Current Issues & Research in Social Sciences, Education and Management (CIRSSEM) Vol. 4 No. 1, pp. 1-15, January, 2026
E-ISSN: 3032-324X
https://www.ejournal.aibpmjournals.com/index.php/ssem/index

Ramalan Skor Kredit Menggunakan Algoritma Pembelajaran Mesin

Khai Wah Khaw*
School of Management, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
*Corresponding Email: khaiwah@usm.my

ARTICLE INFORMATION

Publication information

Research article

HOW TO CITE
Khaw, K. W. (2025). *Ramalan skor kredit menggunakan algoritma pembelajaran mesin*. Current Issues & Research in Social Sciences, Education and Management (CIRSSEM), *4*(1), 1–15

Received: 20 November 2025
Accepted: 22 December 2025
Published: 24 January 2026

ABSTRACT

Skor kredit memainkan peranan yang amat penting dalam industri perbankan dan kewangan kerana ia digunakan sebagai alat utama untuk menilai tahap kebolehpercayaan kredit peminjam serta kebarangkalian kegagalan pembayaran balik pinjaman. Penilaian yang tepat membolehkan institusi kewangan mengurangkan risiko kewangan dan pada masa yang sama memaksimumkan pulangan. Seiring dengan pertumbuhan pesat industri kewangan dan peningkatan jumlah data berskala besar, pendekatan tradisional dalam penilaian skor kredit didapati semakin terhad dan kurang berkesan. Kajian ini menggunakan set data sebenar yang diperoleh daripada Kaggle, yang mengandungi 100,000 rekod pelanggan dengan pelbagai ciri demografi dan kewangan. Enam algoritma pembelajaran mesin digunakan, iaitu K-Jiran Terdekat (KNN), Naif Bayes (NB) (NB), Mesin Vektor Sokongan (SVM), Pokok Keputusan (DT), Regresi Logistik (LR), dan Hutan Rawak (RF). Proses kajian melibatkan prapemprosesan data, analisis data penerokaan, pembahagian data, pemodelan serta penilaian prestasi menggunakan metrik ketepatan, ketepatan ramalan, kepekaan, skor F1 dan ROC-AUC. Keputusan menunjukkan bahawa model Hutan Rawak mencapai prestasi terbaik dengan purata skor tertinggi berbanding model lain. Dapatan ini membuktikan potensi pembelajaran mesin sebagai alat sokongan keputusan yang berkesan dalam penilaian risiko kredit bagi industri kewangan.

Keywords: Pembelajaran Mesin, Skor Kredit, Hutan Rawak, Mesin Vektor Sokongan, Pokok Keputusan.

PENGENALAN

Syarikat kewangan ialah organisasi yang meminjamkan wang kepada individu dan perniagaan. Berbeza dengan bank, syarikat kewangan tidak menerima deposit tunai pelanggan dan tidak menawarkan perkhidmatan perbankan tertentu seperti akaun semak. Keuntungan syarikat kewangan diperoleh daripada kadar faedah yang dikenakan ke atas pinjaman mereka, yang selalunya lebih tinggi berbanding kadar faedah yang dikenakan oleh bank kepada pelanggan mereka.
Skor kredit merupakan salah satu faktor paling penting dalam proses membuat keputusan dalam perbankan dan kewangan. Syarikat pinjaman menggunakan teknik pemarkahan kredit tertentu untuk menilai kebarangkalian pemohon untuk ingkar bagi memaksimumkan pulangan dan mengurangkan risiko kewangan mereka (Boz et al., 2018). Pence (2014) mendefinisikan data raya sebagai maklumat yang besar kuantitinya, pelbagai, kompleks, dan perlu diberikan atau dinilai dengan pantas. Memandangkan pertumbuhan pesat industri perbankan terkini, adalah perlu untuk menggunakan sistem pemodelan yang selamat dan berkesan untuk kelulusan pinjaman. Industri kewangan mendapat manfaat besar daripada keupayaan pembelajaran mesin untuk mengekstrak pengetahuan daripada jumlah data yang besar. Operasi kewangan telah dipengaruhi oleh kemunculan data raya, yang melibatkan pengumpulan dan penyimpanan data dalam kuantiti yang banyak.

Aplikasi model pemarkahan kredit tradisional telah terbatas dengan ketara oleh pertumbuhan data raya di Internet, dan rangka kerja logik perniagaan asal telah hilang di bawah profil data dan situasi perniagaan baharu. Syarikat kewangan perlu melaksanakan pelbagai teknik pembelajaran mesin untuk mengurangkan penglibatan manual dalam proses pemantauan dan pengujian serta menggunakan pendekatan automatik untuk memperbaiku pemberian pinjaman dalam menghadapi puluhan ribu malahan ratusan ribu pengguna yang memohon pinjaman (Wang et al., 2020).

Algoritma pembelajaran mesin menggunakan kaedah statistik, kebarangkalian dan pengoptimuman untuk belajar daripada pengalaman lepas dan mengesan corak berguna dalam set data yang besar, tidak berstruktur dan kompleks (Uddin et al., 2019). Kajian empirikal pembelajaran mesin biasanya mempunyai dua fasa utama. Fasa pertama memberi tumpuan kepada pemilihan pemboleh ubah dan model yang relevan untuk ramalan, dengan memisahkan sebahagian data untuk latihan dan pengesahan model seterusnya mengoptimumkannya. Fasa kedua melibatkan aplikasi model yang dioptimumkan ke atas data ujian dan pengukuran prestasi ramalan.

Satu kelebihan menggunakan pembelajaran mesin untuk pemarkahan kredit adalah ia membolehkan penilaian risiko kredit yang lebih terperinci dan diperibadikan. Sebagai contoh, model pembelajaran mesin boleh mempertimbangkan pelbagai faktor yang lebih luas, seperti pekerjaan peminjam, tahap pendidikan dan maklumat demografi lain, selain faktor pemarkahan kredit tradisional. Ini berpotensi membawa kepada keputusan kredit yang lebih tepat dan adil.

Banyak firma pembiayaan menyediakan pinjaman kepada pelanggan yang tidak dapat mendapatkan pinjaman daripada bank disebabkan sejarah kredit yang buruk. Peminjam sedemikian memberikan cagaran kepada perniagaan kewangan untuk menjamin pinjaman mereka. Justeru, kajian ini bertujuan membantu syarikat kewangan meramal skor kredit pelanggan mereka dengan pelbagai teknik pembelajaran mesin, membolehkan mereka membuat keputusan sama ada untuk menawarkan pinjaman kepada pelanggan tersebut. Syarikat juga dapat menggunakan model pembelajaran mesin terbaik untuk mengesan skor kredit.

SOROTAN KAJIAN

Kajian Lepas

Terdapat beberapa kajian lepas yang telah meneroka penggunaan model pembelajaran mesin untuk meramal skor kredit pelanggan, yang diklasifikasikan sama ada sebagai buruk, standard atau baik. Model-model ini boleh menganalisis set data sedia ada untuk mengenal pasti corak dan menggunakan pengetahuan tersebut untuk membuat ramalan tentang hasil masa hadapan. Beberapa kajian lepas yang relevan disenaraikan di bawah.

Rudra Kumar dan Kumar Gunjan (2020) menjalankan kajian tentang organisasi dengan mengaplikasikan teknik pembelajaran mendalam dan pembelajaran mesin untuk menentukan skor kredit seseorang. Pengujian model pembelajaran mesin yang dicadangkan menunjukkan ia lebih berkesan dan cekap berbanding alternatif bukan pembelajaran mesin dalam proses analisis. Sekiranya sistem direka untuk menjadi lebih praktikal bagi analisis, ia akan memperbaiki proses analisis profil pelanggan, membolehkan penciptaan penyelesaian yang lebih komprehensif dan mampan untuk pengurusan sistem kredit.

Seterusnya, menurut Kumar et al. (2021), sebuah model ramalan yang menggunakan rangkaian neural mendalam dan pengelas pokok keputusan untuk membezakan antara pelanggan ingkar dan tidak ingkar dalam industri kewangan. Fokus utama penyelidikan ini adalah untuk menilai dan mengkategorikan teknik pemarkahan kredit berdasarkan tingkah laku pelanggan. Model ini mempunyai ketepatan anggaran 87% dalam mengenal pasti pelanggan ingkar dan tidak ingkar. Kelebihan utama skema ini adalah keupayaannya menyediakan skor kredit dan ramalan risiko kredit yang tepat untuk pemberi pinjaman dengan menggunakan penunjuk prestasi seperti ketepatan, min, sisihan piawai, ralat punca min kuasa dua, kebarangkalian ingkar dan luas di bawah lengkung, yang telah diuji pada pelbagai set data pemarkahan kredit kehidupan sebenar. Di samping itu, Elsalamony (2014) menjalankan kajian perbandingan Rangkaian Neural Persepsi Berbilang Lapis (MLPNN) dengan kaedah pengelasan lain seperti Tree Augmented Naive Bayes (TAN), Regresi Logistik (LR) dan C5.0. Ketepatan, sensitiviti dan kekhususan setiap model dinilai. Kajian mendapati MLPNN mempunyai ketepatan tertinggi pada 90.49%, LR mempunyai sensitiviti tertinggi pada 65.53% dan C5.0 mempunyai kekhususan tertinggi pada 93.23%. Kajian mendapati C5.0 menunjukkan prestasi sedikit lebih baik daripada MLPNN, TAN dan LR.

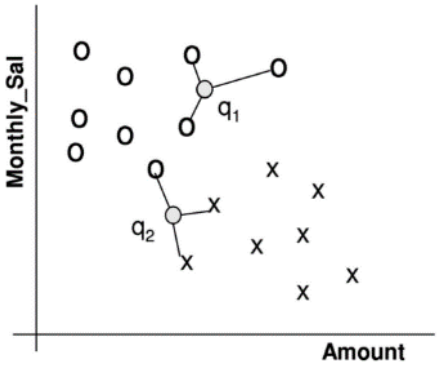
Ampountolas et al., (2021) menjalankan penyelidikan membandingkan pelbagai algoritma pembelajaran mesin pada data mikro-pinjaman sebenar untuk menguji keberkesanannya dalam mengelaskan peminjam ke dalam pelbagai kategori kredit. Algoritma pembelajaran mesin yang digunakan ialah DT, Extra Tree, RF, XGboost, Adaboost, KNN dan Multilayer Perceptron. Skor ketepatan tertinggi ialah 81.21% daripada Adaboost. Ketepatan terendah ialah Multilayer Perceptron yang hanya mencapai 71.59%.

Model Pembelajaran Mesin Terselia

K-Jiran Terdekat (KNN)

Konsep di sebalik Pengelasan Jiran Terdekat adalah mudah: contoh dikelaskan berdasarkan kelas jiran terdekat mereka. Memandangkan selalunya berguna untuk mempertimbangkan lebih daripada satu jiran, teknik ini lebih dikenali sebagai Pengelasan k-Jiran Terdekat (k-NN), di mana k jiran terdekat digunakan untuk menentukan kelas (Taunk et al., 2019). Oleh kerana contoh latihan diperlukan pada masa runtime, iaitu, mereka mesti berada dalam ingatan pada masa runtime, ia juga

dikenali sebagai Pengelasan Berasaskan Ingatan. Oleh kerana induksi ditangguhkan sehingga runtime, ia dikelaskan sebagai teknik Pembelajaran Malas. Oleh kerana pengelasan adalah berdasarkan secara langsung kepada contoh latihan, ia juga dipanggil Pengelasan Berasaskan Contoh atau Pengelasan Berasaskan Kes (Koohang et al., 2022).



Rajah 1. Pengelasan K-Jiran Terdekat dalam ruang ciri 2D (Gaji_Bulanan dan Jumlah)

Rajah 1 menggambarkan Pengelas 3-Jiran Terdekat pada masalah dua kelas dalam ruang ciri dua dimensi. Keputusan untuk q_1 dalam contoh ini adalah mudah---ketiga-tiga jiran terdekatnya adalah dari kelas O, jadi ia dikelaskan sebagai O. Situasi untuk q2 sedikit lebih rumit, kerana ia mempunyai dua jiran kelas X dan satu jiran kelas O. Ini boleh diselesaikan menggunakan sama ada pengundian majoriti mudah atau pengundian berwajaran jarak (lihat di bawah). Hasilnya, pengelasan k-NN mempunyai dua peringkat: yang pertama adalah menentukan jiran terdekat, dan yang kedua adalah menentukan kelas berdasarkan jiran tersebut. Andaikan kita mempunyai set data latihan D yang terdiri daripada (xi)i [1, n] sampel latihan (di mana n = |D|). Contoh diterangkan oleh satu set ciri F, dengan sebarang ciri berangka dinormalkan kepada julat [0,1]. Setiap contoh latihan diberikan label kelas yj Y. Matlamat kami adalah untuk mengkategorikan contoh tidak diketahui q. Kita boleh mengira jarak antara q dan xi untuk setiap xi D seperti berikut: $d(q, \mathrm{x}_i) = \sum_{f \in F} w_f \delta(q_f, x_{if})$. Ini adalah hasil tambah semua ciri dalam F, dengan wf mewakili berat bagi setiap ciri. Metrik jarak ini mempunyai pelbagai aplikasi; versi asas untuk atribut selanjar dan diskret ialah:

$$\delta(q_f, x_{if}) = \begin{cases} 0 & f \text{ discrete and } q_f = x_{if}, \\ 1 & f \text{ discrete and } q_f \neq x_{if}, \\ |q_f - x_{if}| & f \text{ continuous}, \end{cases}$$

Metrik jarak ini digunakan untuk memilih k jiran terdekat. k jiran terdekat tersebut kemudiannya boleh digunakan dalam pelbagai cara untuk menentukan kelas q. Pendekatan paling langsung adalah dengan memberikan pertanyaan kepada kelas majoriti dalam kalangan jiran terdekat (Maleki et al., 2020). Dalam kebanyakan kes, adalah wajar untuk memberikan lebih berat kepada jiran terdekat pertanyaan apabila menentukan kelasnya. Pengundian berwajaran jarak adalah teknik umum untuk mencapai tujuan ini, di mana jiran mengundi kelas kes pertanyaan, dengan undian diberi berat berdasarkan songsangan jarak mereka kepada pertanyaan tersebut.

$$Vote\ (y_j) = \sum_{c=1}^{k} \frac{1}{d(q, x_c)^p}\ 1\ (y_j, y_c)$$

Hasilnya, undian yang diberikan kepada kelas yj oleh jiran xc adalah 1 dibahagikan dengan jarak di antara mereka, iaitu, 1(yj,yc) mengembalikan 1 jika label kelas sepadan dan 0 jika sebaliknya. Dalam persamaan di atas, p biasanya bernilai 1, tetapi nilai yang lebih besar daripada 1 boleh digunakan untuk mengurangkan lagi pengaruh jiran yang lebih jauh (Cunningham & Delany, 2022).

Algoritma Naif Bayes (NB)

Pengelas Naif Bayes adalah pengelas kebarangkalian mudah yang menggunakan teorem Bayes dengan andaian bebas yang kuat (Sarker, 2021). Model ciri penentuan sendiri adalah istilah yang lebih deskriptif untuk model kebarangkalian asas. Intinya, pengelas Naif Bayes mengandaikan bahawa kewujudan satu ciri bagi sesuatu kelas tidak berkaitan dengan kewujudan mana-mana ciri lain. Walaupun andaian asas ini adalah palsu, pengelas Naif Bayes menunjukkan prestasi yang agak baik.

Pengelas Naif Bayes mempunyai kelebihan kerana hanya memerlukan sedikit data latihan untuk menganggar min dan varians pemboleh ubah yang diperlukan untuk pengelasan. Memandangkan pemboleh ubah bebas tidak diketahui, hanya varians pemboleh ubah bagi setiap label perlu dikira, dan bukannya keseluruhan matriks kovarians. Berbeza dengan operator Naif Bayes, operator Naif Bayes (Kernel) boleh digunakan pada atribut berangka (Vijayarani & Dhayanand, 2015).

Algoritma Naif Bayes adalah pengelas kebarangkalian yang mudah dan langsung yang mengira satu set kebarangkalian dengan mengira kekerapan dan kombinasi nilai dalam set data yang diberikan. Algoritma ini menggunakan teorem Bayes dan mengandaikan semua pemboleh ubah adalah bebas daripada nilai pemboleh ubah kelas. Oleh kerana andaian kebebasan bersyarat ini jarang benar dalam aplikasi dunia sebenar, ia dilabelkan sebagai Naif, tetapi algoritma ini belajar dengan pantas dalam pelbagai masalah pengelasan terkawal. Teorem Bayes adalah formula matematik yang digunakan untuk mengira kebarangkalian bersyarat, dinamakan sempena ahli matematik British abad ke-18, Thomas Bayes.

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

P(A|B) adalah kebarangkalian peristiwa A berlaku apabila peristiwa B berlaku, P(A) adalah kebarangkalian A berlaku, P(B|A) adalah kebarangkalian peristiwa B berlaku apabila peristiwa A berlaku, dan P(B) adalah kebarangkalian B berlaku. Set data dikenakan analisis Naif Bayes, dan matriks kekeliruan untuk kelas jantina dihasilkan, dengan dua nilai yang mungkin: kawalan sihat atau pesakit. Matriks Kekeliruan: 1 2 ← dikelaskan sebagai, 9 1 | 1 = Kawalan sihat, 2 11 | 2 = Pesakit.

Bagi matriks kekeliruan yang tersebut di atas, positif benar untuk kelas 1='Kawalan sihat' adalah 9 dan positif palsu adalah 1, manakala positif benar untuk kelas b='Pesakit' adalah 11 dan positif palsu adalah 2, iaitu, unsur pepenjuru matriks 9+11 =20 mewakili instans yang betul dikelaskan dan unsur lain 1+2 = 3 mewakili instans yang salah. Ketepatan pengelasan sebanyak 86.95 peratus diperoleh hasil daripada proses ini. Masa yang telah berlalu kini adalah 0.204262 saat.

Mesin Vektor Sokongan (SVM)

SVM adalah kaedah pembinaan pengelas yang berkuasa. Ia bertujuan untuk mewujudkan sempadan keputusan antara dua kelas yang membolehkan ramalan label daripada satu atau lebih vektor ciri. Sempadan keputusan ini, yang dikenali sebagai hiperplane, diorientasikan supaya ia sejauh mungkin dari titik data terdekat daripada setiap kelas. Titik terdekat ini dirujuk sebagai vektor sokongan (Huang et al., 2018). Vapnik membangunkan SVM sebagai model pembelajaran mesin berasaskan kernel

untuk tugas pengelasan dan regresi. Dalam tahun-tahun kebelakangan ini, komuniti perlombongan data, pengecaman corak dan pembelajaran mesin tertarik kepada keupayaan pengitlakan luar biasa SVM, serta penyelesaian optimum dan kuasa diskriminatifnya. SVM telah terbukti sebagai alat yang berkesan untuk menyelesaikan masalah pengelasan binari praktikal. SVM telah menunjukkan prestasi yang mengatasi kaedah pembelajaran terselia lain (Arnroth & Dennis, 2016). SVM telah menjadi salah satu kaedah pengelasan yang paling luas digunakan dalam tahun-tahun kebelakangan ini, disebabkan asas teori yang kukuh dan kapasiti pengitluan yang tinggi (Cervantes et al., 2020).

Dalam bahagian ini, kita akan melihat SVM dalam konteks pengelasan binari. Kita diberikan data latihan vektor x1... xn dalam ruang X Rd. Kita juga diberikan label mereka y1...yn, di mana yi 1,1 digunakan. SVM, dalam bentuk yang paling asas, adalah hiperplane yang memisahkan data latihan dengan margin maksimum. Semua vektor di satu sisi hiperplane dilabelkan 1, manakala semua vektor di sisi lain dilabelkan 0. Vektor sokongan adalah instans latihan yang paling dekat dengan hiperplane. Secara umum, SVM membolehkan seseorang menggunakan operator kernel Mercer K untuk memproyeksikan data latihan asal dalam ruang X ke ruang ciri berdimensi lebih tinggi F. Dalam erti kata lain, kita mempertimbangkan set pengelas dalam bentuk:

$$f(x) = \left(\sum_{i=1}^{n} a_i K(x_i, x \right)$$

Apabila K memenuhi syarat Mercer, kita boleh menulis: K(u,v) = Φ(u) · Φ(v) di mana Φ: X→F dan "·" menandakan hasil darab dalam. Kita kemudiannya boleh menulis semula f sebagai:

$$f(x) = w \cdot \Phi(x), where\ w = \sum_{i=1}^{n} a_i \Phi(x_i)$$

Oleh itu, dengan menggunakan K, kita secara tersirat memproyeksikan data latihan ke ruang ciri F yang berbeza (sering kali berdimensi lebih tinggi). SVM kemudiannya mengira ai yang sepadan dengan hiperplane dengan margin terbesar dalam F. Kita boleh secara tersirat memproyeksikan data latihan dari X ke ruang F dengan menggunakan fungsi kernel yang berbeza, di mana hiperplane dalam F sepadan dengan sempadan keputusan yang lebih kompleks dalam ruang asal X.

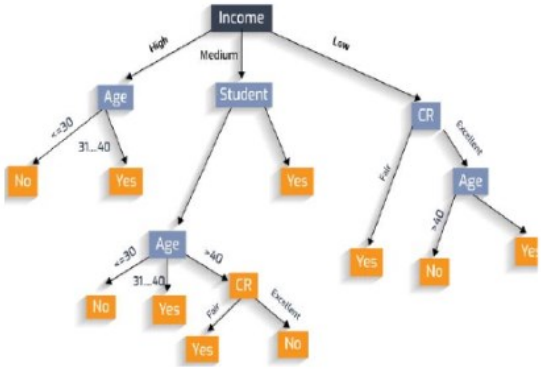
Kernel polinomial $K(u,v) = (u \cdot v + 1)^p$ mendorong sempadan polinomial berdarjah p dalam ruang asal X dan kernel fungsi asas jejarian $K(u,v) = \exp(-\gamma||u-v||^2)$ mendorong sempadan dengan meletakkan Gaussian berwajaran pada instans latihan utama. Untuk kebanyakan kertas ini, kita akan mengandaikan bahawa modulus vektor ciri data latihan adalah malar, iaitu, ||Φ(xi)|| = C untuk beberapa C tetap bagi semua instans latihan xi. Untuk kernel fungsi asas jejarian, kuantiti ||Φ(xi)|| sentiasa malar, jadi andaian ini tiada kesan ke atas kernel ini. Kita memerlukan xi malar untuk ||Φ(xi)|| malar dengan kernel polinomial. Kekangan ke atas ||Φ(xi)|| ini boleh dilonggarkan (Tong & Koller, 2001).

Pokok Keputusan (DT)

Sebuah pokok biasa merangkumi akar, dahan dan daun. Struktur yang sama diikuti dalam Pokok Keputusan. Ia mengandungi nod akar, dahan, dan nod daun (Nasteski, 2017). Pengujian atribut dilakukan pada setiap nod dalaman, hasil ujian berada pada dahan dan label kelas sebagai hasil berada pada nod daun (Mat Amin et al., 2018). Nod akar ialah ibu bapa semua nod dan, seperti namanya, ialah nod tertinggi dalam sebuah pokok. Pokok keputusan ialah pokok di mana setiap nod mewakili satu ciri (atribut), setiap pautan (dahan) mewakili satu keputusan (peraturan), dan setiap daun mencerminkan hasil (nilai kategori atau selanjar) (Jadhav & Channe, 2016).

Memandangkan pokok keputusan meniru pemikiran peringkat manusia, adalah agak mudah untuk mengumpul data dan membuat interpretasi yang baik. Matlamat keseluruhan ialah membina pokok seperti ini untuk semua data dan memproses satu hasil pada setiap daun (Patel & Prajapati, 2018).

Pokok Keputusan adalah salah satu kaedah paling berkuasa dalam banyak bidang, termasuk pembelajaran mesin, pemprosesan imej dan pengecaman corak. Pokok Keputusan ialah model berjujukan yang menyatukan secara cekap dan padu satu siri ujian asas di mana satu ciri berangka dibandingkan dengan nilai ambang dalam setiap ujian. Peraturan konseptual jauh lebih mudah dibina berbanding pemberat berangka dalam rangkaian neural sambungan nod. DT terutamanya digunakan untuk pengelompokan. Tambahan pula, DT ialah model pengelasan popular dalam perlombongan data. Setiap pokok terdiri daripada nod dan dahan. Setiap nod mewakili satu ciri dalam kategori yang hendak dikelaskan, dan setiap subset menentukan satu nilai yang boleh diambil oleh nod tersebut. Pokok keputusan telah menemui banyak bidang pelaksanaan kerana analisis mudah dan ketepatannya pada pelbagai bentuk data. Rajah 2 menggambarkan satu contoh DT (Charbuty & Abdulazeez, 2021).



Rajah 2. Contoh Pokok Keputusan

Entropi digunakan untuk menilai ketidakmurnian atau keacakan suatu set data. Nilai entropi sentiasa berada di antara 0 dan 1. Nilainya lebih baik apabila sama dengan 0 dan lebih teruk apabila sama dengan 1, yang membayangkan bahawa semakin hampir nilainya dengan 0, semakin baik. Seperti yang digambarkan dalam "Rajah 3.", entropi bagi pengelasan set S berkenaan dengan c keadaan [15] jika sasaran adalah G dengan nilai atribut yang berbeza. Ini ditunjukkan oleh persamaan di bawah:

$$Entropy(S) = \sum_{i=1}^{c} p_i \ \log 2^{p_i}$$

Di mana p_i ialah nisbah bilangan sampel subset kepada nilai atribut ke-i.

Satu metrik untuk segmentasi ialah keuntungan maklumat, juga dikenali sebagai maklumat bersama. Ini secara intuitif memberitahu berapa banyak pengetahuan tentang nilai pemboleh ubah rawak yang ada. Ia adalah songsangan bagi entropi; semakin besar nilainya, semakin baik. Berdasarkan definisi entropi, keuntungan data Gain(S,A) ditakrifkan seperti berikut, seperti yang ditunjukkan dalam persamaan di bawah:

$$Gain(S, A) = \sum_{v \in V(A)} \frac{|S_V|}{|S|} Entrypy(S_V)$$

Di mana V(A) ialah julat atribut A dan Sv ialah subset bagi set S yang sama dengan nilai atribut bagi atribut v (Charbuty & Abdulazeez, 2021).

Regresi Logistik (LR)

LR adalah teknik statistik untuk memodelkan hubungan antara pemboleh ubah bersandar dengan satu atau lebih pemboleh ubah tak bersandar (Park, 2013). Ia secara berkesan memodelkan hasil dikotomi dan menilai hubungan antara pemboleh ubah sasaran kategori dengan aspek berangka, kategori, atau kedua-duanya. LR juga merupakan model statistik yang popular yang membenarkan analisis multivariat dan pemodelan bagi pemboleh ubah bersandar binari; regresi linear adalah model yang setanding untuk pemboleh ubah bersandar selanjar. Analisis multivariat memperoleh pekali bagi setiap peramal dalam model akhir dan melaraskannya relatif kepada peramal lain dalam model. Pekali-pekali tersebut mengukur sumbangan setiap peramal kepada pengiraan risiko hasil (Shipe et al., 2019).

Algoritma regresi logistik adalah berdasarkan model regresi linear yang diberikan sebagai $y = h_\theta(x) = \theta^t x$. Persamaan ini tidak cekap untuk meramal nilai binari kita ($y(i) \varepsilon \{0, 1\}$), jadi kami memperkenalkan fungsi dalam persamaan berikut untuk meramalkan kebarangkalian bahawa sasaran tertentu (dengan atribut tertentu) tergolong dalam kelas "1" (positif) berbanding kebarangkalian ia tergolong dalam kelas "0" (negatif) (Zhu et al., 2019).

$$P(y = 1|x) = h_\theta(x) = \frac{1}{1 + \exp(-\theta^t x)} = \sigma(\theta^t x)$$

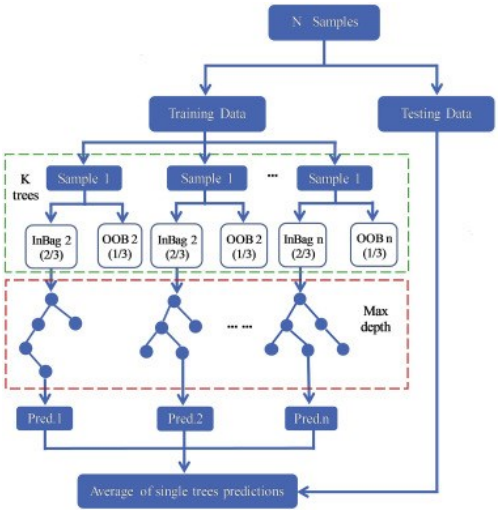
$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - h_\theta(x)$$

Dengan menggunakan persamaan fungsi sigmoid, kita dapat mengekalkan nilai $\theta^t x$ dalam selang [0, 1]. Kemudian, kita mencari nilai θ supaya kebarangkalian $P(y=1 \mid x)=h_\theta(x)$ adalah besar apabila x berada dalam kelas "1" dan kecil apabila x berada dalam kelas "0" (iaitu, $P(y=0 \mid x)$ adalah besar) (Zhu et al., 2019). LR membantu menentukan kemungkinan bahawa satu titik data baru tergolong dalam kelas tertentu (Uddin et al., 2019).

Hutan Rawak (RF)

Hutan rawak adalah sejenis algoritma pembelajaran mesin yang telah terbukti berkesan dalam pelbagai aplikasi (Breiman, 2001). Ia beroperasi dengan membina banyak pokok keputusan semasa latihan dan kemudian menggabungkan ramalan mereka untuk membuat ramalan akhir. Pokok-pokok keputusan dibina menggunakan subset rawak ciri-ciri dan sampel rawak data latihan, yang membantu mengurangkan keterlaluan penyesuaian dan meningkatkan pengitlakan model kepada data baharu. Satu kajian oleh Lee et al. (2010) mendapati bahawa hutan rawak mengatasi algoritma lain dari segi ketepatan ramalan dan masa latihan pada set data kelekatan ikatan protein-ligan. Hutan rawak juga telah digunakan dalam bidang seperti pengesanan penipuan kredit (Kumar et al., 2019) dan ramalan keciciran pelanggan (Xie et al., 2009). Ia secara umumnya dianggap sebagai teknik pembelajaran mesin yang boleh dipercayai dan mudah digunakan.

Hutan rawak adalah model pembelajaran mesin statistik yang berprestasi baik pada set data bersaiz kecil dan sederhana. Ia boleh digunakan untuk meramalkan sama ada pemboleh ubah respons kategori atau selanjar, bergantung pada sama ada tugas itu adalah pengelasan atau regresi. Pemboleh ubah peramal yang digunakan dalam model hutan rawak juga boleh sama ada kategori atau selanjar (Cutler et al., 2011).



Rajah 3. Carta alir regresi hutan rawak

METODOLOGI

Data untuk kajian ini diperoleh dalam format CSV daripada Kaggle. Langkah pertama ialah mengimport pelbagai pustaka termasuk pandas, NumPy, matplotlib, seaborn dan scikit-learn. Pustaka-pustaka ini akan digunakan untuk tugas seperti mencipta bingkai data, bekerja dengan tatasusunan dan matriks multidimensi, memvisualisasikan data, melaksanakan model pembelajaran mesin, pensampelan berstrata, penskalaan dan pemiawaian ciri, menilai prestasi pengelasan, dan mengira metrik ralat untuk pelbagai model. Seterusnya, set data "Credit Score Train.csv" diimport menggunakan fungsi read_csv() daripada pandas. Set data ini, yang mengandungi 20 pemboleh ubah input, 1 pemboleh ubah output (sasaran) dan 100,000 pemerhatian, adalah daripada sebuah bank Portugis. Enam pemboleh ubah yang tidak relevan dengan pembelajaran mesin telah dikeluarkan daripada set data. Pemboleh ubah dalam set data ini digambarkan seperti berikut.

Pra-Pemprosesan Data

Memproses data adalah langkah penting dalam membina model pembelajaran mesin yang cekap. Dalam kajian ini, langkah pra-pemprosesan data termasuk mengenal pasti pemboleh ubah yang tidak perlu, mengendalikan nilai yang hilang dan pengekodan pemboleh ubah kategori. Langkah-langkah ini membantu memastikan data bersih dan sedia untuk digunakan dalam pemodelan.

Pandas menyediakan fungsi isna() dan isnull() untuk mengesan nilai yang hilang dalam data. Fungsi isna().sum() dalam pandas boleh digunakan untuk mengembalikan bilangan nilai yang hilang dalam setiap lajur. Fungsi fillna() dalam pandas boleh digunakan untuk mengisi nilai null. Bergantung pada lajur, nilai yang hilang mungkin diisi dengan min atau mod lajur tersebut. Ini boleh membantu memastikan data lengkap dan sedia untuk digunakan dalam pemodelan pembelajaran mesin.

Set data berstruktur selalunya mengandungi lajur berangka dan kategori. Walau bagaimanapun, algoritma pembelajaran mesin hanya boleh memproses data berangka, bukan teks. Oleh itu, adalah perlu untuk mengekod data kategori ke dalam lajur berangka. Dalam kajian ini, fungsi replace() Python digunakan untuk melabelkan lajur sasaran ("Credit_Score") dengan integer unik. Skim pengekodan ini memberikan nilai 0 (Lemah), 1 (Standard) dan 2 (Baik) untuk mewakili skor kredit pelanggan. Fungsi get_dummies() daripada pustaka pandas digunakan untuk melakukan pengekodan satu-

panas pada pemboleh ubah "Occupation", "Credit_Mis", "Payment_of_Min_Amount" dan "Payment_Behaviour". Pengekodan satu-panas menukar data kategori kepada bentuk berangka, membolehkannya digunakan dalam algoritma pembelajaran mesin.

Penskalaan dan pemiawaian ciri digunakan dalam kajian ini. Penskalaan dan pemiawaian ciri adalah teknik pra-pemprosesan yang digunakan untuk melaraskan nilai ciri berbeza kepada skala yang sama. Ini memastikan semua ciri dalam set data berada pada tahap yang saksama, dan tiada satu ciri pun mempunyai kesan tidak seimbang terhadap hasil model. Dengan menggunakan penskalaan dan pemiawaian ciri, prestasi dan kestabilan algoritma pembelajaran mesin tertentu boleh diperbaiki. Selain itu, ia juga boleh membantu mengurangkan masa yang diperlukan untuk penumpuan apabila menggunakan algoritma pengoptimuman.

Analisis Data Eksploratori (EDA)
Penganalisisan sejagat merujuk kepada amalan mencipta visualisasi yang mewakili dan meringkaskan satu pemboleh ubah atau ciri set data. Jenis visualisasi ini boleh membantu memahami taburan, julat dan corak asas satu pemboleh ubah, dan sering digunakan sebagai langkah pertama dalam analisis data eksploratori (EDA). Analisis dua pemboleh ubah merujuk kepada amalan menganalisis hubungan antara dua pemboleh ubah. Jenis analisis ini digunakan untuk menyiasat perkaitan antara dua pemboleh ubah dan boleh membantu mendedahkan corak dan trend yang mungkin tidak ketara serta-merta apabila mengkaji setiap pemboleh ubah secara individu.

Pembahagian Data
Pembahagian data ialah proses membahagikan set data kepada set latihan dan ujian yang berasingan. Matlamat pembahagian data adalah untuk mencipta satu set data yang digunakan untuk melatih model dan set data berasingan yang digunakan untuk menilai prestasi model. Fungsi train_test_split daripada pustaka scikit-learn digunakan untuk membahagikan set data kepada empat bahagian, iaitu data latihan untuk pemboleh ubah tidak bersandar (X_train), data ujian untuk pemboleh ubah tidak bersandar (X_test), data latihan untuk pemboleh ubah bersandar (y_train) dan data ujian untuk pemboleh ubah bersandar (y_test). 75% set data digunakan untuk latihan dan baki 25% untuk ujian. Parameter random_state juga ditentukan untuk memastikan keputusan konsisten setiap kali fungsi dilaksanakan.

Pemodelan Data
Pemodelan data ialah proses mencipta perwakilan matematik sistem atau masalah dunia sebenar, biasanya dalam bentuk model statistik atau pembelajaran mesin. Perwakilan matematik ini kemudiannya boleh digunakan untuk membuat ramalan, memahami hubungan antara pemboleh ubah berbeza dan membuat keputusan. Enam model pembelajaran mesin iaitu KNN, NB, SVM, DT, LR dan RF dibina dalam peringkat ini.

Penilaian Model
Dalam kajian ini, pelbagai kaedah akan digunakan untuk menilai prestasi model pembelajaran mesin. Satu teknik yang digunakan ialah penggunaan matriks kekeliruan, yang akan diimport daripada pustaka metrik Sklearn. Matriks ini akan digunakan untuk mendapatkan gambaran tentang prestasi setiap model. Tambahan pula, metrik penilaian piawai untuk masalah pengelasan seperti ketepatan, kejituan, dapatan semula dan skor F1 akan dikira. Pada akhirnya, kajian akan membuat kesimpulan dengan memplot lengkung ROC-AUC untuk semua enam model dan membentangkan skor mereka
.

KEPUTUSAN

Dalam bahagian ini, ketepatan tinggi secara umumnya bermaksud model pengelasan menunjukkan prestasi yang baik dan membuat ramalan yang betul untuk banyak titik data. Model ketepatan tertinggi ialah RF yang mempunyai ketepatan 77.86%. NB mempunyai ketepatan terendah hanya mencapai 62.69%. Kejituan sahaja tidak semestinya ukuran terbaik prestasi model, kerana kejituan tinggi juga boleh dicapai oleh model yang membuat sangat sedikit ramalan, yang mungkin tidak berguna. Semua model pembelajaran mesin mempunyai nilai kejituan yang hampir sama. RF mempunyai kejituan tertinggi iaitu 78% dan LR mempunyai kejituan terendah iaitu 64%. Dapatan semula ialah metrik yang biasa digunakan dalam masalah pengelasan, terutamanya apabila matlamatnya adalah untuk meminimumkan bilangan negatif palsu. Dapatan semula tinggi bermaksud model mempunyai bilangan negatif palsu yang rendah. Skor dapatan semula tertinggi (78%) dicapai oleh RF manakala NB menunjukkan skor terendah 63%. Skor F1 bernilai 1 menunjukkan model mempunyai kejituan dan dapatan semula yang sempurna. Skor 0 bermaksud model tidak mempunyai positif benar (iaitu tidak dapat meramalkan sebarang kes positif) dan skor F1 0.5 bermaksud kejituan dan dapatan semula model adalah sama. Secara umumnya, skor F1 tinggi adalah diingini kerana ia bermaksud model mempunyai keseimbangan kejituan dan dapatan semula yang baik. RF mempunyai skor F1 tertinggi iaitu 78% dan NB mempunyai skor F1 terendah iaitu 63%.

Model Hutan Rawak (RF) mengatasi lima model lain dalam kajian ini, dengan skor purata tertinggi 77.94%. Seperti yang ditunjukkan dalam Jadual 3, prestasi RF adalah lebih baik berbanding Mesin Vektor Sokongan (SVM) dan Pokok Keputusan (DT), yang masing-masing mempunyai skor purata 68.97% dan 68.01%. K-Jiran Terdekat (KNN) dan Naif Bayes (NB) juga mempunyai skor purata lebih rendah iaitu 64.70% dan 64.17%. Secara keseluruhan, RF muncul sebagai pemenang jelas dalam kalangan model yang dinilai dalam kajian ini.

Jadual 4.2. Keputusan Prestasi Enam Model

Model	Keputusan				
	Ketepatan	Ketepatan Ramalan	Kepekaan	Skor F1	Skor Purata
KNN	64.80	65.00	65.00	64.00	64.70
NB	62.69	68.00	63.00	63.00	64.17
SVM	68.89	69.00	69.00	69.00	68.97
DT	68.03	68.00	68.00	68.00	68.01
LR	64.29	64.00	64.00	64.00	64.07
RF	77.76	78.00	78.00	78.00	77.94

IMPLIKASI DAN BATASAN

Penemuan kajian ini boleh dijadikan rujukan bagi pengamal dalam meramal skor kredit pelanggan di sebuah syarikat kewangan berdasarkan skor kredit pelanggan terdahulu. Penggunaan syarikat kewangan secara khusus boleh diaplikasikan secara langsung seperti yang ditunjukkan oleh hasil kajian ini, kerana ia biasanya memerlukan sumber yang banyak seperti kos, usaha dan masa untuk membandingkan pelbagai algoritma ML, yang boleh menjadi mahal dalam suasana perniagaan praktikal. Oleh itu, hasil kajian ini boleh berguna bagi pengamal untuk meminimumkan sumber yang diperlukan bagi

memilih teknik ML yang paling sesuai dan menentukan parameter yang sepatutnya digunakan. Hasilnya, kajian ini membantu mengenal pasti pelanggan yang mempunyai kemungkinan lebih tinggi untuk tidak membayar balik hutang selepas mereka meminjam wang.

Tambahan pula, hasil kajian ini juga dapat memperkukuh syarikat kewangan untuk mengambil langkah proaktif bagi mengurangkan risiko meluluskan pinjaman kepada pelanggan yang mempunyai skor kredit yang lemah. Dengan menganalisis corak pelanggan, model pembelajaran mesin boleh digunakan untuk mengenal pasti individu yang mempunyai skor kredit lemah. Ini akan membolehkan syarikat kewangan mengambil langkah untuk mengurangkan risiko hutang tidak berbayar, kerana mereka dapat mengenal pasti pelanggan ini pada peringkat yang lebih awal. Ini memberi lebih banyak masa dan peluang kepada syarikat kewangan untuk menetapkan peraturan dan syarat yang lebih ketat bagi pelanggan ini bagi memastikan mereka tidak mengalami kerugian keuntungan.

Kajian ini mungkin mempunyai beberapa batasan, seperti kemungkinan bias dalam data yang digunakan untuk melatih model pembelajaran mesin, atau batasan model itu sendiri. Dalam proses mengkaji semula kajian ini, adalah penting untuk mengetahui dan memahami batasan kajian untuk menambahbaiknya.

Pertama, kajian menggunakan data yang diperoleh daripada Kaggle.com. Set data tersebut bukan yang paling terkini dan mungkin tidak boleh digeneralisasikan kepada populasi lain disebabkan bias demografi. Oleh itu, keputusan yang diperoleh dan strategi yang dicadangkan adalah tidak tepat selepas beberapa tahun. Selain itu, set data adalah tidak seimbang, dengan 53.17% pelanggan mempunyai skor kredit standard, 28.99% mempunyai skor kredit lemah dan 17.82% mempunyai skor kredit baik. Ketidakseimbangan ini boleh menjejaskan keputusan kajian. Model pembelajaran mesin yang paling tepat mungkin bukan model Hutan Rawak dan kaedah lain perlu dipertimbangkan.

Tambahan lagi, satu lagi batasan kajian adalah bahawa ciri-ciri dalam set data mungkin tidak merangkumi semua ciri yang diperlukan yang boleh menjejaskan skor kredit pelanggan. Ciri-ciri penting atau relevan lain mungkin tiada dalam set data semasa dan tidak digunakan untuk melatih model, yang boleh menurunkan prestasi pengelas. Walau bagaimanapun, walaupun ciri-ciri dalam set data dipertimbangkan secara keseluruhan, pengelas masih dapat mencapai keputusan yang memuaskan dalam meramal skor kredit pelanggan untuk syarikat kewangan.

KESIMPULAN

Kajian ini membentangkan visualisasi satu pemboleh ubah dan dua pemboleh ubah serta beberapa strategi yang dicadangkan dalam Bab 4. Dengan membandingkan keputusan prestasi enam model, KNN menunjukkan skor purata tertinggi, iaitu 77.94%. Ini bermakna RF adalah model yang berprestasi terbaik diikuti oleh SVM, DT, KNN, NB dan LR. Dapat disimpulkan bahawa RF mempunyai hubungan yang paling kuat untuk ramalan skor kredit pelanggan syarikat kewangan. Oleh itu, ia dipilih sebagai model terunggul untuk kajian ini.

Visualisasi satu pemboleh ubah dan dua pemboleh ubah telah ditunjukkan dan beberapa strategi yang dicadangkan boleh didapati dalam Bab 4. Dengan membandingkan keputusan prestasi enam model iaitu KNN, NB, SVM, DT, LR dan RF, KNN menunjukkan skor purata tertinggi iaitu 77.94%. Ini bermakna RF adalah model yang berprestasi

terbaik diikuti oleh SVM, DT, KNN, NB dan LR. Dapat disimpulkan bahawa RF mempunyai hubungan yang lebih kuat untuk ramalan kajian. Justeru, ia dipilih sebagai model terunggul untuk meramal skor kredit pelanggan syarikat kewangan.

Akhir sekali, adalah penting untuk diperhatikan bahawa terdapat beberapa batasan dalam kajian ini. Walau bagaimanapun, beberapa cadangan dan saranan telah diberikan untuk menangani batasan ini dan memastikan keputusan yang lebih baik boleh diperoleh dalam kajian akan datang. Cadangan dan saranan ini diberikan bagi meningkatkan ketepatan dan kebolehgeneralisasian penemuan kajian.

RUJUKAN

A.Elsalamony, H. (2014). Bank direct marketing analysis of data mining techniques. International Journal of Computer Applications, 85(7), 12-22. doi:10.5120/14852-3218

Ampountolas, A., Nyarko Nde, T., Date, P., & Constantinescu, C. (2021). A machine learning approach for Micro-Credit scoring. Risks, 9(3), 50. doi:10.3390/risks9030050

Boz, Z., Gunnec, D., Birbil, S. I., & Öztürk, M. K. (2018). Reassessment and monitoring of loan applications with Machine Learning. Applied Artificial Intelligence, 32(9-10), 939-955. doi:10.1080/08839514.2018.1525517

Breheny, P. (n.d.). Kernel density classification. In Nonparametric Statistics. STA 621:.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:10.1023/a:1010933404324

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support Vector Machine Classification: Applications, challenges and Trends. Neurocomputing, 408, 189-215. doi:10.1016/j.neucom.2019.10.118

Charbuty, B., & Abdulazeez, A. (2021). Classification based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20-28. doi:10.38094/jastt20165

Cunningham, P., & Delany, S. J. (2022). K-nearest neighbour classifiers - a tutorial. ACM Computing Surveys, 54(6), 1-25. doi:10.1145/3459665

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xx, W. (2018, January 02). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics & Proteomics, 15(1), 41-51. doi:10.21873/cgp.20063

Jadhav, S. D., & Channe, H. P. (2016, August). Efficient Recommendation System Using Decision Tree Classifier and Collaborative Filtering. International Research Journal of Engineering and Technology (IRJET), 03(08), 2016th ser., 2113-2118.

Koohang, A., Sargent, C. S., Nord, J. H., & Paliszkiewicz, J. (2022). Internet of things (IOT): From Awareness to continued use. International Journal of Information Management, 62, 102442. doi:10.1016/j.ijinfomgt.2021.102442

Kumar, A., Shanthi, D., & Bhattacharya, P. (2021). Credit Score Prediction System using deep learning and K-means algorithms. Journal of Physics: Conference Series, 1998(1), 012027. doi:10.1088/1742-6596/1998/1/012027

Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E., & Aswini, E. (2019). Credit card fraud detection using random forest algorithm. 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), 149-153. doi:10.1109/iccct2.2019.8824930

Kwon, Y., Shin, W., Ko, J., & Lee, J. (2020). AK-score: Accurate protein-ligand binding affinity prediction using an ensemble of 3D-Convolutional Neural Networks. International Journal of Molecular Sciences, 21(22), 8424. doi:10.3390/ijms21228424

Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., & Forghani, R. (2020). Overview of Machine Learning Part 1. Neuroimaging Clinics of North America, 30(4). doi:10.1016/j.nic.2020.08.007

Mat Amin, M., Yep Ai Lan, J., Makhtar, M., & Rasid Mamat, A. (2018). A decision tree based recommender system for backpackers accommodations. International Journal of Engineering & Technology, 7(2.15), 45. doi:10.14419/ijet.v7i2.15.11210

Nasteski, V. (2017). An overview of the supervised machine learning methods. HORIZONS.B, 4, 51-62. doi:10.20544/horizons.b.04.1.17.p05

Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing, 43(2), 154. doi:10.4040/jkan.2013.43.2.154

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78. doi:10.26438/ijcse/v6i10.7478

Rudra Kumar, M., & Kumar Gunjan, V. (2020). Review of Machine Learning Models for credit scoring analysis. Ingeniería Solidaria, 16(1). doi:10.16925/2357-6014.2020.01.11

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and Research Directions. SN Computer Science, 2(3). doi:10.1007/s42979-021-00592-x

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: An overview. Journal of Thoracic Disease, 11(S4). doi:10.21037/jtd.2019.01.25

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). doi:10.1109/iccs45141.2019.9065747

Tong, S., & Koller, D. (2001, February 11). Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 45-66.

Tsai, C., & Chen, M. (2010). Credit rating by Hybrid Machine Learning Techniques. Applied Soft Computing, 10(2), 374-380. doi:10.1016/j.asoc.2009.08.003

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1). doi:10.1186/s12911-019-1004-8

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1). doi:10.1186/s12911-019-1004-8

Vijayarani, S., & Dhayanand, S. (2015, April). Liver Disease Prediction using SVM and Naïve Bayes Algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), issn: 2278 – 7798, 816-820.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A comparative assessment of Credit Risk Model based on machine learning ——a case study of Bank Loan Data. Procedia Computer Science, 174, 141-149. doi:10.1016/j.procs.2020.06.069

Weng, C., & Huang, C. (2021). A hybrid machine learning model for credit approval. Applied Artificial Intelligence, 35(15), 1439-1465. doi:10.1080/08839514.2021.1982475

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449. doi:10.1016/j.eswa.2008.06.121

Yin, H. (2019). Bank globalization and Financial Stability: International evidence. Research in International Business and Finance, 49, 207-224. doi:10.1016/j.ribaf.2019.03.009

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers, 12(1), 469-477. doi:10.1016/j.gsf.2020.03.007

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179. doi:10.1016/j.imu.2019.100179

MENGENAI PENGARANG

KHAI WAH KHAW memperoleh ijazah Ph.D. dari Pusat Pengajian Sains Matematik, Universiti Sains Malaysia (USM). Beliau kini merupakan Profesor Madya di Pusat Pengajian Pengurusan, USM. Bidang penyelidikannya merangkumi kawalan proses statistik dan analisis data lanjutan.