3rd E-proceeding SENRIABDI 2023
Seminar Nasional Hasil Riset dan Pengabdian kepada Masyarakat
Universitas Sahid Surakarta Vol. 3 Edisi Desember 2023 Hal. 664-671

Implementasi Support Vector Machine dalam Deteksi Diabetes Melalui Indikator Kesehatan

Nofrian Deny Hendrawan[1], Arif Saivul Affandi[2], Rizqullah Fani Fadhilrifat[3]

[1,2]Fakultas Teknologi Informasi/Universitas Merdeka Malang
e-mail: [1]nofrian.hendrawan@unmer.ac.id, [2]fandi@unmer.ac.id
[3] Fakultas Teknologi Informasi/Universitas Merdeka Malang
e-mail: 22083000044@student.unmer.ac.id

Abstrak

Diabetes, sebagai masalah kesehatan global, memerlukan deteksi awal untuk pengelolaan efektif. Penelitian ini mengembangkan model deteksi diabetes menggunakan algoritma Support Vector Machine (SVM) yang diintegrasikan dengan antarmuka Tkinter. Model ini melibatkan data dari survei "Behavioral Risk Factor Surveillance System" 2021, mencakup indikator seperti BMI, tekanan darah tinggi, kolesterol tinggi, dan umur. Model SVM, dilatih dan diuji dengan data ini, menunjukkan akurasi dalam memprediksi risiko diabetes. Proses pengembangan meliputi pra-pemrosesan data, pemilihan fitur, dan normalisasi. SVM dengan kernel linear dipilih berdasarkan karakteristik data. Performa model dievaluasi menggunakan subset data, dan akurasinya menunjukkan efektivitasnya dalam deteksi diabetes. Setelah validasi, model diintegrasikan ke antarmuka Tkinter yang memungkinkan pengguna memasukkan data kesehatan dan menerima prediksi risiko diabetes secara real-time. Hasil menunjukkan potensi SVM sebagai alat bantu deteksi dini diabetes. Penelitian ini menyarankan penerapan SVM dalam analisis data kesehatan sebagai pendekatan efektif untuk deteksi awal diabetes, dengan rekomendasi penelitian lebih lanjut menggunakan dataset yang lebih luas dan variabel tambahan untuk meningkatkan akurasi model. Implementasi teknologi ini berpotensi maju dalam pencegahan dan pengelolaan diabetes.

Kata Kunci: Diabetes, Support Vector Machine, Tkinter, Deteksi Diabetes, Indikator Kesehatan.

Abstract

Diabetes, as a global health issue, requires early detection for effective management. This study developed a diabetes detection model using the Support Vector Machine (SVM) algorithm integrated with the Tkinter interface. The model involves data from the 2021 "Behavioral Risk Factor Surveillance System" survey, including indicators such as BMI, high blood pressure, high cholesterol, and age. The

SVM model, trained and tested with this data, demonstrates accuracy in predicting diabetes risk. The development process includes data preprocessing, feature selection, and normalization. SVM with a linear kernel was chosen based on data characteristics. The model's performance was evaluated using a data subset, and its accuracy indicates its effectiveness in diabetes detection. After validation, the model was integrated into the Tkinter interface, allowing users to enter health data and receive real-time diabetes risk predictions. Results show the potential of SVM as an early detection tool for diabetes. This research suggests the application of SVM in health data analysis as an effective approach for early diabetes detection, with recommendations for further research using broader datasets and additional variables to enhance model accuracy. The implementation of this technology has the potential to advance in diabetes prevention and management.
Keywords: Diabetes, Support Vector Machine, Tkinter, Diabetes Detection, Health Indicators.

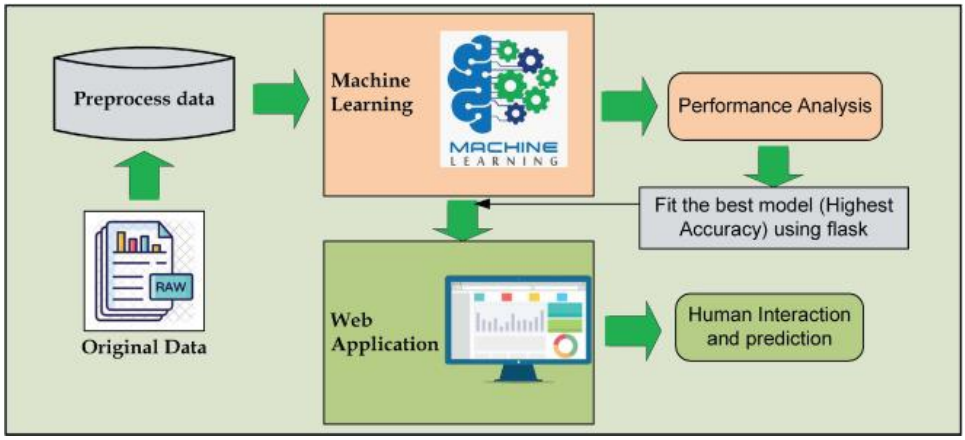
Pendahuluan

Diabetes, sebagai salah satu isu kesehatan global, menuntut inovasi terus-menerus dalam deteksi dan pengelolaan yang efektif. Penting untuk diingat bahwa diabetes adalah penyakit yang bisa dicegah, dan deteksi dini memiliki peran krusial dalam pencegahan. Dengan memiliki model yang dapat memprediksi risiko diabetes berdasarkan indikator kesehatan tertentu, individu dapat mengambil tindakan preventif yang tepat, seperti perubahan gaya hidup dan pengawasan lebih ketat terhadap kesehatan mereka. Selain itu, model ini juga memberikan dukungan bagi profesional kesehatan dalam memberikan intervensi yang lebih tepat waktu kepada pasien mereka. SVM, sebagai salah satu algoritma machine learning yang telah terbukti dalam banyak aplikasi, menawarkan keunggulan dalam hal kemampuan klasifikasi yang tinggi. Kemampuan SVM untuk menemukan pola kompleks dalam data membuatnya menjadi pilihan yang menarik dalam deteksi diabetes. Dalam dunia yang semakin terhubung, data kesehatan yang besar dan beragam menjadi sumber informasi yang berharga untuk menganalisis pola dan tren terkait diabetes. Penggunaan SVM sebagai algoritma utama dalam penelitian ini merupakan langkah maju dalam memanfaatkan potensi data kesehatan untuk mendeteksi diabetes secara lebih efisien. Dengan mengintegrasikan SVM dan antarmuka pengguna yang mudah digunakan, penelitian ini menciptakan sebuah alat bantu yang dapat memberikan manfaat nyata dalam upaya deteksi dini diabetes, yang pada akhirnya dapat membantu mengurangi dampak negatif penyakit ini terhadap individu dan masyarakat secara luas. Support Vector Machine (SVM) adalah salah satu metode machine learning yang digunakan dalam deteksi diabetes. SVM bekerja dengan menemukan suatu bidang pemisah yang disebut dengan hyperplane terbaik yang membagi data menjadi 2 kelas, yaitu kelas yang memiliki risiko diabetes dan kelas yang tidak memiliki risiko diabetes. Kemampuan SVM untuk menemukan pola kompleks dalam data membuatnya menjadi pilihan yang menarik dalam deteksi

diabetes. Dalam penelitian deteksi diabetes menggunakan SVM, data kesehatan yang besar dan beragam menjadi sumber informasi yang berharga untuk menganalisis pola dan tren terkait diabetes. Dengan mengintegrasikan SVM dan antarmuka pengguna yang mudah digunakan, penelitian ini menciptakan sebuah alat bantu yang dapat memberikan manfaat nyata dalam upaya deteksi dini diabetes, yang pada akhirnya dapat membantu mengurangi dampak negatif penyakit ini terhadap individu dan masyarakat secara luas [1], [2]. Kelebihan menggunakan Support Vector Machine (SVM) dalam deteksi diabetes antara lain adalah kemampuannya dalam menemukan pola kompleks dalam data, yang membuatnya menjadi pilihan yang menarik dalam deteksi diabetes. SVM juga menawarkan keunggulan dalam hal kemampuan klasifikasi yang tinggi, sehingga mampu memberikan hasil prediksi yang akurat. Selain itu, keluaran dari model SVM dapat mendiagnosis pasien yang menderita diabetes ataupun yang tidak menderita diabetes, sehingga dapat memberikan manfaat nyata dalam upaya deteksi dini diabetes [1]–[11].

Metode

Metode penelitian yang digunakan dalam analisis data diabetes adalah machine learning, khususnya dengan menggunakan algoritma Support Vector Machine (SVM) untuk membangun model prediktif deteksi diabetes berdasarkan variabel-variabel yang ada dalam dataset. Data yang diberikan sudah dalam format CSV yang siap digunakan dalam model SVM. Selain itu, metode penelitian lain yang dapat digunakan adalah statistik deskriptif, analisis regresi, dan analisis asosiasi untuk memahami karakteristik dasar dari dataset dan hubungan antara variabel-variabel dalam dataset. Dalam penelitian analisis data diabetes, terdapat beberapa metode yang digunakan untuk menjelajahi dataset ini. Salah satu metode yang paling umum digunakan adalah machine learning, khususnya dengan menggunakan algoritma Support Vector Machine (SVM).



Gambar.1 Metode pengolahan data Diabetes menggunakan *machine learning*

SVM adalah algoritma yang kuat untuk membangun model prediktif, dan dalam konteks ini, digunakan untuk mendeteksi diabetes berdasarkan variabel-variabel yang ada dalam dataset. Dataset yang diberikan telah disiapkan dalam

format CSV, sehingga mudah diimpor dan digunakan dalam model SVM. Ini memungkinkan kita untuk melatih model dengan data yang ada dan kemudian menggunakannya untuk memprediksi kemungkinan diabetes berdasarkan informasi yang ada dalam dataset tersebut. Selain dari machine learning, ada juga metode penelitian lain yang dapat digunakan dalam analisis data diabetes. Pertama, statistik deskriptif dapat digunakan untuk memberikan gambaran tentang karakteristik dasar dari dataset ini. Ini termasuk statistik seperti mean, median, dan deviasi standar untuk setiap variabel dalam dataset. Selanjutnya, analisis regresi dapat digunakan untuk memahami hubungan antara variabel-variabel dalam dataset. Ini membantu kita menentukan apakah ada hubungan yang signifikan antara faktor-faktor tertentu dan diabetes. Terakhir, analisis asosiasi dapat digunakan untuk mengidentifikasi pola atau keterkaitan antara variabel-variabel dalam dataset. Dengan menggabungkan berbagai metode penelitian ini, kita dapat mendapatkan pemahaman yang lebih dalam tentang dataset diabetes dan mengembangkan model prediktif yang akurat untuk deteksi diabetes. Ini adalah langkah penting dalam upaya mencegah dan mengelola kondisi ini secara lebih efektif.

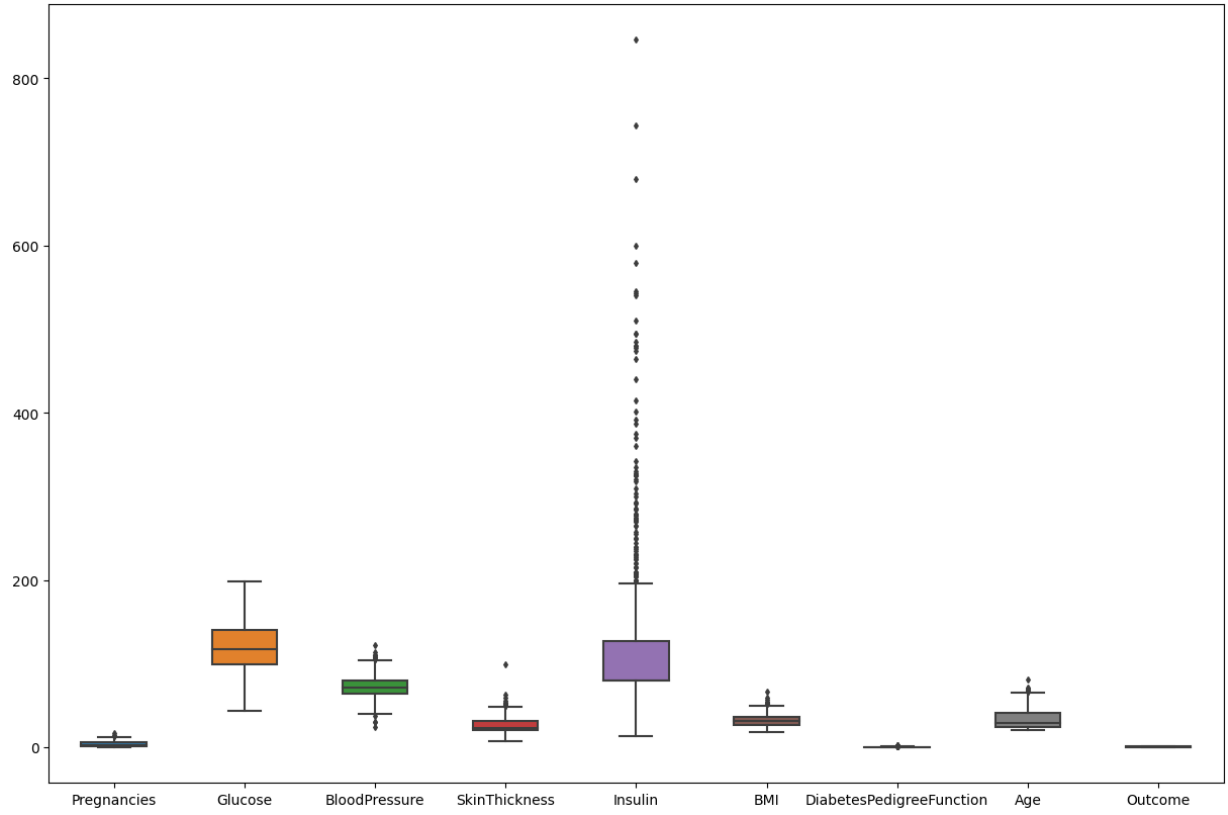
Hasil dan Pembahasan

Tabel 1. *Exploratory Data Analyst* data indicator Kesehatan

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Dalam dataset yang terdiri dari 768 observasi, masing-masing dari delapan variabel biomedis dan satu variabel hasil (Outcome) telah diukur dan diringkas melalui statistik deskriptif. Variabel-variabel tersebut termasuk jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah (BloodPressure), ketebalan kulit (SkinThickness), insulin (Insulin), indeks massa tubuh (BMI), fungsi keturunan diabetes (DiabetesPedigreeFunction), dan usia (Age). Variabel hasil (Outcome) tampaknya merupakan variabel biner yang menandakan keberadaan atau ketiadaan diabetes, dengan '1' mungkin menunjukkan keberadaan dan '0' untuk ketiadaan. Rata-rata (mean) dan simpangan baku (standard deviation, std) masing-masing variabel memberikan wawasan tentang pusat dan penyebaran data. Misalnya, rata-rata kehamilan adalah sekitar 3.85 dengan simpangan baku 3.37, menunjukkan variasi yang moderat dalam jumlah kehamilan di antara subjek. Demikian pula, rata-

rata kadar glukosa adalah sekitar 120.89 mg/dL, yang secara kasar sesuai dengan range yang dianggap normal atau pra-diabetes menurut beberapa standar klinis. Nilai minimum (min) dan maksimum (max) memberikan wawasan tentang rentang data. Beberapa variabel seperti tekanan darah dan ketebalan kulit memiliki nilai minimum 0, yang mungkin menunjukkan keberadaan nilai yang tidak tercatat atau kesalahan pengukuran. Kuartil pertama (25%) dan kuartil ketiga (75%) menandakan batas bawah dan atas dari rentang interkuartil (IQR), yang merupakan ukuran penyebaran data di tengah distribusi. Misalnya, 50% nilai insulin berada di bawah 30.5 IU/mL, sedangkan 25% teratas memiliki nilai yang jauh lebih tinggi, hingga 127.25 IU/mL, menunjukkan sebaran yang lebar untuk variabel ini. Median (50%) memberikan nilai tengah dataset, di mana setengah dari nilai lebih rendah dan setengah lebih tinggi dari titik ini. Misalnya, median usia adalah 29 tahun, yang menandakan bahwa setengah dari subjek lebih muda dari 29 dan setengah lebih tua. Dari statistik ini, kita dapat menyimpulkan bahwa terdapat variasi yang signifikan dalam beberapa variabel yang dapat mempengaruhi risiko diabetes. Variabilitas yang tinggi dalam kadar insulin menunjukkan adanya perbedaan yang signifikan dalam respons insulin antar individu. Penggunaan analisis lanjutan, seperti model regresi logistik atau analisis multivariat, dapat memberikan wawasan lebih lanjut tentang bagaimana variabel-variabel ini secara kolektif mempengaruhi risiko diabetes dalam populasi yang dipelajari.



Gambar.2 Boxplot data analyst

Boxplot yang digunakan untuk memvisualisasikan distribusi, pusat data (median), variasi (kuartil), dan outlier dari kumpulan data yang berhubungan dengan parameter kesehatan dan diabetes. Boxplot adalah alat yang sangat efektif untuk menyajikan ringkasan lima angka dari data: minimum, kuartil bawah (Q1), median (Q2), kuartil atas (Q3), dan maksimum. Analisis boxplot dimulai dengan memeriksa median, yang ditandai oleh garis dalam kotak dan mengindikasikan nilai tengah data. Jika garis median tidak berada di tengah kotak, ini menunjukkan skewness dalam distribusi data tersebut. Kuartil bawah dan atas, yang membentuk tepi kotak, menunjukkan rentang interkuartil (IQR). IQR adalah ukuran dispersi statistik dan memberikan informasi tentang variasi tengah data. Panjang 'kumis' dari boxplot, yang merepresentasikan nilai antara Q1-1.5 IQR dan Q3+1.5 IQR, menunjukkan variasi di luar kuartil tengah dan dapat mencakup nilai-nilai ekstrem yang masih dianggap tidak abnormal. Outlier adalah titik data yang berada di luar 'kumis' dan dapat dianggap sebagai anomali. Dalam konteks ini, variabel seperti Insulin menunjukkan sejumlah besar outlier, yang mungkin menandakan adanya variabilitas yang signifikan dalam respons insulin antar subjek atau potensi kesalahan dalam pengambilan atau pencatatan data. Dalam analisis lebih lanjut, kita akan mencari hubungan antara variabel-variabel tersebut dengan 'Outcome' untuk memahami faktor-faktor yang berkontribusi terhadap hasil akhir, yang mungkin berupa keberadaan atau tidaknya diabetes. Korelasi antara tingginya kadar glukosa dan BMI dengan hasil yang positif untuk diabetes akan sangat menarik untuk diteliti lebih lanjut. Penggunaan boxplot dalam analisis statistik ini penting karena memberikan visualisasi yang jelas dan mudah dipahami tentang struktur data. Namun, interpretasi yang akurat dari boxplot memerlukan pemahaman yang mendalam tentang konteks di mana data tersebut dikumpulkan dan karakteristik dari populasi yang dipelajari.

Simpulan

Dari analisis statistik deskriptif yang disajikan dalam dataset ini, kita bisa menyimpulkan bahwa terdapat variasi yang signifikan dalam faktor-faktor yang berpotensi berkaitan dengan diabetes mellitus. Berikut beberapa poin simpulan:

1. **Variabilitas Data**: Terdapat variasi yang besar dalam variabel seperti Insulin dan Glucose, yang mengindikasikan perbedaan biologis yang signifikan di antara subjek yang diukur.
2. **Potensi Outlier**: Nilai maksimum untuk Insulin sangat tinggi dibandingkan dengan kuartil ketiga, menunjukkan kemungkinan outlier yang dapat mempengaruhi analisis.
3. **Pengukuran dan Pencatatan Data**: Beberapa variabel menunjukkan nilai minimum 0, yang mungkin tidak realistis dalam konteks biomedis (misalnya, tekanan darah atau ketebalan kulit dengan nilai 0), menandakan bahwa data mungkin mengandung kesalahan pengukuran atau pencatatan.

4. **Konsistensi Data**: Variabel Outcome menunjukkan bahwa hampir 35% dari subjek di dataset memiliki diabetes, berdasarkan parameter yang digunakan dalam penelitian ini.

Saran

Berdasarkan simpulan tersebut, berikut beberapa saran untuk penelitian atau analisis selanjutnya:

1. **Pembersihan Data**: Lakukan pembersihan data untuk mengidentifikasi dan, jika memungkinkan, memperbaiki atau menghilangkan outlier dan nilai yang tidak realistis.
2. **Analisis Lebih Lanjut**: Gunakan metode statistik yang lebih canggih, seperti analisis regresi logistik, untuk menentukan faktor-faktor yang paling signifikan mempengaruhi risiko diabetes.
3. **Uji Klinis**: Melakukan uji klinis untuk mengkonfirmasi temuan dari data. Hal ini akan membantu dalam mengembangkan profil risiko yang lebih akurat untuk diabetes.
4. **Edukasi dan Pencegahan**: Fokus pada edukasi pasien mengenai faktor risiko seperti BMI dan kadar glukosa yang tinggi sebagai bagian dari program pencegahan diabetes.
5. **Kebijakan Kesehatan**: Kembangkan kebijakan kesehatan yang didasarkan pada temuan untuk mengatasi faktor-faktor risiko diabetes yang paling berpengaruh.

Dengan pendekatan yang sistematis dan analitis, kita dapat lebih mengerti tentang bagaimana mengelola dan mungkin mencegah diabetes mellitus dalam populasi.

Daftar Pustaka

[1] A. W. Mucholladin, F. A. Bachtiar, and ..., "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *… Teknologi Informasi dan …*, 2021, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8573

[2] C. Aldama and M. Nasir, "KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE SUPPORT VECTOR MACHINE PADA RUMAH SAKIT UMUM PRABUMULIH," *Jurnal Ilmiah Betrik*, 2023, [Online]. Available: https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/117

[3] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetol Metab Syndr*, vol. 13, no. 1, 2021, doi: 10.1186/s13098-021-00767-9.

[4] J. Jendle, K. Rinnert, A. Westman, and ..., "Pilots and diabetes technology: functional health," *Journal of diabetes …*, 2017, doi: 10.1177/1932296816680510.

[5] R. Singla, A. Singla, Y. Gupta, and S. Kalra, "Artificial intelligence/machine learning in diabetes care," *Indian J Endocrinol Metab*, vol. 23, no. 4, pp. 495–497, 2019, doi: 10.4103/ijem.IJEM_228_19.

[6] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput Sci*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[7] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front Genet*, vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[8] O. Llaha and A. Rista, "Prediction and detection of diabetes using machine learning," *CEUR Workshop Proc*, vol. 2872, pp. 94–102, 2021.

[9] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[10] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and ..., "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9076634/

[11] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artif Intell Med*, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336571830681X