PENGESANAN AWAL DIABETES MENGGUNAKAN TEKNIK KLASIFIKASI PEMBELAJARAN MESIN

[1]Zahirah Mohd Zain, [1]Salwani Abdullah

[1]Fakulti Teknologi & Sains Maklumat
43600 Universiti Kebangsaan Malaysia

Copyright@FTSM UKM

Abstrak

Diabetes merupakan penyakit kronik yang semakin meningkat di seluruh dunia dan boleh membawa kepada pelbagai komplikasi kesihatan yang serius seperti penyakit jantung, kegagalan buah pinggang, kerosakan saraf dan masalah penglihatan sekiranaya tidak dikesan dengan lebih awal. Pengesanan awal amat penting bagi membolehkan rawatan yang lebih berkesan dalam usaha mengawal penyakit ini serta mengurangkan risiko komplikasi jangka panjang. Projek ini bertujuan membangunkan model klasifikasi pembelajaran mesin bagi pengesanan awal diabetes dengan menggunakan algoritma seperti K-Jiran Terdekat (KNN), Sokongan Vektor Mesin (SVM) dan Hutan Rawak (RF). Data yang digunakan dalam kajian ini ialah Set Data Ramalan Risiko Diabetes Peringkat Awal yang diambil daripada UCI Machine Learning Repository. Data ini mengandungi maklumat gejala dan faktor demografi yang berkaitan diabetes. Antara ciri yang dianalisis termasuk umur, jantina, serta gejala seperti poliuria (kencing berlebihan), polidipsia (dahaga berlebihan), penurunan berat badan secara tiba-tiba, kelemahan badan, polifagia (selera makan berlebihan), Jangkitan faraj atau kelamin, penglihatan kabur, kegatalan, mudah marah, penyembuhan luka yang lambat, kehilangan deria separa, ketegangan otot, alopecia (keguguran rambut) dan obesiti. Prestasi setiap model klasifikasi akan dinilai berdasarkan metrik seperti ketepatan, kepersisan, dapatan semula dan skor F1 bagi memastikan keberkesanan model dalam mengenal pasti individu yang berisiko dengan lebih tepat. Melalui kajian ini, diharapkan pembangunan model klasifikasi pembelajaran mesin dapat membantu meningkatkan keupayaan sistem saringan kesihatan untuk mengesan diabetes pada peringkat awal seterusnya membolehkan pencegahan yang lebih awal dan berkesan dalam mengurangkan impak penyakit ini.

Kata Kunci: Diabetes, Pembelajaran Mesin, Klasifikasi

Abstract

Diabetes is a chronic disease that is increasingly prevalent worldwide and can lead to various serious health complications such as heart disease, kidney failure, nerve damage, and vision problems if not detected early. Early detection is crucial to enable more effective treatment in managing the disease and reducing the risk of long-term complications. This project aims to develop a machine learning classification model for the early detection of diabetes using algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). The dataset used in this study is the Early-Stage Diabetes Risk Prediction dataset obtained from the UCI Machine Learning Repository. This dataset contains information on symptoms and demographic factors related to diabetes. The features analyzed include age, gender, and symptoms such polyuria (excessive urination), polydipsia (excessive thirst), sudden weight loss, weakness, polyphagia (excessive hunger), genital thrush, blurred vision, itching, irritability, delayed wound healing, partial paresis, muscle stiffness, alopecia (hair loss), and obesity. The performance of each classification model will be evaluated based on metrics such as accuracy, precision, recall, and F1-score to ensure the effectiveness of the model in accurately identifying individuals at risk. Through this study, the development of a machine learning classification model is expected to enhance the capability of health screening systems in detecting diabetes at an early stage, thereby enabling earlier and more effective prevention of the disease's impact.

Keywords: Diabetes, Machine Learning, Classification

Copyright@FTSM UKM

1.0 PENGENALAN

Pengesanan awal diabetes telah menjadi tumpuan utama dalam bidang perubatan memandangkan penyakit ini sering tidak menunjukkan simptom pada peringkat awal namun boleh membawa kepada komplikasi serius jika tidak dirawat dengan segera. Walaupun ujian darah dan kaedah konvensional lain boleh mengesan diabetes, pendekatan ini lazimnya memerlukan masa, kos yang tinggi, dan akses kepada fasiliti kesihatan, yang sering kali tidak tersedia di kawasan luar bandar atau bagi golongan yang kurang berkemampuan. Menurut Kementerian Kesihatan Malaysia (KKM), lebih daripada 3 juta rakyat Malaysia menghidap diabetes, dan 1 daripada 5 dewasa tidak menyedari mereka menghidap penyakit ini. Situasi ini menekankan keperluan untuk kaedah pengesanan yang lebih efisien, tepat dan mudah diakses oleh masyarakat umum.

Kemajuan dalam teknologi kecerdasan buatan, khususnya pembelajaran mesin, telah membuka peluang baharu dalam bidang kesihatan awam. Model pembelajaran mesin boleh dilatih menggunakan data sejarah kesihatan pesakit untuk mengenal pasti corak tersembunyi yang sukar dikesan oleh manusia. Oleh itu, penggunaan teknik pembelajaran mesin dalam pengesanan awal diabetes berpotensi besar untuk meningkatkan ketepatan diagnosis dan mengurangkan kadar kes diabetes yang tidak didiagnosis.

Matlamat utama projek ini adalah untuk membangunkan model pembelajaran mesin yang boleh dipercayai dan tepat bagi pengesanan awal diabetes berdasarkan data kesihatan klinikal yang dikumpulkan daripada pesakit. Untuk mencapai matlamat ini, kajian ini menetapkan beberapa objektif utama iaitu menilai dan membandingkan prestasi tiga algoritma pembelajaran mesin iaitu KNN, SVM dan RF serta menghasilkan model terbaik yang boleh digunakan dalam sistem pengesanan awal yang praktikal. Model ini berpotensi membantu pengamal perubatan dalam membuat keputusan awal dan seterusnya menyumbang kepada strategi pencegahan yang lebih proaktif.

2.0 KAJIAN LITERATUR

Terdapat beberapa kajian lepas yang telah dirujuk untuk mendapatkan gambaran mengenai kajian yang dilakukan. Penyelidikan pertama telah dijalankan oleh Rayhan & Ekadinata (2023) menggunakan dataset Pima Indians Diabetes Dataset (PIDD) dari laman web Kaggle, yang mengandungi 768 data pesakit di mana 500 daripadanya adalah pesakit diabetes dan 268 adalah bukan diabetes. Prestasi model ini dinilai berdasarkan metrik seperti ketepatan, kepersisan, dapatan semula, dan skor F1. Model dengan nilai K=14 mencapai ketepatan sebanyak 84.41%, kepersisan 0.76, dapatan semula 0.62 dan skor F1 0.68 selepas diuji sebanyak tiga kali. Keputusan ini menunjukkan bahawa algoritma KNN dengan K=14 berprestasi baik dalam mengklasifikasikan pesakit diabetes, namun terdapat ruang untuk penambahbaikan terutamanya dari segi dapatan semula untuk meningkatkan keupayaan model dalam mengesan kes diabetes sebenar. Penyelidikan kedua oleh Muthu & Suriya. (2023) bertujuan untuk meramalkan diabetes jenis 2 dengan menggunakan set data PIDD daripada Kaggle seperti dalam kajian pertama. Model KNN dengan nilai K=40 mencapai ketepatan sebanyak 70%. Data dipra-pemproses dengan menyingkirkan nilai hilang bagi mendapatkan hasil yang terbaik dan kaedah *Standard Scalar digunakan* untuk penskalaan data. Teknik *K-Fold Cross Validation* dengan 10 lipatan juga digunakan untuk mengelakkan masalah *overfitting* dan untuk meningkatkan kebolehpercayaan hasil. Penyelidik mencadangkan nilai K lebih tinggi untuk dataset yang lebih besar bagi penambahbaikkan lagi dengan menggunakan nilai K yang lebih tinggi untuk dataset yang lebih besar.

Copyright@FTSM UKM

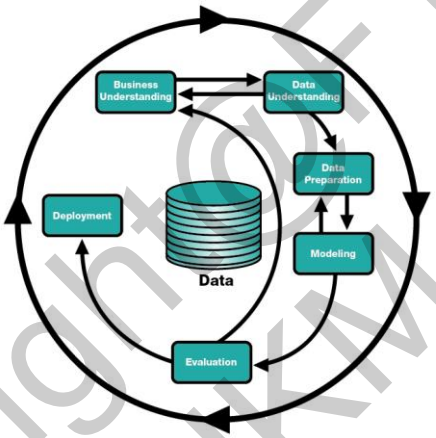
Selain itu, kajian oleh Apriliah, W. et al. (2021) turut menilai prestasi SVM dalam pengesanan awal diabetes, di mana model SVM mencapai ketepatan sebanyak 94.8% serta nilai ROC sebanyak 0.949. Hasil ini menunjukkan bahawa SVM mempunyai keupayaan yang sangat baik dalam membezakan kelas pesakit diabetes dan bukan diabetes dengan ketepatan yang tinggi. Dalam kajian ini, penggunaan SVM dapat mengekalkan prestasi yang stabil walaupun pada dataset yang lebih kecil, menjadikannya pilihan yang kukuh dalam pengesanan awal diabetes terutama apabila jumlah data yang tersedia adalah terhad. Selain itu, kajian lain yang diterbitkan oleh Viloria, A. et al.,(2020) memfokuskan pada penggunaan SVM untuk meramalkan diabetes berdasarkan parameter seperti umur, BMI dan tekanan darah. Dalam kajian ini, model SVM menunjukkan ketepatan sebanyak 95.36%, kepersisan 94.36%, dapatan semula 94.36%, skor F1 94.36% dan nilai ROC sebanyak 0.949, yang menandakan prestasi yang sangat baik dalam mengklasifikasikan pesakit yang berisiko diabetes. Keputusan ini menunjukkan bahawa SVM adalah teknik yang sangat efektif untuk pengesanan awal diabetes, dengan kemampuan untuk memberikan hasil yang stabil dan boleh dipercayai walaupun dalam dataset yang agak kecil.

Seterusnya, Kajian yang menggunakan RF dalam pengesanan diabetes yang diterbitkan oleh Apriliah, W. et al. (2021) menggunakan Sylhet Diabetes Dataset daripada UCI yang mengandungi 520 rekod pesakit dari Rumah Sakit Diabetes Sylhet, Bangladesh. Dalam kajian ini, RF mencapai ketepatan tertinggi iaitu 97.88% dan nilai ROC sebanyak 0.998, yang menunjukkan prestasi yang sangat baik dalam mengklasifikasikan data pesakit diabetes. Model ini telah dibandingkan dengan algoritma lain seperti SVM dan Naive Bayes, di mana RF menunjukkan prestasi yang lebih baik, terutama dari segi nilai ROC yang hampir mencapai 1. Ini menandakan bahawa RF memiliki kemampuan yang sangat baik untuk memisahkan dan mengklasifikasikan kelas-kelas data dengan ketepatan yang lebih tinggi. Kajian lain yang menggunakan RF untuk pengesanan awal diabetes yang ditulis oleh Kopitar, L. et al (2020) bertujuan untuk membandingkan prestasi pelbagai model ramalan berasaskan pembelajaran mesin dalam pengesanan awal diabetes jenis 2. Data yang digunakan dalam kajian ini melibatkan 27,050 individu dewasa yang belum didiagnosis, dengan data yang diambil dari rekod kesihatan elektronik (EHR) yang dikumpulkan antara Disember 2014 hingga September 2017. Antara model-model pembelajaran mesin yang diuji dalam kajian ini adalah Regresi Linear, Model Linear Glmnet, RF, Peningkatan Kecerunan Melampau (XGBoost) dan Mesin Penggalak Kecerunan (LightGBM). Berdasarkan keputusan kajian, model RF menunjukkan prestasi yang mengagumkan dengan RMSE sebanyak 0.842 dan skor R^2 tertinggi iaitu 0.369, menjadikannya model yang sangat berpotensi untuk ramalan yang konsisten dalam pengesanan awal diabetes berbanding dengan model-model lain.

Copyright@FTSM UKM

3.0 METODOLOGI

Metodologi yang diterapkan dalam kajian ini berlandaskan Proses Piawai Cross-Industri untuk Perlombongan Data (CRISP-DM), sebuah model yang diiktiraf secara meluas dalam dunia perlombongan data dan analitik. CRISP-DM terdiri daripada enam fasa utama yang membentuk kitaran hayat projek perlombongan data, menawarkan pendekatan yang sistematik dan berstruktur untuk mencapai hasil yang diinginkan (Chapman et al., 2000). Enam fasa tersebut merangkumi Pemahaman Perniagaan, Pemahaman Data, Penyediaan Data, Pemodelan, Penilaian dan Penggunaan. Rujukan kepada fasa-fasa ini dapat dilihat dalam Rajah 1.1 di bawah:



Rajah 1: Fasa CRISP-DM. (DSLYTICS,2017)

3.1 Fasa Pemahaman Terhadap Perniagaan (Business Understanding)

Fasa pemahaman terhadap perniagaan adalah langkah pertama dalam rangka kerja CRISP-DM, yang bertujuan untuk memastikan pembangunan projek dijalankan dengan memenuhi keperluan dan objektif yang jelas dari perspektif perniagaan. Dalam konteks sistem pengesanan awal diabetes, objektif utama adalah untuk membangunkan model pembelajaran mesin yang berkesan yang mampu meramalkan risiko diabetes pada peringkat awal.

3.2 Fasa Pemahaman Data (Data Understanding)

Fasa Pemahaman Data adalah langkah kedua dalam rangka kerja CRISP-DM. Fasa ini bertujuan untuk memahami data yang ada, mengenal pasti kualiti dan kebolehpercayaan data serta menentukan apa yang perlu dilakukan seterusnya dalam proses perlombongan data.

Copyright@FTSM UKM

Dalam kajian pengesanan awal diabetes, data yang digunakan adalah Set Data Ramalan Risiko Diabetes Peringkat Awal yang diperoleh daripada UCI Machine Learning Repository. Dataset ini mengandungi maklumat perubatan yang berkaitan dengan risiko diabetes, dengan 520 rekod individu yang mengandungi 17 atribut seperti poliuria, polidipsia, penurunan berat badan secara mendadak, kelemahan, polifagia, jangkitan faraj atau kelamin, penglihatan kabur, kegatalan, mudah marah, penyembuhan luka yang lambat, kehilangan deria separa, ketegangan otot, keguguran rambut dan obesiti.

3.3 Fasa Penyediaan Data (Data Preparation)

Fasa Penyediaan Data adalah langkah ketiga dalam pendekatan CRISP-DM yang bertujuan memastikan data yang telah dikumpulkan dalam fasa sebelumnya berada dalam keadaan yang bersih, lengkap, dan sesuai untuk digunakan dalam proses pemodelan. Dalam kajian pengesanan awal diabetes, fasa penyediaan data ini memainkan peranan yang sangat penting dalam meningkatkan kualiti dan kebolehgunaan data untuk membangunkan model pembelajaran mesin yang berkesan. kebolehgunaan data untuk membangunkan model pembelajaran mesin yang berkesan. Langkah-langkah dalam fasa penyediaan data adalah seperti penerokaan data, pembersihan data, penukaran atribut kategori kepada format berangka dan penskalaan data.

3.4 Fasa Pemodelan (Modeling)

Fasa Pemodelan adalah langkah keempat dalam pendekatan CRISP-DM yang bertujuan untuk membangunkan model pembelajaran mesin menggunakan data yang telah diproses dan disediakan. Fasa ini amat penting dalam kajian pengesanan awal diabetes ini kerana ia melibatkan penggunaan algoritma pembelajaran mesin untuk mengenal pasti corak dan hubungan dalam data yang boleh digunakan untuk meramalkan risiko diabetes pada individu. Algoritma yang digunakan dalam penyelidikan ini adalah KNN, SVM dan RF. Fasa pemodelan sering kali melibatkan pendekatan berulang termasuk mencuba pelbagai algoritma dan menilai prestasi model sehingga hasil yang optimum diperoleh. KNN adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi dengan cara membandingkan titik data baru dengan titik data dalam set latihan menggunakan matrik jarak. Algoritma ini berfungsi dengan mencari K jiran terdekat bagi titik data baru dan membuat ramalan berdasarkan purata atau kebanyakan kelas jiran tersebut. Dalam KNN, nilai K memainkan peranan penting dalam menentukan prestasi model. Salah satu matrik jarak yang sering digunakan adalah jarak Euclidean, yang mengukur jarak antara dua titik data berdasarkan perbezaan koordinat mereka (Cover & Hart, 1967). SVM ialah mencari hyperplane terbaik yang memisahkan dua kelas contohnya individu yang berisiko menghidap

Copyright@FTSM UKM

diabetes dan yang tidak berisiko berdasarkan ciri-ciri kesihatan mereka. SVM berfungsi dengan mencari hyperplane yang memaksimumkan margin antara dua kelas. Margin merujuk kepada jarak antara hyperplane dan titik data terdekat dari setiap kelas. RF ialah algoritma pembelajaran mesin yang sering digunakan dalam klasifikasi dan regresi. Dalam RF, keputusan akhir dibuat berdasarkan gabungan ramalan dari pelbagai pokok keputusan. Setiap pokok menyumbang kepada ramalan keseluruhan melalui undian majoriti untuk klasifikasi.

3.5 Fasa Penilaian (Evaluation)

Fasa Penilaian adalah langkah kelima dalam pendekatan CRISP-DM yang bertujuan untuk menilai prestasi model yang telah dilatih menggunakan set data ujian. Set data ujian ini dipisahkan daripada set data latihan semasa fasa latihan model. Dalam kajian pengesanan awal diabetes, fasa ini sangat penting kerana ia melibatkan pengujian sama ada model pembelajaran mesin yang dibangunkan dapat mengesan risiko diabetes dengan tepat. Dalam kajian ini, metrik yang digunakan untuk menilai prestasi model dalam mengesan risiko diabetes melibatkan pengukuran seperti ketepatan, kepersisan, dapatan semula dan skor F1. Nilai ketepatan, kepersisan, dapatan semula dan skor F1 boleh diperolehi daripada nilai True Negative (TN), False Positive (FP), True Positive (TP) dan False Negative (FN) .

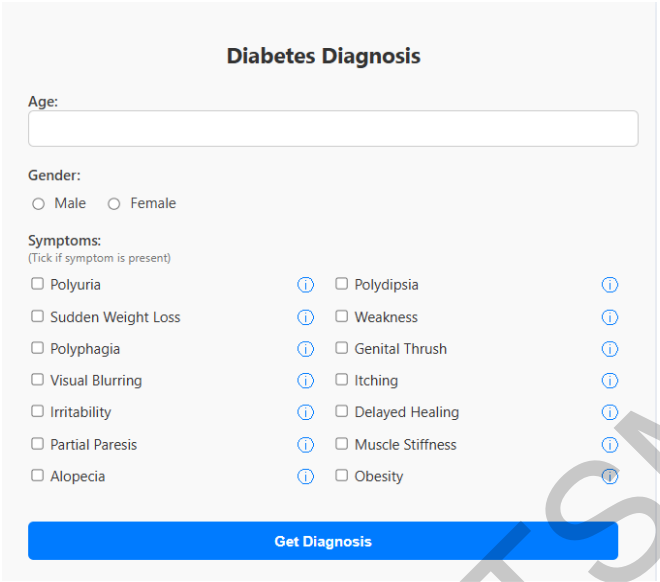
3.6 Fasa Penggunaan (Deployment)

Fasa Penggunaan adalah langkah terakhir dalam pendekatan CRISP-DM yang melibatkan penggunaan model yang telah dibangunkan dan dinilai dalam fasa-fasa sebelumnya. Dalam fasa ini, model pembelajaran mesin yang telah dibangunkan dan diuji akan diimplementasikan untuk tujuan praktikal, memberikan manfaat sebenar kepada pengguna, dan membantu mencapai objektif penyelidikan.

3.7 Reka Bentuk Antara Muka

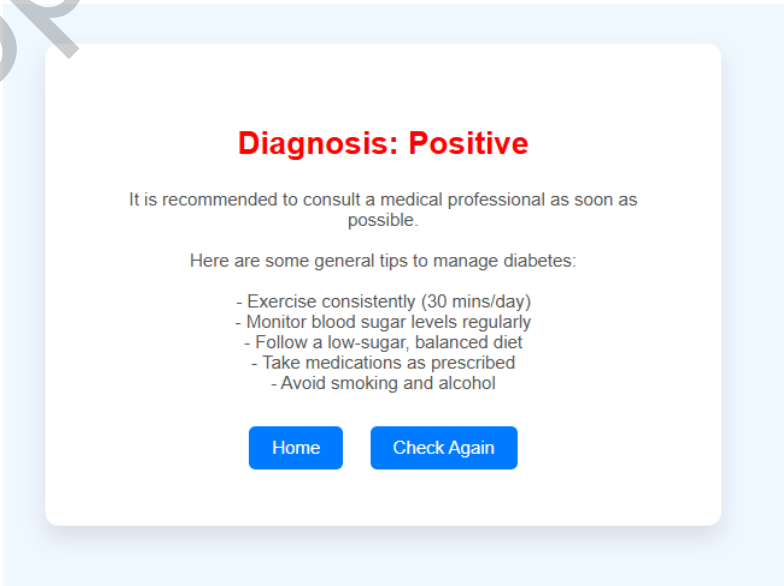
Dalam penyelidikan ini, penyelidik menggunakan Flask, iaitu rangka kerja web berasaskan Python, untuk membangunkan papan pemuka aplikasi web bagi sistem pengesanan awal diabetes. Dengan Flask, penyelidik dapat membina antara muka pengguna yang interaktif dan mesra pengguna untuk membolehkan pengguna memasukkan maklumat kesihatan dan memperoleh keputusan ramalan serta maklumat lanjut berkaitan diabetes. Rajah 2 menunjukkan papan pemuka yang dihasilkan.

Copyright@FTSM UKM



Rajah 2: Antara Muka Pengesesan Diabetes.

Melalui papan pemuka ini, pengguna boleh mengisi maklumat kesihatan seperti umur, jantina, dan beberapa gejala klinikal yang laina. Setelah data dimasukkan dan dihantar, sistem akan memproses maklumat tersebut menggunakan model pembelajaran mesin yang telah dilatih, dan memaparkan keputusan ramalan secara serta-merta. Setelah keputusan dipaparkan, pengguna akan diberikan maklumat lanjut mengenai status ramalan sama ada mereka berisiko menghidap diabetes atau tidak. Sekiranya pengguna didapati berisiko, sistem turut menyediakan cadangan langkah pencegahan dan maklumat penjagaan kesihatan yang bersesuaian. Antara muka keputusan ini ditunjukkan dalam Rajah 3.



Rajah 3: Antara muka keputusan.

Copyright@FTSM UKM

4.0 HASIL

Bahagian ini membentangkan hasil penilaian metrik untuk model klasifikasi, membandingkan Keputusan penilaian metrik bagi model KNN, SVM dan RF, dipaparkan dalam bentuk jadual dan rajah.

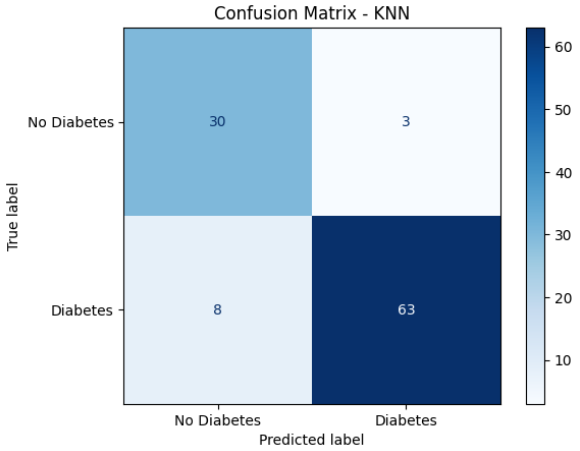
4.1 Penilaian K-Jiran Terdekat (KNN)

Jadual 1 menunjukkan keputusan penilaian metrik bagi model KNN sebelum penalaan termasuk kepersisan, dapatan semula, skor F1 dan sokongan bagi setiap kelas.

Jadual 1: *Keputusan penilaian metrik model KNN*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	0.79	0.91	0.85	33
Diabetes	0.95	0.89	0.92	71
Ketepatan: 89.42%				

Berdasarkan Jadual 1, model KNN mencatatkan ketepatan keseluruhan sebanyak 89.42%, menunjukkan prestasi klasifikasi yang baik. Bagi kelas Tidak Diabetes, model mencapai kepersisan 0.79, yang agak rendah berbanding dapatan semula sebanyak 0.91. Ini menunjukkan bahawa model cenderung untuk mengesan hampir semua kes sebenar Tidak Diabetes, tetapi menghasilkan lebih banyak positif palsu. Skor F1 sebanyak 0.85 menunjukkan prestasi seimbang antara ketepatan dan keupayaan mengesan kelas tersebut. Sementara itu, bagi kelas Diabetes, model menunjukkan prestasi cemerlang dengan kepersisan 0.95, dapatan semula 0.89, dan skor F1 0.92. Ini menunjukkan bahawa model sangat berkesan dalam mengenal pasti pesakit diabetes dengan kadar kesilapan yang rendah. Rajah 4 menunjukkan matriks kekeliruan bagi model KNN.



Rajah 4: Matriks kekeliruan model KNN

Copyright@FTSM UKM

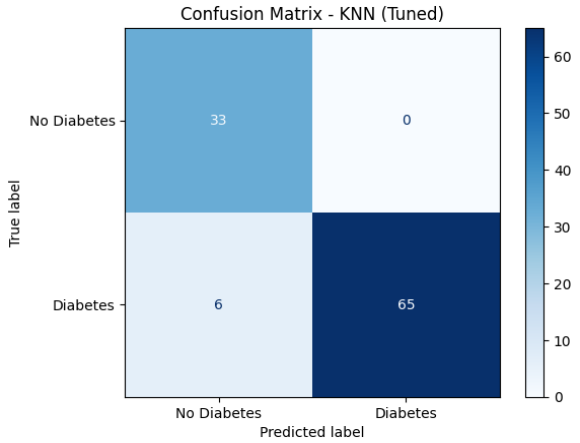
Berdasarkan Rajah 4, model KNN berjaya mengklasifikasikan 63 daripada 71 kes Diabetes dengan betul dan 30 daripada 33 kes Tidak Diabetes dengan betul. Walau bagaimanapun, terdapat 3 kes Tidak Diabetes yang disalah klasifikasikan sebagai Diabetes dan 8 kes Diabetes tersilap dikelaskan sebagai Tidak Diabetes. Walaupun berlaku beberapa kekeliruan antara kelas, model masih mengekalkan tahap prestasi keseluruhan yang baik.

Seterusnya, model KNN telah ditambah baik melalui proses penalaan hiperparameter menggunakan kaedah RandomedSearch bagi mendapatkan konfigurasi model terbaik. Jadual 2 menunjukkan keputusan penilaian metrik model KNN yang telah ditala.

Jadual 2: *Keputusan penilaian metrik model KNN selepas penalaan (Train-Test Split)*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	0.85	1.00	0.92	33
Diabetes	1.00	0.92	0.96	71
Ketepatan: 94.23%				

Berdasarkan Jadual 2, penalaan hiperparameter berjaya meningkatkan prestasi model KNN secara keseluruhan. Ketepatan meningkat kepada 94.23%. Bagi kelas Tidak Diabetes, skor F1 meningkat daripada 0.85 kepada 0.92, menunjukkan prestasi lebih seimbang dan tepat selepas penalaan. Bagi kelas Diabetes, peningkatan kecil dalam semua metrik menunjukkan model menjadi lebih konsisten dan boleh dipercayai dalam pengesanan kes sebenar.



Rajah 5: Matriks kekeliruan model KNN selepas penalaan

Copyright@FTSM UKM

Berdasarkan Rajah 5, model yang telah ditala menunjukkan pengurangan yang ketara dalam kes kekeliruan. Model berjaya mengklasifikasikan 33 kes Tidak Diabetes dan 65 daripada 71 kes Diabetes dengan betul. Hanya 6 kes Diabetes yang dikelaskan secara salah. Ini menunjukkan bahawa proses penalaan hiperparameter memberikan kesan positif terhadap prestasi klasifikasi, khususnya dalam meningkatkan keupayaan generalisasi model terhadap data baharu. Secara keseluruhannya, model KNN menunjukkan prestasi yang baik dalam pengesanan awal diabetes. Penalaan hiperparameter berjaya meningkatkan ketepatan, keupayaan mengesan kes sebenar dan mengurangkan kesilapan klasifikasi. Model ini berpotensi untuk digunakan dalam sistem pengesanan awal dengan keupayaan membuat keputusan yang lebih tepat dan konsisten.

4.2 Penilaian Mesin Sokongan Vektor (SVM)

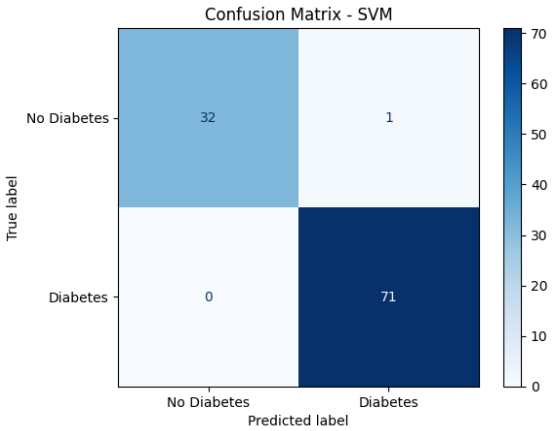
Jadual 3 menunjukkan keputusan penilaian metrik bagi model SVM sebelum penalaan, merangkumi nilai kepersisan, dapatan semula, skor F1, dan sokongan untuk setiap kelas.

Jadual 3: *Keputusan penilaian metrik model SVM*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	1.00	0.97	0.98	33
Diabetes	0.99	1.00	0.99	71
Ketepatan: 99.04%				

Berdasarkan Jadual 3, model SVM menunjukkan prestasi yang sangat tinggi dalam klasifikasi kedua-dua kelas dan mencapai ketepatan sebabyak 99.04%. Untuk kelas Tidak Diabetes, model mencatatkan kepersisan 1.00 dan dapatan semula 0.97, yang membawa kepada skor F1 sebanyak 0.98. Bagi kelas Diabetes, dapatan semula mencapai 1.00, menunjukkan semua kes diabetes berjaya dikenal pasti dengan betul oleh model, dengan skor F1 yang sangat tinggi iaitu 0.99. Rajah 4 mempersembahkan matriks kekeliruan bagi model SVM.

Copyright@FTSM UKM



Rajah 6: Matriks kekeliruan model SVM

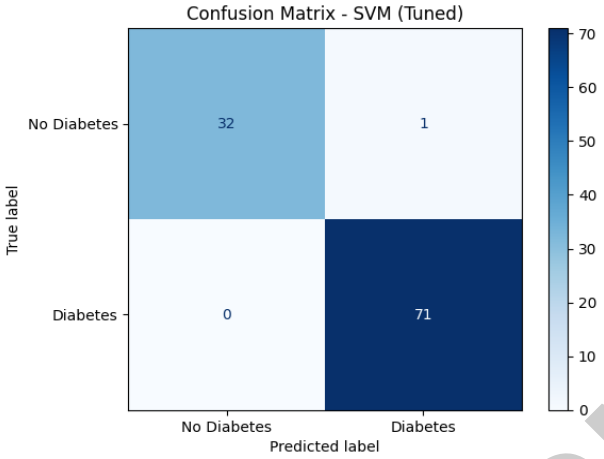
Berdasarkan matriks kekeliruan pada rajah 6, sebanyak 32 daripada 33 sampel kelas Tidak Diabetes diklasifikasikan dengan betul, manakala 1 sampel telah disalah klasifikasikan. Untuk kelas Diabetes, semua 71 sampel berjaya diklasifikasikan dengan betul. Ini menunjukkan model sangat konsisten dan berprestasi tinggi dalam pengesanan diabetes, walaupun sebelum sebarang penalaan parameter dilakukan. Penalaan hiperparameter juga telah dijalankan ke atas model SVM menggunakan kaedah RandomizedSearch untuk membandingkan prestasi. Hasil menunjukkan model SVM selepas penalaan masih menunjukkan keputusan yang sama seperti sebelum ini. Jadual 4 menunjukkan keputusan penilaian metrik bagi model SVM selepas penalaan.

Jadual 4: *Keputusan penilaian metrik model SVM selepas penalaan (Train-Test Split)*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	1.00	0.97	0.98	33
Diabetes	0.99	1.00	0.99	71
Ketepatan: 99.04%				

Berdasarkan jadual 4, prestasi model SVM selepas penalaan masih kekal tinggi. Ini menunjukkan bahawa parameter lalai yang digunakan oleh model SVM telah pun cukup optimum bagi dataset ini. Oleh itu, tiada perubahan ketara yang diperoleh melalui penalaan hiperparameter, namun ia tetap penting sebagai langkah pengesahan model. Struktur matriks kekeliruan kekal sama seperti sebelumnya, dengan 1 kes tersilap untuk kelas Tidak Diabetes dan 0 kesalahan untuk kelas Diabetes. Keputusan ini memperkukuhkan lagi kecekapan model dalam mengklasifikasikan data dengan sangat tepat.

Copyright@FTSM UKM



Rajah 7: Matriks kekeliruan model SVM selepas penalaan

Secara keseluruhan, model SVM menunjukkan prestasi yang cemerlang dalam tugas pengesanan awal diabetes. Dengan ketepatan sebanyak 99.04%, model ini berjaya mengklasifikasikan kedua-dua kelas dengan sangat sedikit kesilapan. Tiada peningkatan prestasi yang diperoleh selepas penalaan hiperparameter kerana model sudah berada pada konfigurasi yang hampir optimum. Kecekapan tinggi dalam mengesan pesakit diabetes menjadikan model SVM sangat sesuai untuk diaplikasikan dalam sistem pengesanan awal diabetes.

4.3 Penilaian Hutan Rawak (RF)

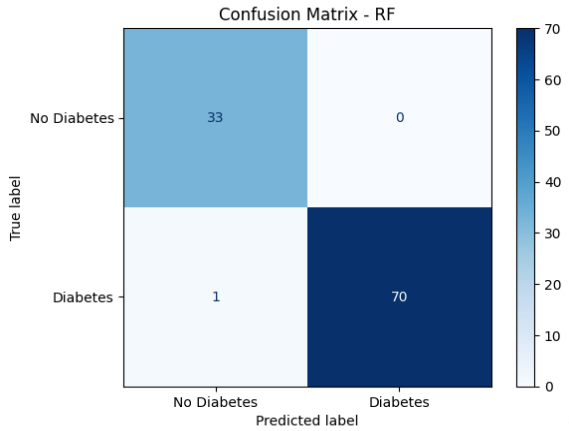
Bagi model RF, Jadual 55 menunjukkan pmenunjukkan keputusan penilaian model Random Forest sebelum penalaan, merangkumi metrik kepersisan, dapatan semula, skor F1, dan sokongan bagi kedua-dua kelas.

Jadual 5: *Keputusan penilaian metrik model RF*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	0.97	1.00	0.99	33
Diabetes	1.00	0.99	0.99	71
Ketepatan: 99.04%				

Berdasarkan jadual 5, model RF menunjukkan prestasi klasifikasi yang sangat tinggi, dengan ketepatan keseluruhan sebanyak 99.04%. Kelas Tidak Diabetes mempunyai kepersisan 0.97 dan dapatan semula 1.00, menunjukkan semua kes sebenar Tidak Diabetes berjaya dikesan. Manakala untuk kelas Diabetes, semua ramalan adalah betul dan model hanya gagal mengesan 1 kes sebenar Diabetes, memberikan dapatan semula 0.99.

Copyright@FTSM UKM



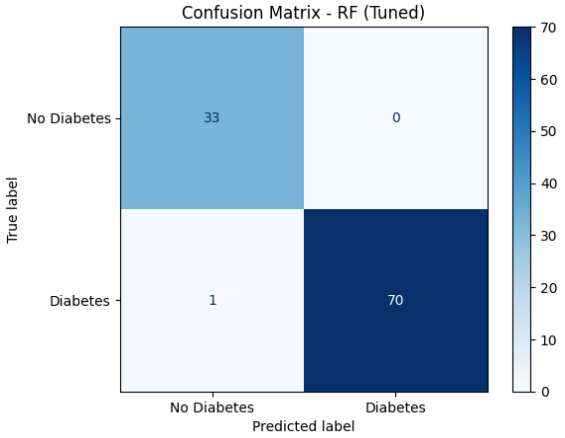
Rajah 8: Matriks kekeliruan model RF

Berdasarkan rajah 8, matriks kekeliruan bagi model RF menunjukkan bahawa semua 33 kes Tidak Diabetes diklasifikasikan dengan betul, manakala daripada 71 kes Diabetes, hanya 1 kes disalah klasifikasikan sebagai Tidak Diabetes. Ini membuktikan model RF sangat efektif dan stabil dalam pengesanan awal diabetes, walaupun tanpa penalaan lanjut. Model RF telah melalui proses penalaan hiperparameter menggunakan teknik seperti RandomizedSearch bagi membandingkan prestasi. Walau bagaimanapun, hasil menunjukkan bahawa prestasi selepas penalaan adalah sama dengan prestasi asal model, membuktikan bahawa konfigurasi asal sudah hampir optimum. Jadual 6 menunjukkan matriks penilaian metrik bagi model RF selepas penalaan dan rajah 7 menunjukkan matriks kekeliruan bagi model RF selepas penalaan.

Jadual 6: *Keputusan penilaian metrik model RF selepas penalaan (Train-Test Split)*

Label	Kepersisan	Dapatan Semula	Skor F1	Sokongan
Tidak Diabetes	0.97	1.00	0.99	33
Diabetes	1.00	0.99	0.99	71
Ketepatan: 99.04%				

Copyright@FTSM UKM



Rajah 9: Matriks kekeliruan model RF selepas penalaan

Tiada perubahan ketara pada prestasi model selepas penalaan. Model masih mengekalkan kebolehan klasifikasi yang sangat tinggi, dengan hanya satu kes Diabetes yang disalahklasifikasikan, manakala semua kes Tidak Diabetes diklasifikasikan dengan betul. Secara keseluruhan, model Random Forest mencatatkan prestasi yang sangat tinggi dan konsisten, dengan ketepatan 99.04% dan skor F1 yang hampir sempurna untuk kedua-dua kelas. Proses penalaan tidak memberikan perubahan signifikan kerana model sudah mencapai tahap optimum sejak awal lagi. Dengan keupayaan mengesan kes diabetes dan bukan diabetes dengan tepat serta kestabilan prestasi walaupun selepas penalaan, RF merupakan salah satu model terbaik yang diuji dalam kajian ini.

4.4 Perbandingan Model Secara Keseluruhan

Perbandingan prestasi antara model KNN, SVM dan RF merupakan langkah penting dalam menentukan pendekatan terbaik bagi sistem pengesanan awal diabetes. Ketiga-tiga model telah melalui proses penalaan hiperparameter bagi mengoptimumkan prestasi masing-masing. Jadual 7 menunjukkan keputusan perbandingan berdasarkan metrik ketepatan, kepersisan, dapatan semula, skor F1, serta ranking keseluruhan model.

Jadual 7: *Keputusan penilaian metrik model KNN, SVM dan RF*

Model	Ketepatan	Kepersisan	Dapatan Semula	Skor F1	Ranking
KNN	0.9423	0.92	0.96	0.94	2
SVM	0.9904	0.99	0.98	0.99	1
RF	0.9904	0.99	0.99	0.99	1

Copyright@FTSM UKM

Berdasarkan jadual 7, model SVM dan RF menunjukkan prestasi tertinggi dengan ketepatan 99.04%, manakala model KNN mencatatkan ketepatan 94.23%. Proses penalaan hiperparameter bagi SVM dan RF tidak membawa perubahan ketara terhadap prestasi kerana kedua-dua model ini telah mencapai konfigurasi hampir optimum sejak awal, menunjukkan keupayaan semula jadi mereka dalam mengendalikan dataset yang digunakan. Sebaliknya, model KNN menunjukkan peningkatan prestasi yang ketara selepas penalaan hiperparameter, khususnya dalam metrik dapatan semula dan skor F1, yang mencerminkan keupayaannya untuk mengenal pasti kes diabetes dengan lebih baik selepas konfigurasi optimum diperoleh. Walaupun ketepatan SVM dan RF adalah sama, perbezaan kecil dapat dilihat pada metrik dapatan semula, iaitu RF (0.99) sedikit mengatasi SVM (0.98). Ini menunjukkan bahawa RF lebih konsisten dalam mengenal pasti semua kes diabetes tanpa tertinggal. Namun, secara keseluruhan perbezaan tersebut adalah kecil dan tidak menjejaskan ranking utama berdasarkan ketepatan keseluruhan. Matrik kekeliruan bagi ketiga-tiga model turut dianalisis untuk memahami corak klasifikasi yang dilakukan. Model SVM dan RF hampir tidak menunjukkan kesilapan klasifikasi, dengan hanya satu atau dua kesilapan minimum, manakala model KNN masih menunjukkan beberapa klasifikasi tidak tepat, terutamanya pada kelas “Tidak Diabetes”. Perbezaan ini menunjukkan bahawa walaupun KNN boleh mencapai prestasi yang baik selepas penalaan, ia masih tidak dapat menandingi ketepatan tinggi dan konsisten yang ditunjukkan oleh SVM dan RF.

Secara keseluruhan, model SVM dan RF terbukti paling berkesan dalam klasifikasi awal diabetes bagi dataset ini, dengan prestasi yang konsisten dalam semua metrik penilaian. SVM dan RF merupakan pilihan utama manakala KNN boleh dipertimbangkan dalam situasi yang memerlukan model yang lebih mudah dan interpretatif. Proses penalaan hiperparameter pula telah membuktikan kepentingannya dalam memperbaiki prestasi model tertentu seperti KNN, dan dalam masa sama mengesahkan keupayaan semula jadi SVM dan RF untuk memberikan hasil terbaik walaupun dengan konfigurasi lalai.

Copyright@FTSM UKM

4.5 Kebolehginaan model pada set data Pima Indian Diabetes (PIDD)

Set data PIDD telah digunakan sebagai satu lagi set data untuk menilai kebolehlaksanaan model pembelajaran mesin dalam pengesanan awal diabetes. Dataset ini mengandungi maklumat perubatan seperti paras glukosa, tekanan darah, BMI dan umur serta status diabetes bagi setiap pesakit. Model KNN, SVM dan RF yang telah dibina dan dilatih menggunakan Set Data Ramalan Risiko Diabetes Peringkat Awal daripada UCI telah diuji semula pada set data PIDD untuk menilai keupayaan dan prestasi. Ketiga-tiga model ini telah dievaluasi berdasarkan empat metrik utama iaitu kepersisan, dapatan semula, skor F1 dan ketepatan.

Jadual 4.10 menunjukkan keputusan penilaian prestasi model klasifikasi untuk set data PIDD. Berdasarkan jadual tersebut, model RF mencatatkan prestasi terbaik secara keseluruhan dengan ketepatan 76.62%, diikuti oleh SVM 75.97% dan KNN 72.08%. Dari segi dapatan semula bagi kelas "Diabetes", RF memperoleh nilai 0.67, diikuti oleh SVM 0.65, dan KNN 0.55. Ini menunjukkan bahawa RF paling cekap dalam mengenal pasti pesakit yang menghidap diabetes berbanding dua model lain.

Jadual 8: *Keputusan penilaian metrik model klasifikasi untuk set data PIDD*

Model	Ketepatan	Kepersisan	Dapatan Semula	Skor F1	Ranking
KNN	0.7208	0.62	0.55	0.58	3
SVM	0.7597	0.67	0.65	0.66	2
RF	0.7662	0.67	0.67	0.67	1

Berdasarkan jadual 8, model RF menunjukkan prestasi paling seimbang antara kedua-dua kelas, dengan dapatan semula dan skor F1 tertinggi untuk kelas “Diabetes”. Model SVM berada di kedudukan kedua, dengan prestasi yang hampir setara, manakala KNN mencatatkan prestasi terendah dalam semua metrik utama, menjadikannya kurang sesuai untuk digunakan pada dataset ini, terutamanya apabila keupayaan untuk mengesan pesakit diabetes adalah keutamaan. Kesimpulannya, bagi set data PIDD, model RF merupakan model paling sesuai untuk tugas klasifikasi awal diabetes. Ia menunjukkan prestasi keseluruhan yang terbaik dan keupayaan paling tinggi dalam mengenal pasti pesakit yang menghidap diabetes. Model SVM boleh dipertimbangkan sebagai alternatif dengan ketepatan dan skor F1 yang kompetitif, manakala KNN kurang sesuai digunakan kerana prestasinya yang lebih rendah dalam semua metrik utama.

Copyright@FTSM UKM

5.0 KESIMPULAN

Kesimpulannya, projek ini telah berjaya mencapai objektif untuk membangunkan model pembelajaran mesin bagi pengesanan awal diabetes. Model RF telah dikenalpasti sebagai model terbaik berdasarkan prestasi metrik yang digunakan, diikuti oleh SVM dan KNN. Kajian ini juga menekankan kepentingan pra-pemprosesan data dalam pembangunan sistem klasifikasi yang berkesan. Walaupun terdapat beberapa kekangan, projek ini telah membuka ruang bagi penyelidikan lanjutan yang lebih mendalam dalam usaha membangunkan sistem ramalan penyakit yang lebih cekap, responsif dan boleh dipercayai.

6.0 PENGHARGAAN

Segala puji dan syukur bagi Allah, Yang Maha Pengasih lagi Maha Penyayang.saya panjatkan rasa syukur ke hadrat Ilahi kerana dengan limpah kurnia dan izin-Nya, saya berjaya menyiapkan projek ini dalam tempoh yang ditetapkan. Segala cabaran dan rintangan yang dihadapi sepanjang penyelidikan ini dapat diatasi dengan kesabaran dan keazaman. Saya ingin merakamkan setinggi-tinggi penghargaan kepada penyelia saya, Prof. Dr. Salwani Abdullah, atas bimbingan dan dorongan sepanjang pelaksanaan projek ini.

Akhir sekali, penghargaan khas buat keluarga, ibu bapa dan adik-beradik serta rakan-rakan yang sentiasa memberi dorongan, motivasi dan semangat sepanjang saya menempuh perjalanan akademik ini. Sokongan mereka amat berharga dalam memastikan kejayaan usulan projek ini. Sekian, terima kasih.

7.0 RUJUKAN

American Diabetes Association. (2023). Gestational diabetes. *American Diabetes Association*. https://www.diabetes.org/diabetes/type-2/gestational-diabetes

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). *Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi random forest*. *JurnalInformatika dan Teknologi Kesehatan*, https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/1129

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121-167. https://doi.org/10.1023/A:1009715923555

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

Centers for Disease Control and Prevention (CDC). (2022). National diabetes statistics report, 2022. *Centers for Disease Control and Prevention*. https://www.cdc.gov/diabetes/data/statistics-report/index.html

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Copyright@FTSM UKM

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., et al. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*, 13, 148. https://doi.org/10.1186/s13098-021-00767-9

Hassan, A. S., Malaserene, I., & Leema, A. A. (2020). *Diabetes mellitus prediction using classification techniques*. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(5), 2795-2800. ISSN: 2278-3075.

Hotz, N. (2022). *What is CRISP DM?* https://www.datascience-pm.com/crisp-dm-2/

Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics*, 13(14), 2383. https://doi.org/10.3390/diagnostics13142383

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *International Journal of Computer Science and Technology Engineering*, 10(2), 112-120. https://doi.org/10.1016/j.icte.2021.02.004

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports, 10*(1), 11981. https://doi.org/10.1038/s41598-020-68771-z

L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. Doi:10.1023/A:1010933404324.

Muthu, J., & Suriya, S. (2023). Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm. *Journal of Trends in Computer Science and Smart Technology*, 5(2). https://doi.org/10.36548/jtcsst.2023.2.007

Rahman, T., Ferdous, J., & Rahman, R. (2020). *Sylhet Diabetes Hospital Dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/529/

Rayhan, I. & Ekadinata, (2023). Aplikasi Pendeteksi Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma K-Nearest Neighbor (K-NN) di Puskesmas Kenanga. *Diploma thesis*, Politeknik Manufaktur Negeri Bangka Belitung.

Copyright@FTSM UKM

Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science, 167*, 706–716. https://doi.org/10.1016/j.procs.2020.03.336

Xue, J., Min, F., & Ma, F. (2020). *Research on diabetes prediction method based on machine learning*. *Journal of Physics: Conference Series*, 1684(1), 012062. IOP Publishing. https://doi.org/10.1088/1742-6596/1684/1/012062

Zahirah Mohd Zain (A195516)
Prof. Dr. Salwani Abdullah
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia

Copyright@FTSM UKM