MERAMALKAN TAHAP PENCEMARAN UDARA DI KUALA LUMPUR MENGGUNAKAN RANGKAIAN NEURAL DAN ALGORITMA MESIN VEKTOR SOKONGAN

Jocelyn Tan Yi Xuan

Khairuddin Omar

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRACT

As we all know air pollution is a public health and environmental concern in Malaysia, especially in cities. The air quality in Malaysia has worsened due to urbanisation and the seasonal haze in the country. Besides that, the increasing industrialisation and the high number of motor vehicles which can increase air pollutant in the city. The levels of ozone have generally shown an increasing in urban area from year by year. To increase the awareness of the citizen to avoid air pollution in the specific city, prediction of the levels of air pollutant is needed by using machine learning technique which is Support Vector Machine and Neural Network. Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges, it is mostly used in classification problem. Neural Network algorithm is a series of algorithm that endeavours to recognize underlying relationship in a set of data through a process that mimics the way the human brain operates. Therefore, we use these two techniques to predict the air pollutant level based on air quality and condition in the area of Kuala Lumpur. In these studies, utilized selected techniques, such as Support Vector Machine (SVM) which has 50.49% of accuracy model and Neural Network which has 54.13% of accuracy model are used to predict ambient air pollutant levels based on data of condition surrounding in Kuala Lumpur which can affect the level of air pollutant whether is high or low.

ABSTRAK

Seperti yang kita tahu, pencemaran udara adalah masalah kesihatan awam dan masalah alam sekitar di Malaysia, terutama di bandar. Kualiti udara di Malaysia bertambah buruk kerana pembandaran dan keadaan jerebu musim di negara ini. Selain itu, peningkatan perindustrian dan jumlah kenderaan bermotor dapat menyebabkan pencemaran udara di bandar. Tahap pencemaran udara secara amnya menunjukkan peningkatan di kawasan bandar dari tahun ke tahun. Untuk meningkatkan kesedaran warganegara untuk mengelakkan pencemaran udara di bandar-bandar tertentu, ramalan tahap pencemaran udara diperlukan dengan menggunakan teknik pembelajaran mesin iaitu Mesin Vektor Sokongan (SVM) dan rangkaian neural (ANN) atau rangkaian neural buatan (ANN). SVM adalah algoritma pembelajaran mesin yang diawasi yang dapat digunakan untuk kedua-dua cabaran klasifikasi atau regresi, kebanyakannya digunakan dalam masalah klasifikasi. Algoritma ANN adalah rangkaian algoritma yang

berusaha untuk mengenali hubungan yang mendasari dalam satu set data melalui proses kecerdasan buatan. Oleh itu, projek ini menggunakan dua teknik ini untuk meramalkan tahap pencemaran udara berdasarkan kualiti dan keadaan udara di Kuala Lumpur. Dalam kajian ini, teknik terpilih yang digunakan, seperti SVM yang mempunyai 50.49% ketepatan model dan ANN pula mempunyai 54.13% ketepatan model yang digunakan untuk meramalkan tahap pencemaran udara ambien berdasarkan keadaan di sekitar Kuala Lumpur yang boleh menpengaruhkan tahap pencemaran udara tinggi atau rendah dari semasa ke semasa.

1 PENGENALAN

Kemerosotan kualiti udara adalah begitu ketara di kawasan-kawasan bandar besar seperti Kuala Lumpur yang mengalami proses perbandaran dan perindustrian (Tan 1982). Pencemaran udara akibat daripada pembakaran bahan fosil bagi keperluan industri serta kenderaan bermotor merupakan ciri masalah perbandaran. Masalah ini telah lama wujud sejak pertengahan abad yang lalu, khususnya di negara-negara maju dan Malaysia juga tidak terkecuali terutamanya di kawasan bandar seperti Kuala Lumpur. Sebelum tahun 1970an, beberapa pemerhati berpendapat bahawa masalah pencemaran udara di negara Malaysia adalah tidak penting kerana ianya menerima 2 jumlah hujan yang rata-rata melebihi 2200 mm setahun, dan ini dikatakan cukup untuk membersihkan udara yang tercemar (Sham dan Jamaluddin 1990).

Kecerdasan buatan menjadi satu kaedah permodelan yang sangat popular bagi menyelesaikan masalah-masalah yang komplek. Salah satu cabang daripada kecerdasan buat yang sering digunakan ialah ANN dan SVM. Pada dasarnya, SVM merupakan sebuah algorithm klasifikasi untuk data linear dan tidak linear. SVM telah menggunakan mapping untuk mentransformasikan training data awal ke dimensi yang lebih tinggi. SVM adalah algoritma pembelajaran mesin yang dapat digunakan untuk kedua-dua cabaran klasifikasi atau regresi, tetapi kebanyakkan digunakan dalam klasifikasi. Dalam algorithm SVM, projek ini memetakan setiap data sebagai titik dalam ruang n-dimensi dengan nilai setiap ciri. SVM boleh melakukan klasifikasi dengan mencari hiper-satah yang membezakan dua kelas dengan sangat baik. ANN pula merupakan aplikasi pemodelan yang akan meramalkan tahap pencemaran udara di Kuala Lumpur. ANN terdiri daripada sekumpulan neuron yang berkaitan secara fungsi. Satu neuron tunggal boleh dihubungkan dengan banyak neuron lain. Kecerdasan

buatan, pemodelan kognitif, dan jaringan saraf adalah paradigma pemprosesan maklumat yang dapat daripada sistem saraf biologi. Ia cuba mensimulasikan beberapa sifat ANN dan telah berjaya diterapkan untuk mengecamkan pertuturan, analisi gambar, dan kawalan adaptif, membina agen perisian atau robot autonomi.

Projek ini bertujuan untuk membangunkan algoritma yang boleh meramalkan tahap pencemaran udara di Kuala Lumpur dengan menggunakan pembelajaran mesin seperti SVM dan ANN. Proses ini dijalankan dengan memasukkan data kenderaan yang didaftar oleh penduduk Kuala Lumpur untuk meramalkan tahap pencemaran udara sama ada meningkat atau menurun pada tahun seterusnya.

2 PEMASALAHAN KAJIAN

Pencemaran udara di bandar-bandar besar seperti Kuala Lumpur semakin teruk dan masyarakat tidak mempunyai kesedaran terhadap masalah ini. Pencemaran udara dipanggil juga jerebu, jerebu ialah satu fenomena yang disebabkan oleh kewujudan banyak pratikel-pratikel kecil yang tidak boleh dilihat oleh mata kasar dan terapungapung di udara. Pratikel ini mungkin berasal dari semula jadi ataupun kesan daripada aktiviti manusia. Punca utama jerebu adalah disebabkan pembakaran secara terbuka, asap dari kilang dan asap dari kenderaan.

Oleh itu, orang ramai memerlukan satu model yang berupaya meramal tahap pencemar udara untuk mengelakkan jerebu berlaku di kawasan mereka. Model pembelajaran mesin akan digunakan dalam membangunkan sistem ramalan tahap pencemar udara ini dengan set data berkenaan. Hal ini dapat memudahkan pihak berkuasa untuk meramalkan pencemaran udara berlaku dengan cepat dan dapat dielakkan dengan menggunakan pembelajaran mesin. Dalam sistem ramalan ini terdapat dua model pembelajaran mesin iaitu SVM dan ANN. Model pembelajaran mesin dipilih untuk meramal tahap pencemaran udara kerana model tersebut boleh mendapat ramalan yang baik mengenai keadaan di sekitar Kuala Lumpur. Kedua-dua model ini sangat mencabar dalam konteks pembangunan dari sudut mengira ketepatan ramalan tahap pencemaran udara.

3 OBJEKTIF KAJIAN

Kajian ini bertujuan untuk meramalkan tahap pencemar udara dan mengelakkan pencemaran udara di kawasan Kuala Lumpur dengan menggunakan pembelajaran mesin. Objektif utama adalah menentukan punca pencemaran udara menggunakan pembelajaran mesin yang tertentu untuk

1. Membangunkan model SVM dan ANN untuk meramal tahap pencemaran dan kualiti udara di Kuala Lumpur.
2. Mengenalpasti model pembelajaran mesin terbaik untuk meramal tahap pencemaran dan kualiti udara di Kuala Lumpur.

4 METOD KAJIAN

Metod yang digunakan dalam process pembangunan algoritma pembelajaran mesin dalam kajian ini adalah agile model. Agile model adalah salah satu proses mudah dan berkesan untuk mengubah visi keperluan perniagaan ke penyelesaian perisian. Model ini tidak memerlukan perancangan yang sempurna untuk memulakan projek kerana model ini memerlukan pengguna untuk mengubah dan memberi tindak balas untuk mengubah dari semasa ke semasa. Terdapat 5 fasa untuk membangunkan algorithma untuk meramalkan tahap pencemar udara iaitu:

1. Fasa keperluan

   Sebelum memulakan merancang projek, projek ini perlu membuat dokumentasi awal yang akan menyenaraikan keperluan awal.

2. Fasa Reka Bentuk

   Terdapat dua cara untuk mendekati reka bentuk dalam pengembangan perisian - satu adalah reka bentuk visual dan yang lain adalah struktur seni bina aplikasi.

3. Fasa Pembangunan

   Selepas reka bentuk akan menukar dokumentasi reka bentuk ke perisian sebenar dalam proses pemgembangan perisian. Proses ini terdapat paling penting dan mengambil banyak masa sepanjang projek dijalankan

4. Fasa Pengujian

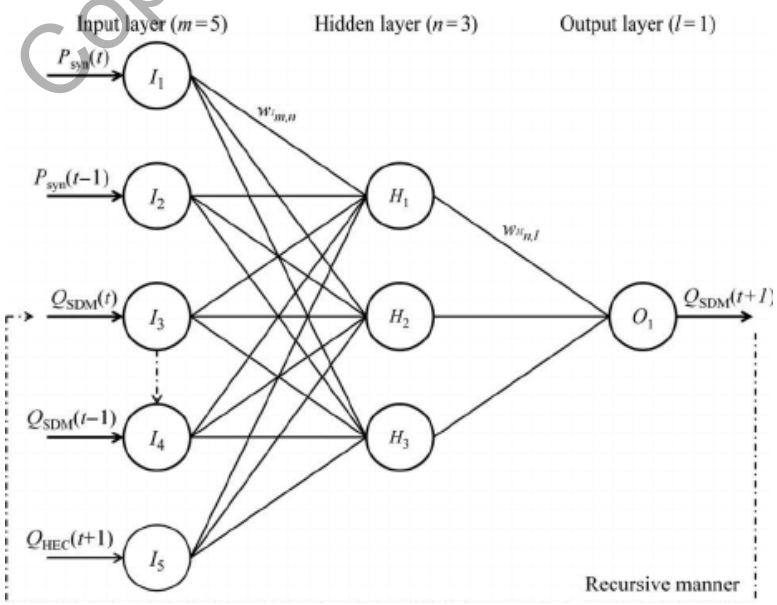
Pengujian ini dijalankan untuk memastikan program yang dijalankan tidak mempunyai pepijat dan serasi. Dalam fasa ini. pengguna boleh membantu menguji program yang telah dibangun untuk memastikan program boleh berlangsung dengan betul.

5. Fasa Penilaian

Setelah semua fasa selesai, projek ini boleh mengumpulkan pandangan orang lain untuk menilai program projek ini sebelum produk dikeluarkan ke pasaran terutamanya pandangan pengguna untuk membuat penambahbaikan program tersebut.

4.1 Fasa Perancangan

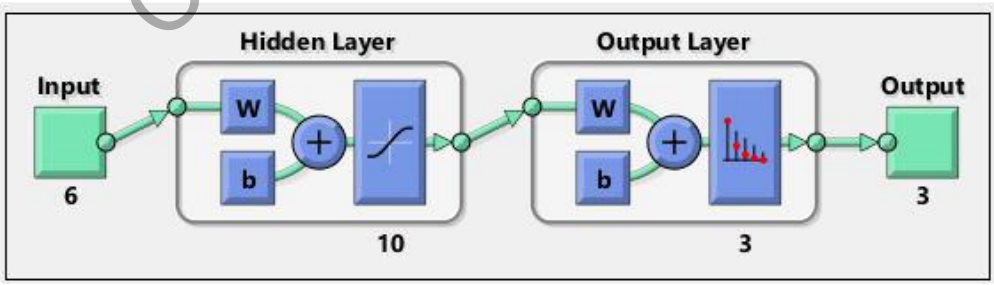
Model pembelajaran mesin iaitu ANN dan SVM akan digunakan dalam projek ini. Perambatan balik adalah singkatan dari "percambahan kesalahan yang progresif," kerana kesalahan diproses pada hasil dan digunakan secara terbalik sepanjang lapisan sistem. Ia selalu digunakan untuk menyiapkan rangkaian saraf yang mendalam dan juga salah satu jenis model ANN. Dalam projek ini, terdapat 9 atribut telah sebagai ciri iaitu lapisan input untuk menjalankan ramalan melalui ANN. Rajah 1 menunjukkan model ANN.



Rajah 1: Rangkaian Neural (ANN)

ANN adalah model biologi yang menarik daripada pemprosesan data kerana dia boleh mengira dan mengambil keputusan dan kesimpulan sama seperti otak manusia. Hal ini juga sebab mengapa manusia sangat pandai menyelesaikannya proses seperti mengenali objek dan wajah, mengenali persekitaran, menyelesaikan masalah situasi yang berbeza, tindakan dan reaksi adalah organisasi otak dan berfungsi. Otak manusia mempunyai neuron berbilion yang saling berkaitan antara satu sama lain dan berkomunikasi atas isyarat elektrokimia. ANN berfungsi seperti salinan otak manusia untuk menyelesaikan masalah kompleks di kawasan pembelajaran mesin (C. M. Bishop,1995). Pembelajaran dalam ANN tidak dilakukan pada setiap neuron, dan dia akan dilakukan sebagai sistem, di peringkat global. Semua neuron belajar bersama membina rangkaian yang dapat menyelesaikan masalah dengan ketepatan yang tinggi. Salah satu ANN yang biasa digunakan ialah perambatan balik (Back propagation) ANN kerana dia digunakan untuk mengurangkan ralat ramalan (C. Zhang, et al., 2017).

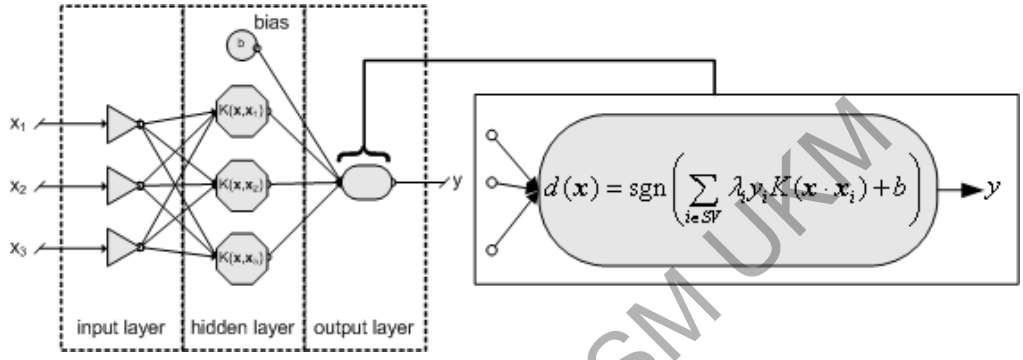
Perambatan balik bermula selepas pengiraan input data. Fungsi aktif pada setiap neuron telah ditentukan oleh nilai output dari lapisan sebelumnya dan berat antara neuron. ANN dalam projek ini memainkan peranan mengelaskan sampel kepada tahap pencemaran udara "rendah", "sederhana" dan "tinggi", bergantung pada set data latihan. Contohnya, terdapat 6 input dan 3 nilai output yang akan keluar sebagai ramalan, dan juga terdapat lapisan tersembunyi yang mengandungi 10 neuron untuk menjalankan model ini. Terdapat banyak peraturan yang menentukan bilangan neuron dalam lapisan tersembunyi ini (M. Bramer, 2007). Rajah 2 menunjukkan model seni bina bagi ANN.



Rajah 2: Model Seni Bina ANN

Kaedah kernel adalah kelas teknik pembelajaran mesin yang telah menjadi alat pembelajaran yang semakin popular tugas seperti pengecaman corak, klasifikasi atau kebaruan pengesanan (S. Canu, 2014). SVM merupakan kaedah kernel yang amat

popular dan kernel mesin dapat digunakan dalam banyak aplikasi yang disediakan antara linear ke tidak linear. SVM merupakan model pembelajaran yang diselia yang berkaitan dengan algoritma pembelajaran analisii data yang digunakan untuk klasifikasi dan analisis regresi. Model SVM adalah contoh dari kategori yang telah berasingan dibahagikan dengan vektor utama (satah hiper) selebar mungkin. SVM dapat menyelesaikan sample kecil, tidak linear, dimensi tinggi dan minimum titik dan masalah praktikal yang lain (Hipni, A., A., 2013). Rajah 3 menunjukkan model seni bina bagi SVM dengan fungsi linear dan tidak linear.



Rajah 3: Model Seni Bina SVM

4.2 Fasa Keperluan

4.2.1 Spesifikasi Keperluan Sistem

a) Keperluan Perkakasan

Jadual 1 menunjukkan spesifikasi keperluan perkakasan semasa membangunkan sistem ramalan projek ini.

Jadual 1: Spesifikasi Keperluan Perkakasan

Perkakasan	Perincian
Memori (RAM)	4GB
Pemproses	CPU Intel core i5
Cakera Keras	HDD
Sistem operasi	Window 10 x64 bit

b) Keperluan Perisian

Jadual 2 menunjukkan spesifikasi keperluan perisian untuk menyelesaikan pembangunan dan kajian projek ini.

Jadual 2: Spesifikasi Keperluan Perisian

Perisian	Perincian
Microsoft Word 2016	Menulis tesis projek tahun akhir.
Microsoft Excel 2016	Membuat data analisis bagi menunjukkan keputusan perjalanan pembangunan sistem.
Anaconda	Sebagai IDE untuk menulis algoritma python semasa membangunkan sistem ramalan ini.
Google Colab	Sebagai IDE untuk menulis algoritma python semasa membangunkan sistem ramalan ini.

c) Keperluan Set Data

Set Data yang digunakan di dalam pembanguanan sistem ini adalah set data indeks kualiti udata di Kuala Lumpir untuk meramal tahap pencemaran udara di Kuala Lumpur. (Gaganjot Kaur, 2018; Gaurav Pandey,2013) Set data terdapat 45283 data dan 13 atribut dan sumber set data daripada Kaggle. Set Data mempunyai rekod keadaan quality udara di sekitar Kuala Lumpur dari tahun 2012 hingga tahun 2018. Jadual 3 menunjukkan ciri-ciri bagi setiap atribut dalam set data yang digunakan pada projek ini.

Jadual 3: Ciri-ciri bagi setiap atribut dalam set data.

Atribut	Ciri-ciri
date_time	Masa dan tarikh pada hari kumpul

is_holiday	Sama ada hari itu bercuti atau tidak
humidity	Kelembapan berangka dalam Celcius
wind_speed	Kelajuan angin berangka
wind_direction	Arah angin kardinal (0-360 darjah)
visibility_in_miles	Penglihatan jarak
dew_point	Titik embun berangka dalam Celcius
temperature	Suhu purata berangka dalam Kelvin
rain_p_h	Jumlah nilai pH hujan yang berlaku dalam satu jam
clouds_all	Peratusan berangka penutup awam
weather_type	Kategori cuaca semasa
traffic_volume	Jumlah trafik setiap jam terikat pada arah tertentu
air_pollution_index	Kualiti Udara (0-500)

4.3 Fasa Reka Bentuk

4.3.1 Reka Bentuk Pangkalan Data

Set Data yang digunakan di dalam pembanguanan sistem ini adalah set data indeks kualiti udata di Kuala Lumpur untuk meramalkan tahap pencemaran udara di Kuala Lumpur. Set data yang akan digunakan untuk menjalankan ramalan ini adalah seperti berikut:

a) Bacaan Indeks Pencemar Udara (IPU) Bagi Negeri Wilayah Persekutuan Kuala Lumpur 2019 (Jabatan Alam Sekitar)

b) Bacaan Indeks Pencemar Udara (IPU) untuk setiap jam pada tahun 2017 bagi Kuala Lumpur. (Open Data Malaysia)

c) Private vehicles, goods vehicles and others registered 2004-2016. (Institut Automotif, Robotik dan IoT Malaysia, MARii)

d) Annual minimum and maximum air pollutant index for selected stations, Malaysia 2000-2017 (Jabatan Perangkaan Malaysia)

4.3.2 Reka Bentuk Algoritma

Dalam proses pembangunan sistem ini menggunakan perisian Google Colab dan bahasa pengaturcaraan digunakan adalah Python. Google Colab dipilih sebagai perisian bagi proses kerana penghasilan model akan bergantung kepada RAM dan GPU. Google Colab menyediakan ruang GPU yang sangat mencukupi untuk menjalankan dan menghasilkan model. Dalam projek ini, terdapat 45283 set data dan 13 atribut akan digunakan untuk meramalkan pencemaran udara di Kuala Lumpur. Sebelum memecahkan set data kepada data latihan dan data ujian, data perlulah dikemaskan dan pastikan tiada pencilan dalam set data. Selepas membunagkan pencilan dalam set data, terdapat hanya meninggal 41318 set data akan dipecahkan untuk menjalankan ramalan tahap pencemaran udara. Rajah 4.1 menunjukkan teknik julat interkuatil (IQR) adalah ukuran penyebaran statistik, sama dengan perbezaan antara persentil ke-75 dab ke-25 atau antara kuartil atas dan bawah, IQR = Q3 - S1.

Removing Outliers

```
[ ]  Q1 = pollution.quantile(0.25)
     Q3 = pollution.quantile(0.75)
     IQR = Q3 - Q1
     print(IQR)
```

```
humidity                26.000
wind_speed               3.000
wind_direction         155.000
visibility_in_miles      4.000
dew_point                4.000
temperature             18.825
rain_p_h                 0.000
clouds_all              89.000
weather_type             5.000
air_pollution_index      2.000
traffic_volume        3750.000
dtype: float64
```

```
[ ]  pollution = pollution[~((pollution < (Q1 - 1.5 * IQR)) |(pollution > (Q3 + 1.5 * IQR))).any(axis=1)]
     pollution.shape
```

```
(41318, 11)
```

Rajah 4: Atur Cara fasa pembersihan data dengan menggunakan IQR

Seterusnya, set data akan dipecahkan kepada dua jenis data iaitu data latihan dan data ujian. Terdapat 27709 data dipecahkan sebagai data latihan dan 13649 data dipecahkan sebagai data ujian. Set data yang digunakan adalah sama bagi kedua-dua model ANN dan SVM. Rajah 5 menunjukkan fasa model latihan bagi aplikasi ramalan pencemaran udara yang akan digunakan dalam model pembelajaran mesin ANN dan SVM.

```
features=pollution
target=pollution['air_pollution_index']
```

```
features=features.drop('air_pollution_index',axis=1)
features.head()
```

	humidity	wind_speed	wind_direction	visibility_in_miles	dew_point	temperature	clouds_all	weather_type	traffic_volume
0	89	2	329	1	1	288.28	40	1	5545.0
1	67	3	330	1	1	289.36	75	1	4516.0
2	66	3	329	2	2	289.58	90	1	4767.0
3	66	3	329	5	5	290.13	90	1	5026.0
4	65	3	329	7	7	291.14	75	1	4918.0

```
# split a dataset into train and test sets
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.33)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(27709, 9) (13649, 9) (27709,) (13649,)
```

Rajah 5: Atur Cara fasa model latihan bagi aplikasi ramalan pencemaran udara

Ramalan dengan menggunakan model ANN. Rajah 6 menunjukkan atur cara untuk meramalkan indek pencemar udara berdasarkan set data menggunakan model ANN.

```
def plot_predicted(predicted_data, true_data):
    fig, ax = plt.subplots(figsize=(17,8))
    ax.set_title('Prediction vs. Actual after 50 epochs of training')
    ax.plot(true_data, label='True Data', color='green', linewidth='3')

    ax.plot(predicted_data, label='Prediction', color='red', linewidth='2')
    plt.legend()
    plt.show()

rmse = np.sqrt(mean_squared_error(y_test, y_pred))

y_test = MinMaxScaler().fit_transform(y_test.values.reshape(-1, 1))
y_pred = MinMaxScaler().fit_transform(y_pred.reshape(-1, 1))
#y_pred = scaler.inverse_transform(y_pred.reshape(-1,1))


plot_predicted(y_pred[:100,], y_test[:100,])
print('Root Mean Squared Error: {:.4f}'.format(rmse))
print("R2 score : %.2f" % r2_score(y_test,y_pred))
```

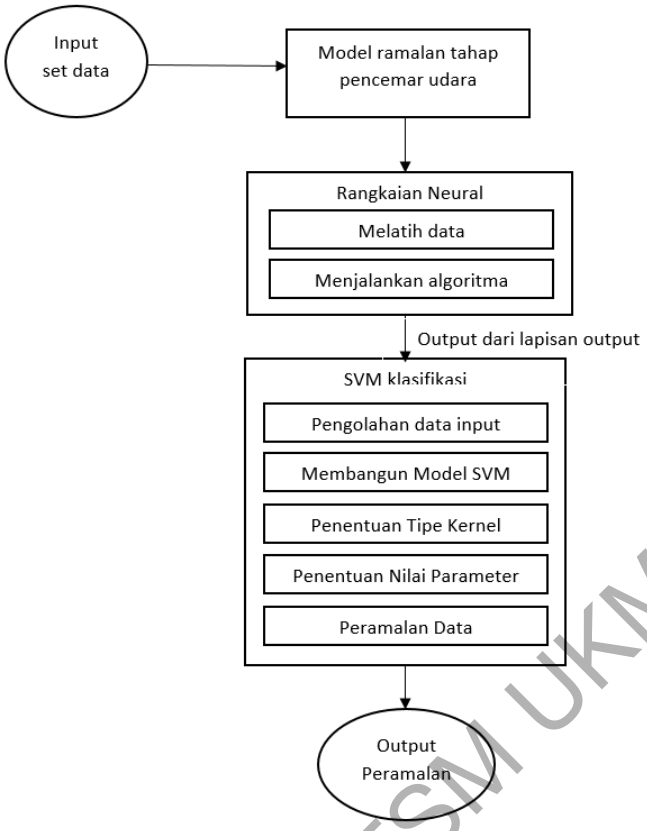
Rajah 6: Atur cara meramalkan indek pencemar udara menggunakan model ANN.

Ramalan dengan menggunakan model SVM. Rajah 7 menunjukkan atur cara meramalkan pencemaran udara menggunkan model SVM.

```
x = train_X
y = train_y

regr = SVR(C = 2.0, epsilon = 0.1, kernel = 'rbf', gamma = 0.5,
           tol = 0.001, verbose=False, shrinking=True, max_iter = 10000)

regr.fit(x, y)
data_pred = regr.predict(x)
y_pred = scaler.inverse_transform(data_pred.reshape(-1,1))
y_inv = scaler.inverse_transform(y.reshape(-1,1))

mse = mean_squared_error(y_inv, y_pred)
rmse = np.sqrt(mse)

def plot_predicted(predicted_data, true_data):
    fig, ax = plt.subplots(figsize=(17,8))
    ax.set_title('Prediction vs. Actual ')
    ax.plot(true_data, label='True Data', color='green', linewidth='3')

    ax.plot(predicted_data, label='Prediction', color='red', linewidth='2')
    plt.legend()
    plt.show()

def run_test_nonlinear_reg(x, y):
    data_pred = regr.predict(x)
    y_pred = scaler.inverse_transform(data_pred.reshape(-1,1))
    y_inv = scaler.inverse_transform(y.reshape(-1,1))

    mse = mean_squared_error(y_inv, y_pred)
    rmse = np.sqrt(mse)
    print('Mean Squared Error: {:.4f}'.format(mse))
    print('Root Mean Squared Error: {:.4f}'.format(rmse))

    #Calculate R^2 (regression score function)
    print('Variance score: {:2f}'.format(r2_score(y_inv, y_pred)))
    return y_pred, y_inv
```


Rajah 7: Atur cara meramalkan pencemaran udara menggunakan model SVM.

Rajah 8 menunjukkan model ramalan tahap pencemar udara menjalankan peramalan dengan menggunakan dua teknik pembelajaran mesin iaitu ANN dan SVM.



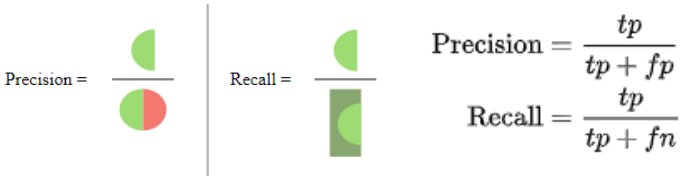
Rajah 8: Carta alir peramalan tahap pencemar udara dengan teknik ANN dan SVM.

4.4 Fasa Pengujian

Pra-pemprosesan menyaringkan kehingaran sperti tanda baca, nombor, ruang lebih dan watak khas. Pada tahap yang satu ini, data yang telah diseleksi akan kembali diseleksi ulang. Seleksi kedua ini berfungsi untuk membuang data yang sekiranya tidak diperlukan. Data yang dibuang ini diantaranya adalah data yang tidak valid, data yang tidak konsisten, dan data ganda. Data yang tidak lengkap, noisy dan tidak konsisten telah dinyatakan sebagai data kotor. Dalam proses ini, semua data kotor ini akan dibuangkan. Teknik IQR juga akan dijalankan untuk memastikan set data yang digunakan tiada mempunyai pencilan yang lebih. Oleh itu, keputusan yang dihasilkan dapat lebih tepat.

Pengesahan model juga salah satu langkah pengujian dalam sistem ini. Pengesahan model dalam sistem ini akan menggunakan matrik kekeliruan dan laporan klasifikasi yang mengandungi precision, recall, f1-score, dan support yang memberikan keputusan terperinci mengenai prestasi model. Precision juga disebut nilai ramalan positif adalah pecahan data yang relevan di antara data yang diambil, recall juga

merupakan kepekaan adalah pecahan data yang relevan yang diambil. F1-score menyampaikan keseimbangan antara precision dan recall. Rajah 9 menunjukkan formula untuk precision dan recall dimana tp adalah true positive, fp adalah false positive dan fn adalah false negative.



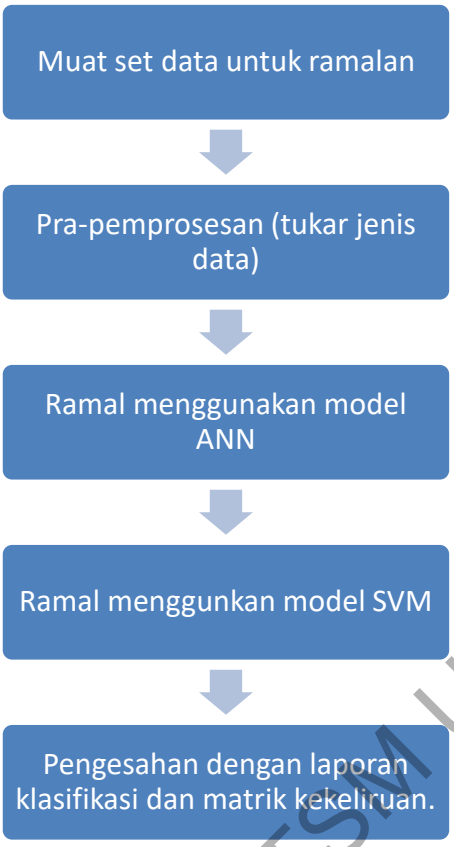
Rajah 9: Formula untuk precision dan recall.

Ini adalah kaedah harmonik yang memberikan ukuran yang lebih baik bagi kes yang dikelaskan secara salah daripada Metrik Ketepatan. Rajah 10 menunjukkan formula untuk f1-score.

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

Rajah 10: Formula untuk f1-score.

Terdapat 5 langkah akan dijalankan dalam proses pengujian iaitu memuatnaik data, data pemprosesan, ramalan dengan menggunakan model ANN, ramalan dengan menggunakan model SVM dan akhirnya membuat pengesahan dengan laporan klasifikasi dan matrik kekeliruan. Rajah 11 menunjukkan proses pengujian model.



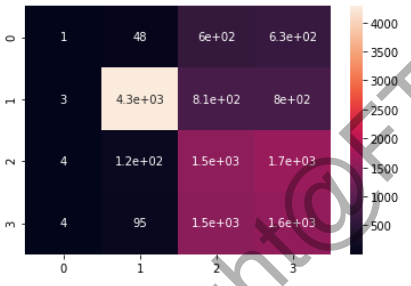
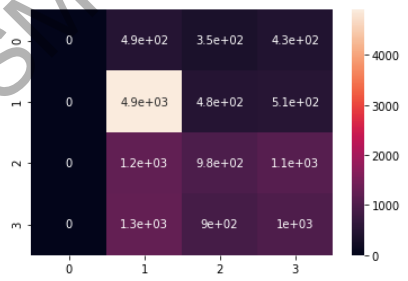
Rajah 11: Proses pengujian model.

Terdapat 45283 set data akan digunakan untuk meramalkan pencemaran udara. Set data akan menjalankan pra-pemprosesan untuk memastikan set data yang digunakan bersih dan tepat untuk membuat ramalan. Set data seterusnya akan menjalankan dua model pembelajaran mesin iaitu ANN dan SVM. Sebelum membuat ramalan, set data akan dipecahkan kepada dua jenis set data iaitu set data latihan dan set data ujian. Terdapat 27709 data akan dipecahkan sebagai data latihan manakala 13649 data akan dipecahkan sebagai data ujian. Melalui proses pengekstrak ciri dan penilaian pengelas, model yang menghasilkan ketepatan yang paling tinggi akan dipilih dan menggunakan matrik kekeliruan dan laporan klasifikasi untuk mengira kualiti ramalan berdasarkan model yang diggunakan.

Laporan klasifikasi digunakan untuk mengukur kualiti ramalan dari algoritma klasifikasi. Berapa ramalan yang benar dan berapa ramalan yang salah. Lebih khusus lagi, Positif Benar, Positif Salah, Negatif Benar dan Negatif Palsu digunakan untuk meramalkan matrik laporan klasifikasi seperti yang ditunjukkan di jadual bawah. Jadual

4 menunjukkan laporan klasifikasi dan matrik kekeliruan bagi kedua-dua model ANN dan SVM.

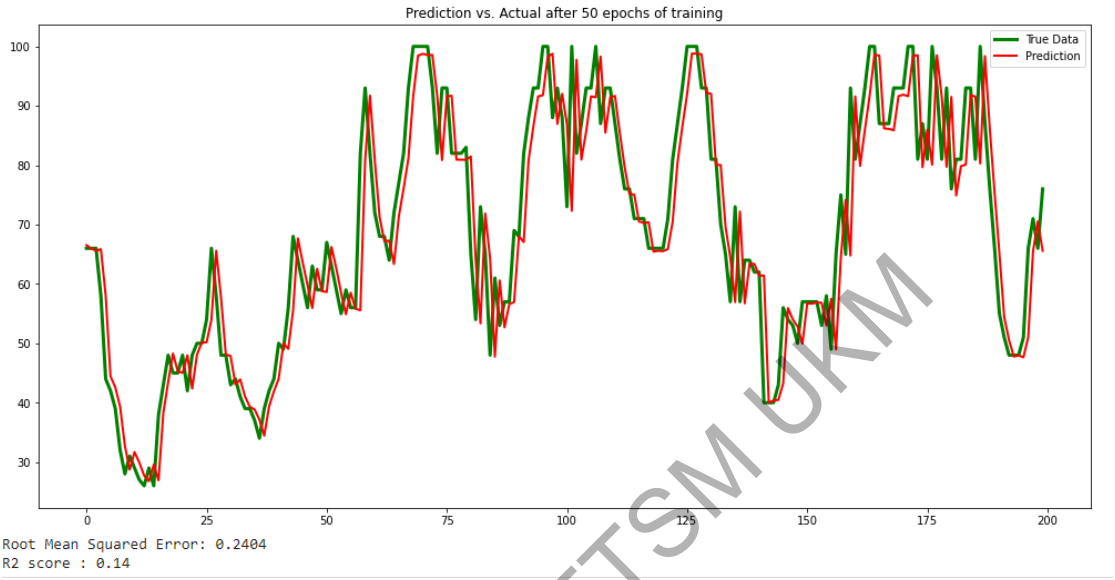
Jadual 4: Laporan klasifikasi dan matrik kekeliruan bagi ANN dan SVM.

Model	ANN	SVM
Laporan klasifikasi	Accuracy of : ANN Classifier 54.12850758297311 precision recall f1-score support 0 0.08 0.00 0.00 1278 1 0.94 0.73 0.82 5897 2 0.35 0.46 0.40 3314 3 0.34 0.50 0.40 3160 accuracy 0.54 13649 macro avg 0.43 0.42 0.40 13649 weighted avg 0.58 0.54 0.54 13649	Accuracy of Support Vector Classifier: 50.48721518059931 precision recall f1-score support 0 0.00 0.00 0.00 1278 1 0.62 0.83 0.71 5897 2 0.36 0.30 0.32 3314 3 0.33 0.32 0.32 3160 accuracy 0.50 13649 macro avg 0.33 0.36 0.34 13649 weighted avg 0.43 0.50 0.46 13649
Matrik kekeliruan		

Berdasarkan matrik kekeliruan dan report klasifikasi, ketepatan model ANN adalah 54.13% manakala ketepatan model SVM adalah 50.49%. Model yang dipilih untuk melakukan ramalan adalah model ANN yang mempunyai ketepatan tertinggi. Ketepatan model tersebut walaupun mempunyai sebanyak 54.13% masih dikira kurang baik bagi ramalan. Hal ini demikian kerana kemungkin pengkelasan model tersebut tidak ada hubungan antara ciri dan kelas serta kekurangan atribut untuk menguji. Set data yang digunankan untuk pecahan kepada set latihan dan set ujian mempunyai ralat yang rendah tetapi ralat pengesahan yang tinggi, maka pengkelasan model ini mempunyai varians yang terlalu tinggi. Dalam kes ini, terdapat juga kemungkinan parameter regularisasi yang agar rendah dalam pengkelasan model ini. Oleh itu, ketepatan model bagi ANN hanya 54.13% dalam kajian ini.

5. HASIL KAJIAN

Rajah 12 menunjukkan hasil dan output dengan menggunakan model pembelajaran mesin ANN untuk membuat ramalan terhadap dataset yang digunakan iaitu air pollution index di Kuala Lumpur.



Rajah 12: Hasil dan output dengan menggunakan model ANN.

Berdasarkan matrik kekeliruan dan report klasifikasi, ketepatan model ANN adalah 54.13% manakala ketepatan model SVM adalah 50.49%. Model yang dipilih untuk melakukan ramalan adalah model ANN yang mempunyai ketepatan tertinggi.

6. KESIMPULAN

Tuntasnya, sistem ramalan tahap pencemar udara ini akan membantu banyak orang terutamanya pihak berkuasa untuk lebih awal mengelakkan pencemaran udara berlaku di kawasan atau tempat tersebut. Sistem ini akan dapat dibangunkan dengan skop kajian dan objektif yang ditentukan. Limitasi dan cadangan peningkatan masa depan akan menjadi panduan untuk menghasilkan sistem ini lebih tepat dan cekap pada masa depan.

7. RUJUKAN

C. M. Bishop, “NNs for Pattern Recognition”, Oxford, 1995

C. Zhang, et al., “Understanding deep learning requires rethinking generalization”, ICLR, 2017

Gaganjot Kaur; Jerry Gao; Sen Chiao. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development 9(1):8-16. 2010-0264

Gaurav Pandey; Bin Zhang. (2013). Predicting submicron air pollution indicators: A machine learning approach. Environmental Science: Processes and Impacts 15(5). 2050-7895

Hipni, A., A. El-shafie, A. Najah, A.O. Karim, A. Hussain and M. Mukhlisin, 2013. Daily forecasting of dam water level: comparing a SVM model with adaptive neuro fuzzy inference system (ANFIS), Water Recourse Management, 27: 3803-3823

M. Bramer, “Principles of data mining”, Springer, 2007

S. Canu, “SVM and kernel machines: linear and nonlinear classification”, OBIDAM, Brest, 2014

UNEP. Status of Fuel Quality and Vehicle Emission Standards Latin America [Internet]. 2016. Available from: https://www.lead.org.au/PCFV/Lead_Matrix_LAC_201206.pdf. [Accessed: 3 November 2020]

Vehicle Registration Statistics for Private Vehicle, Goods Vehicle and Others, Internet: https://www.data.gov.my/data/en_US/dataset/vehicle-registration-statistics-for-private-vehicle-goods-vehicle-and-others [Accessed: 20 December 2020]

World Health Organization, Media Centre. Air pollution levels rising in many of the world's poorest cities [Internet]. 2016. Available from: http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/