

OAJIS

Open Access
Journal of
Information
Systems

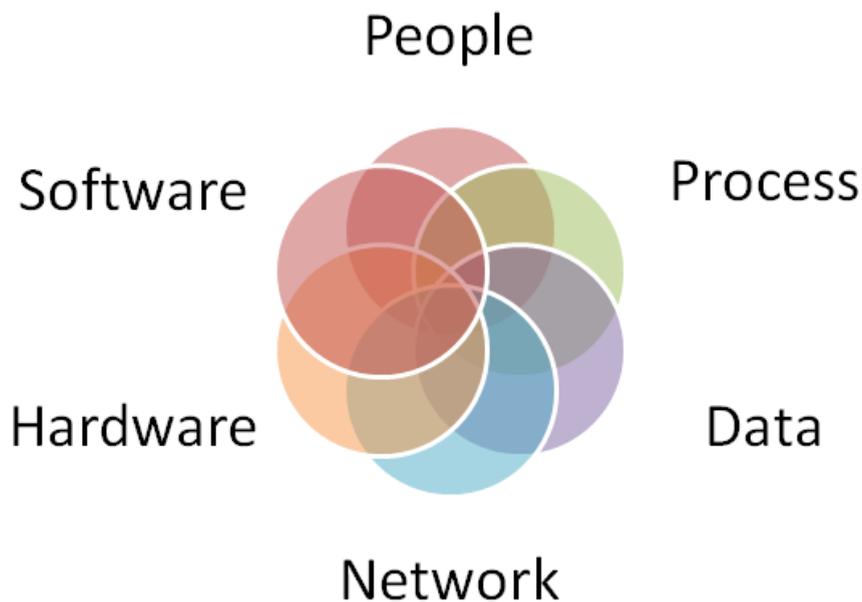
is.its.ac.id/pubs/oajis/

ISSN 1979-3979



jurnal sisfo

Inspirasi Profesional Sistem Informasi

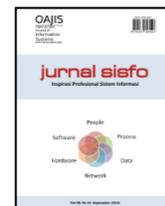


OAJIS

Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 11 No. 01 (2024)



Pimpinan Redaksi

Sholiq

Dewan Redaksi

Reny Nadlifatin

Mudjahidin

Tining Haryanti

Faizal Mahananto

Rizal Risnanda Utama

Radityo Prasetyanto Wibowo

Tata Pelaksana Usaha

Heppy Nuryanti

Sekretariat

Departemen Sistem Informasi – Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember (ITS) – Surabaya

Telp. 031-5999944 Fax. 031-5964965

Email: editor@jurnalsisfo.org

Website: <http://jurnalsisfo.org>

Jurnal SISFO juga dipublikasikan di *Open Access Journal of Information Systems* (OAJIS)

Website: <http://is.its.ac.id/pubs/oajis/index.php>



Mitra Bestari

Prof. Mahendrawathi ER., S.T., M.Sc, Ph.D. (Institut Teknologi Sepuluh Nopember)

Prof. Nur Aini Rakhmawati, S.Kom., M.Sc.Eng., Ph.D. (Institut Teknologi Sepuluh Nopember)

Dr. Muhammad Ainul Yaqin, S.Si., M.Kom. (Universitas Islam Negeri Maulana Malik Ibrahim)

Dr. Apol Pribadi Subriadi, S.T., M.T. (Institut Teknologi Sepuluh Nopember)

Dr. Bambang Setiawan, S.Kom., M.T. (Institut Teknologi Sepuluh Nopember)

Dr. Wiwik Anggraeni, S.Si., M.Kom. (Institut Teknologi Sepuluh Nopember)

Dr. Indra Waspada, S.T., M.T.I. (Universitas Diponegoro)

M. Amirul Haq, S.T., M.Sc., Ph.D. (Universitas Muhammadiyah Surabaya)

Ashr Hafizh Tantri, S.Kom., M.Kom. (Universitas Muhammadiyah Surabaya)

Doddy Ridwandono, S.Kom., M.Kom. (Universitas Pembangunan Nasional “Veteran Jawa Timur”)

Dhiani Tresna Absari, S.T., M.Kom. (Universitas Surabaya)



Daftar Isi

Implementasi Enterprise Resource Planning Odoo 10 Pada PT XYZ Dengan Metode Action Design Research <i>Diajeng Ciptaning Ayu, Mahendrawathi ER, Ghifary Muhammad</i>	1
Penggunaan Explainable Machine Learning untuk Prediksi Pasien Diabetes <i>Muhammad Reza Pahlawan</i>	11
Teori dan Penerapan Backpropagation Neural Networks untuk Internet of Things: Online dan Batch Mode <i>Anisa Dzulkarnain, Mochamad Nizar Palefi Ma'ady</i>	25
Economic Impact of IT-based Business Process Management Improvement Projects: A Systematic Literature Review <i>Nungky Amalia Imran, Muhammad Febrilian Dwi Syahputra</i>	39
Faktor-Faktor Penentu Adopsi Game PUBG Mobile di Kalangan Generasi Z Menggunakan Model Extended TPB: Studi Kasus di Provinsi Papua Barat <i>Ni Komang A S Devi, Dedi I Inan, Ratna Julia</i>	57
Rancang Bangun Sistem Informasi Rumah Sakit Menggunakan Aplikasi AppSheet (Studi Kasus : Rumah Sakit Khusus Ibu dan Anak Permata Bunda Yogyakarta) <i>Abdullah Gymnastiar Abdoerrani, Achmad Holil Noor Ali, Felicia Evelina Soetjipto, Erika Cahya Ningtyas</i>	71
Evaluasi Kematangan Proses Rekayasa Kebutuhan Dengan Mengacu Model REPM (Requirements Engineering Process Maturity) dan CMMI (Capability Maturity Model Integration) <i>Carissa Cindy Febiana, Apol Pribadi Subriadi, Nabila Kumala Gantari, Syamil Rizqy Rayvianda Agil</i>	87

OAJIS

Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 11 No. 1 (2024)



Halaman ini sengaja dikosongkan



Penggunaan Explainable Machine Learning untuk Prediksi Pasien Diabetes

Muhammad Reza Pahlawan*

Jurusan Akuntansi, Sekolah Tinggi Ilmu Ekonomi Indonesia (STIESIA) Surabaya

Abstract

Diabetes has been recorded already affects 424 million people and it was predicted to increase to more than 600 million by 2045. This encourages many researchers to develop a prediction model that can help minimize this phenomenon by identifying the factors that can improve the probability of having diabetes. However, most researchers only focus on increasing accuracy values because we cannot use the same evaluation metrics for all case studies. This research aims to build a model that has high sensitivity and is easy to interpret through explainable machine learning. The algorithms used are Decision Tree, Logistic Regression, Random Forest, and XGBoost. This research also experiments with resampling data to minimize the risk of bias caused by imbalanced class proportions. The research results show that XGBoost can produce a sensitivity value of 73%. The application of explainable machine learning also shows that the factors that increase diabetes are glucose, Body Mass Index (BMI), and age. The model also indicates that high levels of glucose in the body can improve the probability of having diabetes.

Keywords: Diabetes, Explainable Machine Learning, Glucose, Sensitivity

Abstrak

Diabetes tercatat sudah menyerang 424 juta orang dan diprediksi meningkat hingga lebih dari 600 juta pada 2045. Hal ini mendorong banyak peneliti untuk membangun model prediksi yang dapat membantu meminimalisir hal ini dengan mengidentifikasi faktor yang paling berpengaruh terjadinya diabetes. Namun, mayoritas peneliti hanya berfokus pada peningkatan nilai akurasi karena kita tidak bisa menyeragamkan penggunaan metrik evaluasi yang sama untuk semua studi kasus. Penelitian ini bertujuan membangun model yang memiliki sensitivitas yang tinggi serta mudah diinterpretasikan melalui *explainable machine learning*. Model yang digunakan adalah *Decision Tree*, *Logistic Regression*, *Random Forest*, dan *XGBoost*. Penelitian ini juga bereksperimen *data resampling* untuk meminimalisir bias akibat ketidakseimbangan jumlah target kelas (*imbalanced classification*). Hasil penelitian menunjukkan *XGBoost* mampu menghasilkan nilai sensitivitas 73%. Penerapan *explainable machine learning* juga menunjukkan bahwa faktor yang memperbesar terkena diabetes adalah glukosa, *Body Mass Index* (BMI), dan usia. Model juga mengindikasikan tingginya kadar glukosa dalam tubuh akan memperbesar peluang terkena diabetes.

Kata kunci: Diabetes, Explainable Machine Learning, Glukosa, Sensitivitas

*Corresponding Author

Email address: mrezapa@stiesia.ac.id (Muhammad Reza Pahlawan)

<https://doi.org/10.24089/j.sisfo.2024.05.002>

© 2024 Jurnal SISFO.

Histori Artikel: Disubmit 15-03-2024; Direvisi 23-04-2024; Diterima 29-04-2024; Tersedia online 31-05-2024

1. Pendahuluan

Diabetes adalah suatu kondisi kronis yang dipengaruhi oleh kadar gula di dalam darah atau biasa disebut glukosa. Apabila dibiarkan, kondisi ini dapat menyebabkan gangguan pada ginjal, mata, dan jantung [1]. Diabetes menjadi perhatian utama seiring meningkatnya jumlah kasus belakangan ini. Tercatat terdapat 424 juta kasus diabetes pada tahun 2017 dan diprediksi meningkat hingga 628,6 juta pada 2045 [2]. *Internasional Diabetes Federation* (IDF) *Diabetes Atlas 2021* menuturkan bahwa 6,7 juta penderita diabetes meninggal pada 2021 [3]. Hal ini mendorong beberapa peneliti untuk mengembangkan model klasifikasi yang mampu meminimalisir kasus diabetes. Secara garis besar model klasifikasi bertugas melakukan prediksi suatu obyek data untuk mengetahui obyek tersebut masuk ke dalam suatu kelas secara tepat [4], [5].

Beberapa model *machine learning* telah diterapkan untuk mengembangkan sebuah model prediksi pada studi kasus diabetes, diantaranya *Multilayer Perceptron* (MLP), *Logistic Regression*, dan *K-NearestNeighbor* (KNN) [6], [7], [8]. Walaupun begitu, penerapan satu jenis model dianggap masih belum cukup sehingga beberapa peneliti menggunakan beberapa model sekaligus melalui beberapa teknik seperti *bagging*, *voting*, dan *stacking*. Penggunaan teknik *stacking* dengan menggunakan KNN sebagai *meta-learner* mampu mengungguli kinerja model tunggal dengan akurasi sebesar 90% [9]. Penelitian lain menggunakan teknik *boosting* menggunakan *XGBoost* yang mampu menghasilkan akurasi sekitar 76% [10]. Adapun peneliti lain menggunakan *deep learning* seperti *Convolutional Neural Network* (CNN) dengan akurasi sebesar 76% [11]. Mayoritas peneliti hanya berfokus pada peningkatan akurasi di mana metrik ini masih belum mampu menggambarkan seberapa baik kinerja terutama pada bidang kesehatan. Oleh karena itu, pemilihan metrik pengukuran yang tepat perlu menjadi perhatian utama untuk meminimalisir penyampaian informasi yang kurang relevan.

Seperti yang kita tahu bahwa pengukuran kinerja model evaluasi dapat melalui akurasi, *precision*, *recall*, dan *auc score*. Pengukuran tersebut pasti memiliki makna tersendiri sehingga tidak bisa hanya berfokus pada satu pengukuran untuk semua studi kasus. Hal ini juga didukung oleh penelitian terdahulu yang menyebutkan bahwa studi kasus yang berbeda membutuhkan metrik pengukuran yang berbeda pula [12]. Beberapa penelitian di bidang kesehatan menggunakan pengukuran *recall* atau biasa disebut *sensitivity* untuk mengukur kinerja model yang dihasilkan. Pengukuran *sensitivity* telah digunakan di klasifikasi gagal jantung dan kanker prostat [13], [14]. Penelitian lain juga menuturkan bahwa pentingnya meningkatkan sensitivitas model walaupun dalam konteks klasifikasi kanker payudara [15]. Sejauh penelusuran studi literatur, saat ini masih minim penelitian di bidang penyakit diabetes yang berfokus pada peningkatan *sensitivity* pada model. Model yang memiliki kinerja yang tinggi tentu sangat membantu dalam justifikasi terkait kehandalan model dalam melakukan prediksi.

Mayoritas peneliti hanya berfokus membangun model dengan kinerja tinggi tanpa menyajikan interpretasi model tersebut sehingga model yang dibangun tidak bisa sepenuhnya digunakan sebagai sarana pengambilan keputusan. Pada dasarnya, penerapan *data mining* di mana dalam hal ini adalah membangun model *machine learning* seharusnya mampu digunakan untuk hal tersebut. Hal ini dapat dilakukan dengan sebuah algoritma dalam penyelesaian masalah untuk relasi data [16]. Selain itu, tidak adanya interpretasi model biasanya tidak dipercaya oleh para pakar [10]. Oleh karena itu, penelitian ini akan berfokus pada membangun model klasifikasi dengan kinerja yang tinggi yang didukung oleh pemaparan intepretasi model.

Penggunaan model yang mudah diinterpretasikan melalui *explainable machine learning* sangat penting mengingat banyaknya yang dipertaruhkan dalam bidang kesehatan [10]. Tidak dapat dipungkiri bahwa masih banyak orang yang memiliki pengetahuan atau pemahaman yang berbeda terkait diabetes. Penelitian

ini dapat menjawab hal tersebut di mana dapat memberikan informasi yang lebih komprehensif terkait diabetes seperti variabel atau faktor apa saja yang mempengaruhi diabetes. Penelitian ini juga berfokus pada peningkatan kinerja model yang lebih relevan seperti peningkatan tingkan sensitivitas model terhadap ketepatan prediksi suatu kelas tertentu melalui pengukuran *recall*. Harapannya, model yang dibangun mampu memprediksi pasien terdiagnosa diabetes lebih akurat.

2. Tinjauan Pustaka/Penelitian Sebelumnya

2.1 Diabetes

Diabetes memiliki beberapa tipe, yaitu diabetes tipe 1, tipe 2, dan *gestational*. Diantara ketiga tipe tersebut, tipe 2 merupakan yang paling sering terjadi dan biasanya berkaitan dengan gaya hidup, kebiasaan diet, dan obesitas [10]. Pankreas gagal memproduksi insulin dengan cukup sehingga tubuh rentan terhadap insulin [17]. Diabetes telah terbukti meningkatkan kejadian penyakit kronis dan akut di hampir semua wilayah dalam beberapa dekade terakhir. *World Health Orgnaization* (WHO) memprediksi bahwa penyakit ini dapat menjadi penyebab kematian tertinggi ketujuh pada tahun 2030 [18].

2.2 Explainable Machine Learning

Model *machine learning* diharapkan tidak hanya menghasilkan kinerja setinggi mungkin, namun mampu menghasilkan sebuah informasi berdasarkan model yang dihasilkan. *Explainable machine learning* dikembangkan untuk memudahkan proses identifikasi isu atau informasi terbaru pada suatu studi kasus. Hal ini juga dapat membantu dalam menilai efektivitas suatu kebijakan berdasarkan penjelasan model tersebut [19]. Tanpa adanya proses pemaparan interpretasi model, *domain expert* biasanya juga tidak dapat mempercayai hasil prediksi model walaupun memiliki kinerja yang mengesankan [10]. Model *machine learning* berbasis pohon seperti *decision tree* biasanya sudah menyediakan cara interpretasi model yang memadai. Namun, untuk model yang lebih kompleks biasanya memerlukan analisis tambahan seperti *Shapley Additive Explanations* (SHAP) [20]. Analisis ini memudahkan memperoleh informasi bagaimana interaksi antar fitur pada model *machine learning* khususnya model berbasis pohon [21]. Oleh karena itu, penelitian ini akan berfokus menggunakan model berbasis pohon seperti *Decision Tree*, *Random Forest*, dan *XGBoost* untuk mempermudah proses interpretasi model.

2.3 Metriks Evaluasi Model Klasifikasi

Pengukuran kinerja model berbasis *confusion matrix* berperan penting mengenai bagaimana melihat efektivitas model *machine learning* [22]. Metriks ini menjadi dasar dalam memperoleh kinerja model klasifikasi seperti akurasi, presisi, dan *recall*. Persamaan (1-3) merupakan rumus perhitungan tiap metriks.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

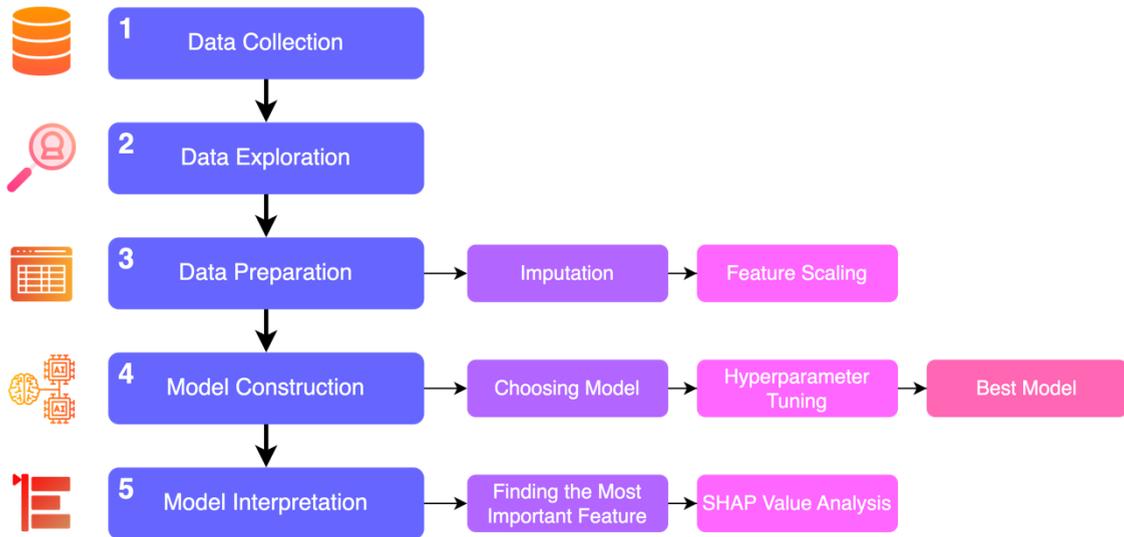
$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \quad (3)$$

Penelitian terdahulu banyak yang menganggap sama semua metriks ini apabila melihat dari pengukuran yang dipakai. Rata-rata hanya berfokus pada peningkatan akurasi tanpa memerhatikan metriks yang lain. Studi kasus yang berbeda membutuhkan metriks pengukuran yang berbeda pula. Nilai *precision* sangat penting ketika *false positive* tidak ditoleransi karena akan berdampak pada biaya. Sebagai contoh adalah studi kasus pesan *spam*, kita tidak bisa kehilangan *email* penting sehingga kita meminimalisir pesan yang bukan *spam* namun diprediksi sebagai *spam*. Namun, di bidang kesehatan *false positive* mengindikasikan alih-alih

kehilangan pasien, penyelidikan lebih lanjut diperlukan untuk menentukan pengobatan terbaik secara tepat waktu [12]. Oleh karena itu, penelitian ini menggunakan nilai *recall* atau *sensitivity* karena ingin berfokus pada meminimalisir nilai *false negative* di mana pasien sebenarnya mengidap diabetes namun diprediksi tidak mengidap. Hal ini tentu dapat membantu pelayanan kesehatan yang lebih efisien dalam memberikan layanan medis terhadap pasien yang dicurigai terkena diabetes.

3. Metodologi

Metodologi yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1 berikut:



Gambar 1. Metodologi Penelitian

3.1 Data Collection

Tahap ini adalah proses pengumpulan data untuk obyek penelitian. Data yang digunakan dalam penelitian ini adalah berasal dari situs *Kaggle* yang dapat diakses [di link berikut](#). Data berisi beberapa fitur atau variabel seperti jumlah kehamilan, tingkat glukosa, BMI, umur, dan kadar insulin dalam tubuh. Data dihimpun dari *National Institute of Diabetes and Digestive and Kidney Diseases* di India.

3.2 Data Exploration

Data exploration merupakan upaya untuk memperdalam karakteristik suatu data. Aktivitas ini akan membantu kita dalam menentukan preproses yang diperlukan. Tahap ini terdiri atas beberapa aktivitas seperti pemeriksaan terkait korelasi antar fitur, distribusi data, dan keberadaan data pecilan. Selain itu, eksplorasi data juga dilakukan secara numerik dan grafik.

3.3 Data Preparation

Tahap ini berfokus pada mempersiapkan data menuju aktivitas membangun model (*modelling*). Secara garis besar, tahap ini terdiri dari beberapa aktivitas seperti mengatasi keberadaan data yang hilang (*missing value*), data duplikat, dan penyeragaman rentang nilai tiap fitur yang biasa disebut *feature scaling*. Penelitian ini mencoba menerapkan teknik *data preparation* yang lebih komprehensif daripada penelitian lain. Penelitian

lain biasanya tidak menyebutkan alasan menggunakan suatu teknik tertentu sehingga berpotensi menimbulkan *misleading* informasi terkait penggunaan teknik tertentu.

Penanganan *missing value* akan menyesuaikan karakteristik data yang dimiliki. Penanganan *missing value* dapat berupa menghilangkan baris data yang memiliki *missing value* atau mengganti dengan suatu nilai seperti nilai rata-rata atau median. *Dataset* ini memiliki suatu nilai tertentu yang dikatakan sebagai *missing value*. Nilai 0 pada fitur *Pregnancy*, *BloodPressure*, *Skin Thickness*, Insulin, dan BMI dianggap sebagai *missing value* karena tidak wajar [2]. Teknik *feature scaling* yang digunakan pada penelitian ini juga menyesuaikan pada karakteristik data yang dimiliki. Apabila data yang dimiliki tidak berdistribusi normal dan memiliki banyak data pecilan, maka teknik yang bisa digunakan adalah *RobustScaler*. Teknik ini dinilai tidak rentan terhadap data pecilan di mana hal ini menjadi kelemahan pada teknik *MinMax*. Selain itu, *StandardScaler* dianggap rentan terhadap data yang tidak berdistribusi normal. Pengujian apakah data berdistribusi normal atau tidak akan menggunakan pengujian *Kolmogorov-Smirnov* karena cocok digunakan pada sampel data yang besar (di atas 50 baris data) [23]. Penelitian ini juga melakukan eksperimen terkait pengaruh *resampling* data untuk memastikan model yang dibangun tidak bias terhadap suatu kelas tertentu.

3.4 Modelling

Tahap ini merupakan membangun model prediksi untuk memprediksi apakah suatu obyek data termasuk ke dalam diabetes atau tidak. Bisa dikatakan bahwa tahap ini kita melatih model untuk bisa membedakan apakah suatu pasien mengidap diabetes atau tidak. Proses ini diawali dengan pemilihan model dasar menggunakan fungsi dasar yang disediakan oleh *library sklearn* pada bahasa pemrograman *Python*. Penelitian ini akan mencoba menggunakan model *Decision Tree*, *Random Forest*, *XGBoost*, dan *Logistic Regression* yang mewakili beberapa teknik model *machine learning* seperti penggunaan model tunggal (*Decision Tree* dan *Logistic Regression*), teknik *bagging* yang diwakilkan *Random Forest*, dan teknik *boosting* yang diwakilkan oleh *XGBoost*. Model terbaik akan dipilih untuk eksperimen selanjutnya, yaitu eksperimen *resampling* data dan *hyperparameter tuning*.

Eksperimen *resampling* dilakukan untuk memastikan bahwa model yang dibangun tidak menjadi bias terhadap kelas atau kelompok tertentu. Penelitian ini akan membandingkan beberapa teknik seperti *undersampling*, *oversampling*, dan *SMOTE*. Proses *hyperparameter tuning* digunakan untuk memperoleh skenario parameter model terbaik sehingga memperoleh model dengan kinerja yang mengesankan. Proses ini akan mencoba beberapa skenario parameter secara *gridsearch* dengan mekanisme *5-fold cross validation*.

3.5 Model Interpretation

Tahap ini adalah proses menginterpretasikan model *machine learning* yang dihasilkan. Tahap ini dapat membantu menghindari manipulasi dari suatu sistem, khususnya penerapan model *machine learning* dalam melakukan prediksi, karena bersifat transparan serta dapat mendeteksi bias yang terjadi pada proses pengambilan keputusannya. Selain itu, tahap ini juga merupakan proses pemanfaatan dari model *machine learning* untuk mengekstraksi pengetahuan yang sesuai dengan hubungan *domain-expertise* yang terdapat pada data. Terdapat beberapa upaya untuk menginterpretasikan model diantaranya identifikasi fitur yang penting dan analisis SHAP. Analisis SHAP mampu mengidentifikasi bagaimana dampak suatu fitur terhadap kinerja model [10]. Harapannya, kita dapat meningkatkan pemahaman terkait proses prediksi penyakit diabetes berdasarkan model *machine learning* yang digunakan.

4. Hasil dan Pembahasan

Bab ini akan membahas mengenai hasil implementasi tahapan metodologi penelitian seperti *Data Collection*, *Data Exploration*, *Data Preprocessing*, *Modelling*, dan *Model Interpretation*. Eksperimen dilakukan menggunakan bahasa pemrograman *Python* melalui aplikasi berbasis website *Google Colab*.

4.1 Data Collection

Penelitian ini menggunakan data pasien diabetes yang telah dihimpun oleh *National Institute of Diabetes and Digestive and Kidney Diseases*, India. Tabel 1 merupakan sampel data yang digunakan dalam penelitian. Data yang digunakan terdiri dari 768 baris data yang terdiri dari delapan variabel independen dan satu variabel dependen. Variabel dependen diwakilkan oleh fitur '*Outcome*' yang terdiri dari nilai 0 dan 1. Nilai 0 merepresentasikan pasien tersebut tidak terdiagnosis diabetes sementara 1 adalah pasien yang terdiagnosis diabetes. Secara khusus, data yang dihimpun adalah pasien wanita.

Tabel 1. *Dataset*

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33,6	0,627	50	1
1	85	66	29	0	26,6	0,351	31	0
8	183	64	0	0	23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1

4.2 Data Exploration

Tabel 2 merupakan deskripsi tiap fitur yang terdapat pada *dataset*. Semua fitur pada *dataset* ini memiliki satu jenis tipe data, yaitu tipe data kontinyu sehingga tidak memerlukan perlakuan khusus baik dalam eksplorasi data maupun *data preprocessing*. Nilai yang tidak wajar seperti keberadaan nilai 0 pada fitur *Pregnancies*, *Glucose*, *SkinThickness*, *BMI*, dan *Insulin* diganti tipe datanya menjadi *np.nan* yang merupakan representasi nilai tersebut tidak memiliki nilai. Fitur *Insulin* dan *SkinThickness* memiliki persentase *missing value* yang cukup besar, yaitu di atas 25%. Gambar 2 menunjukkan bagaimana korelasi dari sisi *missing value* tiap fitur. Terlihat pada gambar tersebut bahwa kedua fitur tersebut memiliki korelasi kuat positif. Hal ini mengindikasikan bahwa *missing value* pada fitur *Insulin* akan diikuti *missing value* pada fitur *SkinThickness*. Begitu sebaliknya, keberadaan *missing value* pada *SkinThickness* akan diikuti *missing value* pada *Insulin*.

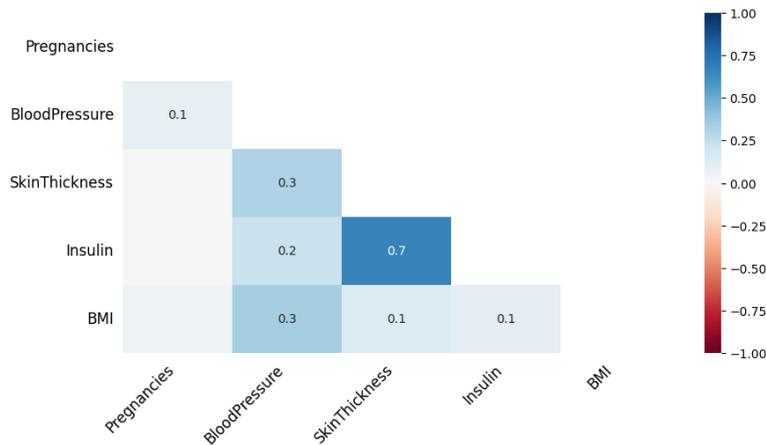
Gambar 3 merupakan korelasi antar fitur terhadap kelas target. Terlihat pada gambar tersebut bahwa fitur *Pregnancies* memiliki korelasi yang cukup kuat terhadap *Outcome*. Hal ini mengindikasikan bahwa semakin sering seorang wanita hamil, maka semakin besar pula peluang untuk mengidap diabetes. Fitur *Glucose* dan *BloodPressure* juga memiliki korelasi kuat positif di mana hal ini wajar karena penyebab utama diabetes adalah kadar glukosa dan tekanan darah diastolik pada tubuh. Peningkatan kedua kadar ini akan memperbesar peluang seseorang untuk terkena diabetes.

Analisis selanjutnya adalah menguji distribusi data yang dimiliki, khususnya pada variabel independen. Gambar 4 (a) merupakan distribusi data tiap fitur. Terlihat pada gambar tersebut bahwa tidak ada distribusi yang berbentuk seperti lonceng tepat di tengah yang mengindikasikan data tersebut berdistribusi normal. Selain penilaian berdasarkan gambar, pengujian juga dilakukan secara statistika melalui pengujian

Kolmogorov-Smirnov. Hasil pengujian mengindikasikan tidak ada data yang berdistribusi normal karena tidak ada yang memiliki *p-value* di atas 0,05 seperti terlihat pada Tabel 3.

Tabel 2. Deskripsi Fitur pada *Dataset*

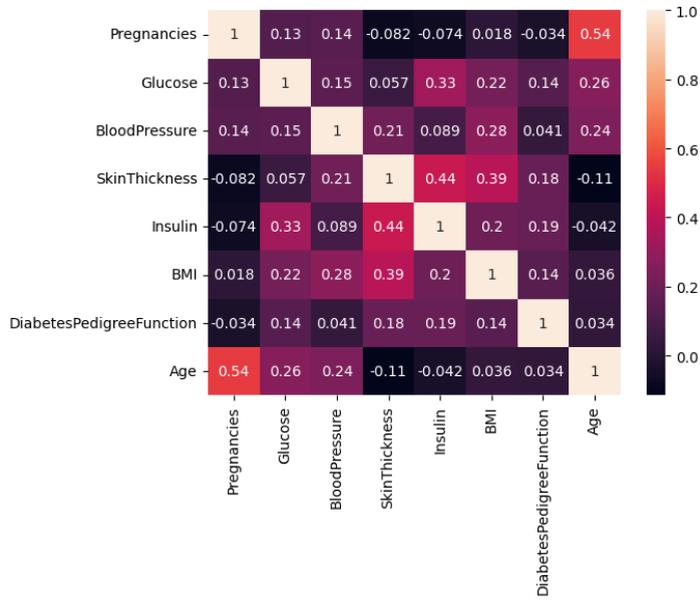
Variabel	Deskripsi	Persentase Data Missing Value (dalam %)
Pregnancies	Jumlah kehamilan pasien	12,33
Glucose	Kadar glukosa	
Blood Pressure	Tekanan darah diastolik dalam satuan mm/Hg	3,88
Skin Thickness	Ketebalan lipatan kulit trisep (dalam mm). Nilai yang digunakan untuk memperkirakan lemak tubuh	25,22
Insulin	Kadar insulin dalam tubuh dalam satuan $\mu\text{U/ml}$	41,56
BMI	<i>Body Mass Index</i> di mana dihitung berdasarkan berat badan (dalam kg) per tinggi badan (dalam satuan m^2)	1,22
Diabetes Pedigree Function	Indikator riwayat diabetes dalam keluarga	0
Age	Umur (dalam tahun)	0
Outcome	Bernilai 0 → mengindikasikan tidak mengidap diabetes Bernilai 1 → mengindikasikan mengidap diabetes	0



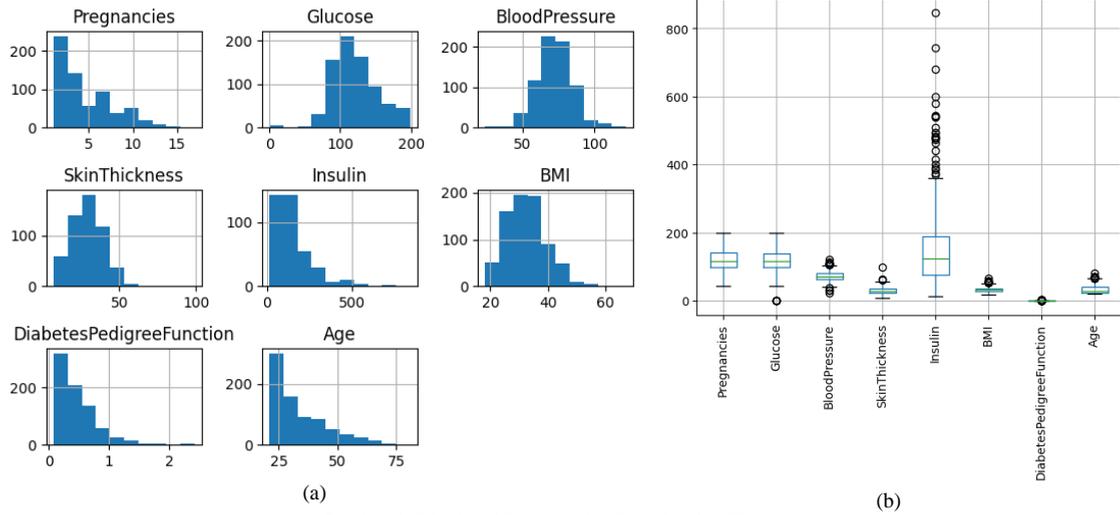
Gambar 2. Korelasi Keberadaan Missing Values antar Fitur

Berdasarkan pemeriksaan data pecilan seperti yang terlihat pada Gambar 4 (b), hanya fitur *Pregnancies* yang tidak memiliki data pecilan. Fitur *Insulin* merupakan fitur yang memiliki paling banyak data pecilan. Banyaknya data pecilan pada tiap fitur menjadi acuan dalam menentukan teknik *feature scaling* yang dilakukan. Penelitian ini akan menggunakan *RobustScaler* karena lebih kuat dalam menangani data pecilan dan dapat digunakan pada data tidak berdistribusi normal.

Gambar 5 merupakan perbandingan jumlah pasien antara yang tidak mengidap diabetes dan pengidap diabetes. Diketahui bahwa proporsi terbesar didominasi oleh pasien yang tidak mengidap diabetes sebesar 65,1%. Hal ini tentu perlu menjadi pertimbangan khusus dalam membangun model agar model yang dibangun tidak menjadi bias karena data latih yang digunakan didominasi oleh kelas tertentu.



Gambar 3. Korelasi antar Fitur



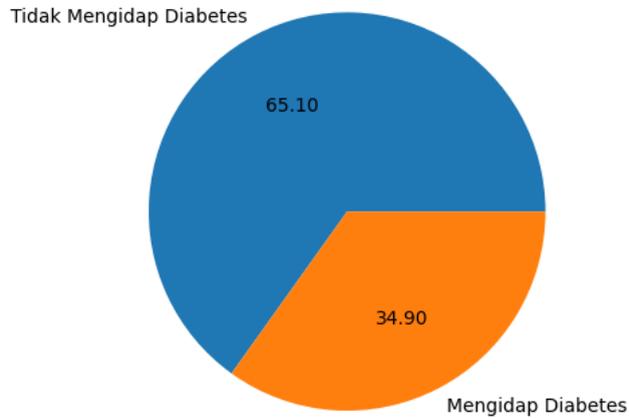
Gambar 4. Distribusi Data (a) dan Boxplot tiap Fitur (b)

Tabel 3 Hasil Pengujian Kolmogorov-Smirnov

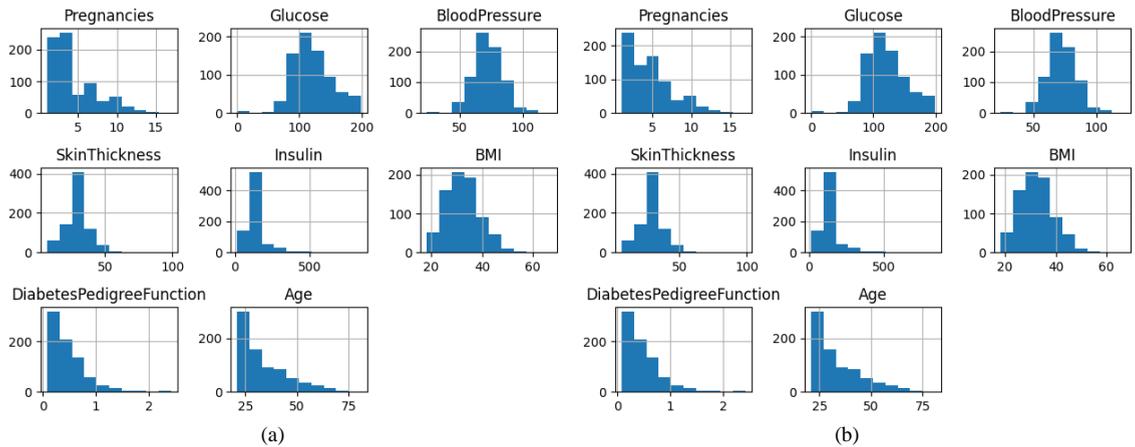
Variabel	p-value
Pregnancies	0,0
Glucose	0,0
Blood Pressure	0,0
Skin Thickness	0,0
Insulin	$3,83 \times 10^{-188}$
BMI	0,0
Diabetes Pedigree Function	$1,35 \times 10^{-203}$
Age	0,0

4.3 Data Preprocessing

Tahap selanjutnya adalah melakukan persiapan data sebelum dilakukan eksperimen permodelan. Persiapan data yang pertama dilakukan adalah menangani *missing value*. Seperti yang terlihat pada Tabel 2, diketahui bahwa beberapa fitur memiliki *missing value*. Penelitian ini menggantikan baris data yang memiliki *missing value* dengan nilai median fitur tersebut. Gambar 6(a) menunjukkan bahwa penggunaan nilai median untuk mengisi *missing value* relatif tidak mengubah distribusi data secara signifikan apabila dibandingkan dengan distribusi data sebelum dilakukan pengisian *missing value*.



Gambar 5. Perbandingan Jumlah Pasien



Gambar 6. Perbedaan Pengaruh Penggunaan Teknik Imputasi menggunakan median (a) dan mean (b) terhadap Distribusi Data

4.4 Modelling

Tahap yang pertama dilakukan adalah pemilihan model untuk eksperimen pembangunan model. Tabel 4 merupakan perbandingan kinerja model menggunakan fungsi dasar yang disediakan oleh *library sklearn*. Terlihat bahwa model *XGBoost* memiliki nilai sensitivitas tertinggi pada data testing yaitu sebesar 58% yang

kemudian disusul oleh *Decision Tree*, *Random Forest*, dan *Logistic Regression* dengan nilai sensitivitas 54%, 53%, dan 51%. Model XGBoost memiliki nilai standar deviasi yang paling rendah di mana ini mengindikasikan model ini memiliki kinerja relatif stabil dari 5 kali pelatihan menggunakan *cross-validation*. Model XGBoost juga relatif memiliki waktu komputasi yang tidak terlalu lama di mana merupakan model ini tercepat kedua di bawah *Decision Tree*. Eksperimen selanjutnya akan menggunakan *XGBoost* karena memiliki kinerja paling tinggi dan waktu komputasi yang relatif cepat. Eksperimen selanjutnya adalah pengaruh penggunaan *resampling* pada data.

Tabel 4. Perbandingan Kinerja Model menggunakan Parameter Default

Model	Recall (5-fold CV)	Recall (std. dev)	Waktu Komputasi Latih Model (dalam detik)	Recall (pada data testing)
XGBoost	0.54	0.07	0.06	0.58
Decision Tree	0.51	0.11	0.01	0.54
Random Forest	0.57	0.12	0.20	0.53
Logistic Regression	0.59	0.09	0.51	0.51

Tabel 5 merupakan perbandingan kinerja model *XGBoost* berdasarkan teknik *resampling* data. Teknik *random undersampling* memiliki kinerja yang lebih tinggi di mana mampu menghasilkan nilai *recall* sebesar 70%. Eksperimen berikut mengindikasikan bahwa menyeimbangkan proporsi kelas target dapat membuat model mempelajari data lebih baik dan meminimalisir bias terhadap kelas tertentu. Kekurangan penerapan teknik *random undersampling* adalah terbuangnya data sehingga berpotensi kehilangan informasi berharga. Sebaliknya, *random upsampling* memiliki risiko model yang dihasilkan mengalami *overfitting* karena data diduplikat secara acak. Penelitian ini menggunakan teknik *SMOTE* di mana berfokus pada menduplikat data melalui mekanisme interpolasi sehingga meminimalisir data yang terduplikasi secara bias.

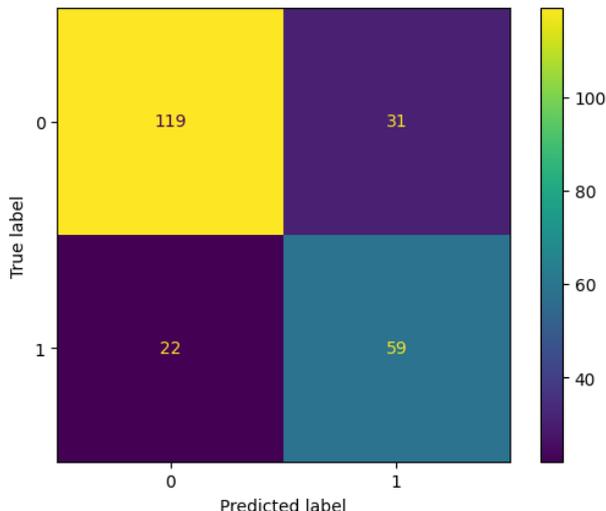
Tabel 5. Perbandingan Penggunaan Teknik Resampling Data

Experimen	Recall (5-fold CV)	Recall (std. dev)	Waktu Komputasi Latih Model	Recall (pada data testing)
Random Undersampling	0,68	0,09	0,06	0,70
SMOTE	0,83	0,12	0,09	0,65
Random Upsampling	0,90	0,09	0,07	0,57

Tahap terakhir adalah melakukan *hyperparameter tuning* yang bertujuan memperoleh kinerja model yang lebih baik melalui eksperimen beberapa skenario parameter. Tabel 6 merupakan skenario parameter yang dilakukan. Eksperimen dilakukan secara *gridsearch* sehingga penelitian ini melakukan eksperimen sebesar 64 skenario. Hasil eksperimen menunjukkan bahwa nilai *recall* meningkat menjadi 73% di mana kombinasi parameter terbaik adalah *n_estimator* 50, *max_bin* 255, dan *learning_rate* 0,05. Gambar 7 merupakan *confusion matrix* dari model yang dihasilkan.

Tabel 6. Skenario Hyperparameter Tuning

Parameter	Skenario Parameter
<i>n_estimators</i>	50, 100, 150, 200
<i>max_bins</i>	230, 255, 275, 300
<i>learning_rate</i>	0,1 , 0,075 , 0,125 , 0,05



Gambar 7. Confusion Matrix

Berdasarkan hasil *classification report* model pada Gambar 8, dapat disimpulkan bahwa model *XGBoost* mampu mengurangi sampai 79% pasien yang sebenarnya tidak mengidap diabetes namun tetap diberi perawatan. Kesimpulan ini didapat dari nilai *recall* sebesar 0,79 pada label “Non-Diabetes”. Hal ini tentu bermanfaat dalam hal efisiensi pemberian perawatan pada orang yang sebenarnya tidak mengidap diabetes. Selain itu, nilai *recall* sebesar 73% mengindikasikan model dapat memprediksi pasien mengidap diabetes dengan tepat sebesar 73%. Bisa dikatakan bahwa 100 orang yang diperkirakan mengidap diabetes, 73 diantaranya diprediksi dengan tepat.

Penelitian ini berfokus pada peningkatan nilai *recall* karena ingin berfokus pada model yang mampu meningkatkan tingkat *True Positive* dan meminimalisir *False Negative*. Artinya, kami ingin menghasilkan sebuah model yang mampu memprediksi pasien yang mengidap diabetes secara tepat dari yang dicurigai/diprediksi terkena diabetes. Menurut tinjauan literatur peneliti, sebenarnya masih belum ada panduan terkait suatu model *machine learning* harus mencapai nilai *recall* tertentu untuk dinilai apakah model tersebut memiliki nilai *recall* rendah atau tinggi. Namun, yang menjadi catatan adalah nilai *recall* yang masih tergolong rendah, yaitu hanya 73%.

Penyebab rendahnya nilai *recall* dapat disebabkan oleh minimnya data pasien yang mengidap diabetes sehingga model kurang mampu mempelajari karakteristik pasien yang mengidap diabetes dengan baik. Seperti yang terlihat dari Gambar 8, nilai beberapa metrik seperti *recall*, *precision*, dan *f1-score* kelas negatif (non-diabetes) yang selalu lebih tinggi daripada kelas positif (diabetes). Walaupun model sudah dilatih menggunakan teknik *data resampling*, bisa dikatakan model masih ada kecenderungan untuk memiliki kinerja yang lebih baik terhadap kelas tertentu. Hal ini bisa saja terjadi karena data *testing* tidak dilakukan *resampling* layaknya pada proses pelatihan model.

4.5 Model Interpretation

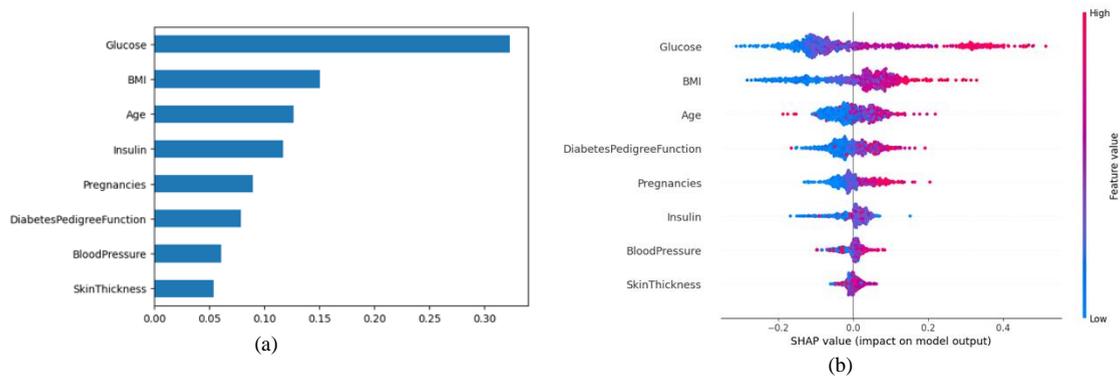
Tahap terakhir adalah menginterpretasikan model *machine learning* yang dibangun untuk mendapatkan informasi lebih mendalam. Gambar 9 merupakan hasil interpretasi model *XGBoost* pada diabetes. Terlihat pada Gambar 9(a) bahwa tiga fitur atau faktor yang paling mempengaruhi orang mengidap diabetes adalah kadar glukosa, BMI, dan usia. Selanjutnya penelitian ini juga menggunakan *SHAP value* untuk mengukur seberapa besar kontribusi tiap fitur yang mempengaruhi orang mengidap pasien diabetes seperti terlihat pada Gambar 9 (b). Titik poin berwarna merah merepresentasikan semakin besar nilai atau kontribusi suatu fitur terhadap kelas target sementara titik poin berwarna biru sebaliknya. Grafik ini mampu melengkapi informasi

yang disampaikan pada Gambar 9(a) di mana hanya menunjukkan faktor atau variabel apa saja yang berpengaruh pada kelas target.

	Precision	Recall	F1-Score	Support
Non-Diabetes	0,84	0,79	0,82	150
Diabetes	0,66	0,73	0,69	81
Accuracy			0,77	231
Macro Avg	0,75	0,76	0,75	231
Weighted Avg	0,78	0,77	0,77	231

Gambar 8. Classification Report

SHAP value menunjukkan bagaimana arah kontribusi fitur yang digunakan. Semakin tinggi kadar glukosa pada tubuh memperbesar peluang seseorang untuk mengidap diabetes. Hal ini bisa menjadi indikasi bahwa kita harus senantiasa menjaga kadar glukosa pada tubuh agar tidak terlalu tinggi. Salah satu upaya yang bisa dilakukan adalah menjaga perilaku hidup sehat seperti mengontrol makanan atau minuman yang manis. Berat badan juga perlu diperhatikan untuk meminimalisir terkena diabetes. Hasil eksperimen menunjukkan bahwa semakin besar BMI akan diikuti semakin besar peluang terkena diabetes. Ada indikasi bahwa semakin bertambahnya umur akan memperbesar peluang terkena diabetes. Namun tidak tertutup kemungkinan bahwa orang yang berusia di usia produktif (sekitar 30-50) juga bisa terkena diabetes apabila melihat kepadatan titik poin pada grafik.



Gambar 9. Interpretasi Model Machine Learning menggunakan Feature Importance (a) dan SHAP values (b)

5. Kesimpulan

Berdasarkan proses eksperimen yang dilakukan, dapat ditarik beberapa poin kesimpulan yang dapat diambil beserta potensi pengembangan untuk penelitian selanjutnya, yaitu:

5.1 Simpulan

Penelitian ini bertujuan menghasilkan model yang berfokus pada tingkat sensitivitas berhubung banyak penelitian terkait prediksi diabetes tidak memperhatikan pengukuran tersebut dalam proses pembangunan model. Selain itu, model yang tidak menghasilkan suatu informasi seperti fitur atau variabel apa yang

memiliki kontribusi terbesar terhadap target kelas juga dinilai kurang intuitif dan justru kurang dipercaya oleh para ahli. Diabetes merupakan penyakit dengan tingkat kesembuhannya sangat kecil, namun bukan berarti tidak ada upaya pencegahan dan pengelolaan gaya hidup dengan baik. Hasil eksperimen menggunakan *XGBoost* mampu menghasilkan nilai *recall*, *precision*, dan akurasi sebesar 73%, 66%, dan 77% di mana ini belum bisa dikatakan baik. Walaupun begitu, penelitian ini berupaya dalam membangun sebuah model yang memiliki sensitivitas yang tinggi terhadap suatu kelas. Hasil interpretasi model menunjukkan bahwa beberapa faktor penyebab orang terkena diabetes adalah kadar glukosa, BMI, dan usia. Hal ini bisa menjadi informasi berharga seperti tingginya kadar gula harus selalu dikontrol untuk meminimalisir terkena diabetes.

5.2 Saran

Penelitian ini ditujukan untuk membangun sebuah model dengan nilai *recall* tertinggi dan memperoleh informasi berharga melalui *explainable machine learning*. Sampel data yang digunakan dihimpun dari situs *kaggle* yang berisi sekitar 700 baris data namun hanya pasien wanita. Oleh karena itu, penelitian selanjutnya dapat menggunakan data yang lebih variatif yaitu adanya pasien pria untuk memperoleh informasi yang lebih komprehensif seperti pengaruh jenis kelamin terhadap risiko mengidap diabetes. Perlu adanya penelitian lebih lanjut seperti apakah ada fitur atau variabel lain di luar *dataset* ini yang dapat digunakan untuk memprediksi apakah seseorang bisa terkena diabetes atau tidak.

6. Daftar Rujukan

- [1] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, Feb. 2023, doi: 10.1016/j.measen.2022.100605.
- [2] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24536.
- [3] B. Md. Ahmed, M. E. Ali, M. M. Masud, and M. Naznin, "Recent trends and techniques of blood glucose level prediction for diabetes control," *Smart Health*, vol. 32, p. 100457, Jun. 2024, doi: 10.1016/J.SMHL.2024.100457.
- [4] S. T. Agatha and R. Tyasnurita, "Perbandingan Klasifikasi Kredibilitas Pengguna Kartu Kredit Menggunakan Decision Tree dan Neural Network," *Jurnal SISFO*, vol. 10, no. 01, pp. 11–24, 2021.
- [5] S. Prayoginingsih and R. P. Kusumawardani, "Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM)," *Jurnal SISFO*, vol. 07, no. 02, pp. 83–98, 2018.
- [6] 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, Dec. 2021, doi: 10.1016/j.ict.2021.02.004.
- [7] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/9930985.
- [8] H. Salem, M. Y. Shams, O. M. Elzeki, M. A. Elfattah, J. F. Al-amri, and S. Elnazer, "Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes," *Applied Sciences (Switzerland)*, vol. 12, no. 3, Feb. 2022, doi: 10.3390/app12030950.
- [9] Z. Zhang *et al.*, "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, Jan. 2024, doi: 10.1016/j.jksuci.2023.101873.
- [10] G. Dharmarathne, T. N. Jayasinghe, M. Bogahawaththa, D. P. P. Meddage, and U. Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2024.100301.
- [11] C. Wachinger, T. N. Wolf, and S. Pölsterl, "Deep learning for the prediction of type 2 diabetes mellitus from neck-to-knee Dixon MRI in the UK biobank," *Heliyon*, vol. 9, no. 11, p. e22239, Nov. 2023, doi: 10.1016/J.HELIYON.2023.E22239.
- [12] S. Szabó, I. J. Holb, V. É. Abriha-Molnár, G. Szatmári, S. K. Singh, and D. Abriha, "Classification Assessment Tool: A program to measure the uncertainty of classification models in terms of class-level metrics," *Appl Soft Comput*, vol. 155, p. 111468, Apr. 2024, doi: 10.1016/j.asoc.2024.111468.
- [13] C. Yang *et al.*, "Deep learning model for the detection of prostate cancer and classification of clinically significant disease using multiparametric MRI in comparison to PI-RADS score," *Urologic Oncology: Seminars and Original Investigations*, Feb. 2024, doi: 10.1016/J.UROLONC.2024.01.021.
- [14] S. Harford *et al.*, "A machine learning based model for Out of Hospital cardiac arrest outcome classification and sensitivity analysis," *Resuscitation*, vol. 138, pp. 134–140, May 2019, doi: 10.1016/J.RESUSCITATION.2019.03.012.
- [15] F. Prinzi, A. Orlando, S. Gaglio, and S. Vitabile, "Breast cancer classification through multivariate radiomic time series analysis in DCE-MRI sequences," *Expert Syst Appl*, vol. 249, p. 123557, Sep. 2024, doi: 10.1016/J.ESWA.2024.123557.
- [16] M. I. Tegar and W. H. Haji, "Penggunaan Algoritma C4.5 untuk Rekomendasi Peminjaman Uang pada Koperasi Sejahtera Mandiri," *Jurnal SISFO*, vol. 09, no. 01, pp. 15–28, 2019.

- [17] C. J. Nolan and M. Prentki, “Insulin resistance and insulin hypersecretion in the metabolic syndrome and type 2 diabetes: Time for a conceptual framework shift,” *Diabetes and Vascular Disease Research*, vol. 16, no. 2. SAGE Publications Ltd, pp. 118–127, Mar. 01, 2019. doi: 10.1177/1479164119827611.
- [18] H. Qi, X. Song, S. Liu, Y. Zhang, and K. K. L. Wong, “KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features,” *Comput Methods Programs Biomed*, vol. 231, p. 107378, Apr. 2023, doi: 10.1016/J.CMPB.2023.107378.
- [19] B. Yan *et al.*, “Explainable machine learning-based prediction for aerodynamic interference of a low-rise building on a high-rise building,” *Journal of Building Engineering*, vol. 82, p. 108285, Apr. 2024, doi: 10.1016/J.JOBE.2023.108285.
- [20] J. P. S. S. Madushani, R. M. K. Sandamal, D. P. P. Meddage, H. R. Pasindu, and P. I. A. Gomes, “Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers,” *Transportation Engineering*, vol. 13, Sep. 2023, doi: 10.1016/j.treng.2023.100190.
- [21] H. Lei, M. Xue, H. Liu, and J. Ye, “Unveiling the driving patterns of carbon prices through an explainable machine learning framework: Evidence from Chinese emission trading schemes,” *J Clean Prod*, vol. 438, p. 140697, Jan. 2024, doi: 10.1016/J.JCLEPRO.2024.140697.
- [22] G. Prabhakar, V. R. Chintala, T. Reddy, and T. Ruchitha, “User-cloud-based ensemble framework for type-2 diabetes prediction with diet plan suggestion,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 7, Mar. 2024, doi: 10.1016/j.prime.2024.100423.
- [23] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, “Descriptive statistics and normality tests for statistical data,” *Ann Card Anaesth*, vol. 22, no. 1, pp. 67–72, Jan. 2019, doi: 10.4103/aca.ACA_157_18.

