Model Predictive Analytics Terhadap Pasien Diabetes Menggunakan *Exploratory Data Analysis* dan Algoritma Random Forest

Hamid Muhammad Jumasa[1*]
[1]Teknologi Informasi, Universitas Muhammadiyah Purworejo, 54111, Indonesia
hamidjumasa@umpwr.ac.id

Abstrak

Diabetes merupakan salah satu penyakit yang masuk dalam kategori penyakit kronis (jangka panjang). Penyakit ini ditandai dengan meningkatnya kadar gula darah (glukosa) yang melebihi ambang batas normal. Sehingga mengakibatkan fungsi hormon insulin dalam tubuh tergaggu.1Tahun 2021, *International Diabetes Federation* (IDF) mencatat ada 537 juta orang dewasa dengan rentang usia 20 – 79 tahun(Reza Pahlevi, 2021). Diabetes juga menyebabkan 6,7 juta kematian. Beberapa faktor penyebab diabetes antara lain kelebihan berat badan, kadar kolesterol yang tinggi, gaya hidup serta tidak berolahraga dan faktor usia. Hingga saat ini belum ditemukan obat yang mampu mengobati penyakit ini secara total, sehingga yang perlu dilakukan ialah mendeteksi diabetes sejak dini guna mengontrol bahaya penyakit diabetes ini.

Penelitian ini akan membuat model predictive analytics dalam memprediksi seseorang terkena diabetes. Teknik analisa data menggunakan *Exploratory Data Analysis* (EDA) dan model *machine learning* menggunakan *Random Forest*. Peneltian ini menggunakan data dari website Kaggle dengan jumlah data 769 orang. Data tersebut terdiri atas 9 kolom dengan 7 data dan 2 data.

Setelah dilakukan analisa terhadap data sampling dan mendapati akurasi pada *data training* sebesar 0.998207 dengan *Mean Squared Error* 0.00179. Data testing diperoleh10.74603 dengan *Mean Squared Error* 0.25396. Hasil prediksi dari 20 data sampel yang diuji, diperoleh 18 kali model memprediksi benar dan 2 kali model salah melakukan prediksi.

Kata kunci: *Exploratory Data Analysis, EDA, Random Forest, Predictive Analytics*.

Abstract

Diabetes is one of the diseases that fall into the category of chronic (long-term) diseases. This disease is characterized by increased blood sugar (glucose) levels that exceed the normal threshold. As a result, the function of the insulin hormone in the body is disrupted.1In 2021, the International Diabetes Federation (IDF) noted that there were 537 million adults aged 20 - 79 years (Reza Pahlevi, 2021). Diabetes also causes 6.7 million deaths. Several factors that cause diabetes include being overweight, high cholesterol levels, lifestyle and not exercising and age. Until now, no medicine has been found that can treat this disease completely, so what needs to be done is to detect diabetes early to control the dangers of diabetes.1

This research will create a predictive analytics model to predict whether someone will develop diabetes. The data analysis technique used Exploratory Data Analysis (EDA) and the machine learning model used Random Forest. This research used data from the website Kaggle with a total of 769 people. The data consists of 9 columns with 7 data and 2 data.

After analyzing the sampling data, the accuracy of the training data was 0.998207 with a Mean Squared Error of 0.00179. Testing data obtained was 0.74603 with Mean Squared Error of 0.25396. The prediction results from 20 sample data tested, obtained 18 times the model made correct predictions and 2 times the model made incorrect predictions.

Keywords: *Exploratory Data Analysis, EDA, Random Forest, Predictive Analytics.*

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit yang perlu diwaspadai. Penyakit ini masuk dalam kategori penyakit kronis (jangka panjang) ditandai dengan meingkatnya kadar gula darah (glukosa) yang melebihi ambang batas normal (Sagita et al., 2021). Akibatnya fungsi hormon insulin dalam tubuh terganggu. Penyakit diabetes terbagi atas dua tipe. Diabetes tipe pertama menyerang sel-sel produksi insulin pada sistem imun yang berakibat tubuh tidak dapat memproduksi insulin sama sekali. Diabetes tipe kedua tubuh tidak akan mampu lagi dalam memproduksi insulin(Nur Ikhromr et al., 2023).

Tahun 2021, *International Diabetes Federation* (IDF) mencatat ada 537 juta orang dewasa dengan rentang usia 20 – 79 tahun(Reza Pahlevi, 2021). Diabetes juga menyebabkan 6,7 juta kematian. Beberapa faktor penyebab diabetes antara lain kelebihan berat badan, kadar kolesterol yang tinggi, gaya hidup serta tidak berolahraga dan faktor usia. Hingga saat ini belum ditemukan obat yang mampu mengobati penyakit ini secara total, sehingga yang perlu dilakukan ialah mendeteksi diabetes sejak dini guna mengontrol bahaya penyakit diabetes ini (Isnaini & Ratnasari, 2018) (Nasution et al., 2021).

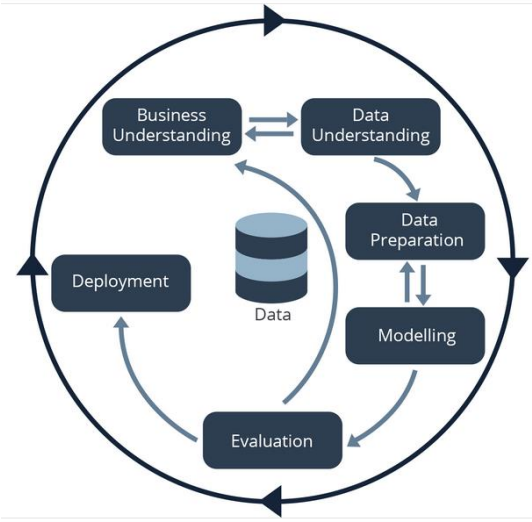
Beberapa penelitian mengenai model prediksi telah banyak dilakukan. Seperti penelitian (Apriliah et al., 2021) mengenai prediksi kemungkinan diabetes pada tahap awal dengan membandingkan model SVM, *Naive Bayes* dan *Random Forest*. Penelitian lain (Sriyanto & Ria Supriyatna, 2023) mengenai prediksi penyakit diabetes menggunakan algoritme *Random Forest*.

Penelitian lain (Rahmi et al., 2023) membandingkan *Random Forest* dengan AdaBoost untuk memprediksi Peserta JKN-KIS yang menunggak. Penelitian lain (Yuliani, 2022) menggunakan *Random Forest* dalam memprediksi kelangsungan hidup pasien gagal jantung dengan seleksi fitur *Bestfirst*. Penelitian lain (Madaerdo Sotarjua & Budhi Santoso, 2022) membandingkan algoritme KNN, *Decision Tree* dan *Random Forest* perbaikan yang efektif terhadap *data imbalanced class* untuk klasifikasi promosi karyawan.

Melihat beberapa penelitian sebelumnya, maka dalam penelitian ini akan membuat model *predictive analytics* dalam memprediksi seseorang terkena diabetes. Teknik analisa data menggunakan *Exploratory Data Analysis* (EDA) dan model *machine learning* menggunakan *Random Forest*. Penelitian ini menggunakan data dari website http://kaggle.com dengan jumlah data 769 orang. Data tersebut terdiri atas 9 kolom dengan 7 data dan 2 data. *Random Forest* dipilih karena algoritme ini mampu memprediksi dengan nilai akurasi yang tinggi.

2. METODE

Pada Gambar 1 disajikan alur penelitian yang dilakukan dalam membangun model. Adapun alur penelitiannya sebagai berikut:



Gambar 1. Alur Penelitian
(https://www.dicoding.com)

Dari alur penelitian diatas dapat dijelaskan sebagai berikut (Ayu Mardhiyah et al., 2020) (Sabariah et al., 2012):

a. Business Understanding

   Fase awal ini memahami masalah bisnis lalu merancang solusi analitik berdasarkan data guna menyelesaikan permasalahan yang akan diselesaikan.Tahapannya mencakup menentukan tujuan bisnis, menilai situasi dan menentukan tujuan data mining.

b. Data Understanding

   Fase ini memahami data yang akan digunakan dalam penelitian.1Seperti mengumpulkan data, mendeskripsikan data, mengeksplorasi data, dan melakukan verifikasi kualitas data.

c. Data Preparation

   Fase ini persiapan data yang mencakup semua kegiatan untuk membangun dataset

akhir dari data mentah awal. Tahapannya dengan mengumpulkan data awal, mendeskripsikan data serta mengeksplorasinya. Dan tahap terakhir melakukan verifikasi kualitas data.

d. Modelling

Fase ini memilih pemodelan yang sesuai untuk dapat dipilih dan diterapkan menggunakan parameter-parameter yang ditentukan. Dalam penelitian ini menggunakan algoritme Random Forest.

e. Evaluation

Fase selanjutnya melakukan evaluasi terhadap hasil tmodel yang sudah dibangun. Selain itu juga perlu memperhatikan setiap langkah yang telah dilakukan. Apabila masalah belum terselesaikan, maka cukup dipertimbangkan. Bagian terakhir dari fase ini mengenai keputusan terkait penggunaan data dan hasil mining yang harus tercapai.

f. Deployment

Fase terakhir ialah meningkatkan pengetahuan dari data kemudian menyajikan hasilnya dapat digunakan oleh banyak orang.

***Exploratory Data Analysis* (EDA)**

Exploratory Data Analysis (EDA) merupakan proses investigasi awal terhadap data untuk menganalisis bentuk karakteristik, menemukan pola, anomali dan memeriksa asumsi pada data. Teknik ini biasanya dalam bentuk pengukuran data statistik dan representasi grafis atau visualisasi (Radhi et al., 2021).

Langkah-langkah yang dilakukan dalam menganalisis EDA data ini antara lain:

a. Menentukan jenis variabel pada dataset yang digunakan
b. Menentukan Distribusi variabel dalam dataset
c. Menemukan missing value
d. Mencari fitur yang tidak berguna (*redudant*)
e. Menentukan korelasi antara fitur dan target1

Random Forest

Random forest merupakan salah satu algoritme klasifikasi. Cara kerjanya dengan membangun beberapa pohon klasifikasi secara paralel kemudian menentukan prediksi berdasarkan suara terbanyak (Hasan et al., 2022)(Karim et al., 2023).

Tahap pembuatan model prediksi menggunakan algoritme ini dengan *n* observasi dan *p* peubah penjelas(Rahmi et al., 2023) (Apriliah et al., 2021). Berikut penjelasannya:

a. Proses bootstrap dengan menarik sampel secara acak sejumlah n (Erdiansyah et al., 2022)

b. Membangun pohon klasifikasi tunggal menggunakan data dari hasil training yang baru setelah proses bootstrap selesai

c. Pohon klasifikasi dibentuk dengan menerapkan random feature selection, dimana secara acak memilih peubah penjelas dengan ketentuan m<p. Nilai m nantinya dipilih berdasarkan nilai peubah penjelas terbaik sebagai pemisah kemudian lanjut dengan pemisahan menjadi dua simpul baru. Proses ini terus berlanjut sampai ukuran minimum dari pengamatan pada simpul tercapai. Nilai m yang direkomendasikan yaitu akar p, dengan catatan perlu mengoptimalkan parameter dengan nilai m untuk mendapatkan hasil terbaik.

d. Langkah selanjutnya mengulangi tahap 1 dan 2 sebanyak x kali guna memperoleh x pohon klasifikasi.1Setiap phon klasifikasi memberikan hasil satu suara. Sehingga kelas klasifikasi ditentukan oleh suara terbanyak dari jumlah x suara.

3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini diperoleh dari1Diabetes Patients Data dengan dataset terdiri atas 769 orang. Data tersebut terdiri atas 9 kolom dengan 7 data bertipe int64 dan 2 data bertipe float6. Adapun variabel yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Variabel Dataset Pasien Diabetes

No	Variabel	Keterangan	Variable
1	Pregnancies	Menyatakan jumlah ibu hamil terkena diabetes	Fitur
2	Glucose	Menyatakan kadar gula dalam darah	Fitur
3	Blood Pressure	menyatakan pengukuran	Fitur

No	Variabel	Keterangan	Variable
		tekanan darah	
4	SkinThickness	mengekspresikan ketebalan kulit	Fitur
5	Insulin	mengekspresikan tingkat insulin dalam darah	Fitur
6	BMI	Menyatakan indeks massa tubuh	Fitur
7	Diabetes Pedigree Function	Menyatakan persentase diabetes	Fitur
8	Age	Menyatakan seseorang terkena diabetes berdasarkan usia	Fitur
9	Outcome	Menyatakan hasil akhir 1 adalah ya dan 0 adalah tidak	Target

Langkah awal dengan melakukan *Exploratory Data Analysis* (EDA). EDA merupakan proses investigasi awal pada data untuk menganalisa data berdasarkan karakteristik, menemukan pola, anomali, dan memeriksa asumsi pada data. Hampir semua kolom ditemukan *missing value* atau data bernilai 0.

```
Nilai 0 di kolom Pregnancies ada:  111
Nilai 0 di kolom Glucose ada:  5
Nilai 0 di kolom BloodPressure ada:  35
Nilai 0 di kolom SkinThickness ada:  227
Nilai 0 di kolom Insulin ada:  374
Nilai 0 di kolom BMI ada:  11
Nilai 0 di kolom DiabetesPedigreeFunction ada:  0
Nilai 0 di kolom Age ada:  0
Nilai 0 di kolom Outcome ada:  500
```

Gambar 2. Menemukan Missing Value Pada Fitur

Missing value dalam penelitian ini diperbaiki dengan cara mengganti dengan menghitung nilai mean dari fitur masing-masing, kecuali fitur age dan pregnancies.

Langkah selanjutnya memastikan data sampel nilainya tidak jauh dari cakupan umum data utama. Artinya apabila data sampel berada sangat jauh maka disebut dengan outlier. Pengamatan1outlier dilakukan dengan menggunakan teknik visualisasi data (boxplot). Hampir semua fitur ditemukan outlier kecuali pada fitur glucose.

```
df.Glucose = df.Glucose.replace(0,int(df.Glucose.mean()))
```

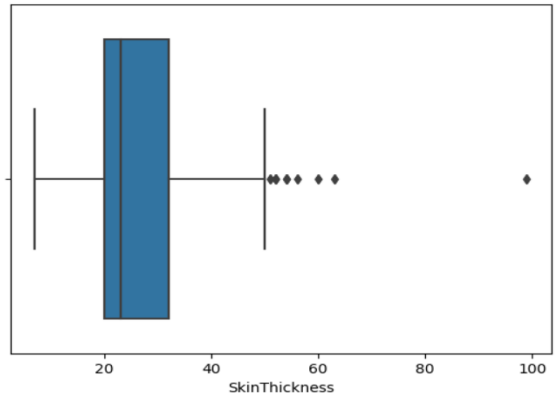
```
#mengisi nilai pada fitur yang bernilai 0
df.Glucose.unique()
```

```
array([148,  85, 183,  89, 137, 116,  78, 115, 197, 125, 110, 168, 139,
       189, 166, 100, 118, 107, 103, 126,  99, 196, 119, 143, 147,  97,
       145, 117, 109, 158,  88,  92, 122, 138, 102,  90, 111, 180, 133,
       106, 171, 159, 146,  71, 105, 101, 176, 150,  73, 187,  84,  44,
       141, 114,  95, 129,  79, 120,  62, 131, 112, 113,  74,  83, 136,
        80, 123,  81, 134, 142, 144,  93, 163, 151,  96, 155,  76, 160,
       124, 162, 132, 173, 170, 128, 108, 154,  57, 156, 153, 188, 152,
       104,  87,  75, 179, 130, 194, 181, 135, 184, 140, 177, 164,  91,
       165,  86, 193, 191, 161, 167,  77, 182, 157, 178,  61,  98, 127,
        82,  72, 172,  94, 175, 195,  68, 186, 198, 121,  67, 174, 199,
        56, 169, 149,  65, 190])
```

Gambar 3. Perbaikan Data Pada Salah Satu Fitur1

```
sns.boxplot(x=df.SkinThickness)
```

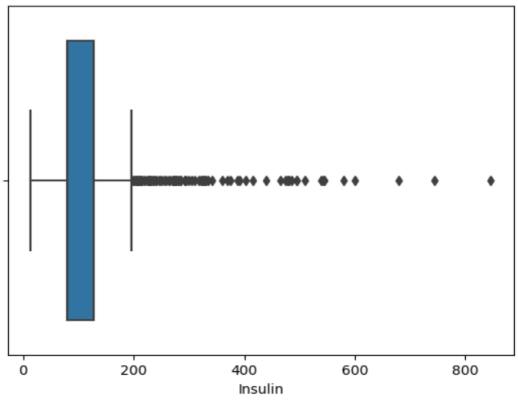
```
<Axes: xlabel='SkinThickness'>
```



Gambar 4. Outlier Pada Fitur SkinThickness

```
sns.boxplot(x=df.Insulin)
```

```
<Axes: xlabel='Insulin'>
```



Gambar 5. Outlier Pada Fitur Insulin

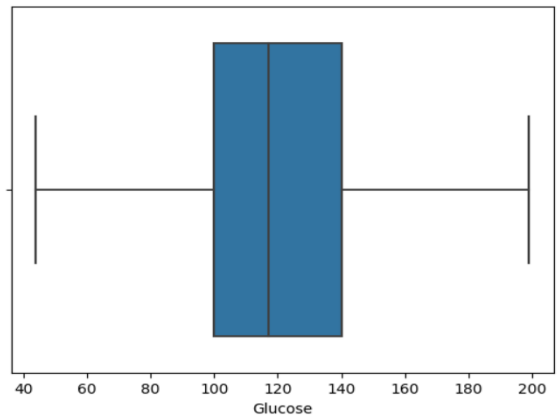
Setelah mengamati hasil dan memvisualisasikan outlier dari masing-masing fitur, langkah selanjutnya mencari nilai outlier menggunakan teknik IQR. Metode *Inter Quartile Range* (IQR) merupakan konsep kuartil, dimana dari suatu populasi terdiri dati tiga nilai yang membagi distribusi data menjadi empat sebaran. Pembagiannya seperempat dari data itu di kuartil pertama (Q1), setengah dari data itu di kuartil

p-ISSN 2620–4843
e-ISSN 2620–4924

kedua (Q2) dan tiga perempat dari data itu di kuartil ketiga (Q3). Maka perhitungannya IQR = Q3 - Q1.

```
#check outliers
sns.boxplot(x=df.Glucose)
```

```
<Axes: xlabel='Glucose'>
```



Gambar 6. Fitur Glucose Tidak Outlier

```
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR=Q3-Q1
df=df[~((df<(Q1-1.5*IQR))|(df>(Q3+1.5*IQR))).any(axis=1)]
# Cek ukuran dataset setelah kita drop outliers
df.shape
```

```
(621, 9)
```

Gambar 7. *Inter Quartile Range* Pada Data Sampel

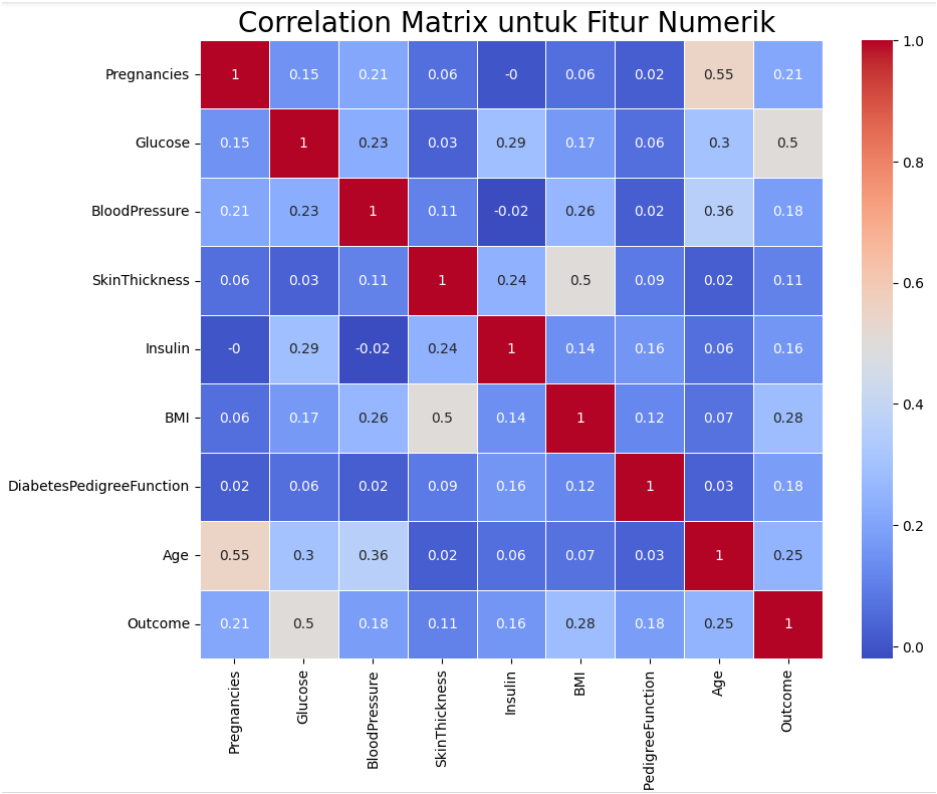
Langkah selanjutnya melakukan proses transformasi pada data sehingga menjadi bentuk yang cocok untuk proses pemodelan. Dataset dibagi ke dalam data training 90% dan data testing 10%.

```
from sklearn.model_selection import train_test_split
X = df.drop(["Outcome","DiabetesPedigreeFunction","Insulin"],axis =1)
y = df["Outcome"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.1, random_state = 123)
```

Gambar 8. Membagi Data Sampel

Langkah selanjutnya menggunakan model Random Forest untuk melakukan prediksi terhadap pasien penderita diabetes. Hasil akurasi pada data training adalah 0.998207 dengan Mean Squared Error 0.00179. Kemudian untuk hasil akurasi pada data testing adalah 0.69841 dengan Mean Squared Error 0.30158.

Nilai akurasi antara training data dengan testing data diperoleh hasil yang tidak akurat dan terindikasi overfitting. Oleh sebab itu maka fitur yang nilai korelasinya mendekati 0 dihapus. Sehingga fitur diabetes pedigree function dan insulin dihapus.1



Gambar 9. Correlation Matriks Semua Fitur

Setelah memperbaiki overfitting, maka hasil akurasi pada data training 0.99820 dengan Mean Squared Error 0.00179. Sedangkan akurasi pada data testing 0.74603 dengan Mean Squared Error 0.25396.

```
Prediksi dengan menggunakan model Random Forest
[0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0]
y_true:
688    0
352    0
342    0
461    0
670    0
501    0
265    0
471    0
70     1
498    1
15     1
372    0
381    0
531    0
59     0
423    0
671    0
443    1
338    1
467    0
Name: Outcome, dtype: int64
```

Gambar 10. Hasil Prediksi Menggunakan Model RF

Hasil prediksi menggunakan 20 data sampel diperoleh 2 kali model salah memprediksi hasil pasien penderita Diabetes.

4. KESIMPULAN

Setelah melakukan penelitian hingga tahap akhir, maka dapat diambil kesimpulan bahwa *Exploratory Data Analysis* (EDA) memperbaiki data sampel sebelum membangun model prediksi menggunakan Random Forest. Ditemukan beberapa fitur missing value dan mengalami outlier. Kemudian fitur pedigreefunction dan diabetes dihapus karena mendekati 0.

Fitur tersebut dihapus karena nilai akurasi pada data training adalah 0.998207 dengan Mean Squared Error 0.00179 sedangkan data testing sebelumnya 0.69841 dengan Mean Squared Error 0.30158. Setelah dihapus maka1nilai akurasi pada data training sama dan data testing 0.74603 dengan Mean Squared Error 0.25396. Hasil prediksi dari 20 data sampel yang diuji, diperoleh 18 kali model memprediksi benar dan 2 kali model salah melakukan prediksi.

DAFTAR PUSTAKA

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). *Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest* (Vol. 10, Issue 1). http://sistemasi.ftik.unisi.ac.id

Ayu Mardhiyah, P., Ruli A. Siregar, R., & Palupiningsih, P. (2020). Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Menggunakan Algoritma C4.5. *Jurnal Teknologia*, *3*(1), 91–101.

Erdiansyah, U., Irmansyah Lubis, A., & Erwansyah, K. (2022). Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. *Jurnal Media Informatika Budidarma*, *6*(1), 208. https://doi.org/10.30865/mib.v6i1.3373

Hasan, I. K., Resmawan, R., & Ibrahim, J. (2022). Perbandingan K-Nearest Neighbor dan Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa. *Indonesian Journal of Applied Statistics*, *5*(1), 58. https://doi.org/10.13057/ijas.v5i1.58056

Isnaini, N., & Ratnasari. (2018). Faktor Risiko Mempengaruhi Kejadian Diabetes Mellitus Tipe Dua. *Jurnal Kebidanan Dan Keperawatan Aisyiyah*,*14*(1),59–68. https://doi.org/10.31101/jkk.550

Karim, A. A., Ary Prasetyo, M., & Saputro, M. R. (2023). *Perbandingan Metode Random Forest, K-Nearest Neighbor, dan SVM Dalam Prediksi Akurasi Pertandingan Liga Italia* (Vol. 2). http://www.football-data.co.uk.

Nasution, F., Azwar Siregar, A., & Tinggi Kesehatan Indah Medan, S. (2021). Faktor Risiko Kejadian Diabetes Mellitus (Risk Factors for The Event of Diabetes Mellitus). *Jurnal Ilmu Kesehatan*, *9*(2), 94–102.

Nur Ikhromr, F., Sugiyarto, I., Faddillah, U., & Sudarsono, B. (2023). Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naives Bayes dan K-Nearest Neighbor. *Journal of Information Technology and Computer Science (INTECOMS)*, *6*(1).

Radhi, M., Ryan Hamonangan Sitompul, D.,

Hamonangan Sinurat, S., & Indra, E. (2021). Analisis BIG DATA Dengan Metode Exploratory Data Analysis (EDA) Dan Metode Visualisasi Menggunakan Jupyter Notebook. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, *4*(2), 23–27.

Rahmi, I. A., Afendi, F. M., & Kurnia, A. (2023). Metode AdaBoost dan Random Forest untuk Prediksi Peserta JKN-KIS yang Menunggak. *Jambura Journal of Mathematics*, *5*(1), 83–94. https://doi.org/10.34312/jjom.v5i1.15869

Sabariah, M. K., Mukharil Bachtiar, A., Dharmayanti, D., & Perdana, I. (2012). *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA) 49 Volume. I Nomor. 2, Bulan Oktober*. http://bit.ly/kuesionerpasar

Sagita, P., Apriliana, E., Mussabiq, S., Soleha, T. U., & Dokter, P. (2021). *Pengaruh Pemberian Daun Sirsak (Annona muricata) Terhadap Penyakit Diabetes Melitus*. http://jurnalmedikahutama.com