PENGELASAN INDEKS KUALITI AIR TASIK BERASASKAN MODEL PEMBELAJARAN MESIN

AMAN SHAH BIN ABDUL SATAR

UNIVERSITI KEBANGSAAN MALAYSIA

PENGELASAN INDEKS KUALITI AIR TASIK BERASASKAN MODEL PEMBELAJARAN MESIN

AMAN SHAH BIN ABDUL SATAR

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

08 July 2024

AMAN SHAH BIN ABDUL SATAR
P121043

PENGHARGAAN

Alhamdulillah syukur ke hadrat Allah atas inayah dan keberkatan yang telah dikurniakan sepanjang proses penyediaan kajian projek akhir ini sehingga disiapkan dengan jayanya. Jutaan terima kasih serta penghargaan disampaikan kepada penyelia saya, Ts. Dr. Nor Samsiah Sani atas bimbingan dan panduan yang diberikan sepanjang kajian ini dibuat. Sekalung penghargaan atas segala tunjuk ajar dan dorongan yang diberikan sepanjang tempoh penyeliaan di bawah beliau.

Tidak ketinggalan juga, ucapan terima kasih kepada semua tenaga pengajar yang terlibat sepanjang sesi pembelajaran Sarjana Sains Data di Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaan Malaysia.

Ucapan terima kasih tidak terhingga kepada pihak penaja biasiswa iaitu Perbadanan Putrajaya yang menganugerahkan pembiayaan penuh bagi mengikuti pengajian dalam bidang ini. Setinggi-tinggi penghargaan juga diberikan kepada Jabatan Perancangan Bandar yang membenarkan kajian dibuat dan menyediakan data-data berkaitan dalam membantu kajian ini sehingga siap.

Akhir sekali, terima kasih dan penghargaan kepada isteri dan anak-anak serta bonda yang sentiasa memberi dokongan, nasihat dan sokongan moral yang berterusan sepanjang pengajian dalam bidang ini. Tidak lupa juga rakan-rakan sekelas yang sama-sama berusaha sedaya upaya sepanjang sesi pengajian dan turut sama membantu menyumbangkan buah fikiran dalam menyiapkan kajian ini.

Sekian, terima kasih.

ABSTRAK

Secara amnya, kajian indeks kualiti air dijalankan menggunakan kaedah konvensional iaitu melalui kajian makmal dan pengiraan statistik yang melibatkan masa yang lama dan kos yang tinggi secara tidak langsung mengakibatkan pemantauan masa-sebenar menjadi kurang efektif. Kajian kualiti air memerlukan kaedah yang lebih praktikal dan menjimatkan kos. Oleh yang demikian, terdapat keperluan mendesak supaya penghasilan model untuk menjangka dan menentukan kualiti air bagi mengawal pencemaran air serta memaklumkan kepada pengguna sekiranya terdapat penurunan kadar kualiti air. Bagi menguruskan kualiti air secara efektif, ketepatan jangkaan dalam menentukan kelas kualiti air amat diperlukan. Sehubungan itu, kelebihan yang terkandung di dalam kaedah pembelajaran mesin digunapakai bagi menghasilkan model yang sesuai untuk menjangka indeks dan kelas kualiti air. Set data indeks kualiti air tasik bagi kajian ini telah diperolehi daripada Seksyen Pengurusan Ekohidrologi Tasik dan Kawasan Tadahan, Bahagian Alam Sekitar, Tasik dan Wetland, Jabatan Perancangan Bandar, Perbadanan Putrajaya. Set data ini mempunyai sebanyak 1020 data dengan 27 fitur dan merupakan data bacaan kualiti air di Tasik Putrajaya dari 17 stesen. Fitur Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Ammoniacal Nitrogen (NH3-N), Suspended Solid (SS) dan pH yang digunakan di dalam pengiraan fomula Indeks Kualiti Air oleh Jabatan Alam Sekitar Malaysia (DOE) diberikan perhatian di dalam kajian ini. Fitur bernama *Class* pula dipilih sebagai pembolehubah bersandar di dalam kajian ini. Setelah pemilihan fitur dan penaalan hiperparameter dibuat, sepuluh fitur dipilih untuk digunakan di dalam model pengelasan indeks kualiti air dengan menggunakan teknik pembelajaran mesin. Tiga (3) algoritma pembelajaran mesin digunakan untuk mengelas indeks kualiti air Tasik Putrajaya ialah Hutan Rawak (RF), Rangkaian Neural Buatan (ANN) dan Mesin Vektor Sokongan (SVM). Bacaan laporan pengelasan, matrik kekeliruan dan ujian statistik ditentukan bagi setiap model pengelasan dan perbandingan dibuat untuk menentukan algoritma yang terbaik. Hasil dapatan yang diperolehi membuktikan bahawa model pengelas Hutan Rawak (RF) berjaya memperoleh prestasi model pengelasan terbaik dengan nilai ketepatan 100% bagi set latihan dan 95.1% bagi set ujian. Berdasarkan pemilihan fitur bagi model terbaik Hutan Rawak (RF), sepuluh fitur telah terpilih iaitu COD mg/l, NH3N mg/l, BOD mg/l, Ammonium mg/l, D.O mg/l, D.O, TSS mg/l, Temperature, Conductivity µS/cm dan Turbidity NTU. Daripada sepuluh fitur tersebut, hanya lima fitur Jabatan Alam Sekitar (DOE) telah terpilih bagi model terbaik iaitu COD mg/l, NH3N mg/l, BOD mg/l, D.O dan TSS mg/l.

LAKE WATER QUALITY INDEX CLASSIFICATION BASED ON MACHINE LEARNING MODEL

ABSTRACT

Generally, water quality index studies are conducted using conventional methods, namely through laboratory studies and statistical calculations involving longer time usage and high costs resulting in real-time monitoring being less effective. The study of water quality requires more practical and cost-effective methods. Therefore, there is an urgent need for the production of models to anticipate and determine water quality to control water pollution as well as to inform public if there is a decline in water quality. In order to effectively manage water quality, accuracy of expectations in determining water quality classes is required. In this regard, the advantages of the machine learning method are used to produce a suitable model for pedicting water quality index and classes. The lake water quality index data set for this study was obtained from the Ecohydrology Management Section of Lakes and Catchment Areas, Environment, Lakes and Wetland Division, Town Planning Department, Perbadanan Putrajaya. The dataset has a total of 1020 data with 27 features water quality reading data in Putrajaya Lake from 17 stations. Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Ammoniacal Nitrogen (NH3-N), Suspended Solid (SS) and pH are features used in the calculation of the Water Quality Index formula by Malaysian Department of Environment (DOE) was focussed during this study. The attribute named *Class* was selected as a dependent variable in this study. After the feature selection and hyperparameter tuning are executed, ten features were selected to be used in the water quality index forecast model using machine learning techniques. Three (3) machine learning algorithms were selected to classify the Putrajaya Lake water quality index namely Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Classification report, confusion matrics and statistical test are determined for each model and comparisons are made to obtain the best algorithm. The findings proved that the Random Forest (RF) classification model achieved the best performance with an accuracy value of 100% for the training set and 95.1% for the test set. Based on feature selection for the best model Random Forest (RF), ten features have been selected, namely COD mg/l, NH3N mg/l, BOD mg/l, Ammonium mg/l, D.O mg/l, D.O, TSS mg/l, Temperature, Conductivity µS/cm and Turbidity NTU. Out of ten features, only five Department of Environment (DOE) features have been selected for the best models, namely COD mg/l, NH3N mg/l, BOD mg/l, D.O and TSS mg/l.

KANDUNGAN

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

ANN	Rangkaian Neural Buatan (*Artificial Neural Network*)
BASTW	Bahagian Alam Sekitar, Tasik dan Wetland
CM	Matrik Kekeliruan (CM)
DOE	Jabatan Alam Sekitar (*Department of Environment*)
IKA	Indeks Kualiti Air (IKA)
IoT	Internet Benda (IoT)
JR	Jabatan Perancangan Bandar, PPJ
MLP	Perseptron Berbilang Lapisan (*Multi Layer Perceptron*)
NLWQS	Piawaian dan Kriteria Kebangsaan Kualiti Air Tasik (National Lake Water Quality Criteria And Standards)
NWQS	Piawaian Kualiti Air Kebangsaan Malaysia (*National Water Quality Standards For Malaysia*)
PPj	Perbadanan Putrajaya
RF	Hutan Rawak (*Random Forest*)
RFE	Penghapusan Fitur Rekursif (RFE)
SHAP	SHapley Additive exPlanations
SVM	Mesin Vektor Sokongan (*Support Vector Machine*)
XLSX	Microsoft Excel Open XML Format Spreadsheet file

BAB I

PENDAHULUAN

1.1 Pengenalan

Bab ini menjelaskan secara umum keseluruhan kajian merangkumi latar belakang kajian, penyataan masalah, objektif kajian, skop kajian, metodologi kajian, ringkasan sumbangan kajian dan diakhiri dengan struktur kandungan bagi tesis ini.

1.2 Latar Belakang Kajian

Tasik Putrajaya sering digunakan untuk pelbagai program dan aktiviti antaranya sukan air dan pertandingan memancing. Tasik Putrajaya berkeluasan 400 hektar merupakan salah satu tarikan utama di Putrajaya. Perbadanan Putrajaya merupakan pihak berkuasa tempatan yang bertanggungjawab untuk menyelenggara dan memastikan kualiti air tasik Putrajaya sentiasa berada dalam tahap indikasi B2 iaitu bersih dan sesuai digunakan untuk aktiviti riadah dan sukan air.

Tasik ini terletak di bahagian selatan wetland. Hampir 60% daripada air tasik masuk melalui Wetland manakala baki 40% daripada persiaran bersempadan dengan Putrajaya. Persiaran selebar 20 meter dibina sebagai penampan sekeliling tebing tasik (BASTW 2023). Ciri-ciri utama Tasik Putrajaya adalah seperti Jadual 1.1 di bawah.

Jadual 1.1 Ciri-Ciri Tasik Putrajaya

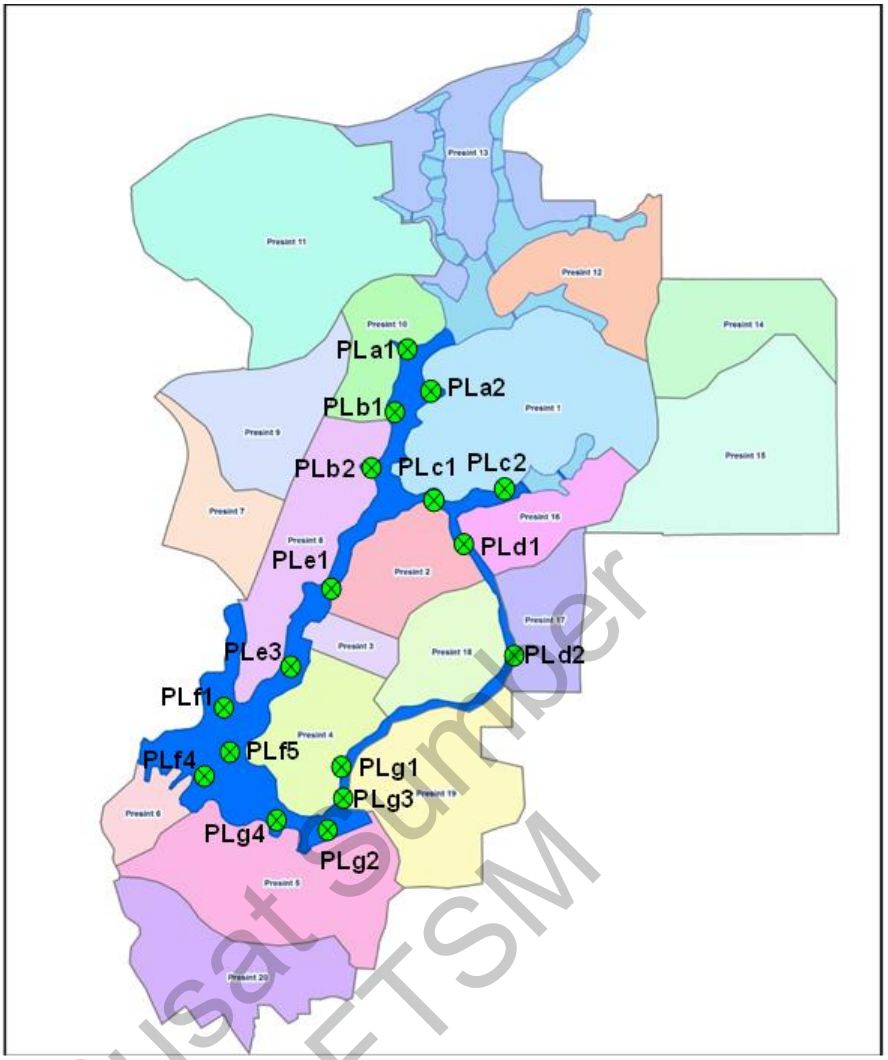
Kawasan Takungan	Paras Air	Luas Kawasan	Isipadu Simpanan	Purata Dalam	Purata Air Masuk	Purata Tempoh Takungan
50.90 KM2	RL21.00 M	400 ha (4 K M2)	23.50 mil. M3	6.60 M	6.60 M	132 hari

Sumber : BASTW 2023

Tasik Putrajaya telah direka untuk memenuhi kelas IIB klasifikasi NWQS untuk tujuan rekreasi. Aktiviti pembangunan dan penggunaan tanah yang pesat di sekitar tasik yang telah mengakibatkan beban pencemaran yang tinggi dilepaskan ke kawasan tanah lembap. Ini telah memberi tekanan kepada sistem tasik dan gagal bertindak sebagai sistem penapisan sebelum air mengalir ke dalam tasik (Asmat et al. 2018).

Terdapat 17 buah stesen persampelan air seperti Rajah 1.1 bagi memantau kualiti air Tasik Putrajaya. Data bagi kajian ini dikumpul secara bulanan bagi tujuan pemantauan dan rujukan Bahagian Alam Sekitar, Tasik dan Wetland (BASTW). Bahagian ini bertanggungjawab dalam mengurus tasik dan wetland Putrajaya dari aspek perancangan dan pembangunan meliputi kawasan persiaran, pengurusan kawasan tadahan, kawasan persisiran, pembangunan kemudahan awam serta infrastruktur yang berkaitan dengan aktiviti tasik. Bahagian ini juga mengawasi operasi penyelenggaraan dan aktiviti di tasik dan wetland termasuk pengawasan kualiti air dan biologi, tumbuhan wetland, burung dan ikan (BASTW 2023).

Data IKA Tasik Putrajaya dipilih berdasarkan beberapa kriteria yang penting bagi mendapatkan hasil kajian yang memberi kesan yang signifikan. Sebagai contoh, sekiranya data tersebut tidak mempunyai fitur yang mencukupi bagi tujuan kajian, dikhuatiri hasil ujian dari model pembelajaran mesin yang bakal diperolehi kurang tepat bagi mencapai sasaran yang ditetapkan. Jumlah data yang diperolehi juga perlu melebihi 1000 baris bagi mendapatkan ketepatan yang lebih tinggi.



Rajah 1.1 Lokasi Tasik & Wetland Putrajaya

Sumber : BASTW 2023

Teknik pembelajaran mesin telah giat dijalankan di seluruh dunia dalam mengkaji kualiti air terutamanya air tasik. Berdasarkan kajian terkini terhadap Tasik Rawal di Pakistan menunjukkan pengiraan IKA secara konvensional bergantung pada data semasa dan pengiraan matematik berbanding algoritma pembelajaran mesin mempertimbangkan data sejarah dan *trend* untuk menentukan pengelasan kualiti air. Dengan menggunakan teknik pembelajaran mesin, seperti Pokok Keputusan, Jiran Terdekat-k, Regresi Logistik, *Multilayer Perceptron* (MLP) dan Teluk Naïve, kualiti air berjaya dikelaskan dengan ketepatan yang tinggi. Sebagai contoh, algoritma Pokok Keputusan mencapai ketepatan pengelasan sebanyak 99%. Pembelajaran mesin didapati membantu menangani batasan IKA konvensional sedia ada seperti terlalu bergantung terhadap fitur tertentu, lokasi dan kekerapan pengumpulan data yang boleh

menyebabkan *bias* dan ketidakpastian dalam penilaian kualiti air (Ahmed, Mumtaz & Zaidi 2021).

Kajian terhadap Tasik Taal di Filipina juga menggunakan algoritma pembelajaran mesin membolehkan penyelidik mengelaskan dan memahami proses alam sekitar dan pengaruh ekosistem manusia. Hasil kajian juga membantu mengekalkan integriti alam sekitar dan melindungi komuniti yang bergantung kepada sumber akuatik. Pembelajaran mesin menawarkan alat yang berkuasa untuk menganalisis sejumlah besar data kualiti air. Ia membolehkan pengelasan kualiti air yang tepat bagi tujuan Indeks Kualiti Air (IKA) dan Klasifikasi Kualiti Air (WQC) yang penting untuk menilai kesihatan ekosistem akuatik dan melindungi kesejahteraan masyarakat (Tanega, Fajardo & Limbago 2023).

1.3 Penyataan Masalah

Antara isu yang dihadapi oleh penganalisis data sebelum ini ialah ketepatan analisis data berdasarkan hubungan antara attibut masih menggunakan kaedah konvensional dan kurang efisyen. Indeks kualiti air kebiasaannya dianalisis dan dikira menggunakan teknik pengiraan menggunakan pelbagai formula yang panjang, mengambil masa yang terlalu lama dan sering kali melibatkan ralat pengiraan yang tidak disengajakan.

Kaedah pengelasan secara konvensional juga mengakibatkan ketepatan jangkaan yang diperolehi dari data dikhuatiri adalah rendah. Teknik pembelajaran mesin diperlukan dalam penilaian dan pengurusan kualiti air atas beberapa sebab. Antaranya kaedah konvensional penilaian kualiti air menggunakan ujian makmal dan prosedur statistik memakan masa yang agak lama, melibatkan kos yang tinggi dan tidak berkesan untuk pemantauan masa nyata. Algoritma pembelajaran mesin mampu bertindak mengautomasikan penciptaan model analisis, mengesan corak dan membuat keputusan secara automatik, menjadikan penilaian kualiti air lebih praktikal dan kos efektif. Selain itu, teknik pembelajaran mesin mempunyai keupayaan untuk mengurangkan masa pengiraan dan kesilapan dalam klasifikasi kualiti air, pengelasan fitur dan pengelasan indeks kualiti air. Algoritma ini telah terbukti berkesan dalam memodelkan hubungan kompleks antara pembolehubah dan memberikan pengelasan yang tepat untuk fitur kualiti air. Dengan menggunakan model pembelajaran mesin,

pengurusan kualiti air dan ketepatan pengelasan indeks kualiti air dapat dipertingkatkan (Shamsuddin, Othman & Sani 2022). Model pembelajaran mesin yang terbaik bergantung kepada data dan prestasi algoritma pembelajaran. Algoritma pembelajaran yang terbaik kemudiannya dilatih menggunakan data sebenar yang dikumpulkan dilengkapi pengetahuan tentang aplikasi yang digunakan sebelum sistem dapat membantu mengeluarkan keputusan yang bijak (Sarker 2021).

Masalah utama dalam pembelajaran mesin adalah bahawa prestasi model dipengaruhi oleh pemilihan hiperparameter yang tepat. Hiperparameter menentukan struktur dan fungsi model, seperti kadar pembelajaran, bilangan lapisan, dan bilangan nod dalam setiap lapisan. Tanpa penalaan hiperparameter yang betul, model mungkin mengalami masalah seperti overfitting atau underfitting, yang mengakibatkan prestasi yang tidak konsisten atau optimum (Suwadi et al. 2022). Oleh itu, penalaan hiperparameter adalah penting untuk memastikan model pembelajaran mesin mencapai prestasi yang maksimum dan memberikan ramalan yang tepat.

Kajian menggunakan model RF untuk ramalan kualiti air Sungai Godavari, India mendapati risiko *overfitting* dapat dikurangkan melalui pemilihan fitur serta membantu membahagikan pokok-pokok dan menangani outliers dengan baik (Satish et al. 2024). Dalam kajian lain terhadap Sistem Pengairan An Kim Hai di utara Vietnam mendapati gabungan kaedah pemilihan fitur dan model terbaik RF sebagai pilihan efektif untuk mengira IKA dengan memperolehi prestasi terbaik dengan pengurangan parameter input (Lap et al. 2023). Oleh yang demikian, kaedah pemilihan fitur didapati sesuai untuk mengurangkan risiko *overfitting* di dalam kajian ini bagi mencapai prestasi terbaik.

Berdasarkan semakan, kajian terhadap kualiti air tasik Putrajaya secara komprehensif dengan menggunapakai model pengelasan pembelajaran mesin masih terhad. Oleh yang demikian, kajian pembangunan model pengelasan pembelajaran mesin untuk mengelas kualiti air tasik Putrajaya adalah penting dan perlu. Secara tidak langsung fitur penting yang mempengaruhi prestasi pengelasan indeks kualiti air tasik Putrajaya dapat dikenalpasti.

1.4 Objektif Kajian

Kajian ini secara khususnya dilaksanakan untuk mencapai objektif-objektif seperti berikut:

1. Mencadangkan model pengelasan dan mengenalpasti model yang terbaik untuk mengelas indeks bacaan kualiti air tasik Putrajaya.
2. Meningkatkan prestasi model pengelasan indeks kualiti air tasik melalui penalaan hiperparameter semasa proses latihan
3. Mengenalpasti fitur penting yang mempengaruhi prestasi pengelasan indeks kualiti air tasik.

1.5 Persoalan Kajian

Bagi mencapai objektif di atas, berikut adalah persoalan kajian:

1. Apakah model pengelasan terbaik yang digunakan bagi tujuan pengelasan indeks kualiti air tasik Putrajaya?
2. Apakah nilai hiperparameter yang digunakan untuk meningkatkan model pengelasan indeks kualiti air tasik Putrajaya?
3. Apakah fitur penting yang mempengaruhi prestasi pengelasan indeks kualiti air tasik Putrajaya?

1.6 Skop Kajian

Sepanjang proses pemilihan projek berjalan, beberapa halangan terpaksa diatasi sebelum keputusan dibuat. Antara halangan yang perlu dilalui ialah soal selidik bagi pemilihan data dan tajuk. Terdapat juga beberapa proses yang perlu dilalui bagi mendapatkan set data. Proses soal selidik telah dijalankan selama enam bulan bermula Januari sehingga Jun 2023 bagi menilai dan mengkaji latar belakang setiap data yang terdapat di dalam organisasi pilihan iaitu Perbadanan Putrajaya.

Sejumlah kakitangan Perbadanan Putrajaya dari pelbagai latarbelakang telah ditemubual bagi mendapatkan data yang sesuai bagi kajian ini. Akhirnya set data kualiti air tasik Putrajaya telah dipilih berdasarkan jumlah kolum atau fitur yang mencukupi bagi tujuan kajian. Setelah set data tersebut dipilih, terdapat beberapa proses perlu dibuat seperti permohonan penggunaan data bagi tujuan kajian.

Kelulusan perlu diperolehi daripada Naib Presiden Jabatan Perancangan Bandar, Perbadanan Putrajaya melalui surat rasmi dengan menyatakan tujuan serta jumlah data yang diperlukan. Perbincangan bersama pemilik set data tersebut iaitu pihak Seksyen Pengurusan Ekohidrologi Tasik dan Kawasan Tadahan, Bahagian Alam Sekitar, Tasik dan Wetland perlu dibuat secara berkala bagi mendapatkan khidmat nasihat berdasarkan kepakaran dan pengalaman yang dimiliki.

Kajian ini menumpukan kepada data yang diperolehi daripada Bahagian Alam Sekitar, Tasik dan Wetland (BASTW). Set data ini mengandungi data-data nombor (*numerical*) dan kategori kelas indeks air terdiri dari I dan II (binari) seperti yang telah ditetapkan mengikut piawaian kebangsaan di Malaysia. Penilaian status kualiti air melibatkan enam fitur utama iaitu DO, BOD, COD, SS, pH dan NH3N pada 17 buah stesen pensampelan air yang telah ditetapkan di sekitar Tasik Putrajaya. Kesemua fitur ini diukur menggunakan Standard Kualiti Air Kebangsaan (NWQS) dan Indeks Kualiti Air (IKA) oleh Jabatan Alam Sekitar (DOE) Malaysia untuk menilai tahap dan status kualiti air.

Set data ini dikumpulkan di Tasik Putrajaya dari Januari 2018 sehingga Disember 2022 mempunyai sebanyak 1020 rekod dengan 27 Fitur. Kajian ini dijalankan menggunakan bahasa pengaturcaraan *Python.* Setiap hasil daripada analisis dan keputusan kajian akan dinilai melalui proses penilaian untuk memilih model pengelasan yang terbaik. Skop kajian terdiri daripada beberapa perkara seperti berikut:

1. Kajian ini akan mengenalpasti fitur penting yang mempengaruhi prestasi pengelasan indeks kualiti air tasik dengan menggunakan data-data dari BASTW.
2. Eksperimen dijalankan dengan membangunkan model menggunakan tiga algoritma pembelajaran mesin iaitu Hutan Rawak (RF), Mesin Vektor Sokongan (SVM) dan Rangkaian Neural Buatan (ANN).

3. Model terbaik dipilih berdasarkan prestasi setiap model.

1.7 Kepentingan Kajian

Adalah diharapkan hasil kajian ini dapat digunakan sebagai batu loncatan bagi memperkenalkan teknologi Sains Data di dalam operasi harian serta kajian berkaitan data di Perbadanan Putrajaya.

Penggunaan Sains Data juga diharap akan mempercepatkan lagi proses membuat keputusan yang tepat bagi menjimatkan masa dan kos selain membantu dalam kelulusan aktiviti tasik yang melibatkan penduduk dan pengunjung Putrajaya sebagai pengguna.

Permulaan yang tidak mudah bagi memberi pendedahan kepada sebuah organisasi yang telah sekian lama menggunakan kaedah konvensional bagi menganalisa data, namun usaha perlu terus digiatkan antaranya melalui dapatan kajian ini.

1.8 Senarai Perisian Kajian

Penggunaan perisian dalam kajian adalah penting untuk memastikan sasaran berjaya dicapai. Perisian memainkan peranan penting dalam pelbagai aspek kajian, menjadikannya komponen penting dalam metodologi. Berikut adalah huraian berkaitan senarai perisian yang digunakan dalam kajian ini.

1.8.1 Microsoft Excel Workbook

Beberapa kajian telah meneroka penggunaan Microsoft Excel dalam konteks pembelajaran mesin. Dragan (2005) dan Zhang (2020) kedua-duanya membincangkan potensi Excel sebagai alat bahagian depan (*front-end*) untuk rangka kerja pembelajaran mesin, di mana Dragan memberi tumpuan secara khusus kepada Machine Learning Framework (MLF) sebuah perisian perlombongan data dan Zhang menggunakan Excel untuk mengajar konsep pembelajaran mesin.

1.8.2 Python

Python adalah bahasa yang dominan untuk pembelajaran mesin, kerana ia menawarkan pelbagai perpustakaan yang berkuasa dan API peringkat tinggi yang bersih (Raschka, Patterson & Nolet 2020). *Python* juga dianggap bahasa terbaik untuk mengautomasikan tugas pembelajaran mesin (Mshvidobadze 2021). *Python* telah muncul sebagai bahasa yang dominan untuk pembelajaran mesin kerana gabungan ciri-ciri yang unik seperti kesederhanaan iaitu sintaks *Python* amat ringkas dan jelas, menjadikannya mudah untuk belajar dan digunakan, walaupun untuk pengguna biasa yang tiada latar belakang pengaturcaraan. *Python* juga merupakan kod yang mudah dibaca dan kurang terdedah kepada kesilapan berbanding bahasa pengaturcaraan yang lain. Selain dari itu, *Python* mempunyai perpustakaan dan rangka kerja yang luas seperti *Scikit-learn, TensorFlow, PyTorch, NumPy* dan *Pandas*. Dari segi fleksibiliti dan kebolehskalaan (*scalability*), *Python* boleh disesuaikan dengan pelbagai tugas di luar pembelajaran mesin, termasuk analisis data, pembangunan web, dan pengkomputeran saintifik. Kebolehskalaannya membolehkan pengendalian set data yang besar dan model yang kompleks.

a. Scikit-Learn (SKLearn)

Scikit-learn adalah perpustakaan Python yang amat popular dengan menyediakan antara muka mesra pengguna untuk melaksanakan pelbagai algoritma pembelajaran mesin (Bisong 2019; Pedregosa et al. 2011). Ia direka supaya pengaturcaraan menjadi mudah, cekap, dan boleh diakses oleh pengguna biasa, dengan memberi tumpuan kepada kemudahan penggunaan, prestasi, dokumentasi, dan konsistensi API (Chary 2020; Pedregosa et al. 2011).

1.9 Organisasi Tesis

Tesis ini mengandungi lima (5) bab secara keseluruhan yang merangkumi:

1. Bab 1: Bab ini menerangkan tentang kajian yang bakal dilaksanakan merangkumi latar belakang kajian, penyataan masalah, objektif kajian, kepentingan kajian dan skop kajian secara keseluruhan.

2. Bab 2: Bab ini menghuraikan aspek utama berkaitan kajian yang akan dilaksanakan. Dalam bab ini huraian lebih terperinci berkenaan tinjauan terdahulu dan kajian-kajian yang telah dibuat berkaitan konsep dan teknik pengelasan terutamanya dalam masalah menentukan model pengelasan menggunakan algoritma pembelajaran mesin yang terbaik.

3. Bab 3: Bab ini merangkumi metodologi kajian yang menerangkan secara mendalam tentang kaedah kajian yang dilaksanakan terdiri daripada fasa data termasuk kaedah mengenalpasti ciri-ciri data, analisis struktur data dan pra-pemprosesan data. Selain dari itu, reka bentuk kajian juga dijelaskan dengan lebih lanjut di dalam bab ini.

4. Bab 4: Bab ini membentangkan hasil yang diperolehi daripada analisis pembangunan model pembelajaran mesin yang telah dibuat. Fokus turut diberikan dalam menghuraikan ketepatan pembangunan model serta Penalaan Hiperparameter dari analisis yang dibuat. Perbandingan prestasi berdasarkan ketepatan model turut dibuat dan dinilai.

5. Bab 5: Bab ini mengandungi kesimpulan dan rumusan tentang kajian yang telah dilaksanakan. Beberapa aspek lain juga diberi tumpuan seperti sumbangan kajian dan cadangan penambahbaikan yang boleh ditingkatkan pada masa akan datang.

BAB II

KAJIAN KESUSASTERAAN

2.1 Pengenalan

Bab ini menghuraikan tentang hasil kajian kesusasteraan yang melibatkan kajian-kajian terdahulu. Kajian ini memfokuskan terhadap penggunaan pembelajaran mesin terhadap indeks kualiti air (IKA) bagi data yang dikumpulkan dari tasik, sungai dan air bawah tanah di Malaysia dan di luar negara. Model-model terdahulu yang telah dibangunkan perlu dikaji dan dianalisa untuk menilai keberkesanan model pengelasan yang telah siap dibangunkan. Perbandingan antara model perlu diambilkira sebagai salah satu kaedah bagi mendapatkan ketepatan pembangunan model dari pelbagai perspektif dan dimensi. Model-model pengelasan berasaskan pembelajaran mesin menghasilkan keupayaan tersendiri dalam menjana pengelasan indeks kualiti air (IKA) untuk tempoh beberapa tahun akan datang. Kesesuaian model-model tersebut dinilai berdasarkan data dan fitur indeks kualiti air. Bab ini mengandungi tujuh (7) bahagian iaitu: (i) Teknik Pengelasan dan Regresi Dalam Penentuan Indeks Kualiti Air; (ii) Pengelasan Indeks Kualiti Air Menggunakan Teknik Pembelajaran Mesin; (iii) Pendekatan Pembelajaran Mesin; (iv) Metrik Prestasi; (v) Analisis Korelasi Data; (vi) Pemilihan Fitur (Feature Selection); (vii) Analisis SHAP.

2.2 Pengelasan Indeks Kualiti Air Menggunakan Pendekatan Pembelajaran Mesin

Kaedah pengelasan dan regresi telah digunakan untuk mengkaji data kualiti air samada dari sumber air sungai, tasik atau air bawah tanah. Indeks kualiti air berbeza mengikut negara, lokasi dan cuaca. Daripada keseluruhan 52 kajian kesusasteraan yang dibuat, sejumlah 26 kajian melibatkan penggunaan kaedah pengelasan bagi

mengkaji IKA, 17 kajian menggunakan kaedah regresi dan sembilan kajian menggabungkan kedua-dua kaedah iaitu pengelasan dan regresi.

Daripada kajian yang dibuat, majoriti kaedah pengelasan menggunakan ukuran dari segi ketepatan, kejituan, *recall* dan Skor-F1 bagi menilai setiap model yang digunakan. Selain itu, beberapa kaedah lain yang telah dibangunkan untuk meramal kualiti air seperti regresi dan gabungan (pengelasan dan regresi). Kaedah regresi diukur ketepatan menggunakan Ralat Mutlak Min (MAE), Ralat Min Kuasa Dua (MSE), Ralat Punca Min Kuasa Dua (RMSE), Ralat Peratusan Mutlak Min (MAPE), R-kuasa dua (R2) dan R-kuasa dua diselaraskan manakala kaedah gabungan menggunakan kedua-dua ukuran bagi pengelasan dan regresi.

Antara model yang digunakan di dalam kajian-kajian tersebut adalah Hutan Rawak (RF), Mesin Vektor Sokongan (SVM), Pokok Keputusan (DT), Rangkaian Neural Buatan (ANN), Perceptron Berbilang Lapisan (MLP), XGBoost, Regresi Linear (LR), Peningkatan Kecerunan (Gradient Boosting), Memori Jangka Pendek-Panjang (LSTM), Mesin Peningkatan Kecerunan Ringan (LGBM), Regresi Vektor Sokongan (SVR), Rangkaian Neural Fungsi Asas Jejari (RBFNN), Regresi proses Gaussian (GPR), K-Jiran Terdekat (KNN), KStar (Lazy), Teluk Naïve (NB), Regresi Rabung (RR), Rangkaian Neural Kabur (FNN), Pokok Rawak Ekstrem (ERT) / Extra Trees dan Perceptron Berbilang Lapisan- K-Jiran Terdekat (MLP-KNN).

Jadual 2.1 menunjukkan jumlah model terbaik menggunakan kaedah pengelasan dari keseluruhan kajian kesusasteraan yang telah dibuat.

Jadual 2.1 Jadual Model Terbaik

Model Terbaik	**Jumlah**
Hutan Rawak (RF)	7
Mesin Vektor Sokongan (SVM)	7
Pokok Keputusan (DT)	6
Rangkaian Neural Buatan (ANN)	6
Peningkatan Kecerunan (GB)	4
Memori Jangka Panjang-Pendek (LSTM)	1
Jiran Terdekat-K (KNN)	1

bersambung…

…sambungan

Regresi Linear (LR)	1
Lazy K-Star	1
Pokok Rawak Extreme (ERT)	1

Tanega, Fajardo dan Limbago (2023) turut memilih RF sebagai algoritma pembelajaran mesin dalam mengelaskan dengan tepat kualiti air Tasik Taal di Filipina. Algoritma Hutan Rawak mencapai kadar ketepatan tertinggi sebanyak 95.0% berbanding model lain yang diuji. Ini menunjukkan bahawa algoritma pembelajaran mesin amat berharga dalam memantau dan mengklasifikasikan kualiti air.

Anbuchezhian, Venkataraman dan Kumuthavalli (2018) menggunakan lima algoritma pengelasan seperti Teluk Naif (NB), Pokok Keputusan (DT), Jiran Terdekat–k (KNN), SVM dan RF untuk mengelaskan kelas kualiti air berdasarkan pelbagai fitur seperti suhu, oksigen terlarut, pH, kekonduksian dan lain-lain. Eksperimen yang dijalankan menggunakan set data sebenar dan sintetik menunjukkan bahawa pengelas RF mencapai hasil yang lebih baik berbanding pengelas lain. Penulis juga menekankan keperluan untuk sistem automatik yang memudahkan pengiraan IKA dan memperluaskan aplikasinya.

Kuthe et al. (2022) menggunakan kaedah pengelasan dan regresi di dalam kajian dan mendapati model regresi menunjukkan trend dan korelasi yang konsisten antara satu sama lain. RF melakukan yang terbaik berbanding model lain. ANN mempunyai prestasi yang agak rendah dan terdedah kepada overfitting. SVM digunakan sebagai model garis dasar dan model Jiran Terdekat-k (KNN) berjaya digunakan untuk imputasi data. Kajian ini menyoroti kepentingan pendekatan berasaskan pembelajaran mesin untuk analisis kualiti air yang cekap dan mengelaskan corak kualiti air.

Abirami, Radhakrishna dan Venkatesan (2023) memilih algoritma pengelas RF sebagai model prestasi terbaik untuk mengelaskan kualiti air dengan mencapai skor ketepatan 91.97%. Model ini kemudiannya digunakan dalam pembangunan aplikasi web di mana pengguna boleh memasukkan nilai fitur untuk meramalkan kelas kualiti air. RF dipilih sebagai salah satu algoritma kerana memerlukan sedikit masa

latihan dan mengelaskan output dengan ketepatan yang tinggi, ia juga dapat mengekalkan ketepatan walaupun sebahagian besar data telah hilang.

Qianqian dan Ying (2015) juga mendapati model RF sesuai untuk menilai kualiti air Tasik Chaohu. Hasil penilaian menunjukkan bahawa model berdasarkan algoritma RF adalah berkesan dan boleh membuat penilaian yang munasabah terhadap kualiti air Tasik Chaohu. Algoritma RF mempunyai kelebihan seperti ketepatan tinggi, keupayaan klasifikasi yang kuat, dan kestabilan yang lebih baik berbanding algoritma lain. Model RF juga menyelesaikan masalah seperti keteguhan dan pembelajaran berlebihan yang dihadapi oleh algoritma lain. Secara keseluruhan, penggunaan algoritma RF dalam analisis kualiti air semakin meluas dan mempunyai prospek penggunaan lebih jauh.

Shamsuddin, Othman dan Sani (2022) mengetengahkan potensi model pembelajaran mesin, terutamanya SVM, untuk mengelaskan Indeks Kualiti Air (WQI) dengan ketepatan yang tinggi, menyumbang kepada peningkatan pengurusan kualiti air. Ini boleh membawa kepada pengurusan sumber air yang lebih baik, mengurangkan kos, dan menjimatkan masa dalam proses pemantauan dan penilaian. Di samping itu, kajian ini mengesyorkan penambahbaikan masa depan seperti menggunakan set data yang lebih besar untuk mengelaskan kelas kualiti air, menjalankan penyiasatan yang lebih terperinci mengenai hubungan antara langkah-langkah kualiti air, dan meneroka algoritma pemilihan fitur untuk menguji ketepatan model pengelasan berdasarkan senario fitur kualiti air yang pelbagai. Keputusan kajian dan penilaian model-model yang dibentangkan berpotensi mengubah cara kelas kualiti air dikelaskan dan dipantau, memberikan pandangan berharga bagi pembuat dasar, pakar alam sekitar, dan orang awam untuk meningkatkan pengurusan kualiti air.

Ebron et al. (2020) membangunkan sistem pemantauan kualiti air berasaskan web, yang menggabungkan model dan membolehkan pemantauan kualiti air masa nyata. Ini menekankan kepentingan menggunakan teknik pemodelan pengiraan dan berangka untuk menilai dan menguruskan sumber kualiti air. Kajian ini menekankan kepentingan menggunakan model pengelasan dan alat visualisasi data untuk mengelaskan kualiti air dan memantau sumber air dengan berkesan. Kajian juga

membincangkan cabaran mengendalikan data besar dan keperluan untuk teknik pemodelan yang tepat. Model berprestasi terbaik dalam meramalkan kelas kualiti air ialah Mesin Vektor Sokongan (SVM) dalam kebanyakan fitur.

Derdour et al. (2022) mengesahkan kepentingan menilai dan mengelaskan kualiti air dengan tepat untuk memastikan penggunaan air yang sesuai dan menentukan remedi atau langkah berjaga-jaga yang betul. Ini menyerlahkan keperluan untuk model yang boleh dipercayai yang dapat menilai kualiti air dengan berkesan dan menyumbang kepada perlindungan dan kemampanan sumber air. Model pembelajaran mesin, seperti SVM, boleh menjadi alat yang berkuasa untuk mengelaskan kualiti air minuman pada skala yang lebih besar juga menunjukkan bahawa model-model ini berpotensi untuk menyokong kawalan kualiti air yang cekap dan membolehkan pemakluman keputusan pengurusan sumber air di kawasan gersang.

Jalal dan Ezzedine (2019) menjalankan analisis prestasi algoritma pembelajaran mesin untuk sistem pemantauan kualiti air menggunakan set data sebenar dari stesen rawatan air Tunisia. Penyelidikan ini bertujuan untuk mencadangkan sistem masa nyata yang memantau dan mengelaskan kualiti air berdasarkan fitur fizikokimia dan mikrobiologi. Keputusan menunjukkan bahawa SVM adalah algoritma yang paling sesuai untuk sistem pemantauan kualiti air. Kajian ini menyimpulkan bahawa teknik pembelajaran mesin memberikan pendekatan yang lebih dipercayai dan cekap untuk pemantauan kualiti air, terutama di negara-negara membangun. Kajian ini menyerlahkan potensinya untuk meningkatkan pemantauan kualiti air di negara-negara yang mempunyai sumber terhad mengurangkan penyakit bawaan air.

Li et al. (2013) berpendapat model pengelasan SVM dilakukan lebih baik daripada kaedah statistik tradisional untuk mengklasifikasikan fitur kualiti air. Model komprehensif ini berguna untuk membuat keputusan pengurusan dan dasar berdasarkan sumber air. Walau bagaimanapun, model ini mempunyai batasan dalam menganalisis taburan spatial dan temporal fitur-fitur kualiti air di seluruh kawasan tadahan air.

Danades et al. (2017) membandingkan dua algoritma, K-Jiran Terdekat (KNN) dan Mesin Vektor Sokongan (SVM), untuk mengklasifikasikan status kualiti air. Algoritma SVM mempunyai kadar ketepatan yang lebih tinggi sebanyak 92.40% menggunakan kernel linear, manakala algoritma KNN mempunyai ketepatan purata hanya 71.28% pada K = 7. Ini menunjukkan bahawa algoritma SVM lebih tepat dalam menentukan status kualiti air. Hasil kajian ini boleh digunakan untuk membangunkan sistem yang lebih cekap untuk mengelaskan kualiti air, yang akan menjadi peningkatan berbanding kaedah pengiraan manual semasa. Kajian ini mempunyai implikasi yang signifikan dalam menguruskan kualiti sumber air dan memastikan penggunaannya yang mampan. Data daripada Badan Pusat Statistik (BPS) menunjukkan bahawa, kira-kira 3% isi rumah di Indonesia menjadikan sungai itu sebagai sumber minuman.

Ahmed et al. (2019) menggunakan kaedah regresi dan pengelasan di dalam mengkaji potensi menggunakan algoritma pembelajaran mesin yang diselia untuk pengelasan kualiti air yang cekap. Kajian ini melihat batasan kaedah tradisional untuk menganggarkan kualiti air yang mahal dan memakan masa. Metodologi yang dicadangkan dalam kajian ini menggunakan empat fitur input (suhu, kekeruhan, pH, dan jumlah pepejal terlarut) untuk meramalkan IKA dan Kelas Kualiti Air (WQC) melalui pelbagai algoritma pembelajaran mesin. Keputusan menunjukkan bahawa algoritma peningkatan kecerunan dan regresi polinomial adalah cekap dalam mengelaskan IKA, manakala algoritma perceptron berbilang lapisan (MLP), salah satu daripada pecahan ANN, berkesan dalam mengklasifikasikan Kelas Kualiti Air (WQC). Para penyelidik mencadangkan supaya mengintegrasikan penemuan penyelidikan ini ke dalam sistem pemantauan dalam talian berasaskan Internet Benda (IoT) berskala besar yang akan membolehkan pemantauan masa nyata dan pengenalpastian air berkualiti rendah.

Abuzir dan Abuzir (2022) menerapkan algoritma pembelajaran mesin dalam meramalkan klasifikasi kualiti air. Kajian ini memberi tumpuan kepada tiga algoritma iaitu J48, Naïve Bayes dan Perceptron Berbilang Lapisan (MLP) dan menilai ketepatan mereka dalam mengklasifikasikan kualiti air. Algoritma MLP didapati mempunyai ketepatan tertinggi berbanding algoritma lain. Di samping itu, kajian ini

menyoroti kegunaan pembelajaran mesin dalam menganalisis dan mengelaskan kualiti air dan mencadangkan arah penyelidikan masa depan seperti mengintegrasikan teknologi IoT dan meneroka algoritma lain seperti pembelajaran ensemble, SVM, dan K-NN. Secara keseluruhan, kajian ini menekankan tentang potensi pembelajaran mesin untuk analisis dan pengelasan kualiti air serta menawarkan kaedah yang lebih cepat dan lebih murah untuk mengesan pencemaran air.

Nair dan Vijaya (2022) membincangkan bagaimana algoritma pembelajaran mesin dapat mengesan anomali dalam data kualiti air. Dengan menganalisis corak sejarah dan korelasi antara pembolehubah, algoritma ini dapat mengenal pasti titik data yang tidak normal, membantu dalam pengesanan awal isu kualiti air yang berpotensi. Kajian juga melibatkan pengelasan dan ramalan kualiti air mendapati perceptron berbilang lapisan (MLP) mencapai keputusan tertinggi di dalam kedua-dua ujikaji.

Sulaiman et al. (2019) memberi tumpuan kepada Selat Melaka di Malaysia, yang menghadapi masalah pencemaran kerana lokasinya sebagai laluan perkapalan utama dan kepekatan pertanian, industri, dan perbandaran. Para penyelidik mengumpul data alam sekitar dari tiga lokasi sungai yang berbeza dan fitur yang digunakan seperti pH, jumlah pepejal terampai, oksigen terlarut, permintaan oksigen kimia, permintaan oksigen biologi, dan ammonia. Selepas latihan dan ujian rangkaian ANN, hasilnya menunjukkan klasifikasi ketepatan 80% dengan ralat kuasa dua min akar (RMSE) 0.468. Penemuan ini menunjukkan bahawa ANN dapat menyediakan cara yang cekap dan mudah untuk mengklasifikasikan kualiti air, yang penting untuk pentadbiran kualiti air yang berkesan untuk mencegah kesan negatif terhadap tamadun manusia.

Northep, Srijiranon dan Eiamkanitchat (2020) mengkaji teknik perlombongan data yang dapat digunakan dengan berkesan untuk mengklasifikasikan fitur kualiti air di sungai. Kajian ini memberi tumpuan kepada Sungai Wang di Thailand dan membincangkan penggunaan pelbagai kaedah dan model perlombongan data, seperti MLP-ANN, SVM, dan regresi polinomial evolusi untuk mengelaskan dan menilai kualiti air. Para penyelidik juga mencadangkan model yang dipanggil MLP-kNN,

yang menggabungkan algoritma perceptron berbilang lapisan (MLP) dan k-jiran terdekat (kNN) untuk klasifikasi. Kajian ini mementingkan klasifikasi kualiti air untuk memberi amaran kepada orang ramai mengenai masalah kualiti air dalam waktu yang sama dapat membawa kepada perubahan tingkah laku. MLP-kNN menunjukkan kadar klasifikasi dan skor F tertinggi.

Chou, Ho dan Hoang (2018) berpendapat bahawa pendekatan pembelajaran mesin, khususnya model kecerdasan buatan (AI) boleh digunakan untuk mengelaskan Indeks Keadaan Tropik Carlson (CTSI) untuk penilaian kualiti air di takungan. CTSI adalah metrik yang digunakan untuk menilai keadaan eutrophication, dengan mengambil kira faktor-faktor seperti kepekatan klorofil-a, kepekatan fosforus total, dan kedalaman Secchi. Dengan membandingkan empat model AI tunggal, empat model ensemble, dan model regresi metaheu, didapati Rangkaian Neural Buatan (ANN) mencapai ketepatan tertinggi dari segi langkah-langkah prestasi seperti R, RMSE, MAE, dan MAPE. Ia mempunyai nilai R 0.718, RMSE 3.941, MAE 3.131, dan MAPE 6.786. Selain itu, ia juga memperoleh nilai SI keseluruhan terendah iaitu 0.250, menunjukkan bahawa ia adalah model yang paling tepat untuk mengelaskan indeks kualiti air dalam takungan ristik hibrid untuk menentukan kaedah yang paling tepat untuk mengelaskan CTSI.

Selain itu, terdapat pelbagai kajian terdahulu menggunakan teknik pembelajaran mesin dalam pengelasan kualiti air. Jadual 2.2 menunjukkan ringkasan kajian penyelidikan terdahulu menggunakan teknik pengelasan dan regresi dengan menggunapakai pelbagai algoritma.

Jadual 2.2 Kajian-Kajian Terdahulu Menggunakan Teknik Pembelajaran Mesin Terhadap Kualiti Air

Penulis, Tahun	**Objektif Kajian**	**Model Perbandingan**	**Data**	**Dapatan**
Illa Iza Suhana Shamsuddin, Zalinda Othman dan Nor Samsiah Sani (2022)	Menilai prestasi tiga algoritma pembelajaran mesin dalam mengelaskan indeks kualiti air Lembangan Sungai Langat.	Rangkaian Neural Buatan, Pokok Keputusan, dan Mesin Vektor Sokongan.	Set data kualiti air mentah Lembangan Sungai Langat yang terdiri daripada 560 rekod dan 46 fitur.	**Model SVM** mencapai kadar ketepatan (Accuracy) tertinggi sebanyak 96.35%, ketepatan (precision) 91.97% dan *recall 84.89*%.
Nur Afyfah Suwadi, Morched Derbali, Nor Samsiah Sani, Meng Chun Lam, Haslina Arshad, Imran Khan dan Ki-Il Kim (2022)	Mengelaskan fitur-fitur kualiti air menggunakan teknik pembelajaran mesin dan mengkaji fitur-fitur penting indeks kualiti air (WQI) melalui pengelasan pelbagai petunjuk	Rangkaian Neural Buatan (ANN), Mesin Vektor Sokongan (SVM), Hutan Rawak (RF) dan Teluk Naif (Naïve Bayes)	Set data dari Lembangan Sungai Langat di Selangor, Malaysia terdiri daripada 29 fitur dengan 907 sampel	**Model Hutan Rawak** yang dioptimumkan dengan fitur IKA mencapai prestasi tertinggi, mempunyai ketepatan yang lebih tinggi menggunakan ujian pecahan 70%-30% dalam DOE-WQI dengan data enam fitur memperolehi ketepatan 95.64% dan *recall* dan ketepatan hampir kepada 1.00.
DN Khoi, NT Quan, DQ Linh, PTT Nhi dan NTD Thuy (2022)	Menilai prestasi 12 model pembelajaran mesin dalam mengelaskan indeks kualiti air (WQI) di Sungai La Buong di Vietnam.	Lima Algoritma berasaskan peningkatan (boosted) seperti Adaboost, GBM, HGBM, LightGBM, dan XGBoost. Tiga algoritma berasaskan pokok keputusan seperti DT, ExT, dan RF. Empat algoritma berasaskan ANN (MLP, RBF, DFNN, dan CNN).	Set data kualiti air dari 4 stesen pemantauan di Sungai La Buong di Vietnam. Data ini merangkumi tempoh 2010 hingga 2017 untuk mengira indeks kualiti air (WQI).	**Model XGBoost** mencapai R2 0.989 dan RMSE 0.107 dalam proses ujian, menjadi algoritma ML yang paling sesuai di kawasan kajian diikuti oleh GBM, LightGBM, RF, ExT, MLP, CNN, DT, DFNN, AdaBoost, HGBM, dan RBF
Nur Hanisah Abdul Malek, Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir dan Norshahida Shaadan (2022)	Meramalkan pengelasan kualiti air di Lembangan Sungai Kelantan di Malaysia menggunakan teknik pembelajaran mesin.	Pokok Keputusan, Rangkaian Neural Buatan, Jiran K-Terdekat, Teluk Naif, Mesin Vektor Sokongan, Hutan Rawak (RF), dan Peningkatan Kecerunan.	Set data yang mempunyai 13 fitur fizikal dan kimia kualiti air dari tahun 2005 hingga 2020.	**Model *Ensemble Gradient Boosting*** dengan kadar pembelajaran 0.1 mempamerkan prestasi pengelasan terbaik berbanding algoritma laindengan ketepatan tertinggi (94.90%), sensitiviti (80.00%) dan ukuran f (86.49%), dengan ralat pengelasan terendah.

bersambung...

…sambungan

Shuo Wang, Hui Peng dan Shengkang Liang (2022)	Menilai keberkesanan empat model pembelajaran mesin dalam meramalkan nitrogen ammonia estuarine (NH4 +-N) di muara Sungai Xiaoqing di China.	Regresi Linear Berganda (MLR), Rangkaian Saraf Buatan (ANN), Hutan Rawak (RF), dan Peningkatan Kecerunan Melampau (XGBoost).	Set data yang digunakan berkaitan dengan pengelasan nitrogen ammonia bulanan (NH4 + -N) di muara Sungai Xiaoqing di China	**Model XGBoost** memberikan hasil terbaik berbanding model lain.
Monika Kulisz, Justyna Kujawska, Bartosz Przysucha dan Wojciech Cel. (2021)	Mengoptimumkan model rangkaian neural buatan (ANN) untuk meramalkan indeks kualiti air (WQI), menentukan pembolehubah utama dan membangunkan model rangkaian saraf yang dapat meramalkan kualiti air bawah tanah bagi rawatan dan pengurusan air di kawasan perindustrian di Lublin, Poland.	Model ini dicipta menggunakan hanya lima fitur input: EC, pH, Ca, Mg, dan K. Data input optimum untuk pemodelan rangkaian saraf buatan (ANN) dipilih berdasarkan hasil regresi berbilang (MRP).	Set data yang digunakan dalam kajian ini terdiri daripada sampel air yang dikumpulkan dari 19 telaga yang terletak di sekitar tapak pengekstrakan gas syal di Poland.	**Model ANN** dapat meramalkan kualiti air bawah tanah dengan tahap ketepatan yang diingini. Nilai WQI yang diramalkan ANN sangat kuat dan positif setelah dibandingkan dengan nilai WQI sebenar dengan bacaan RMSE = 0.651258, R = 0.9992 and R2 = 0.9984.
Xiaoping Wang, Fei Zhang dan Jianli Ding (2017)	Mewujudkan indeks kualiti air (WQI) untuk penilaian dan pengelasan kualiti air permukaan di kawasan gersang, membangunkan peta WQI menggunakan GIS, mengekstrak gelang gelombang sensitif dan mewujudkan model indeks spektrum untuk menganggarkan WQI, menilai ketepatan model anggaran berbanding nilai WQI.	Model regresi vektor sokongan pengoptimuman zarah zarah (POS-SVR) dan model regresi berwajaran geografi (GWR)	Set data yang digunakan dalam kajian ini termasuk 48 tapak pensampelan dan 20 fitur kualiti air untuk pemantauan dan analisis di Perairan Tasik Ebinur di China.	**Model POS-SVR** didapati lebih baik berbanding GWR dengan bacaan R2 bersamaan 0.92, RMSE bersamaan 58.4 dan Residual Prediction Deviation (RPD) bersamaan 2.81 dan kecerunan dengan lengkungan bersesuaian 0.97.

bersambung…

…sambungan

Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi dan Abbas Parsaie (2018)	Meramalkan komponen kualiti air di Sungai Tireh di Iran menggunakan teknik kecerdasan buatan.	Rangkaian saraf buatan (ANN), mesin vektor sokongan (SVM), dan kaedah kumpulan pengendalian data (GMDH)	Set data yang digunakan dalam kajian ini terdiri daripada komponen kualiti air Sungai Tireh di Iran.	**Model SVM** diambilkira sebagai model terbaik untuk meramalkan komponen kualiti air di Sungai Tireh di Iran.
Mohammed Hameed, Saadi Shartooh Sharqi, Zaher Mundher Yaseen, Haitham Abdulmohsin Afan, Aini Hussain dan Ahmed Elshafie (2016)	Mengkaji dan mengadaptasi hubungan WQI dengan pembolehubah kualiti air dalam persekitaran tropika (Malaysia)	Dua algoritma ANN yang berbeza, iaitu radial basis function neural network (RBFNN) dan model rangkaian saraf penyebaran belakang atau back propagation neural networks (BPNN)	Set data terdiri daripada 5233 sampel pembolehubah kualiti air dan indeks kualiti air (WQI) yang sepadan dari Januari 2001 hingga Oktober 2010 daripada stesen pemantauan di Lembangan Klang dan Lembangan Langat.	**Model rangkaian saraf fungsi asas radial (RBFNN)** mencapai R2 0.9872, RMSE 0.0157, dan NE 0.9871.
Ismail I Aminu (2022)	Membandingkan prestasi model yang berbeza dalam meramalkan WQI dan menilai ketepatan mereka menggunakan pelbagai metrik penilaian. Matlamatnya adalah untuk menentukan kaedah dan model yang paling sesuai untuk meramalkan kualiti air dan memberi panduan kepada penyelidik masa depan dalam bidang ini	Rangkaian saraf tiruan (ANN), sistem inferens neuro-kabur adaptif (ANFIS), Mesin Vektor Sokongan / regresi (SVM), kaedah linear dan statistik. Secara khususnya, model ANN seperti BPNN, FFNN, ENN, RBFEL. Model ANFIS seperti ANFIS-FCM, ANFIS-GP, dan ANFIS-SC.	Set data diperolehi daripada kertas penyelidikan yang mengkaji penggunaan model pembelajaran mesin dalam meramalkan Indeks Kualiti Air (WQI) sepanjang 1 dekad meliputi pelbagai negara seperti India, Malaysia, Pakistan, Mexico, Iraq, Iran, China, Ethiopia, Sudan, dan Afrika Selatan.	**Tidak ada satu model boleh dianggap terbaik** untuk meramalkan Indeks Kualiti Air (WQI). **Pilihan kaedah atau model yang paling sesuai bergantung kepada rantau atau negara** tertentu di mana kajian dijalankan. Di samping itu, prestasi pelbagai model telah dinilai menggunakan metrik penilaian yang berbeza, seperti R2, RMSE, MAE, NSE, dan lain-lain.

bersambung…

…sambungan

B. Malarkodi, P. Tarakeswari dan Jobin Tomy (2021)	Mencadangkan strategi pembelajaran mendalam untuk pemantauan kualiti air menggunakan teknologi IoT.	Model Memori Jangka Pendek Panjang (LSTM), Rangkaian Neural Buatan (ANN), dan model Purata Pergerakan Bersepadu Autoregressif (ARIMA)	Data yang dikumpul dari sensor disimpan di pelayan dan boleh digunakan untuk menganalisis kualiti air.	**Model LSTM** (Memori Jangka Pendek Panjang) dengan R2 0.984, MSE 0.0068 dan RMSE 0.0782.
Md Galal Uddin, Stephen Nash, Azizur Rahman dan Agnieszka I. Olbert (2022)	Menangani kekurangan ketepatan dan kebolehpercayaan dalam model indeks kualiti air (WQI) dengan menggunakan teknik pembelajaran mesin.	Pokok keputusan (DTs), model regresi linear (LR), model regresi proses Gaussian (GPR), mesin vektor sokongan (SVM), rangkaian neural (NN), dan ensembles model pokok (*Ensembles of Tree*)	Data kualiti air yang dikumpulkan di Cork Harbour, yang terletak di pantai selatan Ireland.	**Model regresi proses Gaussian (GPR)** menggunakan *aggregation function* WQM - *weighted quadratic mean* dengan RMSE 4.094, MAE 2.855 dan MSE 16.763.
Savita Mohurle dan Manoj Devare (2019)	Menggunakan algoritma pengecaman corak dan pembelajaran mesin, khususnya algoritma pengelas KNN, untuk meramalkan dan mengukur ketepatan fitur kualiti air dari segi indeks kualiti air minuman.	Pengelas jiran terdekat K (KNN) dan analisis komponen utama (PCA). Algoritma pengelas KNN digunakan untuk masalah pengelasan dan regresi, manakala PCA digunakan untuk mengubah data dimensi tinggi menjadi ruang dimensi rendah.	Data sampel untuk menganalisis kualiti air minuman. Ia terdiri daripada 20 fitur yang berkaitan dengan kualiti air diperolehi dari sungai, tasik dan empangan.	Hanya menumpukan algoritma pengecaman corak dan pembelajaran mesin, khususnya **Model jiran terdekat K (KNN)** dengan ketepatan 98.8%
Preethi Nanjundan, Jossy P George dan Aabhas Vij (2022)	Membentangkan kaedah yang dipercayai dan kos efektif untuk meramalkan kualiti air menggunakan model pembelajaran mesin yang diselia bertujuan menangani isu kemerosotan kualiti air dengan mencadangkan metodologi yang menganggarkan IKA	Lapan regresi dan sepuluh teknik pengelasan (tidak dinyatakan nama)	Data untuk kajian ini berasal dari ENVIS (Kerajaan India) yang mengandungi 950 sampel yang merangkumi 3 tahun (2017-2019).	**Model Regresi Linear (LR)** mencapai Ralat Mutlak Min (MAE) 0.495, Ralat Min Kuasa Dua (MSE) 0.27, RMSE = 0.52 dan R Squared = 0.28.

bersambung…

…sambungan

Mehreen Ahmed, Rafia Mumtaz dan Zahid Anwar	Mencadangkan dan membangunkan Indeks Kualiti Air Dipertingkat (EWQI) yang mengatasi batasan indeks kualiti air tradisional. EWQI bertujuan untuk menyediakan kaedah yang lebih dipercayai dan tepat untuk memantau kualiti air dengan mempertimbangkan pelbagai fitur.	CatB (CatBoost), LGBM (LightGBM), Hutan Rawak (RF), dan AdaB (AdaBoost).	Set data terdiri daripada pelbagai fitur yang dikumpulkan dari kawasan tadahan air Rawal bagi setiap bulan tengkujuh dari Julai 2018 hingga Ogos 2022.	**Model LightGBM (LGBM)** dengan ketetapatan 99%.
Mehreen Ahmed, Rafia Mumtaz dan Syed Mohammad Hassan Zaidi (2021)	Menganalisis indeks kualiti air dan teknik pembelajaran mesin untuk menilai pencemaran air. Kajian ini bertujuan untuk menilai dan memperbaiki indeks kualiti air sedia ada untuk menghapuskan ketidakpastian dan potensi manipulasi kelas kualiti air	Pokok Keputusan (DT), Jiran K-Terdekat (KNN), Regresi Logistik (LogR), Multilayer Perceptron (MLP), dan Naive Bayes (NB).	Set data pertama dari Empangan Rawal di Pakistan dari Jun hingga Disember 2019 menggunakan sensor Internet of Things (IoT). Set data kedua adalah 1,114 sampel dari tahun 2013 hingga 2018 menggunakan persampelan grab (GIS).	**Model Pokok Keputusan (DT)** dengan ketepatan 99%.
Michelle C. Tanega, Arnel Fajardo dan Jomel S. Limbago (2023)	Menganalisis kualiti air di Tasik Taal, Filipina menggunakan algoritma pengelasan pembelajaran mesin	Hutan Rawak, Pokok Keputusan, dan Mesin Vektor Sokongan.	Set data terdiri daripada 299 rekod yang dikumpul dari lima stesen pemantauan yang terletak di bahagian yang berlainan di Tasik Taal dari Januari 2018 hingga Disember 2022.	**Model Hutan Rawak (RF)** dengan ketepatan keseluruhan tertinggi iaitu 95.0% berbanding Pokok Keputusan 95.0% dan SVM 93.33%. Walaupun ketepatan RF dan DT sama, RF mendapat precision, *recall*, dan skor F1 tertinggi.

bersambung…

…sambungan

Salisu Yusuf Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz dan Azrul Amri Jamal (2015)	Mencadangkan model pengelasan yang sesuai untuk menilai kualiti air menggunakan algoritma pembelajaran mesin bagi mengklasifikasi kualiti air Sungai Kinta di Malaysia	Model Bayes menggunakan Teluk Naif (NB), Model Peraturan (Rules) menggunakan Peraturan Konjunktif (Conjunctive rule), Model Pokok menggunakan algoritma J48, Model Malas (Lazy) menggunakan Kstar dan Model Meta menggunakan Bagging.	Set data daripada Jabatan Alam Sekitar Malaysia, merupakan rekod bulanan dari tahun 2002 hingga 2006 terdiri daripada 135 baris dan 54 fitur.	**Model malas (Lazy)** menggunakan algoritma **KStar** mempunyai ketepatan yang paling cemerlang iaitu 86.67%
Neha Radhakrishnan dan Anju S Pillai (2020)	Menentukan model yang paling berkesan untuk menilai dan meramalkan kualiti air berdasarkan fitur tertentu.	Mesin Vektor Sokongan (SVM), Pokok Keputusan (DT), Teluk Naif (NB).	Set data yang dikumpulkan dari pelbagai wilayah di Uttar Pradesh, India. Set data pertama terdiri daripada 28 fitur kualiti air Sungai Narmada, manakala set data kedua mengandungi data kualiti air bersejarah dari lokasi tertentu di India dari tahun 2003 hingga 2014	**Model Pokok Keputusan (DT**) dengan accuracy 98.5%
Jonalyn G. Ebron, Rommel Ivan D. De Leon, Arviejhay D. Alejandro dan Basaron A. Amoranto (2020)	Membangunkan model pengiraan dan berangka untuk pengelasan kualiti air di Tasik Laguna de Baý, Filipina.	Regresi Linear Multivariate (MLR), Rangkaian Neural Buatan (ANN), Jiran Terdekat-K (kNN), dan Mesin Vektor Sokongan (SVM).	Set data terdiri daripada fitur kualiti air dari Tasik Laguna de Baý, Filipina antara 2009 hingga 2016 dan termasuk pengukuran dari 9 stesen pemantauan utama di tasik.	**Model Mesin Vektor Sokongan (SVM)** mendapat ketepatan paling tinggi.

bersambung…

…sambungan

Abdessamed Derdour, Antonio Jodar-Abellan, Miguel Ángel Pardo, Sherif S. M.Ghoneim, Enas E. Hussein (2022)	Membangunkan model pengelasan yang cekap dan mampan untuk Indeks Kualiti Air (WQI) menggunakan algoritma pembelajaran mesin	Pokok Keputusan, Jiran K-Terdekat, Analisis Diskriminant, Mesin Vektor Sokongan dan Pokok Ensemble	Set data dikumpulkan daripada 169 sampel air bawah tanah dari 12 majlis perbandaran di Wilayah Naâma	**Mesin Vektor Sokongan (SVM)** mengklasifikasikan kualiti air bawah tanah dengan ketepatan yang tinggi (95.4%) dengan data standard dan ketepatan yang lebih rendah (88.88%) untuk data mentah.
Elias Dritsas dan Maria Trigka (2023)	Menggunapakai pendekatan pembelajaran yang diselia untuk merekabentuk model pengelasan bagi mengenalpasti kesesuaian air untuk penggunaan harian atau kegunaan lain.	Teluk Naif (NB), Regresi Logistik (LR), Jiran terdekat–k (kNN), pengelas berasaskan pokok dan teknik ensemble.	Set data mengandungi 7986 rekod dan bilangan kolum ialah 20 bersama kelas sasaran (selamat)	Model pengelasan Menyusun (Stacking) (pengelas: **Hutan Rawak (RF)** dan Teluk Naif (NB), pengelas meta: LR) selepas SMOTE dengan pengesahan silang sepuluh kali ganda mengatasi model yang lain dengan Ketepatan dan *Recall* 98.1%, Ketepatan 100% dan AUC 99.9%.
Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan dan José García-Nieto (2019)	Meneroka algoritma pembelajaran mesin yang diselia untuk menganggarkan indeks kualiti air (WQI), yang merupakan indeks tunggal untuk menggambarkan kualiti umum air dan kelas kualiti air (WQC), yang merupakan kelas yang ditakrifkan berdasarkan WQI.	Lapan Regresi : Regresi Linear, Regresi Polinomial, Hutan Rawak, Peningkatan Kecerunan, SVM, Regresi Rabung, Regresi Lasso, Regresi Bersih Anjal. Sepuluh Pengelas : MLP, Guassian Naïve Bayes, Regresi Logistik, Keturunan Kecerunan Stokastik, KNN, Pokok Keputusan, Hutan Rawak, SVM, Pengelas Meningkatkan Kecerunan, Pengelas Bagging	Set data yang dikumpul daripada PCRWR mengandungi 663 sampel daripada 13 sumber yang berbeza dari Tasik Air Rawal, Pakistan yang dikumpul sepanjang 2009 hingga 2012 dan mengandungi 51 sampel dari setiap sumber dan 12 fitur.	**Model Regresi Polinomial** dengan tahap dua dan **Peningkatan Kecerunan** (Gradient Boosting) dengan kadar pembelajaran 0.1, mengatasi algoritma regresi lain dengan meramalkan WQI paling cekap, manakala **MLP** dengan konfigurasi (3, 7) mengatasi algoritma pengelasan lain dengan mengklasifikasikan WQC paling cekap.

bersambung…

…sambungan

S. K. Bhoi, C. Mallick dan C. R. Mohanty (2021)	Membangunkan model pembelajaran mesin yang dapat mengklasifikasikan kualiti air terusan Taladanda dengan tepat di Odisha, India.	Pokok Keputusan, Rangkaian Neural, k-NN (Jiran Terdekat-k), Teluk Naif, Mesin Vektor Sokongan, dan Hutan Rawak.	Set data kualiti air dari terusan Taladanda, Odisha, India diperolehi daripada Lembaga Kawalan Pencemaran Negeri bermula 2013 hingga 2020 dari 6 stesen yang berbeza di sepanjang terusan.	**Model Pokok Keputusan (DT)** adalah terbaik untuk meramalkan kelas kualiti air dengan ketepatan pengelasan tertinggi sebanyak 96.6%.
Muhammad Sani Gaya, Sani Isah Abba, Aliyu Muhammad Abdu, Abubakar Ibrahim Tukur, Mubarak Auwal Saleh, Parvaneh Esmaili dan Norhaliza Abdul Wahab. (2020)	Menganggarkan indeks kualiti air (WQI) menggunakan pendekatan kecerdasan buatan dan regresi berbilang linear	Regresi berbilang linear (MLR), Rangkaian Neural Buatan (ANN) dan Sistem Inferens Neuro-Fuzzy Adaptif (ANFIS)	Data kualiti air harian diperolehi daripada Lembaga Kawalan Pencemaran Pusat (CPCB) bagi tahun 1999 hingga 2012 dari stesen Palla di sepanjang lembangan Sungai Yamuna di India.	**Model Rangkaian Neural Buatan (ANN)** mencapai ketepatan tertinggi, dengan ralat kuasa dua min (MSE) 9.0E-8, pekali korelasi (DC) 0.9974, dan ralat kuasa dua min akar (RMSE) 0.0003.
Jesmeen Mohd Zebaral Hoque, Nor Azlina Ab. Aziz, Salem Alelyani, Mohamed Mohana dan Maruf Hosain (2022)	Menilai prestasi model pembelajaran regresi yang berbeza dalam meramalkan indeks kualiti air (WQI) menggunakan data terdahulu dari sungai-sungai India	DT, LR, Ridge, Lasso, SVR, RF, ET, dan ANN	Set data daripada Kaggle mempunyai fitur kualiti air dari pelbagai lokasi di India, yang dikumpulkan antara tahun 2003 dan 2014. Set data mengandungi 1991 sampel digunakan oleh kerajaan India untuk menentukan pematuhan air minuman dengan piawaian yang diperlukan.	**Model Regresi Linear (LR) dan Regresi Rabung (RR)**. Kedua-dua algoritma regresi ini menawarkan prestasi terbaik, mencapai ralat persegi min sifar (MSE) dan pekali korelasi tertinggi (r) 1.

bersambung…

…sambungan

Yinshan Yu, Yan Qu, Hongyun Zhang, Lingjie Jiang, Mingzhen Shao, Dandan Wei dan Dongyang Zhang. (2022)	Mencadangkan kaedah untuk meramalkan petunjuk kualiti air dalam air tasik menggunakan model rangkaian saraf kabur (FNN) dan spektrofotometri.	Model yang digunakan dalam dokumen itu ialah model rangkaian saraf kabur (FNN) dan pekali korelasi Pearson	Set data penunjuk kualiti air dari tasik. 89 set data. 73 set data dipilih secara rawak sebagai set latihan, manakala baki 16 set digunakan sebagai set ujian.	**Model Rangkaian saraf kabur (FNN)** dengan ketepatan pengelasan yang tinggi dan keupayaan generalisasi untuk penunjuk kualiti air. Masa pengelasan lebih singkat dan pekali korelasi model didapati antara 0.8-0.95, dan ralat relatif dikawal dalam 15% untuk data set ujian.
Saleh Y. Abuzir dan Yousef S. Abuzir (2022)	Menganalisis prestasi algoritma pembelajaran mesin dalam meramalkan pengelasan kualiti air dan menyediakan kaedah yang lebih cepat dan lebih murah untuk mengesan pencemaran air.	DT, Naïve Bayes dan MLP.	Set data mengandungi metrik kualiti air untuk 3,276 badan air yang berbeza terdiri daripada sepuluh fitur. Fitur pengelasan adalah potability yang mempunyai nilai sama ada kosong atau satu.	**Model MLP** mendapat peratusan pengelasan terbaik bagi kesemua fitur yang berbeza.
Xiaohang Li, Jianli Ding dan Nurmemet Ilyas (2021)	Menggunakan kaedah pembelajaran mesin dan teknologi remote sensing untuk mengenal pasti dan menilai kualiti air di Lembangan Tasik Ebinur dengan cepat di kawasan gersang.	SVM, RF, PLSR dan PLSR-SVM.	Data multispectral dari satelit Sentinel-2A, khususnya Instrumen Multispectral (MSI). Satelit mempunyai 13 jalur. Set data juga termasuk fitur kualiti air (WQPs) yang diukur di Lembangan Tasik Ebinur di kawasan gersang.	**Model PLSR-SVM** mempunyai pekali korelasi (R2v) 0.87 dan ralat kuasa dua min akar (RMSEv) sebanyak 62.927.

bersambung…

…sambungan

Jefferson Lerios dan Mia Villarica (2019)	Menerapkan teknik perlombongan data untuk menganalisis dan meramalkan corak kualiti air di Laguna De Bay, badan air pedalaman terbesar di Filipina.	Naive Bayes, Decision Tree, Random Forest, Gradient Boost dan Deep Learning	Set data terdiri daripada hasil pemantauan kualiti air dari 9 stesen di Tasik Laguna, Filipina dari 2015 hingga 2017 termasuk maklumat tentang stesen dari mana sampel air dikumpulkan dan tarikh pengumpulan.	**Model Pokok Keputusan (DT)** mencapai ketepatan 87.69% dan ketepatan 87.7%.
Farid Hassanbaki Garabaghi, Semra Benzer dan Recep Benzer (2022)	Menilai prestasi model pembelajaran mesin dengan pendekatan pembelajaran ensemble dalam pengelasan indeks kualiti air.	LogitBoost, Hutan rawak, AdaBoost dan XGBoost	Set data kualiti air yang dikumpulkan dari 10 stesen di Büyük Menderes Basin di Turki, antara tahun 2004 dan 2014.	**Model XGBoost** mempunyai ketepatan 96.97%
Dziri Jalal, Tahar Ezzedine (2019)	Mencadangkan sistem pemantauan kualiti air masa nyata menggunakan algoritma pembelajaran mesin	DT dan SVM	Set data dari stesen rawatan air Tunisia Ghadir El Golla terdiri daripada fitur fizikokimia dan mikrobiologi berkaitan kualiti air.	**Mesin Vektor Sokongan (SVM)** mempunyai ketepatan tertinggi, mencapai sehingga 98% untuk satu perempat sampel.
Yafra Khan dan Chai Soo See (2016)	Membangunkan model pengelasan kualiti air menggunakan Rangkaian Neural Buatan (ANN) dan analisis siri masa untuk mengoptimumkan kualiti air.	Rangkaian Neural Buatan (ANN) dengan model siri masa Nonlinear Autoregressive (NAR).	Set data daripada sumber dalam talian Kajian Geologi Amerika Syarikat (USGS) iaitu Sistem Maklumat Air Kebangsaan (NWIS) dari stesen pemantauan di New York dari tahun 2014.	**Model ANN-NAR** berjaya meramalkan empat faktor kualiti air: Kekeruhan, Kepekatan Oksigen Terlarut, Klorofil dan Pengaliran Tertentu (Specific Conductance). Keputusan model dinilai menggunakan ukuran prestasi seperti Regresi, Ralat Min Squared (MSE) dan Ralat Punca Min Kuasa Dua (RMSE).

bersambung…

…sambungan

Ali N Hasan dan Khawla M. Alhammadi (2021)	Meneroka aplikasi pengelas pembelajaran mesin untuk memantau kualiti air minuman di Abu Dhabi	Regresi logistic (LR), Mesin vektor sokongan (SVM), Teluk Naif (NB), Pokok Keputusan (DT) dan jiran terdekat k (KNN)	Set data diperolehi daripada Jabatan Tenaga Abu Dhabi terdiri daripada 7 fitur fizikal dan kimia berkaitan kualiti air minuman di Abu Dhabi.	**Model Pokok Keputusan (DT)** dengan ketepatan 97.7011%.
Mohamed Torky, Ali Bakhiet, Mohamed Bakrey, Ahmed Adel Ismail dan Ahmed I. B. EL Seddawy (2023)	Membangun dan melaksanakan rangka kerja pembelajaran mesin untuk mengklasifikasikan sampel air minuman dan meramalkan indeks kualiti air (WQI)	Model Pengelas seperti Penggalak Kecerunan Melampau (XGBoost), Mesin Penggalak Kecerunan Cahaya (Light GBM), Pokok Keputusan (DT), Pokok Tambahan (ET), Perceptron berbilang lapisan (MLP), Peningkatan Kecerunan (GB), Mesin Vektor Sokongan (SVM), Rangkaian Neural Buatan (ANN) dan Hutan Rawak (RF). Model Regresi seperti Regresi LGBM, Regresi XGB, Regresi Pokok Tambahan ET, Regresi DT, Regresi RF dan Regresi linear	Set data daripada 7996 sampel air dengan 19 fitur (pembolehubah). Set data dibahagikan kepada data latihan, 6396 sampel dan data ujian, 1600 sampel.	Pengelasan: **Light Gradient Boosting Machine (Light GBM)** dengan ketepatan 97% Regresi: **Regresi LGBM dan Regresi Pokok Tambahan** dengan ketepatan data latihan 99.9% (ET), 99% (LGBM) dan ketepatan data ujian 95.5% (ET & LGBM)
Jitha P Nair dan Vijaya M S (2022)	Membangunkan model ramalan kualiti air sungai menggunakan algoritma pembelajaran mesin.	Ramalan menggunakan Regresi Linear, Regresi MLP, Regresi Vektor Sokongan, Hutan Rawak. Pengelasan menggunakan SVM, Teluk Naif (NB), Pokok Keputusan (DT) dan MLP.	Set data kualiti air Sungai Bhavani yang dikumpulkan dari sebelas stesen persampelan dari 2016 hingga 2020 mengandungi 33 fitur dan 10560 *instance*.	Ramalan: **Regresi MLP** memperolehi RMSE terendah 2.432 dan MSE terendah 1.914. Pengelasan: **Pengelas MLP** dengan ketepatan 0.8, Precision 0.6, Recall 0.6 dan F1-Score 0.59.

bersambung…

…sambungan

Md Galal Uddin, Stephen Nash, Mir Talas Mahammad Diganta, Azizur Rahman dan Agnieszka I. Olbert (2022)	Mengenal pasti algoritma pembelajaran mesin yang paling boleh dipercayai dan mantap untuk meramalkan indeks kualiti air pantai (WQI) di titik pemantauan di Cork Harbour, Ireland	Hutan Rawak, Pokok Keputusan, K-Jiran Terdekat, Peningkatan Kecerunan Ekstrem (XGB), Pokok Tambahan, Menyokong Mesin Vektor, Regresi Linear dan Teluk Naïve Gaussian (GNB).	Data kualiti air dari pangkalan data pemantauan kualiti air Agensi Perlindungan Alam Sekitar Ireland (EPA) untuk Cork Harbour di 29 lokasi pemantauan dengan 11 pembolehubah kualiti air untuk pengiraan WQI.	**Model Peningkatan Kecerunan Ekstrem (XGB)** menunjukkan prestasi terbaik, dengan kesilapan latihan terendah (RMSE = 3.3, MSE = 10.91, MAE = 1.67, dan R2 = 1.0) dan ujian (RMSE = 0.0, MSE = 0.0, MAE = 0.02, dan R2 = 1.0)
A. Najah, F. Y. Teo, M. F. Chow, Y. F. Huang, S. D. Latif, S. Abdullah, M. Ismail dan A. El-Shafe (2021)	Menilai kesan Perintah Kawalan Pergerakan (PKP) yang dilaksanakan di Malaysia semasa pandemik COVID-19 terhadap kualiti air sungai bandar (Sungai Klang dan Sungai Pulau Pinang) dan tasik (Tasik Putrajaya)	Perceptron berbilang lapisan (MLP), mesin vektor sokongan (SVM), hutan rawak (RF), dan algoritma pembelajaran mesin pokok keputusan yang dirangsang (BDT).	Set data berkaitan kualiti air sungai bandar (Sungai Klang dan Sungai Pulau Pinang) dan tasik (Tasik Putrajaya)	**Model Multi-layer Perceptron (MLP)** menunjukkan tahap ketepatan yang tinggi, dengan peratusan ralat relatif tidak melebihi 5%
Huu-Du Nguyen, Nguyen Tai Quang Dinh, Hien Nguyen dan Thi-Thu-Hong Phan (2022)	Menjalankan kaji selidik mengenai penerapan algoritma pembelajaran mesin dalam menganggarkan Indeks Kualiti Air (WQI) dan untuk memberikan perspektif dan kajian kes yang berkaitan	Pokok Sangat Rawak (Extremely Randomized Trees) atau Pokok Tambahan (ET) terhadap sepuluh, tujuh dan empat fitur.	Set data daripada sistem pengairan An Kim Hai di dua wilayah besar di Dataran Pantai Utara Vietnam. Sebanyak 657 sampel dari stesen yang berbeza untuk 12 tahun berturut-turut (dari tahun 2007 hingga 2020). Setiap sampel terdiri mempunyai sembilan fitur kualiti air.	**Model Pokok Rawak Extreme (*Extremely Randomized Trees*)** atau Pokok Tambahan (ET) Penggunaan empat fitur memberi hasil ramalan WQI yang lebih baik berbanding tujuh dan sepuluh. Sim meningkat kepada 0.898, MAE menurun kepada 9.46, tetapi RMSE meningkat kepada 14.11. Nilai Sim dan MAE sedikit bertambah baik apabila mengurangkan bilangan fitur, RMSE meningkat, menunjukkan model yang sedikit kurang tepat.

bersambung…

…sambungan

M.Anbuchezhian, R.Venkataraman dan V.Kumuthavalli. (2018)	Mengukur kualiti badan air tertentu menggunakan Indeks Kualiti Air (WQI) dan membangunkan model ramalan untuk pengelasan kualiti air menggunakan algoritma pembelajaran mesin.	Teluk Naif (NB), Pokok Keputusan (DT), Jiran Terdekat – k (KNN), Mesin Vektor Sokongan (SVM) dan Hutan Rawak (RF).	Set data kualiti air mengandungi dua jenis set data sebenar dan sintetik. Set data sebenar dikumpulkan dari pelbagai tempat di Tamil Nadu dan set data sintetik dihasilkan daripada julat fitur kualiti air. Kedua-dua set mempunyai lapan fitur kualiti air.	**Model Hutan Rawak** mencapai ketepatan tertinggi 0.9992, sensitiviti 0.998, kekhususan 0.9995, precision 0.998, *Recall* 0.998, dan Skor-F1 0.998
Annaji Kuthe, ChaitanyaBhake, Vaibhav Bhoyar, Aman Yenurkar, Vedant Khandekar dan Ketan Gawale (2022)	Membangunkan model ramalan kualiti air menggunakan algoritma Pembelajaran Mesin. Tujuannya adalah untuk menilai trend kualiti air, meramalkan corak kualiti air, dan menentukan sama ada air itu sesuai untuk tujuan minuman.	Rangkaian Neural Buatan (ANN), Gaussian Teluk Naif (GNB), Pokok Keputusan, Mesin Vektor Sokongan (SVM), Jiran Terdekat K (KNN) dan Hutan Rawak	Data latihan dan ujian dari repositori data dalam talian United States Geological Survey (USGS).	**Model Hutan Rawak** mencapai ketepatan sehingga 94%
Khadijah Sulaiman, Lokman Hakim Ismail, Mohd Adib Mohammad Razi, Mohd Shalahuddin Adnan, Rozaida Ghazali (2019)	Menyiasat teknik yang secara automatik dapat mengklasifikasikan kualiti air menggunakan Rangkaian Neural Buatan (ANN)	Rangkaian Neural Buatan (ANN)	Set data merangkumi fitur kualiti air yang dikumpulkan dari tiga lokasi berbeza di Selat Melaka di Malaysia iaitu Sungai Pontian, Sungai Batu Pahat dan Sungai Muar.	**Model ANN** dengan pengelasan ketepatan 80% dengan RMSE 0.468

bersambung…

…sambungan

Xianhe Wang, Ying Li, Qian Qiao, Adriano Tavares dan Yanchun Liang (2023)	Menggunakan teknik pembelajaran mesin, bersama-sama dengan kaedah berat entropi dan kaedah pekali korelasi Pearson untuk pemilihan fitur, untuk mencapai ramalan ketepatan tinggi penunjuk kualiti air utama seperti Oksigen Terlarut (DO), Nitrogen Ammonia (NH3-N), Total Phosphorus (TP) dan Total Nitrogen (TN)	Memori Jangka Pendek Panjang (LSTM), Mesin Vektor Sokongan (SVM), Multilayer Perceptron (MLP), Hutan Rawak (RF) dan XGBoost.	Set data diperoleh dari Stesen Umum Pemantauan Alam Sekitar China, khususnya di Bahagian Shijiaoju di Lembangan Sungai Pearl dikumpulkan pada setiap empat jam dari 8 November 2020 hingga 28 Februari 2023. Merangkumi sejumlah 5058 sampel dan sembilan fitur kualiti air.	**Model LSTM** mempunyai nilai R2 0.882 dan nilai RMSE 1.827
K Abirami, Priyadarshini Changail Radhakrishna dan Monisha A Venkatesan (2023)	Menganggarkan dan meramalkan kualiti air menggunakan teknik pembelajaran mesin berdasarkan Indeks Kualiti Air (WQI)	K-Jiran Terdekat (K-NN), Teluk Naif, Mesin Vektor Sokongan (SVM), Pokok Keputusan dan Hutan Rawak.	Set data daripada Kaggle terdiri daripada unsur-unsur utama yang mempengaruhi potability air diperolehi dari beberapa sungai di pelbagai wilayah di India.	**Model Hutan Rawak** mencapai skor ketepatan 91.97%.
Kittichai Northep, Krittakom Srijiranon dan Narissara Eiamkanitchat (2020)	Mengaplikasikan teknik perlombongan data untuk mewujudkan model pengelasan bagi isu-isu kualiti air di Sungai Wang di Thailand.	SVM, DT, MLP-1 dan MLP-kNN	Set data terdiri daripada sembilan fitur kualiti air Sungai Wang di Thailand dari enam stesen pemantauan dengan sejumlah 446 rekod.	**Model MLP-kNN** mempunyai kadar pengelasan melebihi 95% Skor F untuk kedua-dua kelas (baik dan buruk) juga di atas 90%.

bersambung…

…sambungan

Wei Li, Muyi Yang, Zhiwei Liang, Yao Zhu, Wei Mao, Jiyan Shi dan Yingxu Chen (2013)	Menilai kualiti air di Lembangan Sungai Tiaoxi, China menggunakan mesin vektor sokongan (SVM).	Mesin Vektor Sokongan (SVM) dibandingkan dengan Analisis Diskriminant Linear (LDA) dan Analisis Diskriminan Kuadratik (QDA).	Set data kualiti air permukaan dari perairan Sungai Tiaoxi di China untuk 172 tapak dalam tempoh 12 bulan yang dibahagikan kepada 360 sampel latihan dan 240 sampel ujian.	**Model Mesin Vektor Sokongan (SVM)** dengan ketepatan melebihi 95%
Kangyang Chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen , Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng dan Hongqiang Ren (2019)	Menyiasat sama ada data besar dapat meningkatkan prestasi ramalan model pembelajaran mesin dalam menentukan kualiti air permukaan.	Regresi Logistik (LR), Analisis Diskriminan Linear (LDA), Mesin Vektor Sokongan (SVM), Pokok Keputusan (DT), Pokok Rawak Sepenuhnya (CRT), Teluk Naif (NB), K-Jiran Terdekat (KNN), Hutan Rawak (RF), Hutan Pokok Rawak Sepenuhnya (CTF) dan Hutan Lata Dalam (DCF)	Set data dikumpulkan dari sungai dan tasik utama di China dari tahun 2012 hingga 2018 dan diperolehi daripada 124 stesen pemantauan kualiti air terletak di sepuluh sungai dan tasik besar negara China dengan jumlah keseluruhan sejumlah 33,612 rekod.	**Model Pokok keputusan (DT)** dengan bacaan ketepatan tertinggi menghampiri 99%.
Jui-Sheng Chou, Chia-Chun Ho dan Ha-Son Hoang (2018)	Membangunkan pendekatan pembelajaran mesin untuk meramalkan Indeks Keadaan Trofik Carlson (CTSI) untuk penilaian kualiti air di takungan dan menyediakan pendekatan pemodelan kualiti air serba boleh menggunakan model tunggal, ensemble, dan hybrid.	Rangkaian Neural Buatan (ANNs), Mesin Vektor Sokongan (SVM), Regresi Linear (LR) dan Regresi (CART).	Data set dikumpulkan dari pengukuran yang dibuat di 20 stesen takungan di Taiwan dalam tempoh 22 tahun dari tahun 1995 hingga 2016 terdiri daripada 1,636 titik data mengenai 27 fitur.	**Model Rangkaian Neural Buatan (ANN)** dengan R 0.718, RMSE 3.941, MAE 3.131, MAPE 6.786 dan nilai SI keseluruhan terendah iaitu 0.250.

bersambung…

…sambungan

GAO Qianqian dan ZHANG Ying (2015)	Menggunakan algoritma pengelasan Hutan Rawak untuk menilai kualiti air di Tasik Chaohu	Hutan Rawak (RF), Mesin Pembelajaran Ekstrem (ELM), dan Mesin Vektor Sokongan Algoritma Genetik (GA-SVM)	Data pemantauan automatik kualiti air mingguan di lembangan sungai utama negara yang diterbitkan oleh pusat data jabatan perlindungan alam sekitar Republik Rakyat China dari tahun 2010 hingga 2014. Terdiri daripada dua bahagian pemantauan utama Tasik Chaohu iaitu Hefei Hubin dan Yu Xikou.	**Model Hutan Rawak (RF)** dengan ketepatan pengelasan kualiti air di bahagian pemantauan Hefei Hubin, 96.15% dan Yu Xikou, 100%
Amri Danades, Devie Pratama, Dian Anggraini dan Diny Anggriani (2017)	Membandingkan ketepatan dua algoritma yang berbeza, K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) dalam mengklasifikasikan status kualiti air	Jiran Terdekat - K (KNN) dan Mesin Vektor Sokongan (SVM)	Set data air sungai daripada Badan Pusat Statistik (BPS), Indonesia.	**Model Mesin Vektor Sokongan (SVM)** dengan kadar ketepatan yang tinggi sebanyak 92.40%
Mohamed Ladjal, Mohamed Bouamar, Mohamed Djerioui dan Youcef Brik (2016)	Menilai dan mentafsir data kualiti air permukaan di empangan Tilesdit di Algeria untuk meningkatkan prestasi pemantauan dan pengelasan kualiti air.	Rangkaian Neural Buatan (ANN) dan Mesin Vektor Sokongan (SVM).	Set data data kualiti air permukaan di empangan Tilesdit di Algeria terdiri daripada data sebenar yang dikumpulkan dalam tempoh 2009 hingga 2011 dan mengandungi sejumlah 1800 sampel.	**Model Mesin Vektor Sokongan (SVM)** dengan ketepatan 98.76%.

bersambung…

…sambungan

Bachir Sakaa, Ahmed Elbeltagi, Samir Boudibi, Hicham Chafaï, Abu Reza Md. Towfqul Islam, Luc Cimusa Kulimushi, Pandurang Choudhari, Azzedine Hani, Youssef Brouziyne dan Yong Jie Wong (2022)	Meramalkan kualiti air di lembangan sungai Saf-Saf di Algeria menggunakan model pembelajaran mesin.	Hutan rawak (RF) dan algoritma Pengoptimuman Minimum Berurutan (SMO) - Mesin Vektor Sokongan (SVM)	Set data daripada lembangan sungai Saf-Saf terdiri daripada 70 pemerhatian dan 15 fitur kualiti air.	**Model Hutan Rawak (RF)** dengan prestasi ramalan yang tinggi dan dekat untuk pembolehubah kualiti air, baik dalam peringkat latihan dan ujian.

2.2.1 Pendekatan Pembelajaran Mesin

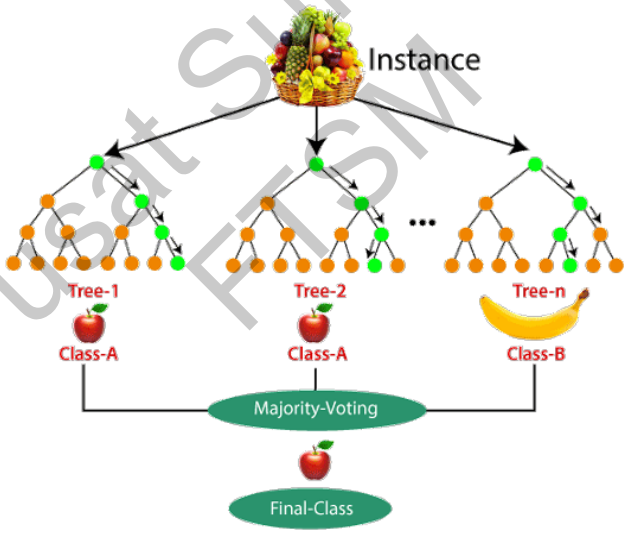
Daripada kajian dan pemerhatian kesusasteraan yang telah dijalankan, tiga algoritma digunakan dalam kajian IKA Tasik Putrajaya ini iaitu algoritma Hutan Rawak (RF), Mesin Vektor Sokongan (SVM) dan Rangkaian Neural Buatan (ANN). Fungsi algoritma-algoritma tersebut diterangkan secara ringkas seperti berikut:

a. Algoritma Hutan Rawak (RF)

Hutan Rawak (RF) adalah gabungan peramal pokok supaya setiap pokok bergantung kepada nilai vektor rawak yang disampel secara bebas dan dengan pengedaran yang sama untuk semua pokok di hutan. Kesalahan umum untuk hutan menumpu (*converges*) hampir pasti menghampiri had kerana bilangan pokok di hutan menjadi semakin besar. Kesalahan umum untuk pengelas pokok hutan bergantung kepada

kekuatan pokok secara individu di hutan dan korelasi di antara mereka (Breiman 2001). RF adalah algoritma adalah kaedah pembelajaran *ensemble* yang dibangunkan oleh Leo Breiman dan Adele Cutler. Leo Breiman adalah seorang ahli statistik dan profesor di University of California, Berkeley. Breiman pertama kali membentangkan konsep RF dalam kertas kerja bertajuk "Hutan Rawak".

RF digunakan untuk kedua-dua tugas pengelasan dan regresi. Ia beroperasi dengan membina banyak pokok keputusan semasa latihan dan output mod atau pengelasan min (regresi) pokok individu untuk setiap input. “Hutan” dibina secara rawak hasil gabungan Pokok Keputusan (*Decision Tree*). Hutan secara rawak merupakan kaedah berasaskan Pokok Keputusan yang terlatih dalam sampel rawak bagi pemerhatian dan ciri-ciri. Rajah 2.3 menunjukkan ilustrasi bagaimana hutan rawak (RF) berfungsi.



Rajah 2.1 Hutan Rawak

Sumber : Sruthi 2023

Secara analogi fungsi RF digunakan terhadap pemilihan majoriti buah di dalam sebuah bakul buah. Anggap sebakul buah sebagai data seperti yang ditunjukkan dalam Rajah 2.3. Kemudian bilangan sampel, n diambil dari bakul buah dan pokok keputusan individu dibina untuk setiap sampel. Setiap pokok keputusan akan menghasilkan output. Output akhir dipertimbangkan berdasarkan pengundian majoriti. Kesimpulannya ialah pokok keputusan majoriti memberikan output sebagai epal jika

dibandingkan dengan pisang. Oleh itu, output akhir diambilkira sebagai epal (Sruthi 2023).

Secara praktikal, RF merupakan algoritma pembelajaran yang paling tepat sehingga kini. *Pesudocode* bagi algoritma RF (Breiman 2001) dijelaskan seperti Jadual 2.4 berikut.

Jadual 2.3 Pseudocode RF

	Algoritma RF
1	function RANDOM_FOREST(Dataset, NumberOfTrees):
2	Forest = []
3	for i = 1 to NumberOfTrees do
4	BootstrapSample = create_bootstrap_sample(Dataset)
5	Tree = DECISION_TREE(BootstrapSample)
6	add Tree to Forest
7	end for
8	return Forest
9	function DECISION_TREE(BootstrapSample):
10	// Mencipta pohon keputusan daripada sampel *bootstrap*
11	// Fungsi ini merangkumi langkah-langkah seperti memilih pecahan terbaik berdasarkan
12	// kriteria seperti kekotoran Gini atau keuntungan maklumat dan berulang
13	// memisahkan nod sehingga kriteria berhenti dipenuhi
14	return Tree
15	function PREDICT(Forest, NewSample):
16	Predictions = []
17	for each Tree in Forest do
18	add Tree.predict(NewSample) to Predictions
19	end for
20	return mode(Predictions) or mean(Predictions)

Berdasarkan Jadual 2.4 di atas, baris pertama hingga kelapan bertujuan untuk latihan hutan rawak (RANDOM_FOREST) yang digunakan untuk mencipta hutan rawak dimulai dengan senarai kosong hutan (Forest) untuk menyimpan pokok keputusan. Kemudian fungsi ini secara berulang akan mencipta pokok keputusan dengan mensampel *bootstrap* dan menggunapakai fungsi DECISION_TREE. Setiap pokok ditambah ke hutan sehingga fungsi ini mengembalikan hutan terlatih.

Baris sembilan hingga 14 bertujuan untuk melatih pokok keputusan (DECISION_TREE). Fungsi ini digunakan untuk mencipta pokok keputusan daripada sampel *bootstrap.* Perlaksanaan ini merangkumi langkah-langkah seperti memilih pecahan terbaik berdasarkan kriteria contohnya *impurity* Gini dan perolehan

maklumat (IG), seterusnya memisahkan nod secara berulang sehingga berhenti apabila kriteria dipenuhi. Fungsi ini kemudiannya mengembalikan pokok keputusan yang terlatih.

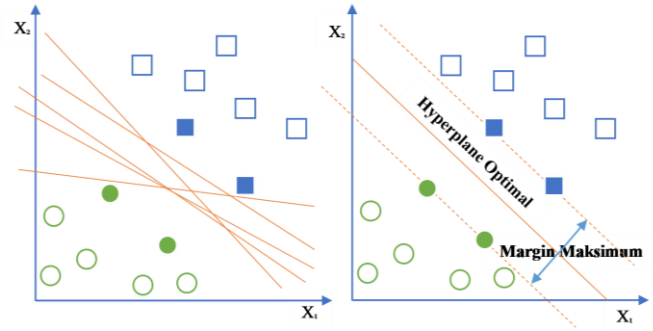
Baris 15 hingga 20 bertujuan meramal model menggunakan fungsi ramalan (PREDICT). Fungsi ini akan meramal menggunakan hutan (Forest) yang terlatih. Bermula dengan senarai ramalan kosong untuk menyimpan ramalan pokok secara individu, fungsi ini secara berulang kali melalui setiap pokok di dalam hutan untuk membuat ramalan bagi sampel Baharu (NewSample). Ramalan kemudiannya ditambah ke dalam senarai sehingga fungsi ini mengembalikan sama ada mod (untuk pengelasan) atau *mean* (untuk regresi) bagi ramalan tersebut.

b. Mesin Vektor Sokongan (SVM)

Mesin Vektor Sokongan (SVM) adalah teknik pengelasan dan regresi yang mantap dengan memaksimumkan ketepatan pengelasan model walaupun tidak terlalu sesuai dengan data latihan. SVM amat sesuai untuk menganalisis data dengan jumlah yang sangat besar contohnya melibatkan beribu-ribu medan peramal. SVM mempunyai aplikasi dalam pelbagai disiplin, termasuk pengurusan perhubungan pelanggan (CRM), pengecaman imej wajah, bioinformatik, pengekstrakan konsep perlombongan teks, pengesanan pencerobohan, ramalan struktur protein dan pengecaman suara dan pertuturan (IBM 2021).

Mesin Vektor Sokongan (SVM) adalah teknik pembelajaran yang diselia yang dicipta oleh Vapnik (Vapnik 1995). SVM boleh digunakan untuk kedua-dua masalah pengelasan dan regresi (Chou, Ho & Hoang 2018). Mesin Vektor Sokongan untuk pengelasan adalah teknik menentukan satah hiperoptimum untuk memisahkan dua kelas dari set vektor latihan. Masalah ini juga dikenali sebagai pengelasan binari, di mana pembolehubah sasaran melibatkan kategori data (Shamsuddin, Othman & Sani 2022).

Objektif algoritma SVM adalah untuk mencari *hyperplane* yang optimal dalam ruang N-dimensi (N - bilangan ciri) yang mengelaskan titik data seperti Rajah 2.3 di bawah:



Rajah 2.2 Mesin Vektor Sokongan

Untuk memisahkan dua kelas titik data, terdapat banyak kemungkinan *hyperplane* yang boleh dipilih. Tujuan utama adalah untuk mencari satah yang mempunyai margin maksimum, iaitu jarak maksimum antara titik data kedua-dua kelas. Memaksimumkan jarak margin ialah dengan menyediakan beberapa pengukuhan supaya titik data selepas ini akan dapat dikelaskan dengan lebih meyakinkan.

Vektor sokongan adalah titik data yang lebih dekat dengan *hyperplane* dan mempengaruhi kedudukan dan orientasi *hyperplane*. Dengan menggunakan vektor sokongan ini, margin pengelas dapat dimaksimumkan. Dengan memadam vektor sokongan pula akan mengubah kedudukan *hyperplane*. Ini akan dapat membantu dalam membina SVM. Jadual 2.5 menunjukkan *Pseudocode* bagi algoritma SVM (C. Cortes & Vapnik 1995).

Jadual 2.4 Pseudocode SVM

Algoritma SVM
Prasyarat: Set latihan (X, y), di mana X ialah matrik fitur dan y ialah vektor label yang sepadan.
function SVM_train(X, y):
Memulakan pemberat (W) dan bias (b) kepada sifar.
Set learning rate (η).
Repeat until convergence:
for each training example (x_i, y_i) in (X, y):
if $y_i * (W * x_i + b) < 1$:
$W = W + \eta * (y_i * x_i)$ # Kemas kini pemberat untuk contoh *misclassified*
$b = b + \eta * y_i$ # Kemas kini bias
else:
$W = W + \eta * 0$ # Kemas kini pemberat untuk contoh *classified* yang betul
$b = b + \eta * 0$ # Kemas kini bias
Return W and b.

bersambung…

…sambungan

```
13      function SVM_predict(X_test, W, b):
14          for each test example x_i in X_test:
15              if W * x_i + b >= 0:
16                  Predict class +1
17              else:
18                  Predict class -1
19      Return the predicted labels.
20      Melatih model SVM menggunakan fungsi SVM_train pada set latihan (X_train, y_train).
21      Buat ramalan pada set ujian (X_test) menggunakan fungsi SVM_predict.
22      Menilai prestasi model
```

Baris pertama *Pseudocode* SVM memulakan definisi fungsi latihan untuk SVM. Fungsi ini mengambil matrik fitur *X* dan melabel vektor *y* sebagai input. Dalam baris kedua, berat (*W*) dan bias (*b*) SVM dimulakan kepada sifar. Dalam baris ketiga, kadar pembelajaran (η) ialah hiperparameter yang menentukan saiz langkah semasa proses pengoptimuman. Baris keempat menunjukkan permulaan gelung yang berterusan sehingga algoritma menumpu. Penumpuan biasanya bermaksud bahawa pemberat dan bias telah stabil atau algoritma telah mencapai bilangan lelaran yang telah ditetapkan. Baris kelima ialah permulaan gelung bersarang yang melalui setiap contoh latihan (fitur vektor x_i dan label sepadan y_i) dalam set latihan. Dalam baris keenam, SVM menyemak sama ada contoh terkini telah salah dikelaskan. Jika produk label (y_i) dan output fungsi keputusan kurang daripada 1, contoh itu salah dikelaskan. Baris ketujuh akan mengemas kini pemberat (*W*) untuk membetulkan kesalahan pengelasan. Ia menyesuaikan pemberat ke arah yang mengurangkan ralat pengelasan. Begitu juga dalam baris kelapan, bias (*b*) dikemaskini untuk membetulkan kesalahan pengelasan.

Untuk baris kesembilan hingga kesebelas, jika contoh dikelaskan dengan betul, pemberat dan bias dikemaskini untuk mengekalkan pengelasan yang betul. Fungsi dalam baris 12 mengembalikan pemberat yang dipelajari (*W*) dan bias (*b*) selepas menyelesaikan proses latihan. Baris 13 memulakan definisi fungsi ramalan untuk SVM. Ia mengambil set ujian ($X_{\_test}$) dan pemberat yang dipelajari (*W*) dan bias (*b*) sebagai input. Baris 14 memulakan gelung yang berulang melalui setiap contoh ujian. Untuk baris 15 hingga 16, SVM meramalkan kelas +1 jika output fungsi keputusan lebih besar daripada atau sama dengan 0. Jika tidak, baris 17 hingga 18 menunjukkan SVM meramalkan kelas -1. Fungsi dalam baris 19 akan mengembalikan label yang

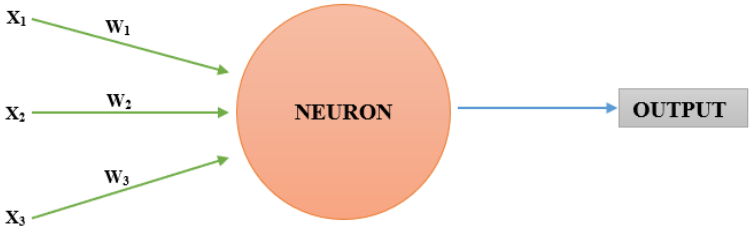
diramalkan untuk set ujian. Baris ke 20 melatih model SVM menggunakan fungsi SVM_train pada set latihan (X_{train}, y_{train}). Baris ke 21 membuat ramalan pada set ujian (X_{test}) menggunakan fungsi ramalan SVM. Baris ke 22 pula akan menilai prestasi model.

c. Rangkaian Neural Buatan (ANN)

Rangkaian Neural (NN), juga dikenali sebagai Rangkaian Neural Buatan (ANN) atau rangkaian neural simulasi (SNN), adalah subset pembelajaran mesin dan berada di pertengahan algoritma pembelajaran mendalam. Nama dan struktur tersebut telah diilhamkan dari otak manusia, meniru cara neuron biologi memberi isyarat antara satu sama lain. Rangkaian Neural Buatan (ANN) terdiri daripada lapisan nod, mengandungi lapisan input, satu atau lebih lapisan tersembunyi, dan lapisan output.

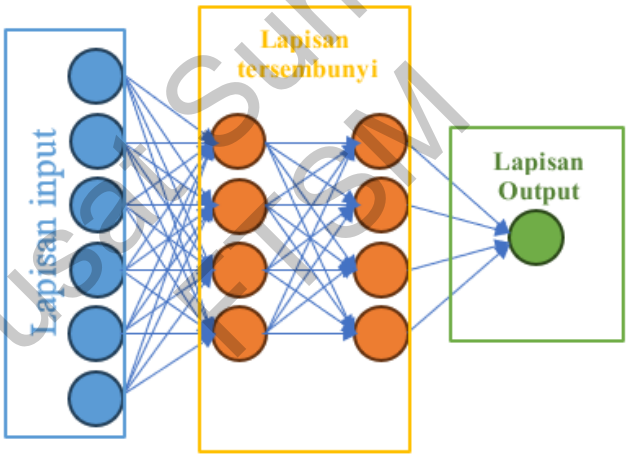
Setiap nod, atau neuron buatan, menghubungkan kepada yang lain dan mempunyai pemberat dan ambang yang berkaitan. Jika output bagi mana-mana nod individu berada di atas nilai ambang yang ditentukan, nod tersebut diaktifkan, menghantar data ke lapisan rangkaian seterusnya. Jika tidak, tiada data dihantar ke lapisan rangkaian seterusnya. Rangkaian saraf bergantung kepada data latihan untuk belajar dan meningkatkan ketepatan mereka dari semasa ke semasa. Apabila algoritma pembelajaran ini dilaksanakan untuk memperolehi ketepatan, ANN akan menjadi alat yang berkuasa dalam bidang sains komputer dan kecerdasan buatan dan membolehkan pengelasan dan pengumpulan data dilaksanakan pada halaju yang tinggi (IBM 2021)

Perceptron adalah satu bentuk mudah Rangkaian Neural dan terdiri daripada satu lapisan di mana semua pengiraan matematik dilakukan seperti Rajah 2.4. Multilayer Perceptron (MLP) juga dikenali sebagai Rangkaian Neural Buatan (ANN) terdiri daripada lebih daripada satu persepsi yang dikumpulkan bersama untuk membentuk neural lapisan berganda.



Rajah 2.3 Perceptron

Rajah 2.5 di bawah menunjukkan Rangkaian Neural Buatan (ANN) terdiri daripada empat lapisan yang saling berkaitan antara satu sama lain. Lapisan pertama iaitu lapisan input terdiri daripada enam nod input, lapisan kedua iaitu lapisan tersembunyi pertama terdiri daripada empat nod tersembunyi (empat perceptron) diikuti lapisan tersembunyi kedua terdiri daripada empat nod tersembunyi dan yang terakhir iaitu lapisan output terdiri daripada satu nod output.



Rajah 2.4 Rangkaian Neural Buatan

Umumnya ANN merupakan algoritma pembelajaran yang paling meluas digunakan kini. *Pesudocode* bagi algoritma ANN (Rumelhart, Hinton & Williams 1986) terkandung dalam Jadual 2.6 berikut.

Jadual 2.5 Pseudocode ANN

	Algoritma ANN
1	function ANN_train($X_{\_train}$, $y_{\_train}$, hidden_layer_sizes, learning_rate, num_epochs):
2	Initialize weights and biases for input and hidden layers randomly.
3	for epoch in range(num_epochs):
4	for each training example (x_i, y_i) in ($X_{\_train}$, $y_{\_train}$):
5	Perform forward propagation:

bersambung…

…sambungan

```
6                   Compute the weighted sum and apply activation functions for each hidden layer.
7                   Compute the output of the output layer.
8               Perform backward propagation:
9                   Compute the error (difference between predicted and actual output).
10                  Update weights and biases using gradient descent and the chain rule.
11          function ANN_predict(X_test, weights, biases):
12              For each test example x_i in X_test:
13                  Perform forward propagation as in steps 5-7.
14                  Return the predicted output.
15          Melatih model ANN menggunakan fungsi ANN_train pada set latihan (X_train, y_train).
16          Buat ramalan pada set ujian (X_test) menggunakan fungsi ANN_predict.
17          Menilai prestasi model menggunakan metrik yang sesuai.
```

Baris pertama Pseudocode ANN memulakan definisi fungsi latihan untuk ANN. Fungsi ini mengambil matrik fitur *X*, label vektor *y*, saiz lapisan tersembunyi, kadar pembelajaran, dan bilangan latihan epochs sebagai input. Dalam baris kedua, pemberat dan bias untuk input dan lapisan tersembunyi dimulakan dengan nilai rawak. Dalam baris ketiga, gelung dimulakan untuk mengulangi bilangan zaman latihan yang ditentukan. Dalam baris keempat, gelung bersarang berulang melalui setiap Latihan epochs. Dalam baris kelima hingga ketujuh, penyebaran ke hadapan mengira output rangkaian neural yang diberikan input. Dalam baris kelapan hingga kesepuluh, penyebaran ke belakang mengira kecerunan kesilapan mengambilkira pemberat dan bias seterusnya mengemaskini kedua-duanya untuk meminimumkan kesilapan.

Baris kesebelas memulakan definisi fungsi ramalan untuk ANN. Fungsi ini mengambil set ujian $X_{\_test}$, pemberat belajar dan pemberat sebelah sebagai input. Dalam baris 12 hingga 14, fungsi ramalan melakukan penyebaran ke hadapan untuk meramalkan output untuk setiap contoh ujian. Baris 15 hingga 17 memberikan gambaran keseluruhan peringkat tinggi mengenai langkah-langkah untuk melatih ANN, membuat ramalan, dan menilai prestasinya pada set ujian. Metrik penilaian bergantung pada tugas (contohnya, ketepatan, kejituan, *recall* dan skor F1).

d. Kelebihan dan Kelemahan Algoritma

Memilih model terbaik untuk sesuatu kajian bergantung kepada beberapa faktor antaranya adalah ciri-ciri set data seperti saiz, jenis, fitur dan *noise*. Faktor kedua adalah keperluan kajian seperti ketepatan, kebolehtafsiran dan kos pengiraan.

Faktor ketiga adalah pengetahuan terhadap domain iaitu pemahaman terdahulu tentang masalah kajian. Jadual 2.7 di bawah menghuraikan kelebihan dan kelemahan bagi setiap model yang digunakan di dalam kajian ini.

Jadual 2.6 Kelebihan dan Kelemahan Model

Model	**Kelebihan**	**Kelemahan**
RF	▪ Ketepatan Tinggi: RF sering mencapai ketepatan yang tinggi dalam pelbagai tugas kerana sifat *ensemble* yang dimilikinya, yang mengurangkan varians dan overfitting. ▪ Ketahanan: RF mempunyai ketahanan terhadap *outliers* dan *noise* dalam data kerana keupayaan merentasi pelbagai pokok keputusan. ▪ Pengendalian Data Berdimensi Tinggi: RF boleh mengendalikan set data dengan efektif melalui pelbagai fitur tanpa memerlukan pemilihan fitur yang signifikan. ▪ Pentafsiran: Skor kepentingan fitur memberikan pandangan tentang kaedah setiap fitur menyumbang kepada pengelasan model.	▪ Sifat Kotak Hitam: Walaupun pokok individu boleh ditafsirkan, keseluruhan model kemungkinan sukar difahami kerana interaksi kompleks di dalam hutan. ▪ Penalaan Hiperparameter: Menala banyak hiperparameter boleh memakan masa dan memerlukan pertimbangan yang teliti. ▪ Kos Pengiraan: Latihan RF boleh dianggap mahal untuk set data yang besar.
ANN	▪ Ketepatan Tinggi: SVM cemerlang dalam memisahkan titik data daripada kelas yang berbeza dan sering mencapai ketepatan yang tinggi pada pelbagai tugas pengelasan. ▪ Efektif untuk Set Data Kecil: SVM boleh berfungsi dengan baik dengan data terhad kerana fokusnya untuk memaksimumkan margin antara kelas. ▪ Pengendalian Data Bukan Linear: SVM boleh mengendalikan data bukan linear dengan berkesan menggunakan fungsi kernel. ▪ Ketahanan: SVM berketahanan tinggi untuk menangani *noise* dan *outliers*.	▪ Kebolehtafsiran: Model SVM kadangkala sukar untuk ditafsirkan, terutamanya apabila menggunakan fungsi kernel kompleks. ▪ Penalaan Hiperparameter: Memilih fungsi kernel yang sesuai dan penalaan hiperparameter menjadi tugasan yang mencabar. ▪ Kos Pengiraan: Latihan SVM boleh dikira mahal, terutamanya untuk set data yang besar dengan banyak fitur. ▪
SVM	▪ Ketepatan Tinggi: ANN boleh mencapai ketepatan yang tinggi pada tugas yang kompleks, terutamanya untuk hubungan bukan linear. ▪ Fleksibiliti: ANN menawarkan fleksibiliti yang hebat dalam mempelajari corak dan hubungan kompleks dalam data. ▪ Pengendalian Data Dimensi Tinggi: ANN boleh mengendalikan set data dengan banyak fitur dan struktur kompleks. ▪ Pembelajaran Berterusan: ANN boleh dipertingkatkan secara berterusan dengan menggabungkan data baru dan memperhalusi model.	▪ Sifat Kotak Hitam: ANN sering dianggap kotak hitam, menjadikannya sukar untuk memahami bagaimana ramalan muncul. ▪ Overfitting: ANN boleh terdedah kepada overfitting jika tidak dilatih dengan betul dan teratur. ▪ Penalaan Hiperparameter: Menala sebilangan besar hiperparameter kemungkinan menjadi kompleks dan memakan masa. ▪ Kos Pengiraan: Latihan ANN boleh menjadi mahal, terutamanya untuk seni bina yang kompleks dan set data yang besar.

e. Matrik Kekeliruan (CM)

Matrik Kekeliruan (CM) ialah jadual yang menggambarkan prestasi model klasifikasi. Ia meringkaskan bilangan ramalan yang betul dan salah yang dibuat oleh model untuk setiap kelas. Contoh CM digambarkan seperti Rajah 2.6 di bawah. Berikut adalah pecahan elemen utama:

1. Baris: Mewakili kelas sebenar titik data.

2. Lajur: Mewakili kelas yang diramalkan bagi titik data.

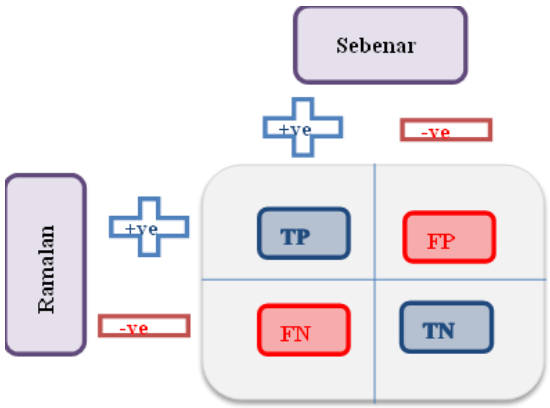
3. Sel: Mengandungi bilangan titik data yang jatuh ke dalam kategori tertentu:

   a. Positif Benar (TP): Kelas positif yang diramalkan dengan betul.

   b. Negatif Benar (TN): Kelas negatif yang diramalkan dengan betul

   c. Positif Palsu (FP): Kelas positif yang diramalkan dengan tidak betul (ralat Jenis I).

   d. Negatif Palsu (FN): Kelas negatif yang diramalkan dengan tidak betul (ralat Jenis II).



Rajah 2.5 Matrik Kekeliruan

Kelebihan dan limitasi bagi Matrik Kekeliruan seperti Jadual 2.8 seperti di bawah.

Jadual 2.7 Kelebihan dan Kelemahan Matrik Kekeliruan

Kelebihan	Limitasi
▪ Visualisasi: Menyediakan gambaran keseluruhan prestasi model yang jelas dan ringkas. Metrik Penilaian: Boleh digunakan untuk mengira pelbagai metrik prestasi seperti ketepatan, ketepatan, *recall* dan skor F1. ▪ Kenal pasti Bias Model: Membantu mengenal pasti jika model itu berat sebelah terhadap kelas tertentu. ▪ Ralat Debug Model: Menyediakan cerapan tentang jenis ralat yang dibuat oleh model.	▪ Kebolehtafsiran: Boleh menjadi sukar untuk mentafsir untuk set data kompleks dengan berbilang kelas. ▪ Normalisasi: Memerlukan normalisasi apabila membandingkan model dengan pengedaran kelas yang berbeza.

2.2.2 Metrik Prestasi

Penilaian prestasi pengelas memainkan peranan penting dalam pembinaan dan pemilihan model pengelasan. Walaupun banyak metrik prestasi telah digunakan di dalam bidang pembelajaran mesin, tiada ketetapan umum ditentukan di kalangan pembangun model berkaitan metrik menjadi pilihan untuk menilai prestasi pengelas. Ketepatan biasanya digunakan untuk mengukur peratusan set ujian yang dikelaskan dengan betul. Keberkesanan algoritma pengelasan dinilai dengan membandingkan ketepatan, kejituan, *recall*, dan keputusan skor f1 dengan algoritma lain (Tanega, Fajardo & Limbago 2023). Terdapat pelbagai metrik yang digunakan namun bergantung kepada data dan tujuan kajian. Jadual 2.8 di bawah menunjukkan metrik prestasi bagi model-model terbaik yang digunakan dalam tempoh dua (2) tahun terdahulu dari segi Ketepatan, Kejituan, *Recall* dan Skor-F1.

Jadual 2.8 Metrik Prestasi Dalam Pengelasan

Metrik	Shamsuddin, Othman & Sani 2022	Suwadi et al. 2022	Tanega, Fajardo & Limbago 2023	Dritsas & Trigka 2023	Nair & Vijay 2022	Abirami, Radhakrishna & Venkatesan 2023
Model	SVM	RF	RF	RF	ANN	RF
Ketepatan (Accuracy)	96.35	95.64	95	98.1	80	91.97
Kejituan (Precision)	91.97	95.6	96	100	60	100
Recall	84.89	95.7	95	98.1	60	83
Skor F1 (F1-Score)	87.98		95		59	93

a. Ketepatan (Accuracy)

Bagi kajian yang melibatkan pengelasan, skor terbaik untuk ketepatan ialah 100%. Walau bagaimanapun, skor ini jarang dicapai kerana semua model ramalan lazimnya mempunyai kesilapan ramalan. Ketepatan prestasi model adalah di antara garis dasar dan skor prestasi terbaik (A. Malek et al. 2022). Ia ditakrifkan seperti berikut:

$$\textbf{Ketepatan} = \frac{TP + TN}{TP + TN + FP + FN}$$

b. Kejituan (Precision)

Kejituan adalah ukuran di mana jumlah ramalan positif yang dibuat adalah betul (positif sebenar). Formula Kejituan adalah seperti berikut:

$$\textbf{Kejituan} = \frac{TP}{TP + FP}$$

c. *Recall*

Recall atau sensitiviti adalah ukuran jumlah kes positif yang diramalkan oleh pengelas dengan betul, ke atas semua kes positif dalam data. Seringkali juga dirujuk sebagai Sensitiviti, formula *Recall* adalah seperti berikut:

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

d. Skor F1 (F1-Score)

Skor F1 adalah metrik penilaian pembelajaran mesin alternatif yang menilai kemahiran ramalan model dengan menghuraikan prestasi bijak kelasnya dan bukannya prestasi keseluruhan seperti yang dilakukan oleh ketepatan. Skor F1 menggabungkan dua metrik yang bersaing iaitu ketepatan dan skor *recall* model, yang membawa kepada penggunaan yang lebih meluas.

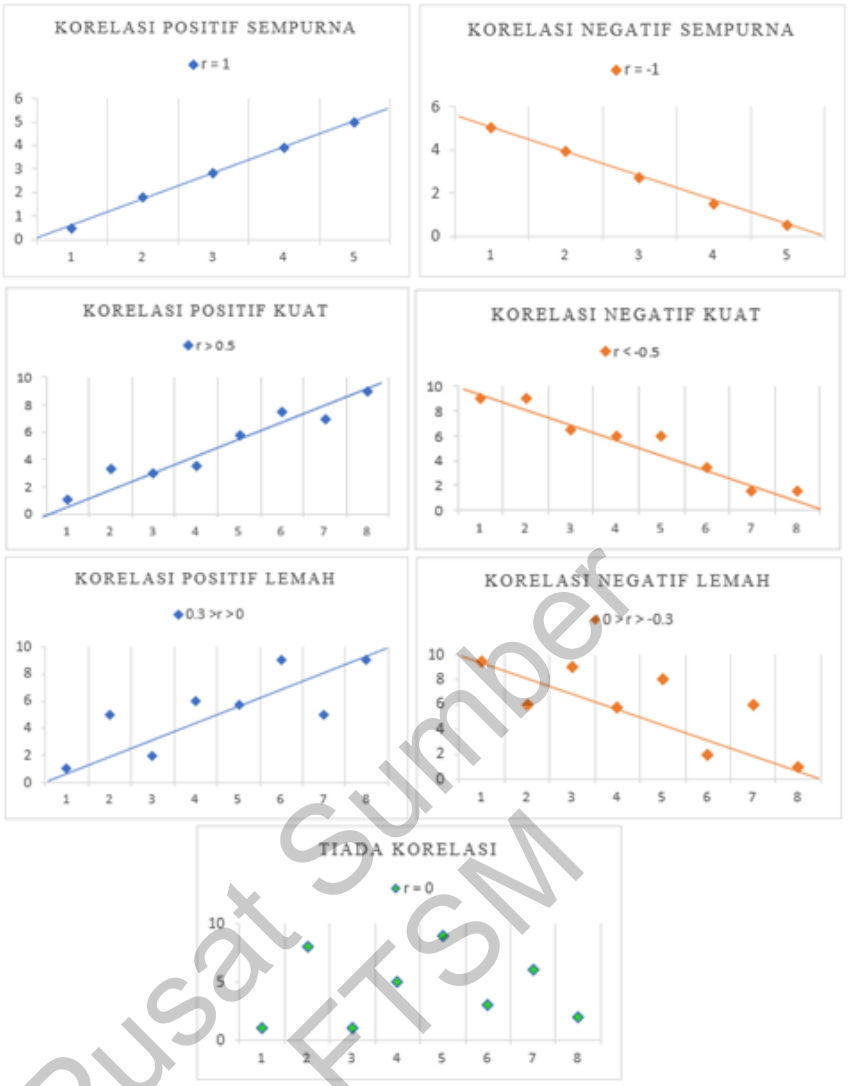
Skor F1 adalah satu lagi ukuran ketepatan model pada set data. Kejituan dan *recall* tidak merangkumi semua aspek ketepatan, min harmonik diambil untuk mencerminkan skor F1, yang merangkumi kedua-dua aspek dan menghasilkan ukuran ketepatan keseluruhan dengan lebih baik. Ia berkisar antara 0 dan 1. Semakin tinggi skor, semakin baik ketepatannya (Umair Ahmed et al. 2019). Kiraan Skor F1 diperolehi seperti berikut:

$$\textbf{Skor F1} = \frac{\textbf{2} \times \textbf{Kejituan} \times \textbf{\textit{Recall}}}{\textbf{Kejituan} \times \textbf{\textit{Recall}}}$$

2.2.3 Analisis Korelasi

Pekali korelasi adalah pengukuran menggunakan angka yang digunakan untuk mengukur tahap dua pembolehubah yang berkaitan dengan kaedah linear. Ia menyediakan cara untuk menyatakan kekuatan dan arah hubungan linear antara dua pembolehubah. Pekali korelasi yang paling biasa digunakan ialah pekali korelasi Pearson, dilambangkan oleh simbol *r*.

Pekali Korelasi Pearson, dinamakan sempena Karl Pearson yang membangunkannya, adalah ukuran statistik yang mengukur tahap hubungan linear antara dua pembolehubah berterusan. Ia dilambangkan dengan simbol *r* dan mengambil nilai antara −1 dan 1, di mana $r = 1$ bersamaan dengan korelasi linear positif yang sempurna, $r = -1$ bermaksud korelasi linear negatif yang sempurna dan $r = 0$ menunjukkan tiada sebarang korelasi linear. Rajah 2.1 di bawah menunjukkan graf bagi menerangkan hubungan pekali korelasi.



Rajah 2.6 Graf Hubungan Pekali Korelasi

Formula bagi pengiraan pekali korelasi adalah seperti berikut:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

Merujuk kepada formula di atas, x mewakili nilai pembolehubah tidak bersandar manakala y mewakili nilai pembolehubah bersandar. Hubungan korelasi dalam pengelasan IKA biasanya diwakilkan dengan fitur kimia bagi menentukan kekuatan hubungan. Kelas IKA diwakili oleh pembolehubah bersandar manakala fitur lain iaitu fitur-fitur IKA diwakili oleh pembolehubah tidak bersandar.

Nilai berkisar antara 0 hingga 1, dengan 0 menunjukkan "tiada maklumat" dan 1 mewakili "maklumat maksimum". Korelasi yang ditentukan antara setiap fitur dan pemboleh ubah output akan mendedahkan fitur mana yang lebih tinggi daripada korelasi positif atau negatif sederhana hingga tinggi (hampir -1 atau 1). Fitur dengan korelasi rendah (nilai hampir sifar) boleh dikeluarkan daripada pemilihan (Suwadi et al. 2022).

Sebagai contoh, Oksigen terlarut (DO) telah menghasilkan hubungan yang kuat. Peningkatan suhu juga menyebabkan penurunan oksigen terlarut (DO) di dalam air, ini disebabkan oleh air panas mengandungi oksigen yang kurang berbanding air sejuk. Tahap BOD yang lebih tinggi membawa kepada kekurangan oksigen yang cepat dalam aliran yang boleh menyebabkan hidupan akuatik menjadi tertekan dan mati lemas. Sumber tahap BOD yang tinggi termasuk daun, tumbuhan, bangkai haiwan dan baja haiwan. Bagi TSS, telah menghasilkan hubungan yang kukuh antara suhu, BOD dan DO untuk musim panas dan hujan (Asmat et al. 2018).

2.2.4 Pemilihan Fitur (Feature Selection) dan Penalaan Hiperparameter

Algoritma pemilihan fitur berkesan dalam mengklasifikasikan fitur yang paling relevan untuk mengelaskan kualiti air, Pengelas Hutan Rawak yang dioptimumkan, menggunakan fitur WQI yang dipilih oleh kaedah pemilihan fitur keuntungan maklumat, mencapai prestasi tertinggi (Suwadi et al. 2022). Garabaghi, S. Benzer dan R. Benzer (2022) meneroka kesan kaedah pemilihan fitur dan kaedah pemisahan set data mengenai prestasi algoritma pembelajaran mesin dalam klasifikasi kualiti air. Untuk mencapai matlamat ini, tiga kaedah pemilihan fitur yang berbeza telah digunakan untuk mengenal pasti fitur yang paling penting untuk proses klasifikasi. Kaedah pemilihan fitur ini termasuk Penghapusan Fitur ke Belakang, Maklumat Bersama dan Hutan Rawak Terbenam.

Srivastava, Joshi dan Gaur (2014) membincangkan peranan pemilihan fitur sebagai proses menghapuskan fitur daripada set data yang tidak relevan dengan tugas yang akan dilaksanakan. Pemilihan fitur adalah penting dari segi penyederhanaan, prestasi, kecekapan pengiraan dan pentafsiran fitur. Pemilihan fitur boleh digunakan untuk kedua-dua metodologi pembelajaran yang diselia dan tanpa penyeliaan. Teknik

sedemikian dapat meningkatkan kecekapan pelbagai algoritma pembelajaran mesin termasuk latihan. Pemilihan fitur mempercepatkan masa pembelajaran, meningkatkan kualiti data dan pemahaman data.

Liu et al. (2011) menggunakan algoritma pemilihan fitur bersama Mesin Vektor Sokongan dengan kernel RBF berdasarkan Penghapusan Fitur Rekursif (SVM-RBF-RFE). SVM-RBF-RFE bermula dengan semua fitur dan kemudian menghapuskan satu fitur dengan berat kuasa dua paling sedikit pada setiap langkah sehingga semua fitur yang terbaik disenaraikan.

Zhu et al. (2019) pula, pemilihan fitur ialah proses memilih fitur yang relevan dan mengeluarkan fitur yang tidak relevan dan bertindan berdasarkan kriteria khusus tertentu dalam set data yang asal. Kajian dibuat menggunakan algoritma hutan rawak (RF) berdasarkan penghapusan fitur rekursif (RFE) untuk memilih fitur-fitur dalam sistem pemantauan beban yang tidak mengganggu (NILM). Dengan menggunakan algoritma hutan rawak sebagai pendekatan asas, model dan fitur penapis berjaya dibina untuk mengenal pasti subset fitur terbaik.

Maharjan (2020) mengkaji fitur terbaik yang dipilih menggunakan *SelectKBest* seterusnya digunakan untuk membandingkan prestasi. Fitur-fitur terpilih yang dikurangkan juga sangat memberi kesan kepada prestasi model. Dalam kajian tersebut, algoritma *SelectKBest* daripada perpustakaan *sklearn* digunakan. Penalaan hiperparameter juga membantu meningkatkan ketepatan manakala penalaan hiperparameter algoritma individu pula membantu mengoptimumkan prestasi. MLP memberikan hasil yang lebih baik daripada algoritma lain apabila menggunakan fitur yang dipilih.

W. Chimphlee dan S. Chimphlee (2022) menggunakan MLP untuk pengelasan pencerobohan dengan menggunakan set data CIC-IDS2018. Hasilnya mendapati bahawa MLP dengan fitur *SelectKbest* menghasilkan prestasi yang tinggi. Kaedah ini mampu mengurangkan bilangan fitur dan memaksimumkan kadar pengesanan. MLP dan *SelectKBest* yang digunakan dalam kajian tersebut memperoleh hasil pengesanan pencerobohan yang baik.

Alshammri et al. (2023) menggunakan *SelectKBest* untuk memilih lapan fitur terbaik set data. SelectKBest merupakan teknik pengurangan dimensi kedua yang paling biasa digunakan, menyumbang kepada 29.1% daripada jumlah keseluruhan penggunaan. *GridSearchCV* dan *SelectKBest* juga dilaksanakan bagi memilih fitur terbaik untuk menentukan hiperparameter terbaik untuk model. MLP mencapai ketepatan yang lebih berbanding kerja-kerja sebelumnya dengan menggunakan set data yang sama.

Alkadi, Al-Ahmadi dan Ismail (2023) membincangkan *SelectKBest* sebagai penyelesaian terkenal mewakili kaedah berasaskan penapis digunakan untuk mengekstrak fitur k pertama dengan markah tertinggi. Pemilihan fitur menggunakan kaedah *SelectKBest* untuk mengenal pasti fitur-fitur yang berkaitan. Sebanyak 30 fitur terbaik yang dicapai menggunakan kaedah *SelectKbest* untuk tugas klasifikasi. Dari segi penilaian prestasi, ketepatan tertinggi sebanyak 99.97% dicapai oleh pengelas MLP, diikuti oleh RF dan KNN.

2.2.5 Analisis SHAP

Analisis SHapley Additive exPlanations atau SHAP merupakan teknik tafsiran model yang memaparkan ramalan secara individu bagi mana-mana model pembelajaran mesin sekalipun untuk model yang kompleks seperti model kotak hitam. SHAP menggunakan konsep teori permainan (nilai *Shapley*) untuk mengagihkan kredit kepada ramalan fitur-fitur yang menyumbang secara adil. SHAP juga menyediakan fitur keutamaan global dan penjelasan secara lokal untuk ramalan tertentu.

Nilai *Shapley* boleh dikira untuk memberikan keputusan pesimis (optimistik). Pendekatan SHAP boleh digunakan untuk menjelaskan klasifikasi yang bersifat model kotak hitam regresi serta membangunkan kaedah baru yang boleh dilihat sebagai hala tuju untuk penyelidikan akan datang. SHAP juga boleh digunakan untuk penjelasan tempatan dan global. Namun kelebihan utamanya ialah ia mengurangkan masa pengiraan untuk menyelesaikan masalah penjelasan dengan ketara (Utkin & Konstantinov 2021).

Nilai *Shapley* telah diterokai dalam pelbagai tugasan melibatkan pembuat keputusan, topik yang diperdebatkan oleh banyak pihak (Kumar et al. 2020). SHAP melaksanakan tugas yang hebat dalam menyahkod kekuatan pembolehubah input yang mempengaruhi ramalan. Nilai SHAP mengira kepentingan fitur dengan membandingkan apa yang diramalkan oleh model dengan fitur dan tanpa fitur (García & Aznarte 2020)

2.3 Kesimpulan

Berdasarkan Jadual 2.1, didapati Hutan Rawak (RF) dan Mesin Vektor Sokongan (SVM) muncul sebagai model terbaik untuk klasifikasi kualiti air berbanding algoritma lain, diikuti oleh Pokok Keputusan (DT) dan Rangkaian Neural Buatan (ANN). Walaupun Pokok Keputusan (DT) berkedudukan ketiga di dalam jadual tersebut, DT menggunakan struktur pokok keputusan yang sama seperti RF, yang berpotensi membawa kepada redundansi atau pertindihan dalam analisis yang bakal dibuat. ANN menawarkan lebih banyak fleksibiliti dan potensi untuk mengendalikan hubungan bukan linear kompleks yang mungkin wujud dalam data kualiti air. Pemilihan ANN juga dijangka mampu memberikan keputusan berharga mengenai seni bina model alternatif dan berpotensi meningkatkan prestasi. Lazimnya RF boleh dianggap model yang paling boleh ditafsirkan berbanding SVM dan ANN yang membolehkan proses membuat keputusan dan kepentingan fitur mudah difahami. SVM menawarkan pentafsiran sederhana melalui fungsi kernel dan vektor sokongan tetapi kurang memberi impak berbanding RF. ANN adalah algoritma yang paling tidak dapat ditafsirkan kerana sifat kotak hitamnya (*Black Box*). Dari segi pemilihan fitur pula, Penghapusan Fitur Rekursif (RFE) dipilih atas dasar *multicollinearity*, di mana fitur-fitur IKA sangat berkaitan antara satu sama lain. RFE boleh mengenal pasti dan menghapuskan fitur-fitur berlebihan dengan cara yang tidak boleh dilakukan oleh pemilihan ke hadapan (FS) atau penghapusan ke belakang (BE). *SelectKBest* dipilih untuk model ANN (MLP) kerana pelaksanaan RFE *Scikit-learn* tidak menyokong fungsi pemarkahan yang diubahsuai (*custom scoring functions*) untuk penganggar yang tidak mempunyai *atribut coef_* atau *feature_importances_*, termasuk *MLPClassifier*. SHapley Additive exPlanations (SHAP) adalah pilihan yang popular untuk mentafsir model pembelajaran mesin, terutamanya yang kompleks seperti Hutan

Rawak. SHAP menyediakan cara yang mantap dan adil untuk mengedarkan kredit untuk pengelasan di antara semua fitur yang terlibat. SHAP juga memberikan nilai sumbangan individu kepada setiap fitur di dalam kajian ini seterusnya memudahkan untuk memahami bagaimana setiap ciri mempengaruhi output model, tanpa memerlukan kepakaran yang tinggi. Kelebihan ini menjadikan SHAP alat serba boleh dan berkuasa untuk mendapatkan pandangan (*insight*) berkenaan model yang kompleks dan meningkatkan pembangunan model.

BAB III

METODOLOGI KAJIAN

3.1 Pengenalan

Bab ini menerangkan berkenaan metodologi kajian yang digunakan dalam kajian yang dibuat. Metodologi yang jelas amat diperlukan bagi memastikan kesahihan dan kebolehpercayaan kajian yang dibuat. Metodologi yang efektif mampu mengurangkan *bias* dan kesilapan, menghasilkan kajian yang boleh dihasilkan semula, menambahbaik pembuatan keputusan dan meningkatkan pembangunan model.

Setiap fasa metodologi memainkan peranan penting dalam memastikan kejayaan keseluruhan kajian melibatkan 4 fasa:

1. Definisi Masalah dan Pengumpulan Data: Mentakrifkan objektif kajian dan memastikan data yang relevan tersedia untuk analisis.

2. Prapemprosesan dan Penerokaan Data: Menyediakan data untuk latihan yang efisyen dan menghasilkan *insight* yang berharga mengenai karakter asas data.

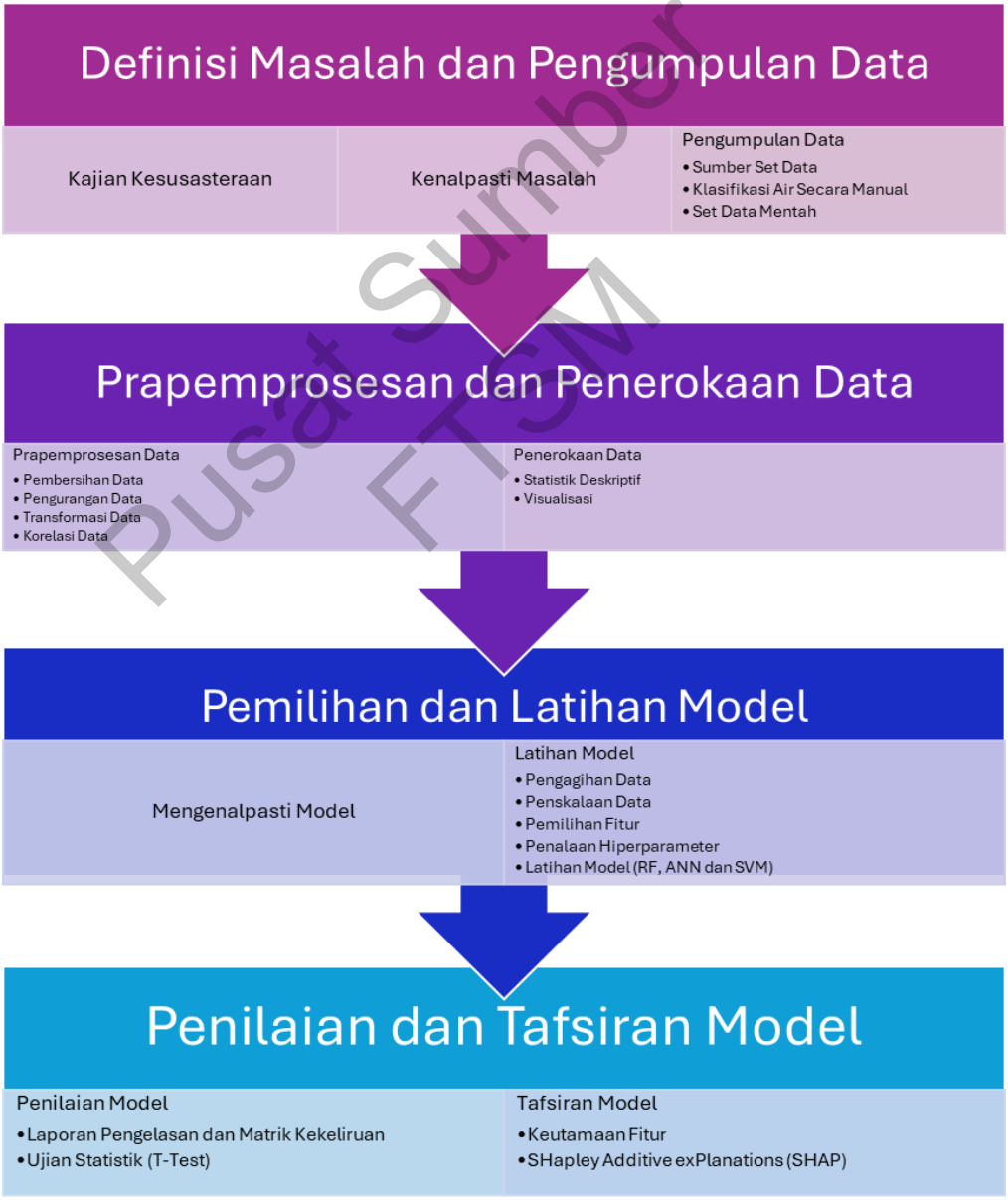
3. Pemilihan dan Latihan Model: Memilih algoritma yang sesuai dan melatih algoritma tersebut bagi mencapai prestasi yang optimum.

4. Penilaian dan Tafsiran Model: Menilai keberkesanan model dan memberikan pandangan yang bermakna terhadap hasilnya.

Beberapa isu utama berkenaan teknik dan strategi yang digunakan akan dibincangkan dalam bab ini.

3.2 Kerangka Metodologi Kajian

Pengumpulan, prapemprosesan, latihan, dan penilaian data adalah peringkat penting dalam pembelajaran mesin (Kotsiantis, Kanellopoulos & Pintelas 2007). Metodologi kajian terbahagi kepada empat fasa iaitu Definisi Masalah dan Pengumpulan Data, Prapemprosesan dan Penerokaan Data, Pemilihan dan Latihan Model serta Penilaian dan Tafsiran Model. Bagi mencapai objektif kajian, setiap fasa metodologi perlu dilaksanakan untuk mendapatkan keputusan yang baik. Kerangka metodologi kajian digambarkan di dalam Rajah 3.1.



Rajah 3.1 Kerangka Metodologi Kajian

3.2.1 Kenalpasti Masalah dan Pengumpulan Data

Kekurangan kepelbagaian dalam pengumpulan data telah menyebabkan kegagalan ketara dalam aplikasi pembelajaran mesin (Hopkins et al. 2023). Kenalpasti masalah dan pengumpulan data merupakan fasa awal bagi mana-mana projek pembelajaran mesin melibatkan kenalpasti masalah dan pengumpulan data. Langkah-langkah penting ini meletakkan asas bagi keseluruhan kajian dan memastikan bahawa model yang dipilih mencapai objektif yang dikehendaki dengan berkesan.

a. Kajian Kesusasteraan

Kajian kesusasteraan adalah antara faktor utama bagi menjayakan kesemua empat fasa metodologi pembelajaran mesin. Kajian terdahulu yang dibuat terhadap jurnal, artikel, buku dan laporan penyelidikan dibuat bagi meningkatkan pemahaman terhadap kajian yang bakal dijalankan secara menyeluruh. Kajian kesusasteraan amat membantu dalam mengenalpasti masalah dan memperoleh data yang relevan, melancarkan strategi prapemprosesan data, membimbing pemilihan model dan latihan, memudahkan penilaian dan tafsiran model serta membolehkan penggunaan dan pemantauan model dilaksanakan secara berkesan. Kajian ini secara tidak langsung membantu di dalam menyelidik masalah yang telah dialami ketika pengelasan dibuat, teknik pengelasan dan regresi yang digunakan dan memastikan model yang digunakan sesuai untuk tujuan kajian dan jenis data serta ketepatan model berdasarkan ralat yang dikeluarkan.

b. Kenalpasti Masalah

Fasa ini memberi tumpuan kepada menentukan dengan jelas masalah yang ingin diselesaikan melalui pembelajaran mesin. Ia melibatkan mengenal pasti masalah tertentu dengan jelas dan menyatakan isu yang ingin diatasi serta hasil yang dikehendaki. Masalah yang dikenalpasti di dalam kajian ini adalah analisis data IKA Tasik Putrajaya masih menggunakan kaedah konvensional dan kurang efisyen iaitu menggunakan teknik pengiraan dan pelbagai formula yang panjang seterusnya mengambil masa yang terlalu lama dan sering kali melibatkan ralat pengiraan yang tidak disengajakan.

Terdapat keperluan untuk menjalankan kajian menggunakan model pengelasan pembelajaran mesin bagi memilih model yang terbaik untuk mengelas IKA Tasik Putrajaya dan menentukan fitur-fitur atau fitur yang mempunyai korelasi terhadap kualiti air tasik Putrajaya. Langkah seterusnya diambil iaitu merumuskan soalan penyelidikan melalui terjemahan masalah ke dalam soalan tertentu yang spesifik, boleh diukur (*measurable*), boleh dicapai (*achievable*), relevan dan terikat dengan masa (time-bound) atau diringkaskan sebagai akronim SMART. Soalan ini telah diajukan kepada pihak BASTW bagi mendapatkan informasi berkaitan kajian yang dibuat.

Langkah seterusnya ialah menentukan pemboleh ubah sasaran yang ingin diramalkan atau dianalisa menggunakan model. Kolum 'Class' telah dipilih sebagai pemboleh ubah sasaran. Pemboleh ubah input juga dikenalpasti dan ditentukan fitur atau titik data yang akan digunakan untuk mengelaskan atau menganalisis pemboleh ubah sasaran. Pemahaman terhadap konteks dan domain dibuat dengan membiasakan diri dengan latar belakang masalah dan pengetahuan domain yang berkaitan. Dalam kajian ini, domain yang dikenalpasti ialah IKA dan segala maklumat diperolehi daripada BASTW telah direkodkan. Akhirnya, kriteria penilaian menggunakan metrik prestasi diwujudkan untuk mengukur kejayaan model yang dikehendaki.

c. Pengumpulan Data

Sebaik sahaja masalah dikenalpasti, langkah seterusnya adalah memperoleh data yang akan digunakan untuk melatih dan menilai model. Proses ini melibatkan pengenalpastian sumber data iaitu dengan mencari set data berkaitan yang mengandungi maklumat yang diperlukan untuk kajian samada melalui pangkalan data awam, set data peribadi atau pengumpulan data dibuat sendiri. Kajian ini menggunakan set data yang dikumpul organisasi PPj bagi memantau kualiti air Tasik Putrajaya.

Data yang diperolehi tersebut seterusnya melalui penilaian kualiti data bagi menilai kesempurnaan, ketepatan dan konsistensi data untuk mengenalpasti dan menangani sebarang kehilangan, ketidakkonsistenan atau bias pada data. Seterusnya pra-pemprosesan data menggunakan perisian *Microsoft Excel* dan *Python* dibuat bagi

membersihkan dan menyediakan data untuk algoritma pembelajaran mesin bagi tujuan pembersihan, normalisasi, kejuruteraan fitur dan transformasi data.

Langkah terakhir ialah pengagihan data menggunakan perpustakaan *Scikit-learn* di dalam *Python* dengan kaedah membahagikan data ke dalam set latihan, pengesahan dan ujian. Set latihan akan digunakan untuk melatih model, set pengesahan akan digunakan untuk memperhalusi hiperparameter model dan set ujian akan digunakan untuk menilai prestasi model akhir iaitu menentukan fitur-fitur data yang diperlukan untuk proses eksperimen di mana hasil kajian akan dinilai pada akhir penyelidikan.

i. Sumber Set Data

Di dalam kajian ini, set data mentah IKA Tasik Putrajaya dari BASTW, JR, PPj telah dipilih dan diperolehi melalui prosedur yang ketat. Permohonan perlu dibuat dan kelulusan peringkat atasan PPj perlu diperolehi bagi mendapatkan kebenaran menggunakan set data bagi tujuan kajian ini. Set data mentah tersebut telah dikutip dari sensor di 17 lokasi-lokasi terpilih seperti Rajah 1.1 secara bulanan. Set data mentah tersebut mempunyai 1020 baris dan 27 fitur (termasuk fitur *Class*) telah dikumpulkan oleh pihak BASTW bagi memenuhi piawaian NLWQS seperti yang ditetapkan oleh DOE.

ii. Klasifikasi Air Secara Manual

Formula yang digunakan bagi tujuan menentukan klasifikasi kualiti air secara manual mengikut NLWQS adalah seperti berikut:

WQI = (0.22* SlDO) + (0.19*SIBOD) + (0.16*SICOD) + (0.15*SIAN) + (0.16 * SISS) + (0.12 * SipH)

di mana:

SIDO = Sublndex DO (Dissolved Oxygen)
SIBOD = Sublndex BOD (Biochemical Oxygen Demand)
SICOD = Sublndex COD (Chemical Oxygen Demand)
SIAN = Sublndex NH3-N (Ammoniacal Nitrogen)
SISS = Sublndex SS (Suspended Solid)
SipH = Sublndex pH
$0 \leq WQI \leq 100$

Fitur yang terlibat di dalam pengiraan di atas ialah Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Ammoniacal Nitrogen (NH3-N), Suspended Solid (TSS) dan pH. Baki fitur lain selain daripada enam fitur yang terlibat dalam pengiraan di atas dikutip bagi tujuan kelas IIB seperti yang ditetapkan di dalam NWQS oleh DOE. Penentuan klasifikasi kualiti air berdasarkan formula NLWQS di atas adalah seperti Rajah 3.1 berikut:

Jadual 3.1 Klasifikasi Air NLWQS

KATEGORI	KETERANGAN
A	Tasik dalam kelas ini yang diuruskan di mana air yang boleh digunakan untuk tujuan rekreasi – hubungan badan utama seperti berenang, menyelam dan berkayak.
B	Tasik dalam kelas ini digunakan untuk tujuan rekreasi – hubungan badan sekunder seperti menaiki bot dan pelayaran. Aktiviti berenang tidak dibenarkan dalam kategori ini.
C	Tasik dalam kelas ini bertujuan untuk memelihara hidupan akuatik dan biodiversiti
D	Tasik dalam kelas ini diuruskan untuk pemeliharaan minimum kehidupan akuatik yang baik di tasik dan bertujuan untuk pemeliharaan hidupan akuatik dan biodiversiti. Ia menggunakan amalan pengurusan tasik yang baik.

Penggunaan pembelajaran mesin untuk pengiraan IKA mampu menawarkan penilaian kualiti air yang lebih tepat, dinamik dan automatik terutamanya bagi set data yang kompleks dan besar. Berbanding pengiraan secara manual di atas, penggunaan pembelajaran mesin mampu mengurangkan masa pemprosesan. Penggunaan pembelajaran mesin menggunakan jumlah pemboleh ubah (*variable*) yang kurang dan lebih memberi impak berbanding pengiraan konvensional. Hanya pemboleh ubah yang paling relevan dipilih melalui kaedah pemilihan fitur untuk mengelaskan kualiti air. Hasilnya tempoh pengiraan menjadi lebih singkat berbanding pengiraan secara manual.

Penggunaan pembelajaran mesin juga lebih berkesan dan pantas dari segi memproses data dalam jumlah yang banyak berbanding penggunaan konvensional. Kaedah pembelajaran mesin mampu memproses jumlah data besar bersifat data raya yang tidak mampu dianalisis menggunakan kaedah konvensional. Sebagai contoh data bersaiz Terabyte (TB) akan menyebabkan perisian Microsoft Excel mengambil masa yang terlalu lama untuk memuatkan (*loading*) dan membuka fail.

iii. Set Data Mentah

Data yang diperolehi adalah berstruktur dan berlabel dari tahun 2018 sehingga 2022. Nama bagi setiap fitur, bentuk data dan keterangan adalah seperti Jadual 3.2 seperti berikut:

Jadual 3.2 Fitur Set Data Mentah

Nama Fitur	Bentuk Data	Keterangan
Sampling_Station	nominal	Stesen Pengumpulan Data
Sampling_Date	angka(format tarikh)	Tarikh Pengumpulan Data
Time	angka(format masa)	Masa Pengumpulan Data
Weather	nominal	Cuaca Ketika Pengumpulan Data (Baik atau Mendung)
Water	nominal	Keadaan Air Tasik (Normal)
Color	nominal	Warna Air Tasik (Jernih, Perang atau Sangat Perang)
Temperature	angka	Suhu dalam Celcius
pH	angka	pH Air Tasik
D.O	angka	Oksigen Terlarut % Ketepuan
D.O_mg/l	angka	Oksigen Terlarut dalam mg/l
Conductivity_µS/cm	angka	Keupayaan bahan untuk membawa arus elektrik, diukur dalam microsiemens per cm.
Salinity_ppt	angka	Garam terlarut dalam air diukur dalam Parts Per Thousand iaitu nisbah jisim garam kepada jumlah jisim larutan (air dan garam).
Transparency_meter	angka	Ketelusan air, iaitu keupayaan air untuk membolehkan cahaya melaluinya dalam ukuran meter.
Ammonium_mg/l	angka	Kepekatan ion ammonium dalam air diukur dalam miligram per liter (mg/l)
Turbidity_NTU	angka	Kekeruhan di dalam air dinyatakan dalam Unit Kekeruhan Nephelometric (NTU). Ukuran kekeruhan atau *haziness* cecair disebabkan oleh sejumlah besar partikel tidak boleh dilihat dengan mata kasar.
BOD_mg/l	angka	Permintaan Oksigen Biokimia dalam miligram per liter.
COD_mg/l	angka	Permintaan Oksigen Kimia dalam miligram per liter.
TSS_mg/l	angka	Jumlah Pepejal Terampai dalam miligram per liter.
NH3N_mg/l	angka	Kepekatan nitrogen Ammonia dalam air dengan ukuran miligram per liter.
T.Phosphorus_mg/l	angka	Kepekatan jumlah Fosforus dalam air dengan ukuran miligram per liter.
T.Nitrogen_mg/l	angka	Kepekatan jumlah Nitrogen dalam air dengan ukuran miligram per liter.
Chlorophyll-a_µg/l	angka	Kepekatan jumlah Chlorophyll-a dalam air dengan ukuran miligram per liter.
E. coli_cfu/100ml	angka	Kepekatan bakteria Escherichia coli (E. coli) dalam air diukur dalam unit pembentukan koloni setiap 100 mililiter.
F.Coliform_cfu /100ml	angka	Kepekatan bakteria Coliform dalam air dan diukur dalam unit pembentukan koloni setiap 100 mililiter.
T.Coliform	angka	Jumlah kumpulan bakteria Coliform dalam air.

bersambung…

…sambungan

WQI	angka	Peratusan indeks kualiti air (IKA) berdasarkan piawaian NLWQS
Class	nominal	Peratusan WQI ditentukan berdasarkan kelas di dalam Rajah 3.1 dengan nilai I mewakili kelas A, II mewakili kelas B, III mewakili kelas C dan IV mewakili kelas D.

Jumlah fitur = 27

3.2.2 Prapemprosesan dan Penerokaan Data

Prapemprosesan data merupakan peringkat asas bagi kaedah pembelajaran mesin oleh kebanyakan penyelidik (Gulati & Raheja 2021). Prapemprosesan dan penerokaan data merupakan fasa permulaan yang kritikal dalam mana-mana kajian pembelajaran mesin. Kedua-duanya merupakan langkah penting dalam analisis data dan aliran proses pembelajaran mesin. Fasa ini perlu bagi memastikan penyediaan dan pemahaman data dibuat sebelum menggunakan sebarang teknik analisis atau pemodelan. Perpustakaan *Python* seperti *Pandas* efektif dalam mengendalikan pembersihan, transformasi dan visualisasi data. Perpustakaan pemplot yang berkeupayaan tinggi seperti *Matplotlib* dan *Seaborn* membantu meneroka data dan mengenal pasti corak.

a. Prapemprosesan Data

Prapemprosesan data adalah langkah penting dalam pembelajaran mesin yang memberi impak besar kepada prestasi model. Kommareddy (2022) menekankan kepentingan fasa ini, di mana beliau menyediakan rangka kerja komprehensif untuk pembelajaran mesin yang diselia. Secara umumnya prapemprosesan data melibatkan empat tunjang iaitu pembersihan data, pengurangan data, transformasi data dan korelasi data seperti Rajah 3.2 seperti di bawah:

Rajah 3.2 Prapemprosesan Data

Proses-proses ini boleh dibuat secara berulang dan dilaksanakan samada menuruti turutan langkah tertentu atau sebaliknya. Tidak semua proses tersebut perlu dilaksanakan kerana setiap langkah bergantung kepada set data yang ingin diproses. Proses ini melibatkan tugasan yang bermula dengan pembersihan data diakhiri dengan korelasi data.

i. Pembersihan Data:

Pembersihan data dibuat untuk membetulkan kesilapan, ketidakkonsistenan dan mengendalikan data yang hilang menggunakan pelbagai kaedah seperti imputasi dan penyingkiran. Perisian *Microsoft Excel* digunakan di dalam proses ini. Di dalam set data mentah kajian ini, semakan mendapati tiada sebarang kehilangan (*null*) atau pengulangan data di dalam setiap fitur, namun terdapat 55 baris kosong yang sengaja dikosongkan di antara bacaan setiap stesen. Kesemua baris kosong tersebut telah dihapuskan kerana tidak mempunyai sebarang kesan terhadap kajian.

Bagi pengendalian data yang tidak releven, tindakan pembetulan telah diambil seperti Jadual 3.3 di bawah.

Jadual 3.3 Pembersihan Data

Data Asal	Kolum	Sebab	Tindakan
96..8 dan 96..9	D.O	Mengandungi 2 titik perpuluhan.	Membuang 1 titik perpuluhan.
6..4	Turbidity_NTU	Mengandungi 2 titik perpuluhan.	Membuang 1 titik perpuluhan.
<0.01	NH3N mg/l	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.005
<0.01	T. Phosphorus mg/l	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.005
<1	T.Nitrogen mg/l	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.5
<0.5	Chlorophyll-a µg/l	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.25
ND(<1)	E. coli cfu/100ml	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.5

bersambung…

…sambungan

ND(<1) dan <1	F. Coliform cfu/100ml	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.5
<1	T. Coliform	Sensor tidak dapat membaca data di luar julat minimum.	Mengganti data kepada 0.5

ii. Pengurangan Data:

Pengurangan data perlu dibuat bagi mempercepatkan proses pembangunan dan penilaian model. Salah satu kaedah yang digunakan di dalam kajian ini ialah menghapuskan kolum yang melibatkan mengalih keluar fitur-fitur yang dianggap tidak relevan, berlebihan atau tidak membantu untuk tugas pembelajaran mesin dan tidak mempunyai sebarang signifikan terhadap kajian.

iii. Transformasi Data:

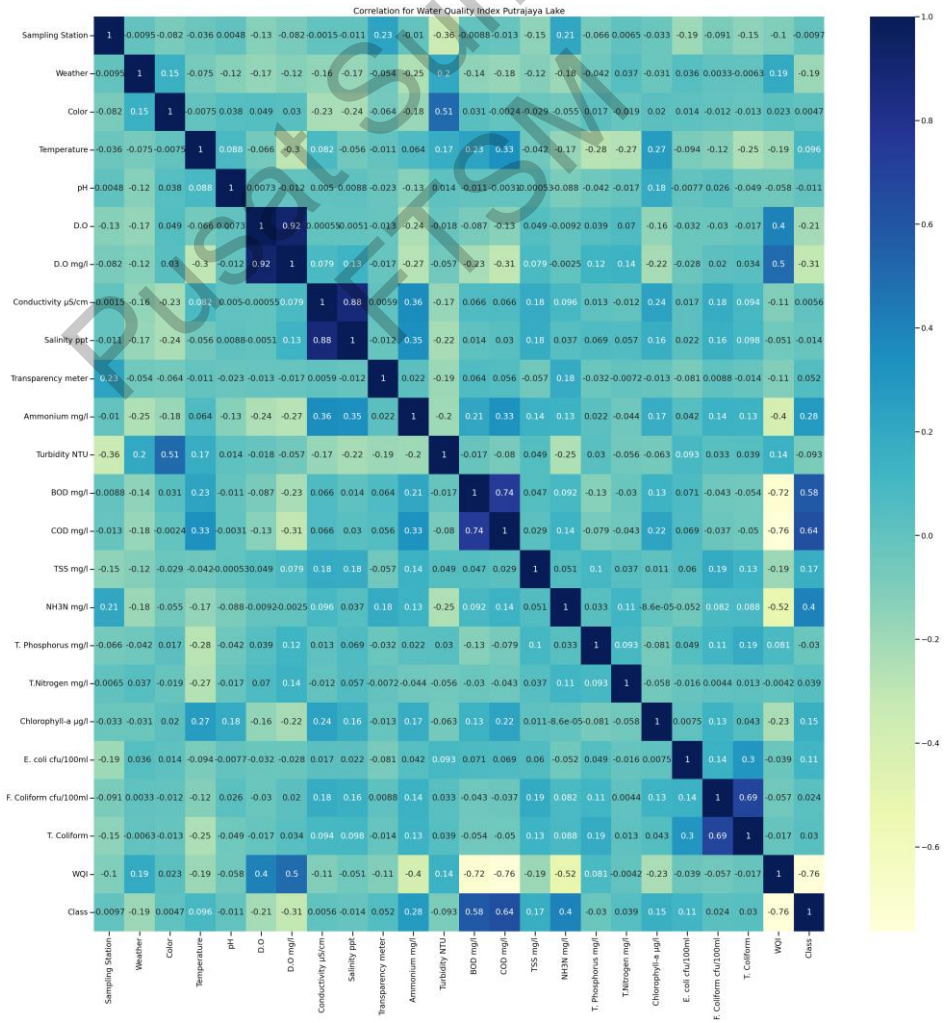
Transformasi data dilaksanakan bertujuan untuk menyeragamkan atau menormalkan fitur berangka. Proses ini melibatkan penukaran fitur kategori kepada fitur berangka menggunakan perpustakaan pemetaan dan merupakan teknik biasa dalam prapemprosesan data untuk menyediakan data untuk algoritma pembelajaran mesin yang memerlukan fitur berangka.

Kaedah Penskalaan Data (*Data Scaling*) untuk pembelajaran mesin bertujuan untuk menentukan kawasan pembelajaran mesin yang terkini di mana penskalaan memainkan peranan utama dan harus dilaksanakan dengan betul untuk mengurangkan ketidakpastian, keputusan yang tidak tepat atau peningkatan kos dan masa pemprosesan (Sharma 2022). Penskalaan data juga merupakan salah satu proses Transformasi Data yang bertujuan untuk menyesuaikan julat fitur berangka ke skala yang sama. *StandardScaler* ialah teknik popular yang menskalakan fitur untuk mempunyai min ‘0’ dan sisihan piawai ‘1’ bagi memastikan semua fitur diperlakukan sama rata dan mengelakkan isu-isu yang berkaitan dengan fitur-fitur dengan julat yang berbeza-beza.

iv. Korelasi Data:

Korelasi Data bertujuan untuk memilih fitur yang paling relevan dan bermaklumat. Kajian ini juga membantu meningkatkan prestasi model, kebolehtafsiran, kecekapan, dan penggunaan sumber. Antara kaedah yang digunakan ialah analisis matrik korelasi dan *Information Gain*. Korelasi adalah ukuran bagaimana dua atau lebih pembolehubah berkaitan dengan satu sama lain, juga dirujuk sebagai pergantungan linear. Perpustakaan *Panda* (.corr) di gunakan di dalam kajian ini.

Melalui kaedah ini, hanya fitur-fitur yang mempunyai kaitan tinggi sahaja dipilih untuk ke peringkat permodelan. Ebron et al. (2020) mendapati korelasi yang lemah antara fitur kualiti air menunjukkan bahawa data tidak boleh dimodelkan dengan berkesan.



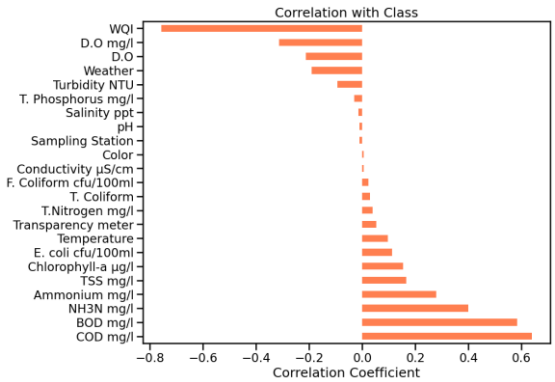
Rajah 3.3 Korelasi Antara Fitur

Rajah 3.3 di atas menunjukkan hubungan korelasi antara fitur. Nilai positif 1 atau negatif 1 menunjukkan tahap korelasi antara fitur. Semakin gelap warna kotak menunjukkan nilai korelasi positif antara fitur semakin tinggi menghampiri '1'. Semakin cerah warna kotak juga menunjukkan nilai korelasi negatif antara fitur semakin tinggi menghampiri '-1'. Korelasi *coefficient* hampir dengan '0' menunjukkan sedikit atau menghampiri tiada hubungan korelasi antara dua fitur iaitu sebarang perubahan pada sesuatu fitur sedikit atau tidak mengubah fitur lain. Bagi memilih fitur yang kuat, setiap fitur dikira menggunakan fungsi samada mempunyai hubungan yang tinggi terhadap pemboleh ubah bersandar yang dipilih iaitu fitur *Class*.

Output nilai korelasi dan graf keseluruhan bagi pengiraan korelasi di atas adalah seperti Jadual 3.4 dan Rajah 3.4 di bawah.

Jadual 3.4 Nilai Korelasi Setiap Fitur terhadap Fitur Class

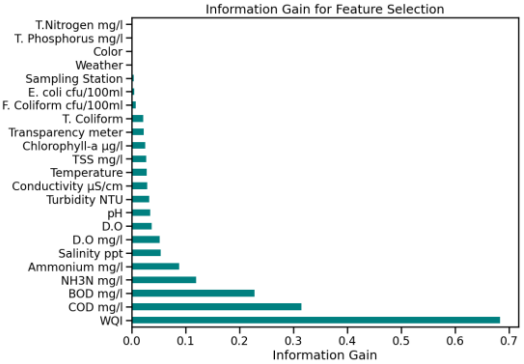
Fitur	**Nilai Korelasi**
Sampling Station	-0.009698
Weather	-0.190540
Color	0.004688
Temperature	0.096409
pH	-0.010602
D.O	-0.213100
D.O mg/l	-0.312885
Conductivity µS/cm	0.005615
Salinity ppt	-0.014127
Transparency meter	0.052298
Ammonium mg/l	0.278204
Turbidity NTU	-0.093284
BOD mg/l	0.584011
COD mg/l	0.638656
TSS mg/l	0.165505
NH3N mg/l	0.400043
T. Phosphorus mg/l	-0.029921
T.Nitrogen mg/l	0.038985
Chlorophyll-a µg/l	0.153288
E. coli cfu/100ml	0.111860
F. Coliform cfu/100ml	0.024104
T. Coliform	0.029987
WQI	-0.756577



Rajah 3.4 Graf Nilai Korelasi Setiap Fitur terhadap Fitur Class

Berdasarkan Jadual 3.4 dan Rajah 3.4 di atas, 14 fitur mempunyai nilai positif adalah COD mg/l, BOD mg/l, NH3N mg/l, Ammonium mg/l, TSS mg/l, Chlorophyll-a µg/l, E. coli cfu/100ml, Temperature, Transparency meter, T.Nitrogen mg/l, T. Coliform, F. Coliform cfu/100ml, Conductivity µS/cm dan Color. Baki sembilan fitur mempunyai nilai negatif iaitu Sampling Station, pH, Salinity ppt, T. Phosphorus mg/l, Turbidity NTU, Weather, D.O, D.O mg/l dan WQI. Hasil daripada kajian di atas juga menunjukkan fitur pH, Sampling Station, Conductivity µS/cm dan Color mempunyai hubungan korelasi *coefficient* yang lemah iaitu menghampiri nilai '0'.

Information Gain (IG) juga merupakan kaedah lain untuk mengukur keberkesanan fitur dalam mengklasifikasikan titik data, dengan menilai keuntungan setiap pembolehubah dalam konteks pembolehubah sasaran. Graf *Information Gain* berdasarkan Rajah 3.5 dihasilkan menggunakan kod *Python* dengan fungsi *mutual_info_classif* dari perpustakaan *sklearnfeature_selection*.



Rajah 3.5 Graf Nilai Information Gain Setiap Fitur terhadap Fitur Class

WQI, COD mg/l, BOD mg/l, NH3N mg/l, Ammonium mg/l dan Salinity ppt merupakan enam fitur yang memperolehi nilai keuntungan (IG) yang tinggi berbanding T.Nitrogen mg/l, T. Phosphorus mg/l, Color, Weather dan Sampling Station yang menghampiri nilai '0'.

b. Penerokaan Data

Fasa ini memberi tumpuan kepada memahami fitur-fitur data, mengenal pasti corak, mendedahkan hubungan yang berpotensi antara fitur-fitur dan menyediakannya untuk analisis lanjut dan pembinaan model. Penerokaan data adalah langkah pertama yang penting dalam mana-mana metodologi pembelajaran mesin dan melibatkan pelbagai teknik termasuk statistik deskriptif, data deskriptif dan korelasi data.

i. Statistik Deskriptif

Statistik deskriptif bertujuan mengira ukuran seperti min, median, sisihan piawai (std), kuartil dan persentil bagi setiap fitur individu dan taburan frekuensi bagi keseluruhan set data untuk memberikan gambaran asas mengenai taburan data. Kaedah ini juga adalah aspek asas penerokaan data, menyediakan ringkasan kuantitatif kecenderungan, penyebaran, dan bentuk pusat data. Gambaran asas ini mencadangkan pilihan teknik penerokaan selanjutnya dan membantu mengenal pasti anomali atau *bias* yang berpotensi dalam data. Namun kaedah ini hanya boleh digunakan pada fitur jenis numerik atau berangka.

Pengukuran Kecenderungan Pusat (*Central Tendency*) dan Penyebaran (*Spread*):

Pengukuran Kecenderungan Pusat (Central Tendency) dan Penyebaran (Spread) adalah dua konsep utama dalam statistik yang membantu kita memahami fitur-fitur set data. Kedua-duanya menunjukkan nilai tipikal dalam data dan jumlah variasi yang terdapat di antara nilai-nilai tersebut. Dalam kajian ini pengukuran *count, min, median, mod* (*max*), standard deviation (std) dan kuartil. dilaksanakan dengan menggunakan kod *python df.describe()* bagi mengeluarkan Jadual 3.5 di bawah:

Jadual 3.5 Pengukuran Kecenderungan Pusat dan Penyebaran Data

Pengukuran	Temperature	pH	D.O	D.O mg/l	Conductivity µS/cm	Salinity ppt	Transparency meter	Ammonium mg/l	Turbidity NTU	BOD mg/l	COD mg/l	TSS mg/l	NH3N mg/l	T. Phosphor	T.Nitrogen mg/l	Chlorophyll-a µg/l	E. coli cfu/100ml	F. Coliform cfu/100ml	T. Coliform	WQI
count	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020	1020
mean	28.99	7.6	97	7.58	89.43	0.04	1.11	0.24	3.19	3.68	12.53	5.22	0.05	0.03	0.63	8.21	174.28	6188.12	17510.29	92.06
std	1.4	0.13	6.76	0.61	10.08	0.01	0.59	0.06	2.48	1.01	3.95	3.41	0.06	0.03	0.51	7.01	457.05	9774.62	20180.36	2.07
min	25.45	6.92	56.4	4.29	52.5	0.02	0.2	0.05	0.1	1	4	0	0.01	0.01	0.5	0.25	0.5	0.5	0.5	84.27
25%	28.05	7.54	94.3	7.31	84	0.04	1	0.22	1.77	3	10	3	0.01	0.01	0.5	2.68	0.5	1100	5200	90.83
50%	29.29	7.59	98	7.63	88	0.04	1.1	0.24	2.5	4	12	4	0.03	0.02	0.5	7.05	0.5	3000	9900	92.42
75%	29.99	7.64	101.1	7.99	94	0.04	1.2	0.27	3.9	4	15	7	0.07	0.04	0.5	12.12	200	7200	23000	93.51
max	31.69	8.64	122.3	9.16	148	0.07	14	0.6	24.86	6	24	21	0.33	0.42	5	56.3	5600	120000	160000	96.43

Kiraan (*count*) merupakan jumlah kiraan keseluruhan nilai yang terdapat di dalam setiap fitur. Min (*mean*) adalah nilai purata semua titik data yang menunjukkan nilai tipikal di dalam set data. Median pula adalah nilai pertengahan apabila data disusun dari terendah ke tertinggi dan kurang sensitif kepada outliers berbanding min. Mod atau max menunjukkan nilai yang paling kerap muncul dalam data dan berguna untuk mengenal pasti trend yang dominan. Kuartil membahagikan set data kepada empat bahagian yang sama, disusun dari nilai terkecil hingga terbesar. Pada dasarnya, kuartil mewakili persentil ke-25 (Q1), ke-50 (Q2), dan ke-75 (Q3) set data. Kuartil dapat menggambarkan bentuk taburan data sebagai contoh sekiranya Q1 dan Q3 menghampiri median, ia menunjukkan pengedaran cenderung ke tengah dan simetri. Sebaliknya jika jauh dari median, ia menunjukkan taburan condong dengan ekor yang lebih panjang.

Julat dan Sisihan Piawai merupakan ukuran bagi penyebaran dan pengagihan data selain daripada Kuartil. Julat merupakan Perbezaan antara nilai terbesar dan terkecil dalam data dan menyediakan pemahaman segera dan intuitif tentang bagaimana data disebarkan. Julat mudah dikira dan ditafsirkan, mampu menetapkan sempadan untuk data dan boleh menunjukkan kehadiran outliers. Julat yang lebih luas menunjukkan kebolehubahan yang lebih besar dalam data.

Sisihan Piawai pula mengukur sisihan purata semua titik data dari min. Sisihan Piawai adalah ukuran penyebaran yang lebih tinggi berbanding julat dengan menganggap semua titik data (bukan hanya yang ekstrem) memberikan gambaran yang lebih tepat tentang bagaimana data disebarkan. Dalam pengagihan biasa (normal distributions), Sisihan Piawai memainkan peranan penting dengan memaparkan berapa banyak data yang berada dalam julat tertentu di sekitar min. Selain itu, nilai Sisihan Piawai yang tinggi juga menunjukkan kehadiran *outliers* yang tinggi.

Berdasarkan Jadual 3.5 di atas, dari segi pengukuran kecenderungan pusat, nilai min bagi pH, D.O, dan BOD menunjukkan keadaan yang sesuai untuk kehidupan akuatik. COD dan TSS pula menunjukkan nilai min yang sederhana, manakala NH3-N sangat rendah yang menunjukkan pencemaran nitrogen yang minimum. Purata WQI 92.06 mengesahkan kualiti air keseluruhan yang sangat baik. Manakala nilai median hampir serupa bagi kebanyakan fitur, kecuali COD dan TSS, yang menunjukkan median yang sedikit rendah mempunyai beberapa *outliers*.

Dari sudut penyebaran dan pengagihan data pula, julat yang luas untuk COD dan TSS menyerlahkan potensi untuk turun dan naik (*fluctuation*) yang ketara dan berkemungkinan memerlukan pemantauan lanjut. Julat yang lebih kecil pada pH, D.O dan NH3-N menunjukkan nilai yang lebih konsisten. Kuartil pula mendedahkan corak penyebaran data. Perbezaan kecil antara Q1, Q2, dan Q3 untuk pH, D.O. dan BOD menunjukkan penyebaran data yang kelihatan berpusat dan simetri. Jurang yang lebih besar untuk COD dan TSS menunjukkan potensi kecondongan ke arah nilai yang lebih tinggi. Sisihan piawai yang lebih tinggi untuk COD dan TSS berbanding fitur lain menunjukkan kebolehubahan yang lebih besar dalam nilai kedua-duanya.

Pengukuran Bentuk (*Shape*):

Indeks Kecondongan (*Skewness*) dan Kurtosis digunakan untuk mengenal pasti normaliti data. Kline (2011) berpendapat bahawa untuk taburan data yang normal, kecondongan hendaklah berada dalam julat nilai ±3, manakala kurtosis hendaklah berada dalam julat nilai ±10. Kecondongan (*Skewness*) merupakan salah satu dari kaedah pengukuran bentuk data. Kaedah ini mengukur asimetri pengagihan data. Kecondongan positif menunjukkan ekor yang lebih panjang di sebelah kanan, manakala kecondongan negatif menunjukkan ekor yang lebih panjang di sebelah kiri.

Jadual 3.6 Ukuran Kecondongan Fitur

Fitur	Nilai Skewness
Transparency meter	20.09
E. coli cfu/100ml	5.79
T.Nitrogen mg/l	4.80
F. Coliform cfu/100ml	4.77
T. Phosphorus mg/l	4.30
Turbidity NTU	3.21
T. Coliform	2.82
NH3N mg/l	2.01
Chlorophyll-a µg/l	1.35
pH	1.19
TSS mg/l	1.19
Conductivity µS/cm	0.83
Salinity ppt	0.71
Ammonium mg/l	0.55
BOD mg/l	0.48
COD mg/l	0.24
Temperature	-0.46
WQI	-0.69
D.O mg/l	-1.23
D.O	-1.47

Berdasarkan Jadual 3.6 di atas didapati fitur E. coli cfu/100ml (5.79), T.Nitrogen mg/l (4.80), F. Coliform cfu/100ml (4.77) dan T. Phosphorus mg/l (4.30) dikategorikan sebagai sangat condong ke kanan (positif). Dua fitur di bawah kategori sederhana condong ke kanan iaitu Turbidity NTU (3.21) dan T. Coliform (2.82). Tiga fitur pula di bawah kategori sedikit condong ke kanan iaitu NH3N mg/l (2.01), Chlorophyll-a µg/l (1.35) dan pH (1.19). Kebanyakan baki fitur lain mempunyai nilai kecondongan menghampiri 0, menunjukkan taburan yang agak simetri.

Hasil analisis kecondongan menunjukkan terdapat beberapa fitur (terutamanya bakteria, nutrien, dan kekeruhan) cenderung ke arah pengedaran yang condong kanan kemungkinan disebabkan *outliers*. Ini menunjukkan terdapat kebarangkalian pencemaran atau kejadian lain memberi kesan ketara kepada fitur ini, yang membawa kepada nilai yang agak tinggi.

Selain dari kecondongan, Kurtosis juga merupakan kaedah pengukuran bentuk data yang kerap digunakan dengan mengukur puncak pengagihan data. Kurtosis tinggi menunjukkan taburan puncak, manakala kurtosis rendah menunjukkan taburan rata. Dalam erti kata lain, Kurtosis yang tinggi mungkin menunjukkan bahawa kebanyakan titik data berkelompok di sekitar min, dengan *outliers* yang lebih rendah pada kedua-dua hujungnya.

Jadual 3.7 Ukuran Kurtosis Fitur

Fitur	**Nilai Kurtosis**
Transparency meter	434.94
E. coli cfu/100ml	45.33
T. Phosphorus mg/l	36.90
F. Coliform cfu/100ml	35.66
T.Nitrogen mg/l	25.86
Turbidity NTU	16.51
T. Coliform	11.27
pH	11.05
Ammonium mg/l	7.00
NH3N mg/l	4.84
Salinity ppt	4.60
D.O	4.58
Conductivity µS/cm	4.01
Chlorophyll-a µg/l	3.59
D.O mg/l	2.77
TSS mg/l	1.49
WQI	0.27
BOD mg/l	-0.09
COD mg/l	-0.22
Temperature	-0.61

Pengukuran Kurtosis dari pada Jadual 3.7 di atas mendapati lapan fitur mempunyai nilai Kurtosis melebihi ‘10’ iaitu Transparency meter (434.94), E. coli cfu/100ml (45.33), T. Phosphorus mg/l (36.90), F. Coliform cfu/100ml (35.66), T.Nitrogen mg/l (25.86), Turbidity NTU (16.51), T. Coliform (11.27) dan pH (11.05). Dua fitur memberikan nilai Kurtosis yang sederhana (antara 3-10) iaitu Ammonium

mg/l (6.99) dan NH3N mg/l (4.83). Baki sembilan fitur pula memulangkan nilai Kurtosis yang rendah iaitu Salinity ppt (4.60), D.O (4.58), Conductivity µS/cm (4.00), Chlorophyll-a µg/l (3.59), D.O mg/l (2.76), TSS mg/l (1.49), WQI (0.27), BOD mg/l (-0.09) dan COD mg/l (-0.22).

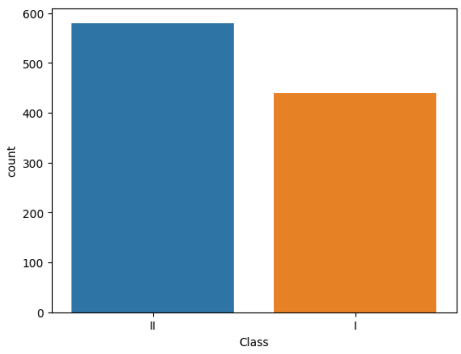
Hasil bacaan Kurtosis mendedahkan bahawa beberapa fitur terutamanya yang melibatkan bakteria, nutrien, dan kekeruhan mempunyai kurtosis yang tinggi iaitu agihan titik data kebanyakannya di puncak graf dengan ekor yang berat. Bacaan ini menunjukkan kebarangkalian terdapat peristiwa atau kejadian berhampiran lokasi pengumpulan data memberi kesan ketara kepada bacaan fitur ini.

ii. Data Deskriptif

Teknik data deskriptif menggunakan visualisasi melalui graf bar, *histogram*, *swarm plots*, dan *box plots* membantu menggambarkan pengedaran data, mengenal pasti hubungan antara fitur dan mengesan *outliers.*

Fitur Berkategori (*Categorical*):

Bagi mendapatkan gambaran taburan data, graf bar diperlukan untuk menunjukkan samada terdapat ketidakseimbangan data. Rajah 3.6 pula menunjukkan output bagi fitur berkategori '*Class*' yang merupakan pemboleh ubah bersandar bagi set data ini.

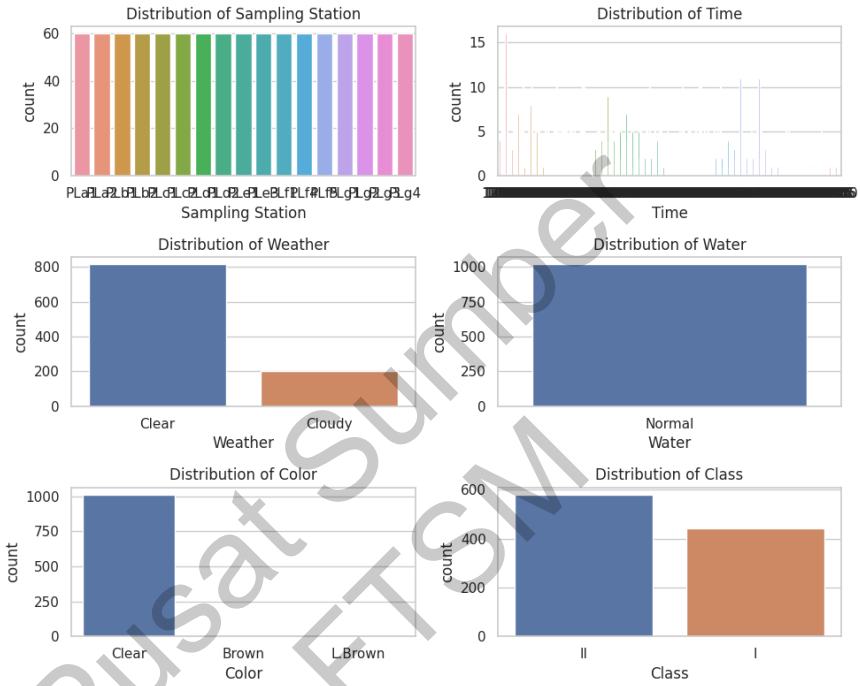


Rajah 3.6 Graf Kiraan Fitur Class

Di dalam Rajah 3.6 di atas, didapati Kelas II iaitu indeks air di bawah kategori B seperti Jadual 3.1 di atas mempunyai kiraan 580 atau 56.86% daripada jumlah keseluruhan manakala baki bagi Kelas I (Kategori A) sejumlah 440 atau 43.14%.

Tiada bacaan bagi Kelas III atau IV menunjukkan Tasik Putrajaya amat sesuai digunakan untuk aktiviti rekreasi air.

Bagi mendapatkan gambaran berkaitan semua fitur jenis kategori, Rajah 3.7 memaparkan output yang dihasilkan bagi fitur jenis kategori seperti Sampling Station, Time, Weather, Water, Color dan Class.

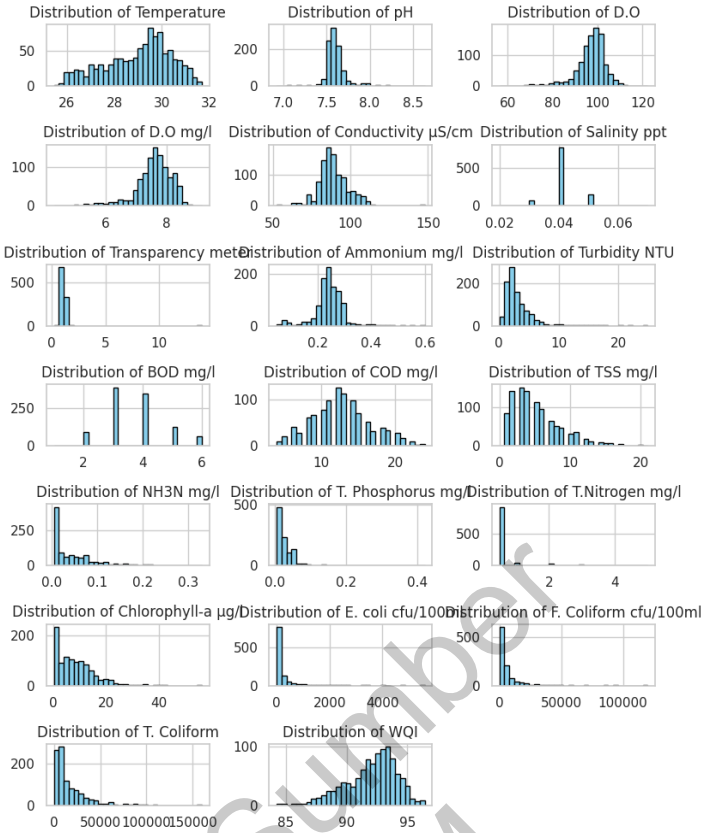


Rajah 3.7 Graf Fitur Berkategori

Daripada Rajah 3.7 di atas, didapati fitur Sampling Station mempunyai kiraan yang sama bagi setiap 17 stesen pengumpulan data. Fitur Weather menunjukkan *Cloudy* memperolehi seperempat daripada jumlah kiraan *Clear*. Fitur Water hanya mempunyai satu nilai iaitu *Normal*. Fitur Color menunjukkan kiraan *Clear* mendominasi dengan nilai tertinggi 1000 berbanding *Brown Color* dan *Light Brown* yang mempunyai kiraan terlalu sedikit.

Fitur Berangka (*Numerical*):

Histogram merupakan perwakilan grafik taburan data yang meringkaskan kekerapan titik data yang berlaku dalam julat tertentu. Histogram seperti Rajah 3.8 menunjukkan taburan data bagi kesemua fitur jenis berangka.

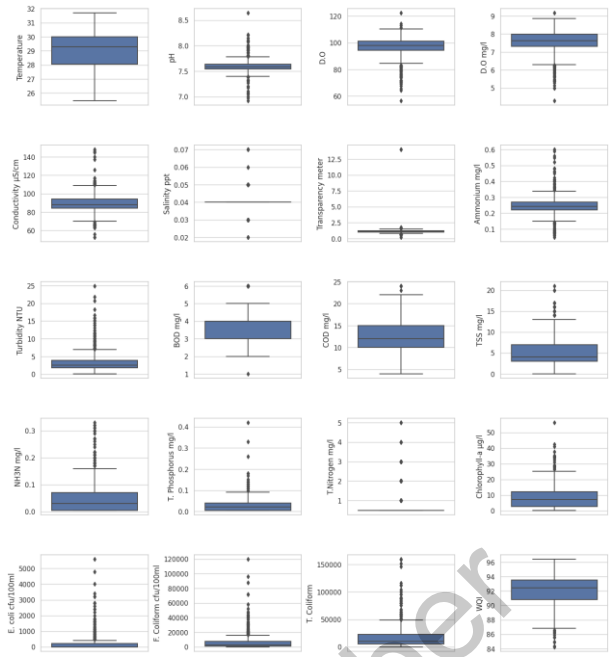


Rajah 3.8 Histogram Fitur Berangka

Berdasarkan Rajah 3.8 di atas, fitur Temperature, D.O, D.O mg/l dan WQI mempunyai kecondongan negatif. Fitur pH, Conductivity µs/cm, Ammonium mg/l, Turbidity NTU, TSS mg/l, NH3N mg/l, T.Phosphorus mg/l, Chlorophyll-a ug/l, E. coli cfu/100ml, F. Coliform cfu/100ml dan T. Coliform cfu/100ml menunjukkan kecondongan positif. Hanya fitur COD mg/l menunjukkan taburan normal.

Plot kotak juga dikenali sebagai plot kotak dan misai, adalah alat grafik atau visualisasi yang sangat berguna untuk menggambarkan beberapa aspek utama taburan set data kuantitatif. Kotak mewakili julat interquartile (IQR), memberi penekanan terhadap pertengahan 50% data. Bahagian bawah kotak menandakan Q1 (kuartil pertama), garis dalam kotak adalah median (Q2), dan bahagian atas kotak menandakan Q3 (kuartil ketiga).

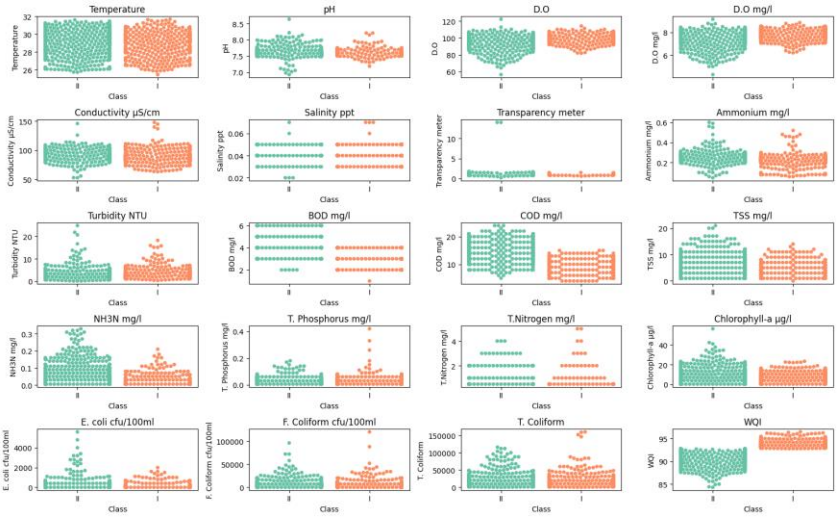
Misai terletak jauh dari kotak dan mewakili baki 25% data. Kedua-dua misai biasanya mencapai sehingga 1.5 kali ganda IQR terletak di atas dan di bawah kotak. *Outliers* merupakan titik data yang terletak di luar misai diplotkan secara individu.



Rajah 3.9 Plot Kotak Fitur Berangka

Rajah 3.9 di atas menunjukkan fitur Temperature mempunyai IQR yang terletak hampir di pertengahan graf begitu juga COD dan BOD. Fitur pH juga mempunyai IQR di pertengahan namun mempunyai *outliers* yang banyak di luar julat samada atas atau bawah misai. Manakala fitur Transparency Meter, Turbidity NTU, TSS mg/l, NH3N mg/l, T. Phosphorus mg/l, Chlorophyll-a ug/l, E. coli cfu/100ml, F. Coliform cfu/100ml dan T. Coliform cfu/100ml mempunyai IQR yang terletak di bawah graf menunjukkan nilai median yang rendah. IQR bagi D.O, D.O mg/l dan WQI terletak sedikit ke atas menunjukkan nilai median yang tinggi. Fitur Salinity ppt menunjukkan IQR yang sangat nipis terletak hampir di pertengahan dengan *outliers* yang banyak di atas dan bawah.

Plot Kawanan merupakan sejenis plot penyebaran bagi kolum kategori yang memaparkan titik data individu di sepanjang paksi tunggal serta mengelakkan titik bertindih dengan melaraskan kedudukan secara mendatar. Setiap titik data diplotkan sebagai titik tunggal di sepanjang paksi kolum kategori dan titik tersebar untuk mengelakkannya daripada bertindih.

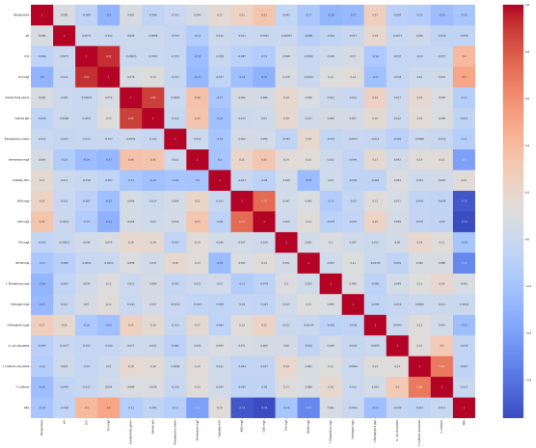


Rajah 3.10 Plot Kawanan

Dari Rajah 3.10 di atas, didapati fitur Temperature, pH, Conductivity µs/cm, Salinity ppt, Ammonium mg/l, Turbidity NTU, T.Phosphorus mg/l, T.Nitrogen mg/l, F. Coliform cfu/100ml dan T. Coliform menunjukkan taburan kawanan titik data yang sekata bagi kedua-dua nilai I dan II. Manakala fitur D.O, D.O mg/l, COD mg/l, TSS mg/l, NH3N mg/l, Chlorophyll-a ug/l, E. Coli cfu/100ml dan WQI menunjukkan sebaliknya iaitu taburan kawanan titik data yang tidak seimbang.

iii. Pekali Korelasi

Pengiraan pekali korelasi adalah bertujuan untuk memadankan pasangan fitur berangka yang mempunyai fitur yang berbeza dan mendedahkan hubungan linear antara fitur berangka.



Rajah 3.11 Matrik Korelasi

Merujuk kepada Rajah 3.11 di atas, berikut disertakan ringkasan bacaan korelasi dari 0.20 hingga 0.92 seperti Jadual 3.8 yang dipilih berdasarkan hubungan linear yang kuat antara fitur yang berbeza.

Jadual 3.8 Korelasi Antara Fitur Dengan Hubungan Kuat

Fitur 1	**Fitur 2**	***Correlation***
D.O mg/l	D.O	0.92
Conductivity µS/cm	Salinity ppt	0.88
COD mg/l	WQI	0.76
BOD mg/l	COD mg/l	0.74
BOD mg/l	WQI	0.72
T. Coliform	F. Coliform cfu/100ml	0.69
NH3N mg/l	WQI	0.52
D.O mg/l	WQI	0.50
Ammonium mg/l	WQI	0.40
WQI	D.O	0.40
Ammonium mg/l	Conductivity µS/cm	0.36
Temperature	COD mg/l	0.33
Ammonium mg/l	COD mg/l	0.33
D.O mg/l	COD mg/l	0.31
T. Coliform	E. coli cfu/100ml	0.30
D.O mg/l	Temperature	0.30
T. Phosphorus mg/l	Temperature	0.28
Ammonium mg/l	D.O mg/l	0.27
Chlorophyll-a µg/l	Temperature	0.27
T.Nitrogen mg/l	Temperature	0.27
NH3N mg/l	Turbidity NTU	0.25
Temperature	T. Coliform	0.25
Ammonium mg/l	D.O	0.24
Conductivity µS/cm	Chlorophyll-a µg/l	0.24
Chlorophyll-a	WQI	0.23
BOD mg/l	Temperature	0.23
D.O mg/l	BOD mg/l	0.23
Chlorophyll-a µg/l	COD mg/l	0.22
Turbidity NTU	Salinity ppt	0.22
Chlorophyll-a µg/l	D.O mg/l	0.22
BOD mg/l	Ammonium mg/l	0.21
Turbidity NTU	Ammonium mg/l	0.20

3.2.3 Pemilihan dan Latihan Model

Fasa pemilihan dan latihan model terdiri daripada dua langkah penting bagi memastikan model yang dipilih sesuai dengan fitur-fitur set data yang digunakan dalam kajian ini. Langkah tersebut adalah mengenalpasti model seterusnya melatih model tersebut bagi melancarkan kajian di fasa selanjutnya.

i. Mengenalpasti Model

Kaedah mengenalpasti model merupakan proses memilih model ramalan terbaik dari sekumpulan model untuk tugas atau masalah tertentu. Pendekatan biasa untuk memilih model terbaik menggunakan satu matrik keseluruhan tidak semestinya berjaya memperolehi model yang paling sesuai untuk aplikasi tertentu (Theissler et al. 2020). Proses ini melibatkan pertimbangan pelbagai faktor seperti jenis masalah, fitur data, keperluan pentafsiran, dan sumber yang ada. Teknik biasa yang kerap digunakan termasuk membandingkan prestasi model pada set latihan dan ujian bagi mengelaskan pembolehubah jenis berkategori (*categorical*) adalah pengelasan atau klasifikasi.

Bagi tujuan pengelasan pembolehubah jenis berterusan (*continuous*) pula, kaedah regresi sering digunakan. Kedua-dua kaedah pengelasan dan regresi ini telah dikaji penggunaanya berdasarkan Jadual 2.2 di atas. Kajian ini menggunakan kaedah pengelasan berikutan fitur Class yang dijadikan sebagai pembolehubah bersandar adalah dari jenis kategori. Nilai yang terkandung di dalam fitur Class ialah I, II, III dan IV yang merupakan label mengikut susunan yang ditetapkan NLWQS seperti di Jadual 3.1 berdasarkan pengiraan IKA.

Selain itu, tahap kerumitan data juga perlu diambilkira seperti menentukan adakah hubungan data tersebut linear atau tidak linear atau sangat kompleks. Jadual 3.9 di bawah menunjukkan hubungan data berdasarkan korelasi pada Rajah 3.11 di atas.

Jadual 3.9 Hubungan Antara Data

Hubungan Antara Data	Penerangan Hubungan
Hubungan Linear	Pemboleh ubah dengan pekali korelasi menghampiri 1 atau -1 menunjukkan hubungan linear yang kuat. Sebagai contoh, D.O dan D.O mg/l. Pemboleh ubah ini sangat dikaitkan secara positif.
Hubungan Bukan Linear	Beberapa pemboleh ubah mempunyai pekali korelasi antara 0.5 dan 0.9, menunjukkan korelasi positif sederhana. Contohnya termasuk Conductivity µS/cm dan Salinity ppt serta D.O mg/l dan WQI. Hubungan ini dikaitkan dengan bukan linear.
Hubungan lemah	Pekali korelasi antara 0.3 dan 0.5 (contohnya, NH3N mg/l dan WQI, D.O mg/l dan WQI) menunjukkan korelasi positif yang lemah.
Hubungan Negatif	Pekali korelasi negatif di bawah 0 hingga -1 (contohnya COD mg/l dan WQI, BOD mg/l dan WQI) menunjukkan korelasi negatif yang kuat.
Hubungan Kompleks	Sesetengah pembolehubah mempunyai pelbagai korelasi dengan pembolehubah lain, dan hubungannya tidak linear dan rumit. Sebagai contoh, BOD mg/l, COD mg/l, dan WQI mempunyai pelbagai korelasi antara satu sama lain, menunjukkan interaksi yang lebih kompleks.
Interaksi Boleh Ubah	Sebilangan pembolehubah, seperti T. Coliform, mempunyai korelasi sederhana dengan pelbagai pembolehubah lain, menunjukkan interaksi yang berpotensi.

Berdasarkan Jadual 2.1 dan langkah-langkah yang telah diambil dalam fasa pertama dan kedua bagi metodologi kajian, algoritma Hutan Rawak (RF), Rangkaian Neural Buatan (ANN) dan Mesin Sokongan Vektor (SVM) dipilih sebagai model yang akan digunakan untuk pengelasan set data IKA Tasik Putrajaya.

ii. Latihan Model

Kemajuan mendadak dalam pembelajaran mesin sejak sepuluh tahun yang lalu telah diterajui oleh ketersediaan data yang sesuai untuk tujuan latihan (Li, Yu & Koudas 2021). Proses pembangunan model yang dipilih untuk tujuan latihan bertujuan untuk mempelajari corak dan hubungan asas.

Pengagihan Data:

Sebelum model dilatih, beberapa langkah awalan perlu diambil seperti pengagihan data menggunakan kaedah *train_test_split* untuk mengagihkan data kepada set latihan dan set ujian. Perpustakaan *train_test_split* juga diimport daripada *sklearn.model_selection.* Saiz bagi set ujian ditetapkan sebanyak 30% manakala set latihan sebanyak 70%.

Penskalaan Data:

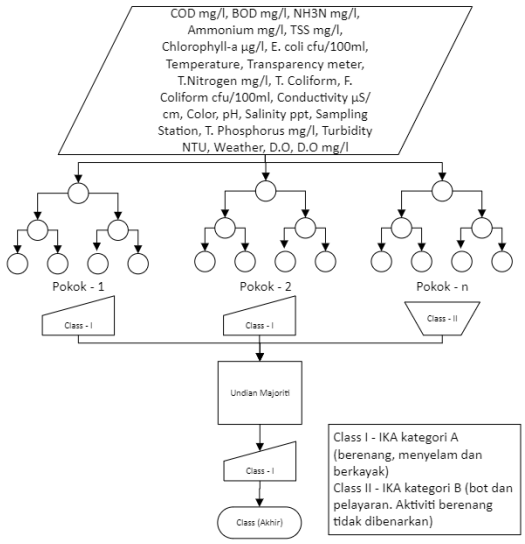
Seterusnya penskalaan data dibuat menggunakan kaedah *StandardScaler* yang merupakan perpustakaan yang diimport daripada *sklearn.preprocessing. X_train_scaled* merupakan nilai X bagi set latihan yang telah diskala manakala *X_test_scaled* merupakan nilai X bagi set ujian yang telah diskala. Kedua-dua nilai X yang telah diskala akan digunakan di dalam fasa latihan dan ujian seterusnya.

Pemilihan Fitur:

Pemilihan fitur merupakan antara proses penting dalam pembelajaran mesin untuk mengenalpasti dan memilih fitur yang paling relevan dan berinformasi dari set data sebagai persediaan pembinaan model. Antara kelebihan penggunaan pemilihan fitur adalah meningkatkan prestasi model, mengurangkan masa latihan, memudahkan permodelan dan mengelakkan *overfitting*. Dalam kajian ini, hasil kajian kesusasteraan yang telah dibuat, *Recursive Feature Elimination* (RFE) dan *SelectKBest* digunakan sebagai alat untuk melaksanakan proses pemilihan fitur. Sejumlah sepuluh fitur akan dipilih bagi fasa penilaian dan tafsiran model. Hanya sepuluh fitur dipilih bagi pemilihan fitur adalah berdasarkan hasil ketepatan yang tinggi diperolehi berbanding jumlah fitur lain.:

Hutan Rawak (RF):

Model pertama dilatih menggunakan algoritma RF dengan penggunaan perpustakaan *RandomForestClassifier* yang diimport daripada *sklearn.ensemble* akan melalui proses pemilihan fitur terlebih dahulu menggunakan kaedah *Recursive Feature Elimination* (RFE), salah satu teknik *Wrapper*. Perpustakaan *RFE* diimport daripada *sklearn.feature_selection* berfungsi untuk mengenalpasti fitur yang paling relevan dalam set data, mengurangkan dimensi dan berpotensi meningkatkan prestasi model. Rajah 3.12 menunjukkan struktur model RF bagi melatih set data melalui teknik pokok sehingga pemilihan akhir melalui undian majoriti.



Rajah 3.12 Struktur Model RF

Penalaan Hiperparameter melibatkan pengoptimuman fitur dalaman model (*hyperparameters*) untuk meningkatkan prestasinya pada data yang tidak kelihatan. Teknik Hiperparameter yang digunakan bagi kajian ini terhadap ketiga-tiga model (RF, ANN dan SVM) dijalankan menggunakan *GridSearchCV* atau *Grid Search Cross-Validation* bagi mendapatkan hasil penalaan hiperparameter yang optimum. *GridSearchCV* diimport daripada *sklearn.model_selection* untuk mencari nilai fitur yang optimum dari grid set data. Jadual 3.10 menunjukkan fitur dan nilai-nilai hiperparameter yang digunakan bagi meningkatkan prestasi model RF.

Jadual 3.10 Nilai Fitur bagi Penalaan Hiperparameter Model RF

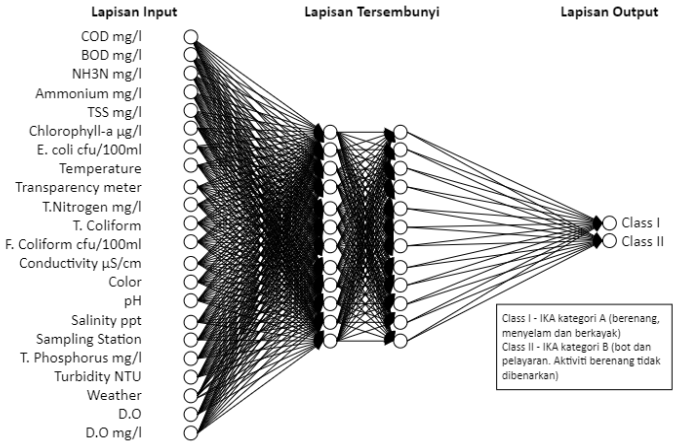
Parameter	**Nilai**
min_samples_split	2,5,10
min_samples_leaf	1,2,5
max_features	’sqrt’
n_estimator	200,300,500
max_depth	5,10,15
n_jobs	-1
random_state	42

Parameter *min_samples_split* mengawal bilangan sampel minimum yang diperlukan untuk memisahkan nod dalaman dengan tiga nilai pilihan (2,5,10). Parameter *min_samples_leaf* pula menetapkan bilangan sampel minimum yang diperlukan untuk berada di nod daun dengan tiga nilai pilihan (1,2,5). Parameter

max_features menentukan bilangan maksimum fitur yang dipertimbangkan untuk memisahkan nod 'sqrt' bersamaan dengan punca kuasa dua jumlah fitur. Parameter *n_estimator* menetapkan bilangan pokok di dalam hutan dengan tiga nilai pilihan (200,300,500). Parameter *max_depth* mengawal kedalaman maksimum setiap pokok di dalam hutan dengan tiga nilai pilihan (5,10,15). Parameter *n_jobs* menentukan bilangan pekerjaan untuk dijalankan selari telah ditetapkan kepada '-1' menggunakan semua pemproses yang ada sesuai untuk latihan yang lebih cepat. Parameter *random_state* disetkan kepada '42' adalah bertujuan memastikan penghasilan semula yang lebih baik.

Rangkaian Neural Buatan (ANN):

Model yang kedua ini dilatih menggunakan algoritma ANN. Dalam kajian ini, model dilatih menggunakan perpustakaan *MLPClassifier* yang diimport daripada *sklearn.neural_network*. Model ini juga melalui proses pemilihan fitur menggunakan kaedah *SelectKBest*, salah satu teknik *univariate* yang menyingkirkan fitur-fitur yang tidak perlu kecuali bilangan fitur-fitur pemarkahan tertinggi. Perpustakaan *SelectKBest* juga diimport daripada *sklearn.feature_selection* berfungsi dengan menggunakan ujian statistik (dalam kajian ini, *mutual information*) untuk mendapatkan skor dan ranking fitur-fitur berdasarkan hubungan fitur-fitur ini dengan pembolehubah sasaran. Bagi melaksanakan pengelasan MLP, penskalaan data dan fungsi classification_report perlu dijalankan semula untuk menggunakan teknik *pipeline*. Rajah 3.13 menunjukkan struktur model ANN bagi melatih set data melalui lapisan tersembunyi sehingga lapisan output.



Rajah 3.13 Struktur Model ANN

Jadual 3.11 menunjukkan fitur dan nilai-nilai hiperparameter yang digunakan bagi meningkatkan prestasi model ANN.

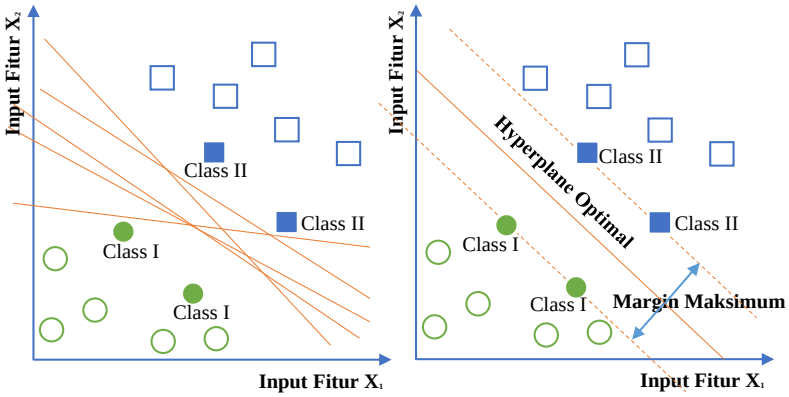
Jadual 3.11 Nilai Parameter bagi Penalaan Hiperparameter Model ANN

Parameter	**Nilai**
mlp__hidden_layer_sizes	(150,) (200,) (250,)
mlp__alpha	0.01, 5, 6, 7

Parameter *mlp__hidden_layer_sizes* mengawal seni bina rangkaian saraf, khususnya bilangan neuron dalam setiap lapisan tersembunyi. Tiga nilai yang dipilih iaitu (150, 200, 250) memberikan pelbagai pilihan sederhana. Atas faktor seni bina yang kurang kompleks, bilangan neuron yang ada sudah memadai. Parameter *mlp__alpha* ialah istilah regularisasi bagi L2, yang membantu mengelakkan *overfitting*. Empat nilai yang dipilih iaitu (0.01, 5, 6, 7) merangkumi julat yang dikira sesuai untuk kajian ini.

Mesin Vektor Sokongan (SVM):

Model terakhir ini dilatih menggunakan algoritma SVM dengan penggunaan perpustakaan *sklearn.svm LinearSVC* yang diimport daripada *sklearn.svm* akan melalui proses pemilihan fitur sama seperti model RF. Rajah 3.14 menunjukkan struktur model SVM bagi melatih set data sehingga nilai *hyperplane* optimum diperolehi.



Rajah 3.14 Struktur Model SVM

Jadual 3.12 menunjukkan fitur dan nilai-nilai hiperparameter yang digunakan bagi meningkatkan prestasi model ANN.

Jadual 3.12 Nilai Parameter bagi Penalaan Hiperparameter Model SVM

Parameter	**Nilai**
C	0.01, 0.1, 1, 10, 100
class_weight	'balanced', None
verbose	0, 1, 2
loss	‘hinge’
max_iter	10, 50, 100
random_state	42

Parameter *C* ialah julat nilai untuk penerokaan awal ketika penalaan hiperparameter dengan empat nilai yang dipilih iaitu (0.01, 0.1, 1, 10, 100). Nilai optimum *C* bergantung kepada fitur-fitur data dan diperlukan bagi mencapai kesilapan latihan dan ujian yang rendah. Parameter *class_weight* diperlukan untuk mengendalikan pengagihan kelas yang seimbang dengan dua nilai pilihan iaitu *balanced* dan *None*. Parameter *verbose* menentukan tahap fleksibel proses latihan SVM dengan dengan tiga nilai yang dipilih iaitu (0, 1, 2).

Parameter *loss* iaitu 'hinge' bermaksud fungsi kehilangan yang sesuai untuk SVM linear yang digunakan dalam tugas pengelasan. Parameter *max_iter* ialah bilangan maksimum pengulangan yang diambil untuk penyelesai bertumpu. Nilai (10,

50, 100) dipilih bagi menentukan tingkah laku penumpuan yang diperhatikan semasa latihan. Sebaliknya jika model tidak bertumpu, kemungkinan memerlukan nilai pengulangan yang lebih tinggi. Sama seperti model RF, parameter *random_state* ditetapkan sebagai '42' bagi memastikan penghasilan semula yang baik.

3.2.4 Penilaian dan Tafsiran Model

Penilaian dan tafsiran model diperlukan untuk menilai keberkesanan model dan memberikan pandangan yang bermakna mengenai hasilnya. Penilaian model mampu membina kepercayaan iaitu dengan memastikan model itu boleh dipercayai dan boleh digunakan dengan yakin dalam aplikasi dunia sebenar. Dalam kajian ini, penilaian metrik prestasi seperti ketepatan, kejituan, *recall* dan skor F1 digunakan serta Ujian-T (T-Test). Tafsiran model pula menunjukkan ketelusan iaitu membolehkan kita memahami penaakulan model, mengenalpasti kecenderungan *bias* dan membuat keputusan melalui penggunaannya. Kajian ini menggunakan teknik kepentingan fitur *feature importance* dan *SHapley Additive exPlanations* (SHAP) bagi menafsirkan model.

i. Penilaian Model.

Penilaian model dilaksanakan untuk menilai sejauh mana prestasi model latihan pada data yang tidak kelihatan, mengenal pasti kekuatan, kelemahan, dan potensi generalisasi data.

Laporan Pengelasan dan Matrik Kekeliruan (CM)

Laporan pengelasan menyediakan ringkasan metrik penilaian utama untuk tugas pengelasan yang dikira daripada matrik kekeliruan (CM). Matrik CM yang digunakan termasuk ketepatan, kejituan, *recall* dan skor F1. Laporan pengelasan membantu menilai keupayaan model untuk mengelaskan kelas yang berbeza dengan betul dan membuat keputusan bermaklumat berkenaan pemilihan dan penambahbaikan model.

CM pula merupakan jadual yang menggambarkan prestasi model dengan membandingkan label kelas yang diramalkan dengan label kelas sebenar. CM juga