PTA-FTSM-2020-134

PERBANDINGAN ANTARA PRESTASI MODEL PEMBELAJARAN MESIN DAN PEMBELAJARAN GABUNGAN DALAM KLASIFIKASI KETOKSIKAN BERBILANG KELAS

Tee Sin Teng

Prof. Madya Dr. Nor Shahnorbanun Sahran

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Dalam arena toksikologi, terdapat beberapa kaedah untuk meramalkan ketoksikan iaitu dalam vivo (*in vivo*), dalam vitro (*in vitro*) dan dalam silico (*in silico*). Kaedah dalam vivo dan dalam vitro telah membekalkan informasi yang berguna dalam ramalan ketoksikan. Walau bagaimanapun, kaedah dalam vivo dan dalam vitro akan dilaksanakan di dalam atau di luar organisma hidup dan melibatkan kos yang tinggi. Oleh itu, kaedah dalam silico dalam ramalan toksikologi telah melakukan usaha yang besar untuk meramalkan ketoksikan dalam bahan kimia. Untuk meneroka lagi kaedah dalam silico dalam ramalan toksikologi, kajian ini dijalankan dengan mengambil kesempatan daripada algoritma pembelajaran mesin dan algoritma pembelajaran gabungan untuk memperoleh prestasi terbaik dalam keputusan ramalan ketoksikan. Kajian ini dijalankan dengan menggunakan data aktiviti reseptor androgen (*Androgen Receptor*, AR) dengan 10K bahan kimia yang dihasilkan dari program Tox21. Dalam kajian ini, kaedah pembelajaran mesin dan pembelajaran gabungan akan digunakan pada set data tox21. Perbandingan prestasi dijalankan terhadap hasil akhir dari model pembelajaran mesin dan model pembelajaran gabungan untuk memperolehi model yang optimum dalam ramalan ketepatan ketoksikan data. Dalam keputusan kajian ini, model pembelajaran gabungan didapati berprestasi lebih cemerlang berbanding dengan model pembelajaran mesin. Model *XGBoost* mempunyai prestasi yang cemerlang dalam ramalan ketoksikan berbilang kelas.

1 PENGENALAN

Pada pandangan ahli toksikologi pengkomputeran, pembelajaran mesin telah menjadi sesuatu strategi yang berkesan dalam ramalan toksikologi *in silico*. Ketoksikan yang

disebabkan oleh pendedahan kimia dapat dilihat secara berturutan pada tahap organisma yang menaik, yang bermula sebagai peristiwa permulaan molekul dan meningkat kepada kesan buruk yang diukur sebagai titik akhir toksikologi untuk sel, tisu, organ, organisma, atau populasi (Ankley et al. 2010; Organization for Economic Co-operation and Development (OECD 2013); Allen et al. 2014). Terdapat tiga kategori strategi dalam pengujian ketoksikan bahan kimia iaitu dalam vivo, dalam vitro, dan dalam silico. Dalam ujian ketoksikan bahan kimia, kaedah dalam silico merupakan cara yang efektif dan kos rendah berbanding dengan kaedah lain seperti dalam vivo dan dalam vitro. Oleh itu, beberapa penyelidik telah menjalankan penyelidikan pemprosesan data kimia dengan menggunakan kaedah dalam silico. Kebelakangan ini, pembelajaran mesin dalam ramalan toksikologi dalam silico telah menjadi sesuatu algoritma yang berkesan. Oleh itu, terdapat beberapa penyelidikan tentang ramalan ketepatan dengan menggunakan pembelajaran mesin telah dijalankan untuk memperoleh model pembelajaran mesin yang menunjukkan prestasi yang cemerlang.

2 PENYATA MASALAH

Ramalan ketoksikan menjadi sesuatu garis panduan untuk menilai keselamatan dadah dan bahan kimia seperti yang dibuktikan (Raies, A.B. & Bajic,V.B. 2016). Dalam ujian ketoksikan bahan kimia mempunyai tiga kaedah iaitu dalam vivo, dalam vitro, dan dalam silico. Dalam ujian ketoksikan bahan kimia, dalam vivo dan dalam vitro merupakan cara yang memerlukan kos yang tinggi dan akan membahayakan kehidupan haiwan (National Research Council 2007). Oleh itu, terdapat banyak kajian menyelidik tentang penggunaan kaedah dalam silico untuk meramalkan ketoksikan bahan kimia. Walau bagaimanapun, pembelajaran mesin merupakan salah satu kaedah dalam silico tetapi pelbagai pengelas pembelajaran mesin menunjukkan prestasi ramalan yang berbeza.

3 OBJEKTIF KAJIAN

Objektif kajian ini adalah menjalankan perbandingan antara prestasi model pembelajaran mesin dan pembelajaran gabungan dalam klasifikasi ketoksikan berbilang kelas. Oleh itu, kajian ini dapat mengenalpasti model klasifikasi yang optimum dalam pemprosesan ramalan ketosikan bahan kimia melalui kaedah pembelajaran mesin dan pembelajaran gabungan.

4 METOD KAJIAN

Metodologi dalam kajian ini yang akan digunakan adalah mengikut amalan CRISP-DM (*Cross-Industry Process for Data Mining*). Amalan ini mempunyai fleksibiliti dan dapat meningkatkan keperluan pengkaji. Metodologi kajian ini mengandungi empat fasa iaitu fasa pemahaman perniagaan, fasa pemahaman data, fasa penyediaan data dan fasa pembangunan model.

4.1 Fasa Pemahaman Perniagaan

Fasa pemahaman perniagaan merupakan memahami dan mengenal pasti matlamat perniagaan. Matlamat perniagaan kajian ini merupakan menjalankan perbandingan prestasi antara pembelajaran mesin yang sering digunakan. Matlamat perniagaan yang seterusnya merupakan mengenal pasti model yang optimum dalam meramalkan aktiviti bahan kimia.

4.2 Fasa Pemahaman Data

Kajian ini menggunakan data *tox21-ar-bla-agonist-p1* yang diperolehi dari laman web *The Toxicology in the 21st Century* (Tox21) yang bekerjasama dengan *NIH's* NCATS dan *National Toxicology Program*. Reseptor androgen (AR) merupakan reseptor hormon nuklear yang memainkan peranan yang

penting dalam barah prostat yang bergantung pada reseptor androgen dan penyakit berkaitan androgen lain-lain. Bahan kimia gangguan endokrin (EDCs) dan interaksinya dengan reseptor hormon steroid seperti reseptor androgen yang berkemungkinan menyebabkan gangguan fungsi endokrin normal serta mengganggu fungsi homeostasis metabolisma, pembiakan, perkembangan dan tingkah laku. Data yang digunakan dalam kajian ini merupakan ujian dalam vitro dijalankan dalam mod agonis untuk menilai sifat agonis bahan kimia. Untuk mengenal pasti kompaun yang mengaktifkan isyarat AR, garis sel *AR-UAS-bla-GripTite* yang mengandungi domain pengikat ligan (LBD) AR tikus dan menyatakan gen pelapor beta-laktamase dengan stabil di bawah kawalan transkrip urutan pengatif hulu (UAS). Data AR beta-laktamase yang dalam mod agonis ini mengandungi sebanyak 31488 data dan 18 attribut.

4.3 Fasa Penyediaan Data

Fasa penyediaan data dijalankan dalam kajian ini mengandungi 4 fasa iaitu pembersihan data, pengurangan data, pelabelan data dan transformasi data. Pembersihan data merupakan proses yang membersihkan attribut yang sama dan menggantikan nilai yang hilang. Seterusnya, pengurangan data merupakan proses yang mengurangkan data yang tidak mempengaruhi hasil akhir. Pelabelan data merupakan proses yang menukarkan label data untuk memudahkan proses analisis. Transformasi data merupakan proses yang mengubahkan struktur data ke struktur yang lain.

a. Pembersihan data

   Pembersihan data merupakan langkah yang awal dalam proses pemprosesan. Pembersihan data merupakan proses penyediaan data dalam data analisis dengan membuang atau mengubah data yang tidak betul, tidak lengkap, tidak berkaitan, duplikasi atau format yang tidak betul. Sebelum pembersihan data dijalankan, data yang diperolehi dari laman web merupakan dalam format txt. Oleh itu, data ini ditukarkan kepada

format csv dengan menggunakan perisian *Excel* untuk memudahkan proses pemprosesan.

i. Kehilangan data

Attribut *AC50* dan *FLAG* mempunyai data yang hilang yang bermaksud tidak menghasilkan keputusan. Kehilangan data ini akan digantikan dengan nilai sifar dan NR dengan mengunakan fungsi *fillna()*. Kehilangan data dalam attribut *AC50* merupakan 91% manakala kehilangan data dalam attribut *FLAG* adalah 89% sebelum menggunakan fungsi *fillna()*.

b. Pengurangan data

i. Pengurangan dimensi

Pengurangan dimensi merupakan proses mengurangkan bilangan pemboleh ubah rawak yang dipertimbangkan dengan mendapatkan satu set pemboleh utama. Pendekatan pengurangan dimensi ini boleh dibahagikan kepada pemilihan ciri dan pengekstrakan ciri.

Dalam kajian ini, pendekatan pengurangan dimensi yang digunakan merupakan pemilihan ciri. Pemilihan ciri merupakan proses memilih ciri-ciri yang paling menyumbang kepada pemboleh ubah ramalan atau output secara automatik atau manual. Model yang belajar berdasarkan ciri yang tidak relevan akan menurunkan ketepatan model. Terdapat 18 attribut dalam data kajian ini dan hanya 9 attribut akan digunakan.

c. Pelabelan data

Pelabelan data merupakan proses yang menukar label data untuk melancarkan proses dan senang difahami. Label data sasaran yang diguna dalam kajian ini iaitu *'active antagonist'* dan *'active agonist'* menukar

kepada *'active'* manakala *'inconclusive antagonist'* dan *'inconclusive agonist'* menukar kepada *'inconclusive'*. Penukaran label data dijalankan dengan mengguna perisian *Excel*.

d. Transformasi data

i. Pengekodan kategori

Pengekodan kategori yang digunakan dalam kajian ini pengekodan label. Pengekodan label menukar pemboleh ubah kategori menjadi perwakilan berangka, sesuatu yang dapat dibaca oleh mesin.

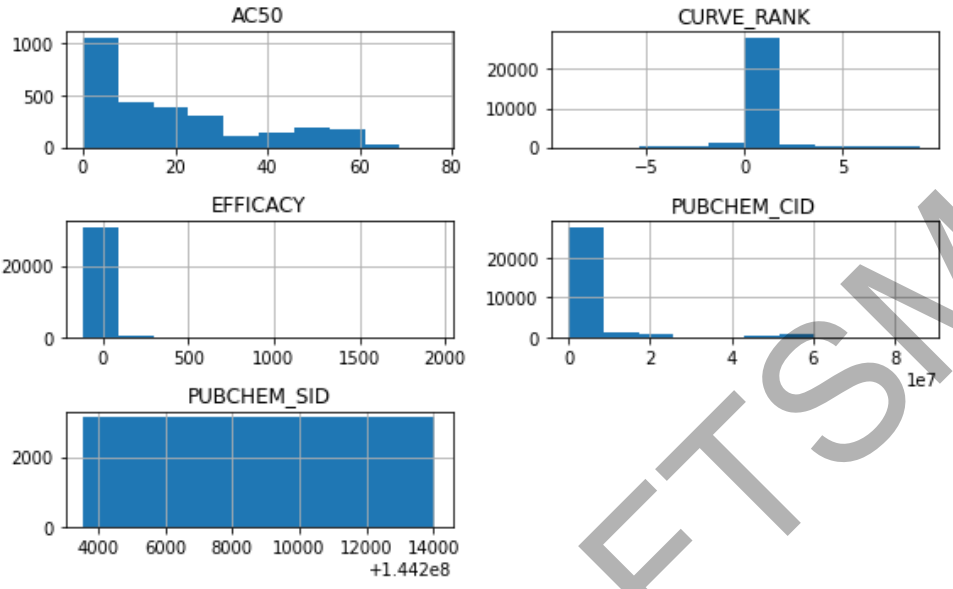
ii. Pembahagian data

Set data perlu dibahagikan kepada dua set iaitu set latihan dan set ujian untuk pembangunan model pembelajaran mesin. Pembahagian data adalah membahagi data kepada dua bahagian untuk tujuan pengesahan silang (*cross validatory*). Sebahagian data iaitu data latihan digunakan untuk membangunkan model ramalan, bahagian yang lain iaitu data ujian digunakan untuk menilai prestasi model. Set data dibahagikan kepada X (*features*) dan y (*target variable*). Seterusnya, fungsi *train_test_split* diimportkan untuk membahagi data kepada 70 peratus set latihan dan 30 peratus set ujian

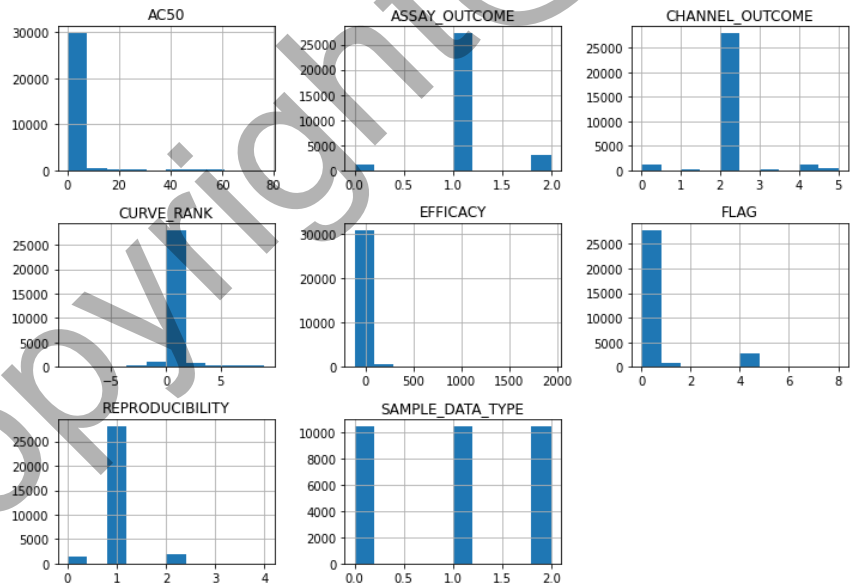
iii. Penyeragaman (*Standardization*)

Penyeragaman data adalah proses untuk menskala semula satu atau lebih attribut sehingga mempunyai nilai min 0 dan sisihan piawai 1. Dalam kajian ini, fungsi *StandardScaler* digunakan untuk menjalankan penyeragaman data.

Fasa penyediaan data ini merupakan fasa yang penting yang menjalankan proses pemprosesan sebelum menjalankan pembangunan model. Rajah 1 dan rajah 2 menunjukkan histogram attribut sebelum dan selepas proses pemprosesan.



Rajah 1 Histogram attribut sebelum proses pemprosesan

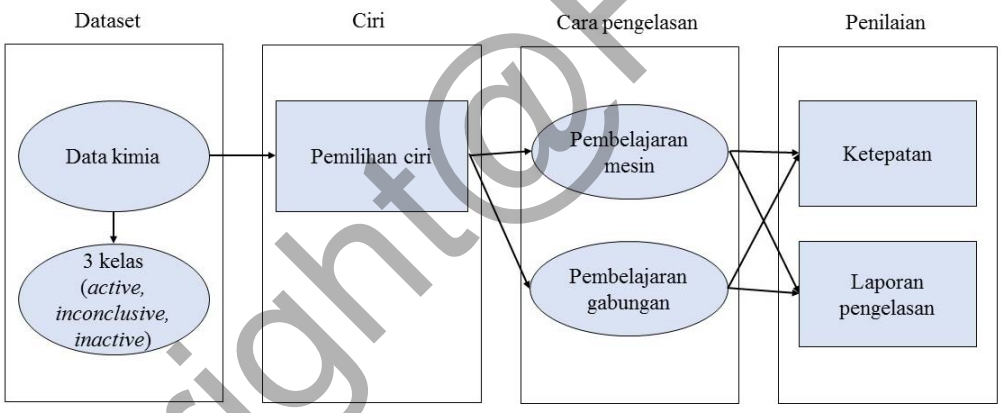


Rajah 2 Histogram attribut selepas proses pemprosesan

4.4 Fasa Pembangunan Model

Fasa pembangunan model ini menerangkan teknik pemodelan akan digunakan dalam kajian ini iaitu klasifikasi. Dalam pembelajaran mesin, klasifikasi merujuk kepada masalah pemodelan ramalan di mana label kelas diramalkan untuk contoh data input yang diberikan.

Dalam kajian ini, pembelajaran mesin dan pembelajaran gabungan dibangunkan untuk menilai prestasi model. Prestasi model dinilai dengan menggunakan ketepatan dan laporan pengkelasan.



Rajah 3 Reka bentuk pembangunan model klasifikasi

a. Pengelas pembelajaran mesin

Pengelas pembelajaran mesin yang digunakan dalam kajian ini merupakan regresi logistik, pengelas *naive bayes*, pengelas k-terdekat jiran, pengelas pelbagai lapisan perceptron, pengelas mesin vektor sokongan dan pengelas pokok keputusan. Model pembelajaran mesin dibinakan dengan mengguna parameter asal yang ditetapkan dalam fungsi sklearn.

b. Pengelas pembelajaran gabungan

Pengelas pembelajaran gabungan yang digunakan dalam kajian ini merupakan pembelajaran gabungan berdasarkan *bagging*, pembelajaran gabungan

berdasarkan *boosting* dan pengelas hutan rawak. Model pembelajaran gabungan dibinakan dengan mengguna parameter asal yang ditetapkan dalam fungsi sklearn.

c. Penilaian

Penilaian kajian ini akan dijalankan dengan membandingkan prestasi ketepatan dan laporan pengkelasan bagi setiap pembelajaran mesin. Perbandingan prestasi antara model pembelajaran mesin dan model pembelajaran gabungan dalam kajian ini digambarkan dengan menggunakan jadual.

5 HASIL KAJIAN

Bahagian ini akan menerangkan hasil kajian daripada proses pembangunan model pembelajaran mesin dan model pembelajaran gabungan. Ramalan keputusan set data tox21 ini dihasilkan dengan mengguna pembelajaran mesin iaitu regresi logistik, pengelas *naive bayes*, pengelas k-terdekat jiran, pengelas pelbagai lapisan perceptron, pengelas mesin vektor sokongan dan pengelas pokok keputusan. Seterusnya, pembelajaran gabungan berdasarkan *bagging*, pembelajaran gabungan berdasarkan *boosting* dan pengelas hutan rawak juga digunakan untuk menilai prestasi ramalan keputusan set data tox21. Set data tox21 mempunyai 7 ciri dan 1 pemboleh ubah sasaran setelah menjalankan proses pemprosesan. Model pembelajaran mesin dan model pembelajaran gabungan dibinakan untuk meramalkan keputusan hasil ujian yang mengandungi nilai *active*, *inactive* dan *inconclusive*. Dalam kajian ini, perbandingan prestasi antara model pembelajaran mesin dan pembelajaran gabungan akan dijalankan dan dibincangkan.

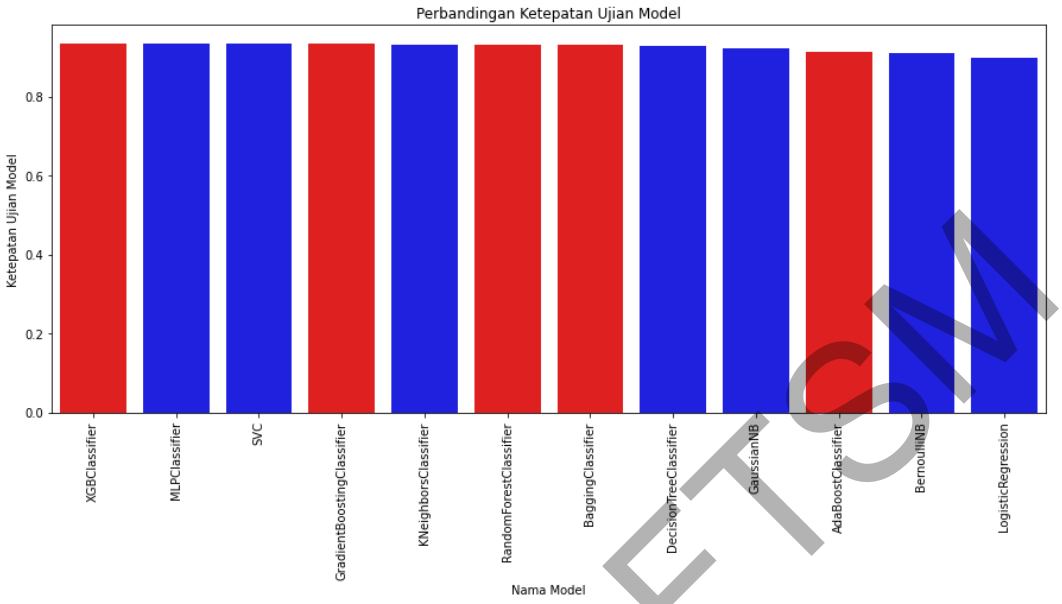
Rajah 4 menunjukkan perbandingan prestasi antara pembelajaran mesin dan pembelajaran gabungan. Rajah 5 menunjukkan graf bar prestasi ketepatan model pembelajaran mesin dan pembelajaran gabungan.

No. Model	Nama Model	Ketepatan Latihan Model	Ketepatan Ujian Model	Ketepatan (Precision) Model	Recall Model	F1 Skor Model
4	XGBClassifier	0.936573	0.934582	0.929783	0.934582	0.928569
9	MLPClassifier	0.937979	0.934265	0.929278	0.934265	0.927910
10	SVC	0.936981	0.933418	0.928311	0.933418	0.926859
2	GradientBoostingClassifier	0.939522	0.932995	0.927742	0.932995	0.927378
8	KNeighborsClassifier	0.938932	0.931936	0.926516	0.931936	0.925835
3	RandomForestClassifier	0.953768	0.931724	0.926072	0.931724	0.925244
1	BaggingClassifier	0.952815	0.931089	0.925419	0.931089	0.924049
11	DecisionTreeClassifier	0.953768	0.927279	0.919905	0.927279	0.919967
7	GaussianNB	0.921691	0.921139	0.915123	0.921139	0.917205
0	AdaBoostClassifier	0.913116	0.912777	0.900366	0.912777	0.901159
6	BernoulliNB	0.911529	0.911400	0.902374	0.911400	0.904992
5	LogisticRegression	0.897328	0.898698	0.881534	0.898698	0.878977

Rajah 4 Keputusan model pembelajaran mesin dan model pembelajaran gabungan

Antara model pembelajaran mesin, model regresi logistik telah menunjukkan ketepatan yang terendah iaitu 89.87%. Model pelbagai lapisan perceptron telah menunjukkan ketepatan yang tertinggi iaitu 93.43%. Selain itu, SVM, KNN, pokok keputusan, *GaussianNB* dan *BernoulliNB* telah menunjukkan ketepatan masing-masing iaitu 93.34%, 93.19%, 92.73%, 92.11% dan 91.14%.

Antara model pembelajaran gabungan yang mengguna parameter asal, model *XGBoost* telah menunjukkan prestasi yang cemerlang iaitu mempunyai ketepatan 93.46%. Model *AdaBoost* telah menunjukkan prestasi yang terendah dengan mempunyai ketepatan yang terendah iaitu 91.28%. Seterusnya, model pengelas *gradient boosting*, model hutan rawak dan pengelas *bagging* telah menunjukkan ketepatan masing-masing iaitu 93.30%, 93.17% dan 93.10%.



Rajah 5 Graf bar keputusan ketepatan model pembelajaran mesin dan model pembelajaran gabungan

Graf bar menunjukkan keputusan ketepatan model pembelajaran mesin dan model pembelajaran gabungan. Dalam graf bar, warna biru mewakili model pembelajaran mesin manakala warna merah mewakili model pembelajaran gabungan.

Jadual 2 Jadual perbandingan prestasi antara model pembelajaran mesin dan pembelajaran gabungan

Model	Pembelajaran Mesin	Pembelajaran Gabungan
Ketepatan tertinggi	0.934265 (Pelbagai lapisan perceptron)	0.934582 (*XGBoost*)
Ketepatan terendah	0.898698 (Regresi logistik)	0.912777 (*Adaboost*)
Purata ketepatan	0.922591	0.928633

Jadual 2 menunjukkan perbandingan prestasi antara model pembelajaran mesin dan model pembelajaran gabungan. Dalam model pembelajaran mesin, model pelbagai lapisan perceptron menunjukkan ketepatan yang tertinggi iaitu 93.43% manakala model regresi logistik menunjukkan ketepatan yang terendah iaitu 89.87%. Seterusnya, model *XGBoost* menunjukkan ketepatan yang tertinggi iaitu 93.46% manakala model *AdaBoost* menunjukkan ketepatan yang terendah iaitu 92.86% dalam

model pembelajaran gabungan. Dalam perbandingan prestasi antara model pembelajaran mesin dan pembelajaran gabungan, purata ketepatan model pembelajaran mesin adalah 92.26% manakala purata ketepatan model pembelajaran gabungan adalah 92.86%. Oleh itu, model pembelajaran gabungan berprestasi cemerlang berbanding dengan model pembelajaran gabungan.

6 KESIMPULAN

Kaedah dalam silico merupakan tren yang baru dan bermakna dalam ramalan ketoksikan dengan tidak membahayakan kehidupan haiwan dan kos rendah. Oleh itu, pembelajaran mesin telah menjadi sesuatu kaedah yang berkesan dalam ramalan ketoksikan. Dalam kajian ini, perbandingan prestasi antara model pembelajaran mesin dan model pembelajaran gabungan telah dijalankan untuk mencari model yang optimum dalam ramalan ketoksikan. Kesimpulannya, model pembelajaran gabungan berprestasi cemerlang dalam perbandingan prestasi antara model pembelajaran mesin dan model pembelajaran gabungan dalam kajian ini. Model *XGBoost* menunjukkan prestasi yang tertinggi dan model regresi logistik menunjukkan prestasi yang terendah. Walau bagaimanapun, bukan setiap model pembelajaran gabungan menunjukkan prestasi yang paling cemerlang berbanding dengan model pembelajaran mesin. Dalam kajian ini, model *AdaBoost* juga salah satu model pembelajaran gabungan tetapi model ini telah menunjukkan prestasi yang lemah berbanding dengan model pembelajaran mesin lain.

7 RUJUKAN

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L. 2010. *Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment*. *Environ. Toxicol. Chem.* 29, 730–741. doi: 10.1002/etc.34

National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and A Strategy. Washington, DC: National Academies Press.*

Organization for Economic Co-operation and Development (OECD 2013). *Guidance Document on Developing and Assessing Adverse Outcome Pathways. Paris: OECD environment, health and safety publications.*

Raies, A.B., Bajic, V.B. 2016. *In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip. Rev. Comput. Mol. Sci.* 6, 147–172. doi: 10.1002/wcms.1240