**t-test**
PM_{10}	1.419	149.8	10.74	23.28	0.707	226.2	9.526	19.67	-3.610	1.217	0.000
$PM_{2.5}$	0.065	116.1	9.04	14.36	0.079	0.079	9.030	15.99	1.621	0.010	0.000
O_3	0.018	130.7	15.72	20.04	0.110	98.50	15.49	19.71	-0.322	0.227	0.181
SO_2	0	6.352	0.679	0.641	0.001	118.0	6.244	2.794	2.153	-5.56	0.000
NO_2	0	44.62	6.861	5.611	0.022	31.75	3.434	3.915	-1.696	3.427	0.000
CO	0.079	2.567	0.243	0.552	0.106	1.353	0.171	0.532	-0.020	0.071	0.000
WSp	0	6.490	0.802	1.300	0	4.51	0.722	1.100	-0.200	0.080	0.000
WDr	1	8	2.641	3.55	1	8	2.562	4.197	0.644	0.079	0.000
Humid	33.19	99	13.67	76.64	0.08	99	23.46	71.79	-4.847	-9.78	0.000
Radiation	0	1373	266.7	181.5	0	1000	261.3	177.2	-4.250	5.466	0.296
Temp	21.21	37.75	3.049	28.40	20.44	36.93	3.107	28.54	0.142	-0.05	0.002
NO	0	0.131	0.007	0.004	0	0.047	0.005	0.002	-0.002	0.002	0.000
NO_x	0	0.000	0.011	0.012	0.000	0.060	0.006	0.006	-0.005	0.004	0.000

Jadual 3.6 Analisis statistik kawasan bandar sebelum dan semasa PKP.

3.4.1 Analisis Korelasi Data Sebelum PKP dan Semasa PKP

Dalam kajian ini, analisis korelasi yang digunakan adalah *t-test* (*two-tailed distribution*) untuk mencari hubungan diantara parameter pencemaran udara dan parameter meteorologi sebelum PKP dan semasa PKP. Untuk melakukan hubungan korelasi ini, kajian menggunakan formula *=T.TEST()* di Microsoft Excel bagi mendapatkan bacaan korelasi.

3.5 PEMBANGUNAN MODEL

Set data yang telah melalui proses pra-pemprosesan data akan digunakan untuk membina model ramalan tahap ozon. Empat algoritma yang berbeza akan dibandingkan dan prestasi algoritma akan dinilai. Empat algoritma tersebut adalah Hutan Rawak, *XGboost*, *Catboost*, dan *Gradient Boosting*. Pembangunan model-model ini akan dibangunkan dengan menggunakan pengaturcaraan Python.

Set data akan dibahagikan ke dalam set latihan dan ujian menggunakan teknik pengesahan silang (*cross validation*) dan 90/10, 80/20, 70/10 sehingga 10/90 sebagai pekadaran pembahgian ujian/latihan. Teknik pengesah silang digunakan untuk mendapatkan pengujian yang lebih baik. Menurut Browne (2000) kaedah ini membahagikan data menjadi k-bahagian yang berbeza atau dikenali sebagai k-lipatan. Set data ini akan dilatih menggunakan 10-lipatan pengesahan silang dan digunakan pada set pengesahan dan merekodkan prestasi ramalan. Proses ini diulang sebanyak 10 kali sehingga setiap bahagian telah digunakan sebagai set pengesahan sekali. Kaedah ini dilakukan dengan menggunakan *cross_val_score* daripada *sklearn library*.

3.5.1 Pembangunan Model

Langkah pertama dalam membangunkan model di persekitaran *jupyter notebook* adalah memuat naik set data yang telah melalui fasa pemprosesan data yang telah di simpan dalam format *CSV* ke dalam kerangka data dan mengimport bebarapa *libraries* yang berkaitan ke dalam *Python*. Set data yang digunakan adalah set data di kawasan bandar, kawasan pinggir bandar, dan kawasan latar. Berikut adalah eksperimen yang dijalankan bagi keempat-empat model.

a. Hutan Rawak

Model pertama yang akan diuji adalah algoritma Hutan Rawak. Algoritma hutan rawak akan diimport daripada *sklearn ensemble library*. Algoritma ini akan menggunakan penalaan parameter seperti *max_depth = 6*, *random_state = 42*, *n_estimator = 1000*. *Max_depth* adalah bilangan tahap maksimum dalam setiap keputusan, *random_state* memastikan bahawa bahagian yang dihasilkan dapat

dihasilkan semula dan *n_estimator* adalah bilangan pokok dalam pengelasan hutan rawak Jadual 3.7 menunjukkan langkah-langkah uji kaji bagi algorima hutan rawak.

Jadual 3.7 Langkah-langkah uji kaji algoritma Hutan Rawak.

Algoritma: Hutan Rawak

Input: UrbanCombined.csv, SubUrbanCombined.csv, Background.csv, dibahagikan kepada data Latihan dan data Ujian. Contoh 70% data Latihan dan 30% data Ujian.
Output: Keputusan penilaian

Kaedah:

(1) Import libraries

```
From sklearn.ensemble import RandomForestClassifier
```

(2) Muatnaik data

```
Import pandas as pd
df = pd.read_csv("/UrbanCombined.csv")
```

(3) Pecahan data latihan dan ujian

```
From sklearn.model_selection import train_test_split
X_training1, X_testing1, y_training1, y_testing1 = train_test_split (X, y,
test_size = 0.3)
```

(4) Melatih Model

```
Model = RandomForestClassifier(max_depth =6, random_state =42,
n_estimators =1000)
Model.fit(X_training1, y_training1)
Predictions = model.predict(X_testing1)
```

(5) Penilaian Model

```
Import cross_val_score, mean_squared_error
cv_results = cross_val_score ( Model, X=X, y=y, cv=10, scoring='accuracy')
mse = cross_val_score( Model, X=X, y=y, cv=10, scoring='neg_mae')
rmse = nq.sqrt(mse)
```

b. XGboost

Algoritma kedua adalah algoritma *XGboost*. Untuk menguji algoritma ini, pengelasan *XGBoost* akan diimport daripada *XGboost library*. Algoritma ini akan menggunakan penalaan parameter sama seperti algoritma hutan rawak seperti *max_depth = 6*, *n_estimator = 1000*, dan *learning_rate = 0.1*. *learning_rate* adalah untuk memperlahankan pembelajaran pembelejaran mesin. Jadual 3.8 menunjukkan langkah-langkah uji kaji bagi algoritma *XGboost*.

Jadual 3.8 Langkah-lankah uji kaji algoritma *XGboost.*

Algoritma: XGboost

Input: UrbanCombined.csv, SubUrbanCombined.csv, Background.csv, dibahagikan kepada data Latihan dan data Ujian. Contoh 70% data Latihan dan 30% data Ujian.

Output: Keputusan penilaian

Kaedah:

(1) Import libraries

```
From XGboost import XGBClassifier
```

(2) Muatnaik data

```
Import pandas as pd
df = pd.read_csv("/UrbanCombined.csv")
```

(3) Pecahan data latihan dan ujian

```
From sklearn.model_selection import train_test_split
X_training1, X_testing1, y_training1, y_testing1 = train_test_split (X, y,
test_size = 0.3)
```

(4) Melatih Model

```
Model =XGBClassifier(max_depth =6, n_estimators =1000, learning_rete=0.1)
Model.fit(X_training1, y_training1)
Predictions = model.predict(X_testing1)
```

(5) Penilaian Model

```
Import cross_val_score, mean_squared_error
```

bersambung…

…sambungan

```
cv_results = cross_val_score ( Model, X=X, y=y, cv=10, scoring='accuracy')
mse = cross_val_score( Model, X=X, y=y, cv=10, scoring='neg_mae')
rmse = nq.sqrt(mse)
```

c. Catboost

Algoritma ketiga adalah algoritma *Catboost*. Untuk menguji algoritma ini, pengelasan *Catboost* akan di import daripada *Catboost library*. Algoritma ini akan menggunakan penalaan parameter sama seperti algoritma *XGboost* seperti *max_depth = 6*, *n_estimator = 1000*, dan *learning_rate = 0.1*. *learning_rate* adalah untuk memperlahankan pembelajaran mesin. Jadual 3.9 menunjukkan langkah-langkah uji kaji bagi algoritma *Catboost*.

Jadual 3.9 Langkah-langkah uji kaji algoritma *Catboost.*

Algoritma: Catboost

Input: UrbanCombined.csv, SubUrbanCombined.csv, Background.csv, dibahagikan kepada data Latihan dan data Ujian. Contoh 70% data Latihan dan 30% data Ujian.

Output: Keputusan penilaian

Kaedah:

(1) Import libraries

```
From Catboost import CatboostClassifier
```

(2) Muatnaik data

```
Import pandas as pd
df = pd.read_csv("/UrbanCombined.csv")
```

(3) Pecahan data latihan dan ujian

```
From sklearn.model_selection import train_test_split
X_training1, X_testing1, y_training1, y_testing1 = train_test_split (X, y,
test_size = 0.3)
```

bersambung…

…sambungan

(4) Melatih Model

```
Model =CatboostClassifier(max_depth =6, n_estimators =1000,
learning_rete=0.1)
Model.fit(X_training1, y_training1)
Predictions = model.predict(X_testing1)
```

(5) Penilaian Model

```
Import cross_val_score, mean_squared_error
cv_results = cross_val_score ( Model, X=X, y=y, cv=5, scoring='accuracy')
mse = cross_val_score( Model, X=X, y=y, cv=5, scoring='neg_mae')
rmse = nq.sqrt(mse)
```

d. Gradient Boosting

Algoritma keempat adalah algoritma *Gradient Boosting*. Untuk menguji algoritma ini, pengelasan *GradientBoosting* akan diimport daripada *sklearn.ensemble library*. Algoritma ini akan menggunakan penalaan parameter sama seperti algoritma RF seperti *max_depth = 6*, *n_estimator = 1000*, dan *random_state = 42*. Jadual 3.10 menunjukkan langkah-langkah uji kaji bagi algoritma *Gradient Boosting*.

Jadual 3.10 Langkah-lankah uji kaji algoritma *GradientBoosting.*

Algoritma: GradientBoosting

Input: UrbanCombined.csv, SubUrbanCombined.csv, Background.csv, dibahagikan kepada data Latihan dan data Ujian. Contoh 70% data Latihan dan 30% data Ujian.

Output: Keputusan penilaian

Kaedah:

(1) Import libraries

```
From sklearn.ensemble import GradientBoostingClassifier
```

(2) Muatnaik data

```
Import pandas as pd
df = pd.read_csv("/UrbanCombined.csv")
```

bersambung…

…sambungan

(3) Pecahan data latihan dan ujian

```
From sklearn.model_selection import train_test_split
X_training1, X_testing1, y_training1, y_testing1 = train_test_split (X, y,
test_size = 0.3)
```

(4) Melatih Model

```
Model =GradientBoostingClassifier(max_depth =6, n_estimators =1000,
random_state=42)
Model.fit(X_training1, y_training1)
Predictions = model.predict(X_testing1)
```

(5) Penilaian Model

```
Import cross_val_score, mean_squared_error
cv_results = cross_val_score ( Model, X=X, y=y, cv=10, scoring='accuracy')
mse = cross_val_score( Model, X=X, y=y, cv=10, scoring='neg_mae')
rmse = nq.sqrt(mse)
```


3.6 PENILAIAN MODEL

Penilaian model ramalan dilakukan bagi menilai tahap ramalan sesuatu model yang dibangunkan untuk mendapatkan model yang terbaik. Dalam kajian ini, ketepatan model dilakukan dengan menambah parameter *scoring='accuracy'* dan *'neg_mean_squared_error'* pada *cross_val_score*.

3.6.1 Mean Square Error dan Root Mean Square Error

Root mean square error adalah nilai punca kuasa dua kepada MSE diantara nilai sebenar dan nilai ramalan. Manakala *square error* adalah purata *square error* di dalam model. Persamaan 3.1 dan persamaan 3.2 menunjukkan formula RMSE dan MAE.

$$RMSE = (1/n(\sum(y_i - x_i)^2))^{1/2} \quad ...(3.1)$$

$$MSE = 1/n(\sum(y_i - x_i)^2 \quad ...(3.2)$$

Berdasarkan persamaan di atas, n adalah bilangan sampel, y_i adalah nilai sebenar, x_i adalah nilai ramalan.

3.6.2 Ketepatan (*Accuracy*)

Ketepatan adalah keupayaan model untuk mengukur bagaimana sesuatu model meramal dengan tepat. Ketepatan ditakrifkan sebagai pecahan ramalan model yang dengan jumlah bilangan ramalan. Namun, ketepatan tidak sesuai digunakan apabila kelas sesuatu set data adalah terhad kepada satu kelas sahaja.

3.7 PEMILIHAN MODEL TERBAIK UNTUK RAMALAN TAHAP OZON

Kesemua model yang telah melalui fasa penilaian akan dibandingkan dan model terbaik akan dipilih. Nilai-nilai seperti ketepatan, "RMSE" dan "MSE" akan direkodkan. Hanya model pembelajaran mesin terbaik, yang mencatatkan nilai RMSE, MAE terendah dan nilai ketepatan yang tinggi akan dipilih. Model terbaik yang dipilih akan digunakan untuk membuat ramalan tahap ozon dan gunakan untuk mecari faktor penting dengan menggunakan nilai SHAP.

3.8 TAFSIRAN FAKTOR PENTING MENGGUNAKAN GINI DAN NILAI SHAP

Apabila model yang terbaik dipilih daripada keempat-empat algortima, model tersebut akan digunakan untuk tafsiran faktor penting peningkatan tahap ozon. Penilaian peringkat menggunakan nilai SHAP dan Gini digunakan untuk mencari faktor penting. Pakej untuk mengira nilai SHAP akan diimport daripada *SHAP library* dan Gini akan di import daripada *scikit-learn*. Jadual 3.11 menunjukkan langkah-langkah pencarian faktor penting menggunakan kedua-dua kaedah.

Jadual 3.11 Langkah-langkah uji kaji menggunakan nilai SHAP dan *Gini*

SHapley Additive exPlanations dan Gini

Input: UrbanCombined.csv, SubUrbanCombined.csv, Background.csv

Output: Plot TreeExplainer, Force Plot, SHAP summary plot, dan SHAP dependence plot

Kaedah:

(1) Import libraries

Import shap

(2) Tetapan untuk nilai SHAP

Row_to_show = 2

Data_for_prediction = X_test.iloc[row_to_show]

Data_for_prediction_array = data_for_prediction.values.reshape(1,-1)

rfc.predict_proba(data_for_prediction_array)

(3) Plot TreeExplainer

Explainer = shap.TreeExplainer(rfc)

(4) Pengiraan nilai SHAP

Shap_values = explainer.shap_values(data_for_prediction)

(5) Penghasilan visual output nilai SHAP

Shap.force_plot(explainer.expected_value[0], shap_values[0], data_for prediction)

(6) SHAP summary plot

Explainer = shap.TreeExplainer()

Shap_values = Explainer.shap_values(X_testing1, approximate = True)

(7) SHAP dependence plot

Shap.dependence_plot (“Attribute”, shap_values[0], X_testing1, interaction_index =inds[i])

(8) Tetapan untuk kaedah Gini

Importances = model.feature_importances_

Indices = np.argsort(importances)

(9) Feature Importances plot

```
features = X_training1.columns
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
print(model _importances_)
```

Setelah plot nilai SHAP dan Gini diekstrak daripada model terbaik, maka faktor penting terhadap peningkatan tahap ozon akan diketahui dengan meggunakan tafsiran seperti yang diterangkan pada Jadual 3.11 diatas.

3.9 KESIMPULAN

Bab ini telah menerangkan proses-proses yang melibatkan metodologi kajian bermula daripada data mentah yang diperoleh sehingga kepada persediaan data bagi menguji model-model bergabung yang diimport. Selain itu, tafsiran data untuk mendapatkan faktor penting yang menyebabkan peningkatan tahap ozon juga dilakukan dengan penilaian peringkat daripada Gini dan nilai SHAP. Kesemua hasil dapatan akan berbentuk visual dan graf yang bersesuian dengan kajian ini.

BAB IV

DAPATAN KAJIAN

4.1 PENGENALAN

Bab ini menerangkan kesemua hasil dapatan kajian yang telah melalui fasa metodologi seperti yang dibentangkan di Bab 3 kajian ini. Bab ini akan bermula dengan dapatan kedua-dua analisis iaitu analisis statistik dan korelasi t-test. Kemudian, bab ini akan membentangkan dapatan terhadap model ramalan di kawasan bandar, pinggir bandar, dan latar dengan membandingkan setiap model dengan menggunakan nilai ketepatan, MSE, dan RMSE. Seterusnya, hasil dapatan daripada model terbaik akan digunakan untuk mengekstrak faktor penting. Kepentingan Gini dan nilai SHAP akan mengeluarkan faktor penting peningkatan tahap ozon dan output daripada dapatan tersebut akan divisualisasikan dengan plot dan graf.

4.2 HASIL ANALISIS STATISTIK

Seperti yang dinyatakan di bab 3, analisis diskriptif dilakukan bagi memvisualisasikan nilai statistik dan menghuraikan ciri-ciri data. Jadual 3.5, Jadual 4.1, dan Jadual 4.2 adalah analisis statistk bagi kawasan bandar, pinggir bandar, dan latar. Manakala, Rajah 4.1, Rajah 4.2, dan Rajah 4.3 menunjukkan graf perbezaan purata sebelum PKP dan semasa PKP bagi setiap parameter pencemaran udara dan meteorologi di kawasan bandar, pinggir bandar, dan latar.

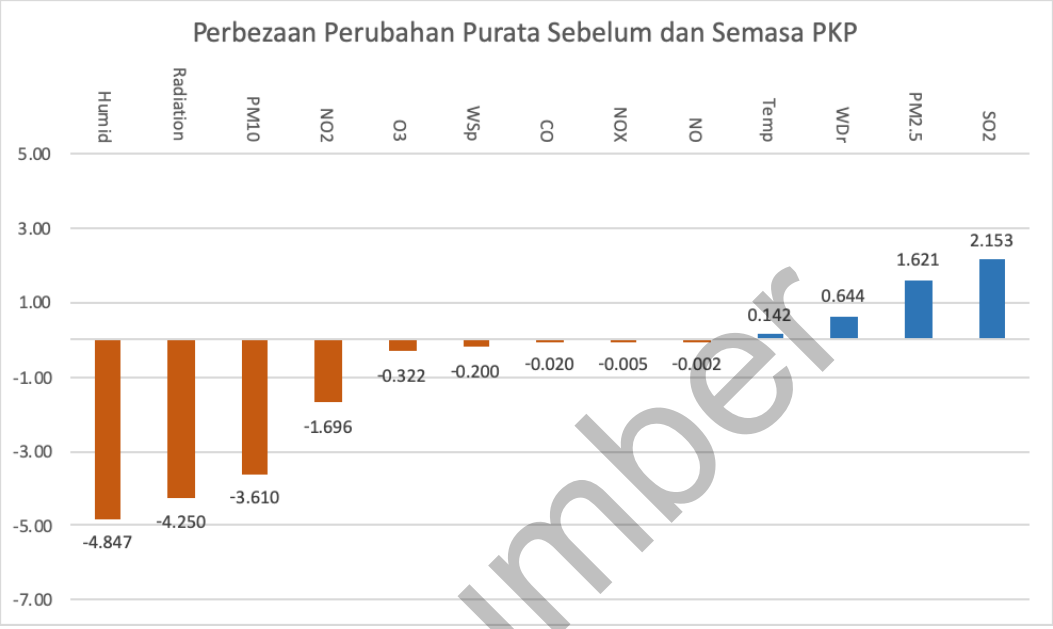
Jadual 4.1 Analisis statistik kawasan pinggir bandar sebelum dan semasa PKP.

	Sebelum PKP				Semasa PKP				Perbezaan		
Atribut	**Min**	**Max**	**SP**	**Purata**	**Min**	**Max**	**SP**	**Purata**	**ΔPurata**	**ΔSP**	**t-test**
PM_{10}	1.436	212.7	11.55	22.86	1.595	203.4	11.01	20.54	-2.319	0.533	0.000
$PM_{2.5}$	0.064	205.0	9.527	14.07	0.087	172.0	10.00	14.95	0.878	-0.48	0.000
O_3	0.094	117.4	15.17	20.40	0.387	85.50	13.50	20.30	-0.100	1.670	0.687
SO_2	0	7.503	0.582	0.950	0.005	9.299	0.352	0.884	-0.066	0.230	0.000
NO_2	0.001	31.83	4.334	5.030	0.007	24.21	2.596	3.217	-1.813	1.737	0.000
CO	0.042	1.939	0.205	0.494	0.046	2.213	0.128	0.435	-0.058	0.077	0.000
WSp	0	7.173	1.100	1.471	0	7.808	0.799	1.137	-0.333	0.300	0.000
WDr	1	8	2.092	3.497	1	8	2.063	3.753	0.256	0.028	0.000
Humid	24.17	98.33	14.57	75.64	23.38	99	14.05	78.16	2.543	0.524	0.000
Radiation	0	1239	289.5	205.9	0	1067	280.4	199.07	-6.832	9.069	0.164
Temp	21.33	37.18	3.335	27.85	20.48	36.94	3.249	28.13	0.275	0.085	0.000
NO	0	0.072	0.003	0.002	0.000	0.022	0.001	0.001	-0.001	0.002	0.000
NO_x	0	0.080	0.006	0.007	0.000	0.030	0.003	0.004	-0.003	0.003	0.000

Jadual 4.2 Analisis statistik kawasan latar sebelum dan semasa PKP.

	Sebelum PKP				Semasa PKP				Perbezaan		
Atribut	**Min**	**Max**	**SP**	**Purata**	**Min**	**Max**	**SP**	**Purata**	**ΔPurata**	**ΔSP**	**t-test**
PM_{10}	1.184	212.1	11.06	18.08	3.796	100.7	10.92	19.30	1.215	0.137	0.023
$PM_{2.5}$	0.094	172.1	9.369	10.71	0.199	88.90	10.16	12.43	1.717	-0.79	0.000
O_3	0.046	33.66	8.335	11.94	0.007	32.26	8.500	12.36	0.419	-0.16	0.308
SO_2	0.399	2.261	0.492	0.867	0.546	1.522	0.139	0.800	-0.067	0.308	0.000
NO_2	0.677	17.29	0.138	3.163	0.307	10.96	1.381	2.011	-1.152	1.113	0.000
CO	0.314	1.409	0.157	0.636	0.409	1.332	0.122	0.666	0.030	0.034	0.000
WSp	0	4.940	1.091	1.278	0.032	4.533	0.976	1.206	-0.072	0.114	0.156
WDr	1	8	1.948	4.582	1	8	2.119	4.582	-0.141	-0.17	0.156
Humid	38.07	95	15.05	74.27	38.45	95	15.74	74.22	-0.047	-0.69	0.950
Radiation	0	939.3	271.0	195.5	0	886.4	270.5	195.2	1.056	0.467	0.936
Temp	18.46	35.03	3.658	27.08	20.85	34.75	3.743	27.08	0.659	-0.08	0.000
NO	0	0.038	0.002	0	0	0.027	0.001	0	-0.001	0	0.000
NO_x	0	0.053	0.003	0.004	0	0.027	0.001	0.002	-0.002	0.002	0.000

4.2.1 Perbandingan Parameter Sebelum PKP dan Semasa PKP

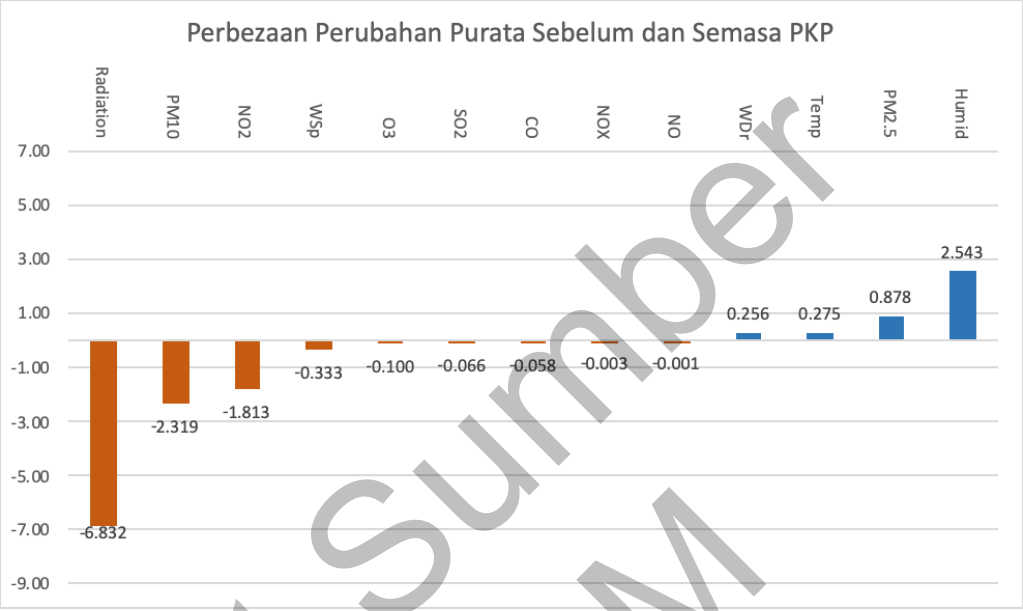


Rajah 4.1 Perbezaan purata sebelum PKP dan semasa PKP di kawasan bandar.

Seperti yang digambarkan pada Rajah 4.1 diatas, SO2 menunjukkan peningkatan purata tertinggi sebanyak 335% di kawasan bandar, diikuti dengan 11%, 18%, 0.5% bagi parameter PM2.5, arah angin, dan suhu. Rajah 4.1 juga menunjukkan penurunan purata sebanyak 15%, 1%, 30%, 3%, 15%, 6%, 2%, 38%, 43% bagi parameter PM10, O3, NO2, CO, kelajuan angin, kelembapan, radiasi, NO, dan NOx. Penurunan ozon sebanyak 1% menjelaskan bahawa ozon tidak banyak berubah setiap hari. Hasil ujian t-test pada Jadual 3.6 menunjukkan bahawa kebanyakan parameter telah meningkat secara signifikan ($p<0.05$) semasa PKP kecuali parameter ozon dan radiasi.

Ozon telah menunjukkan penurunan semasa PKP, namun ujian t-test menunjukkan bahawa ozon tidak menunjukkan perubahan yang signifikan. Perubahan itu tidak selari dengan penurunan parameter pencemaran udara seperti NO, NOx, NO2, PM10, dan CO yang menurun secara signifikan semasa PKP. Selain itu, parameter meteorologi seperti suhu, PM2.5, SO2 dan arah angin telah menunjukkan peningkatan secara signifikan semasa PKP.

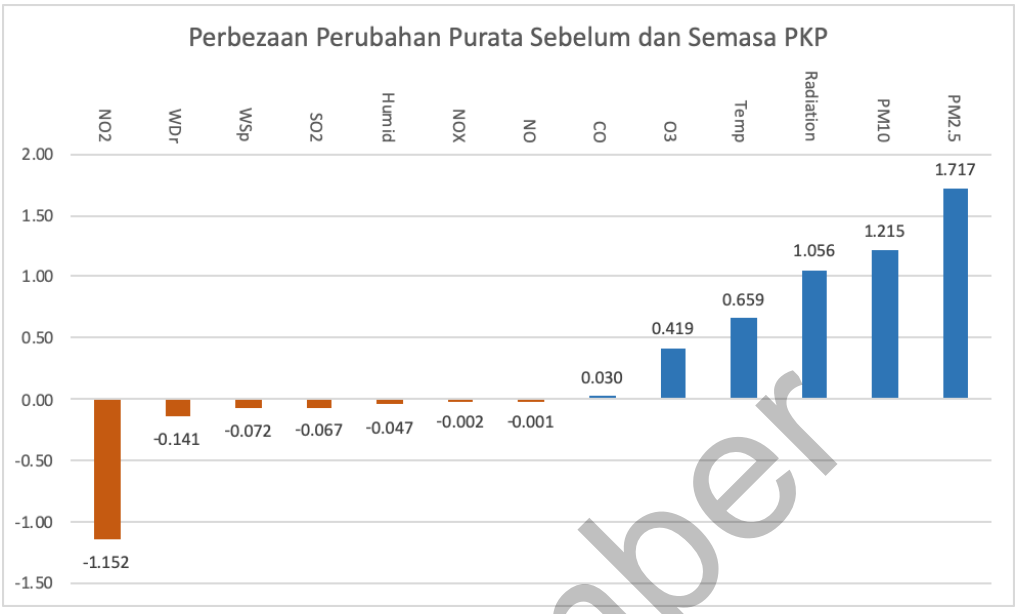
Hal ini telah menunjukkan bahawa penurunan nilai NO dan NOx telah menyebabkan dalam penurunan tahap ozon di kawasan bandar. Namun, oleh kerana suhu telah meningkat dengan signifikan semasa PKP, ia menyebabkan purata penurunan ozon adalah sebanyak 1%.



Rajah 4.2 Perbezaan purata sebelum PKP dan semasa PKP di kawasan pinggir bandar.

Berdasarkan Rajah 4.2, arah angin menunjukkan peningkatan purata tertinggi sebanyak 7% di kawasan pinggir bandar diikuti dengan PM2.5, kelembapan, dan suhu dengan peningkatan purata sebanyak 6%, 3% dan 0.9%. Terdapat juga penurunan purata sebanyak 43%, 38%, 36%, 22%, 11%, 10%, 6%, 3%, dan 0.5% bagi parameter seperti NO, NOx, NO2, kelajuan angin, CO, PM10, SO2, radiasi, dan ozon. Berdasarkan pemerhatian, penurunan purata ozon adalah sangat sedikit iaitu sebanyak 0.5% walaupun terdapat penurunan terhadap parameter pencemaran udara.

Jadual 4.1 menunjukkan hasil ujian t-test bagi kawasan pinggir bandar dan berdasarkan jadual tersebut penurunan parameter ozon dan radiasi tidaklah signifikan ($p>0.05$) semasa PKP. Oleh kerana purata ozon menunjukkan penurunan yang sedikit, hal ini menunjukkan bahawa kenaikan suhu ketika PKP memberi kesan terhadap tahap ozon.



Rajah 4.3 Perbezaan purata sebelum PKP dan semasa PKP di kawasan kawasan latar.

Bagi kawasan latar, PM2.5 menunjukkan peningkatan purata tertinggi diikuti dengan parameter PM10, CO, ozon, suhu, dan radiasi dengan nilai purata 16%, 6.7%, 4.7%, 3.5%, 2.5%, dan 0.5%. Manakala parameter seperti NO, NOx, NO2, SO2, kelajuan angin, arah angin, dan kelembapan telah menunjukkan penurunan purata sebanyak 50%, 39%, 36%, 7%, 5%, 2% dan 0.06%. Nilai purata menjelaskan bahawa peningkatan purata bagi parameter suhu dan radiasi telah menyumbang kepada peningkatan tahap ozon.

Jadual 4.2 menunjukkan hasil ujian t-test bagi kawasan latar dan Jadual 4.2 menunjukkan bahawa beberapa parameter seperti PM2.5, PM10, CO, dan suhu telah meningkat secara signifikan ($p<0.05$) semasa PKP. Selain itu, parameter seperti arah angin, kelajuan angin, kelembapan, dan sinaran radiasi tidak menunjukkan penurunan yang signifikan. Walaupun ozon telah menunjukkan peningkatan semasa PKP, tetapi peningkatan itu tidak signifikan kerana nilai p bagi ozon adalah ($p>0.05$).

Berdasarkan Rajah 4.1, Rajah 4.2, dan Rajah 4.3, peningkatan tahap ozon hanya berlaku di kawasan latar, tetapi peningkatan ozon tersebut tidaklah signifikan. Manakala, penurunan tahap ozon berlaku dikawasan bandar dan pinggir bandar. Walaubagaimanapun, purata penunrunan tahap ozon di kedua-dua kawasan sangatlah sedikit kerana purata suhu di kawasan tersebut meningkat semasa PKP. Oleh itu,

faktor peningkatan tahap ozon di kawasan bandar adalah semasa PKP adalah SO2, PM2.5, arah angin dan suhu. Manakala faktor yang mempengaruhi di kawasan pinggir bandar adalah kelembapan, PM2.5, arah angin dan suhu. Bagi kawasan latar, faktor yang mempengaruhi ozon semasa PKP adalah suhu, arah angin, PM2.5 dan SO2.

4.3 PEMILIHAN MODEL RAMALAN TAHAP OZON

Seperti yang dinyatakan di Bab 3, kajian ini menggunakan 4 pembelajaran mesin bergabung untuk meramal tahap ozon. Pengujian prestasi model bagi keempat-empat model bergabung digunakan untuk meramal tahap ozon di kawasan bandar, pinggir bandar, dan latar. Kesemua model telah diuji dengan teknik pengesahsahihan silang (*cross validation*). Prestasi setiap model dinilai menggunakan ketepatan, MSE, dan RMSE.

4.3.1 Perbandingan Prestasi Model di Kawasan Bandar

Jadual 4.3 dan Jadual 4.4 menunjukkan keputusan prestasi bagi set data kawasan bandar. Hasil penilaian set data menunjukkan model hutan rawak mencatatkan ketepatan yang tertinggi dengan nilai peratusan 89.14% sebelum PKP dan 88.29% semasa PKP. Berdasarkan Jadual 4.3 dan Jadual 4.4, nilai MSE dan RMSE bagi model hutan rawak sebelum PKP adalah terendah dengan nilai 0.111 dan 0.333. Manakala nilai MSE dan RMSE bagi model hutan rawak semasa PKP adalah 0.119 dan 0.360, diikuti dengan model *Gradient Boosting, Catboost,* dan *XGboost*.

Jadual 4.3 Nilai ketepatan, MSE, dan RMSE model bagi kawasan bandar sebelum PKP.

Model	Ketepatan (%)	MSE	RMSE
Hutan Rawak	89.14	0.111	0.333
XGboost	88.61	0.106	0.325
Gradient Boosting	88.23	0.121	0.347
CatBoost	88.75	0.103	0.320

Jadual 4.4 Nilai ketepatan, MSE, dan RMSE model bagi kawasan bandar semasa PKP.

Model	Ketepatan (%)	MSE	RMSE
Hutan Rawak	88.29	0.119	0.360
Gradient Boosting	86.14	0.141	0.375
Catboost	85.29	0.148	0.384
XGboost	85.22	0.149	0.386

4.3.2 Perbandingan Prestasi Model di Kawasan Pinggir Bandar

Jadual 4.5 dan Jadual 4.6 menunjukkan keputusan bagi set data kawasan pinggir bandar sebelum PKP dan semasa PKP. Hasil penilaian set data sebelum PKP menunjukkan algoritma hutan rawak mencatat nilai ketepatan yang tertinggi iaitu sebanyak 88.90% berbanding *XGboost*, *Gradient Boosting,* dan *Catboost*. Berdasarkan hasil penilaian set data semasa PKP, model algoritma menunjukkan perbezaan yang sedikit dengan prestasi model yang terbaik iaitu hutan rawak dengan ketepatan yang dengan nilai peratusan 89.80%, berbanding dengan *XGboost* iaitu 89.70%, *Gradient Boosting* dengan peratus ketepatan 89.69%, dan model *Catboost* dengan peratus ketepatan 89.00%.

Jadual 4.5 Nilai ketepatan, MSE dan RMSE model bagi kawasan pinggir bandar sebelum PKP.

Model	Ketepatan (%)	MSE	RMSE
Hutan Rawak	88.90	0.113	0.336
XGboost	88.22	0.121	0.347
Gradient Boosting	88.30	0.119	0.344
Catboost	87.11	0.130	0.360

Jadual 4.6 Nilai ketepatan, MSE dan RMSE model bagi kawasan pinggir bandar semasa PKP.

Model	Ketepatan (%)	MSE	RMSE
Hutan Rawak	89.80	0.102	0.320
XGboost	89.70	0.103	0.321
Gradient Boosting	89.69	0.104	0.322
Catboost	89.00	0.110	0.331