JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

Penerapan Data Mining dengan Algoritma C4.5 untuk Mengidentifikasikan Prediksi Penyakit Diabetes

Muhammad Ardi Hermansyah[1], Diyas Aditya Adi Saputra[2], Fikri Hamdhan Dwi Saputra[3], Muhammad Rizky Maulana[4], Muhammad Arifin[5]

[1]Teknik, Sistem informasi, Universitas Muria Kudus, Kudus, Indonesia
Email: [1]muhammadardi931@gmail.com, [2]diyassaputraaa@gmail.com, [3]fikrihamdhan13@gmail.com, rm7631835@gmail.com, [5]arifin.m@std.umk.ac.id

Abstrak Diabetes termasuk penyakit kronis dengan angka kematian yang cukup tinggi secara global, dan sering kali baru terdiagnosis pada tahap lanjut karena gejala awalnya yang kurang mencolok. Dua penelitian ini mengembangkan pemanfaatan teknologi *data mining* dengan algoritma Decision Tree C4.5 untuk membentuk model prediksi dini terhadap penyakit diabetes. Data yang digunakan berasal dari repositori publik seperti UCI serta data klinis dari fasilitas kesehatan, dengan atribut-atribut klasifikasi seperti umur, jenis kelamin, berat badan, tekanan darah, detak jantung, dan kadar gula darah. Setelah melalui proses *preprocessing*, perhitungan *entropy* dan *information gain*, serta validasi melalui aplikasi RapidMiner, penelitian ini berhasil menghasilkan aturan klasifikasi yang mampu memprediksi risiko diabetes secara efektif. Kedua studi tersebut menunjukkan akurasi yang tinggi, masing-masing sebesar 95,51% dan 90,00%, yang menandakan bahwa algoritma C4.5 cukup andal dan memiliki potensi untuk dimanfaatkan dalam sistem pendukung keputusan medis guna mendeteksi diabetes sejak dini.

Kata Kunci: Diabetes, Data Mining, Algoritma C4.5, Prediksi Penyakit, Decision Tree.

Abstract Diabetes is one of the chronic diseases with a high mortality rate worldwide and is often only detected at an advanced stage due to its subtle early symptoms. These two studies explored the application of data mining technology using the C4.5 Decision Tree algorithm to develop an early prediction model for diabetes. The data used were obtained from public repositories such as UCI and clinical records from hospitals, with classification attributes including age, gender, body weight, blood pressure, heart rate, and blood sugar levels. After undergoing preprocessing, entropy and information gain calculations, and validation using the RapidMiner application, the studies successfully generated classification rules capable of effectively estimating the risk of diabetes. The results from both studies demonstrated high accuracy levels, at 95.51% and 90.00%, respectively, indicating that the C4.5 algorithm provides reliable predictive performance and holds potential for use in medical decision support systems for early diabetes detection.

Keywords: Diabetes, Data Mining, Algoritma C4.5, Prediksi Penyakit, Decision Tree

1. PENDAHULUAN

Diabetes mellitus adalah jenis penyakit tidak menular yang memiliki tingkat risiko kematian yang cukup tinggi dan kini menjadi salah satu fokus utama dalam isu kesehatan global. Berdasarkan laporan dari Organisasi Kesehatan Dunia (WHO) dan Federasi Diabetes Internasional (IDF), jumlah kasus diabetes terus menunjukkan peningkatan dari tahun ke tahun, khususnya pada kelompok usia produktif antara 49 hingga 59 tahun (WHO, 2023; IDF, 2023). Kondisi ini umumnya berkembang secara bertahap tanpa gejala yang jelas di awal, sehingga sering kali baru terdeteksi setelah muncul komplikasi serius yang memengaruhi organ-organ penting seperti jantung, ginjal, atau sistem saraf pusat. Keterlambatan diagnosis inilah yang menjadi salah satu penyebab utama tingginya angka kematian akibat diabetes.

Dengan kemajuan teknologi informasi, berbagai metode analisis berbasis *data science* kini dimanfaatkan untuk mendukung proses diagnosis dan prediksi penyakit secara lebih cepat dan akurat. Salah satu metode yang cukup banyak diterapkan di bidang medis adalah teknik *data mining*, khususnya menggunakan pendekatan klasifikasi dengan algoritma *Decision Tree C4.5*. Algoritma ini bekerja dengan membangun model prediksi berdasarkan atribut-atribut klinis seperti usia, berat badan, tekanan darah, denyut nadi, serta kadar glukosa dalam darah, untuk menilai tingkat risiko seseorang terhadap diabetes (Quinlan, 1993; Han et al., 2012).

Beberapa penelitian sebelumnya membuktikan bahwa algoritma *C4.5* memiliki tingkat akurasi yang tinggi dalam membangun sistem prediksi penyakit. Dua studi yang menjadi rujukan dalam penelitian ini menunjukkan akurasi prediksi masing-masing sebesar 90,00% dan 95,51%,

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

mengindikasikan bahwa model yang dihasilkan mampu memberikan klasifikasi yang cukup andal. Oleh karena itu, algoritma ini berpotensi besar untuk diterapkan sebagai *decision support tool* dalam sistem informasi kesehatan guna mendukung deteksi dini penyakit diabetes serta membantu pengambilan keputusan klinis secara lebih efisien dan berbasis data (Patil & Kumaraswamy, 2009; Choubey et al., 2019).

2. METODE PENELITIAN

Untuk membangun model prediksi penyakit diabetes menggunakan pendekatan *data mining* dengan algoritma *Decision Tree C4.5*. Metode yang digunakan melibatkan beberapa tahapan, mulai dari studi literatur hingga validasi model, untuk memastikan hasil prediksi yang akurat dan dapat diandalkan.

1. Studi Literatur

   Penelitian ini dimulai dengan studi literatur untuk memahami dasar-dasar yang berkaitan dengan topik, termasuk karakteristik dan penyebab diabetes mellitus, konsep dasar *data mining* dan klasifikasi, penjelasan algoritma *Decision Tree C4.5*, serta metode evaluasi performa model prediksi menggunakan *confusion matrix*, *precision*, *recall*, dan *accuracy*.

2. Pengumpulan dan Pemilahan Data

   Data yang digunakan dalam penelitian ini diambil dari *Pima Indians Diabetes Database* dalam format CSV yang berisi 768 data pasien perempuan berusia 21 tahun ke atas dari suku Indian-Pima dengan atribut yang dianalisis meliputi jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh, riwayat keluarga, usia, serta hasil diagnosis yang menunjukkan kondisi diabetes pasien.

3. Data Preprocessing

   Tahap *preprocessing* dilakukan dengan beberapa langkah seperti mengganti nilai nol pada atribut *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* menggunakan nilai median, melakukan normalisasi pada data numerik jika diperlukan, mengubah label *Outcome* menjadi kelas biner, serta melakukan seleksi fitur dengan menghitung *information gain* untuk menentukan atribut yang paling relevan dalam proses klasifikasi.

4. Pembangunan Model C4.5

   Model klasifikasi Decision Tree C4.5 dibuat dengan cara menghitung entropy data dulu. Setelah itu, dihitung information gain setiap atribut untuk memilih atribut terbaik sebagai akar pohon. Proses ini diulang terus sampai data terbagi dengan baik atau pohon sudah cukup dalam. Semua langkah ini dilakukan menggunakan aplikasi RapidMiner.

5. Validasi Model

   Model yang telah dibangun kemudian divalidasi menggunakan metode split validation dengan membagi data menjadi 70 persen data latih dan 30 persen data uji hasil klasifikasi dievaluasi menggunakan confusion matrix serta dihitung metrik evaluasi seperti accuracy precision recall dan F1-score untuk mengukur seberapa baik model dalam memprediksi risiko diabetes pada data uji.

6. Representasi Data

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

Berdasarkan proses seleksi data yang telah dilakukan, Dataset yang digunakan dalam penelitian ini berasal dari layanan kesehatan di Sylhet, Bangladesh. Dataset ini berisi 520 data pasien dan mencakup berbagai gejala klinis yang berkaitan dengan diabetes, seperti sering buang air kecil (polyuria), sering haus (polydipsia), penurunan berat badan tiba-tiba, kelemahan tubuh, dan gejala lainnya. Setiap data juga memiliki label Class yang menunjukkan hasil diagnosis, yaitu Positive untuk pasien dengan diabetes dan Negative untuk yang tidak.

Tabel 1. Data Diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
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
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31.0	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0.0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38.0	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

3. ANALISA DAN PEMBAHASAN

3.1 Pengujian Manual Algoritma C4.5

Pengujian manual pada algoritma C4.5 dilakukan untuk memahami cara kerja algoritma dalam membentuk pohon keputusan berdasarkan perhitungan entropy dan information gain. Dua konsep utama yang digunakan dalam proses ini adalah entropy, yang mengukur tingkat ketidakpastian suatu data, dan gain, yang mengukur seberapa besar penurunan ketidakpastian setelah data dibagi berdasarkan atribut tertentu.

1. Entropy

   Entropy adalah ukuran ketidakteraturan atau ketidakpastian dalam suatu kumpulan data. Nilai entropy yang tinggi menunjukkan bahwa data sangat acak atau tidak homogen, sedangkan nilai entropy yang rendah menunjukkan data yang lebih teratur atau homogen. Rumus perhitungan entropy adalah sebagai berikut:

$$\text{Entropy}(S) = -i=1\sum npi \cdot \log 2(pi)$$

   Keterangan:

   S : Himpunan data (kasus).

   nnn : Jumlah kelas atau partisi dalam himpunan data.

   pip_ipi : Proporsi data dari kelas ke-iii terhadap total data.

2. Gain

   Information Gain (Gain) adalah selisih antara entropy sebelum dan sesudah data dibagi berdasarkan atribut tertentu. Atribut yang menghasilkan *gain* tertinggi dianggap sebagai

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

atribut terbaik dan dipilih sebagai simpul akar (*root*) pada pohon keputusan. Rumus perhitungan gain adalah sebagai berikut:

$$Gain(S,A)=Entropy(S)-i=1\sum n(|S||Si|\cdot Entropy(Si))$$

Keterangan:

AAA : Atribut yang diuji.

SSS : Himpunan data awal.

SiS_iSi : Subset data setelah dipartisi berdasarkan atribut AAA

|S||S||S| : Jumlah total data.

|Si||S_i||Si| : Jumlah data pada subset ke-iii.

3. Proses Perhitungan Manual

Perhitungan dilakukan berdasarkan data pada Tabel 1, di mana setiap atribut dianalisis untuk menentukan nilai entropy dan gain-nya. Berikut adalah hasil perhitungan manual:

a. Entropy Total

Jumlah data: 364.

Kelas Positif: 224.

Kelas Negatif: 140.

$$\text{Entry Total} = -\left(\frac{224}{364}\log 2\frac{224}{364} + \frac{140}{364}\right) = 0{,}9612$$

b. Atribut Age

≤ 48 tahun: Entropy = 0,9896 (200 data)

≥ 49 tahun: Entropy = 0,9012 (164 data)

$$\text{Gain (Age)} = 0{,}9612\left(\frac{200}{364}\cdot 0,9896\ \frac{164}{364}\cdot 0{,}9012\right) = 0{,}0115$$

c. Atribut Gender

Male: Entropy = 0,9923 (232 data)

Female: Entropy = 0,4395 (132 data)

$$\text{Gain (Gender)} = 0{,}9612 - \left(\frac{232}{364}\cdot 0,9923 + \frac{132}{364}\cdot 0{,}4395\right) = 0,1694$$

d. Atribut: Polyuria

Yes: Entropy = 0,3492 (183 data)

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

No: Entropy = 0,8723 (181 data)

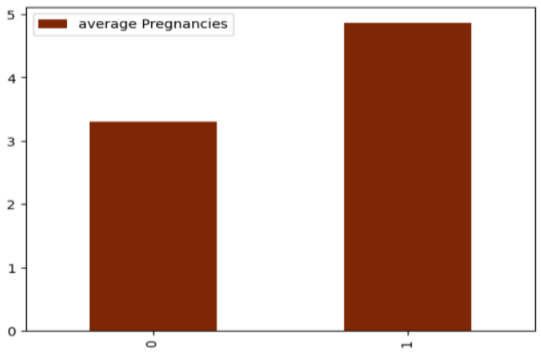
$$\text{Gain (Polyuria)} = 0{,}9612 - \left(\frac{183}{364} \cdot 0,3492 + \frac{181}{364} \cdot 0{,}8723\right) = 0,3519$$

3.2 Pengujian Algoritma C4.5

Pengujian manual dilakukan untuk memahami cara kerja algoritma C4.5 dalam membentuk pohon keputusan melalui perhitungan entropy dan information gain. Dari beberapa atribut yang dihitung, polydipsia menunjukkan nilai gain tertinggi, sehingga paling berpengaruh dalam pemisahan data. Proses ini menunjukkan bagaimana C4.5 secara bertahap memilih atribut terbaik untuk membentuk pohon keputusan yang optimal dan akurat.

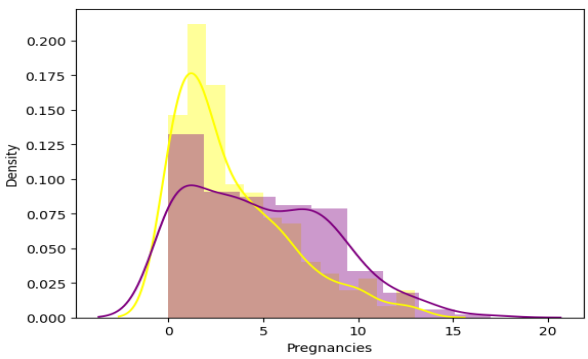
1. Hubungan Jumlah Kehamilan dengan Diabetes

   Analisis awal dilakukan untuk melihat hubungan antara jumlah kehamilan (Pregnancies) dan status diabetes. Variabel ini dipilih karena jumlah kehamilan dapat memengaruhi kondisi metabolik wanita. Rata-rata jumlah kehamilan dihitung berdasarkan status diabetes (Outcome), yaitu 0 untuk tidak diabetes dan 1 untuk diabetes.



Gambar 1. Grafik Batang Rata-rata Jumlah Kehamilan (Bar Chart).

Gambar 1 menunjukkan bahwa rata-rata jumlah kehamilan lebih tinggi pada pasien diabetes (4,87) dibandingkan yang tidak diabetes (3,30), mengindikasikan kemungkinan korelasi positif. Namun, diperlukan analisis lanjutan untuk memastikan hubungan ini bukan dipengaruhi variabel lain.



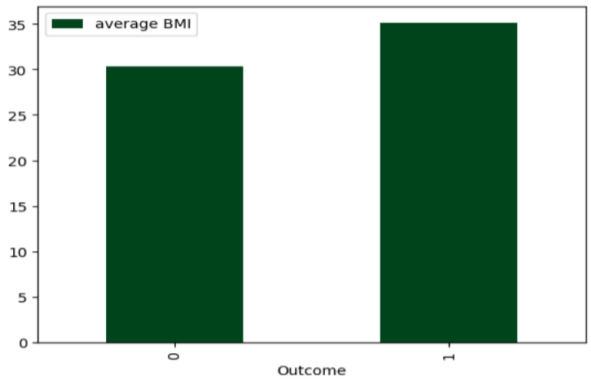
JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

Gambar 2. grafik distribusi kepadatan **(density plot)**

Gambar 2 menunjukkan bahwa pasien diabetes memiliki rata-rata dan rentang jumlah kehamilan yang lebih tinggi, dengan distribusi lebih merata hingga >10 kehamilan. Sebaliknya, pasien sehat cenderung terkonsentrasi pada 0–3 kehamilan, mengindikasikan kecenderungan jumlah kehamilan yang lebih besar pada pasien diabetes.

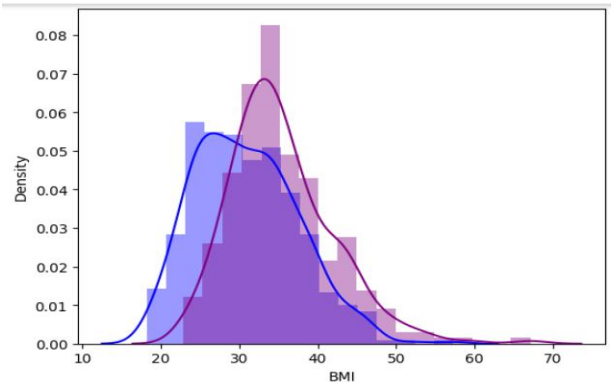
2. Hubungan antara BMI dan Diabetes

Analisis ini bertujuan mengetahui hubungan antara nilai BMI dan status diabetes. BMI dipilih karena merupakan indikator umum obesitas, yang dapat meningkatkan risiko resistensi insulin dan diabetes tipe 2. Rata-rata BMI dihitung berdasarkan status Outcome (0 = tidak diabetes, 1 = diabetes). Hasil ditampilkan dalam diagram batang, dengan sumbu Y menunjukkan rata-rata BMI dan batang hijau tua mewakili nilai tiap kategori Outcome.



Gambar 3. Grafik Batang Rata-rata nilai BMI (Body Mass Index).

Gambar 2 menunjukkan bahwa rata-rata BMI pasien diabetes (35,14) lebih tinggi dibandingkan pasien non-diabetes (30,30). Perbedaan ini mengindikasikan bahwa peningkatan BMI berhubungan dengan risiko diabetes, sehingga menjaga berat badan penting untuk pencegahan diabetes tipe 2.



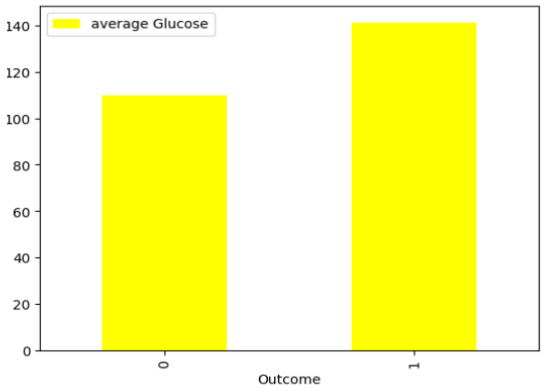
Gambar 4. Grafik distribusi nilai BMI **(Body Mass Index)**

Gambar 4 menunjukkan bahwa distribusi BMI pasien diabetes bergeser ke kanan, dengan puncak kurva pada nilai BMI lebih tinggi dibanding pasien non-diabetes. Ini menguatkan bahwa obesitas berhubungan erat dengan risiko diabetes, sehingga BMI dapat menjadi indikator penting dalam memprediksi kemungkinan seseorang mengidap diabetes.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

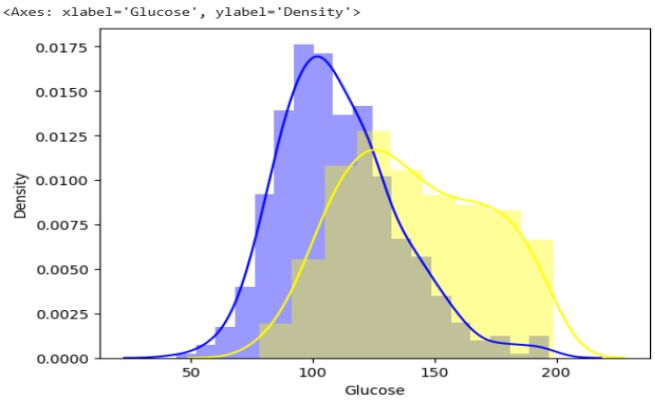
3. Hubungan antara Glukosa dan Diabetes

Analisis ini bertujuan mengkaji hubungan antara kadar glukosa rata-rata dan status diabetes. Hasilnya menunjukkan bahwa pasien diabetes memiliki kadar glukosa rata-rata lebih tinggi (141,26) dibandingkan pasien non-diabetes (109,98), sehingga kadar glukosa dapat menjadi indikator penting dalam mendeteksi risiko diabetes.



Gambar 5. Grafik membandingkan rata-rata kadar glukosa

Visualisasi menunjukkan bahwa rata-rata kadar glukosa pasien diabetes (141,26) jauh lebih tinggi dibandingkan non-diabetes (109,98). Ini mengindikasikan hubungan kuat antara kadar glukosa dan risiko diabetes.



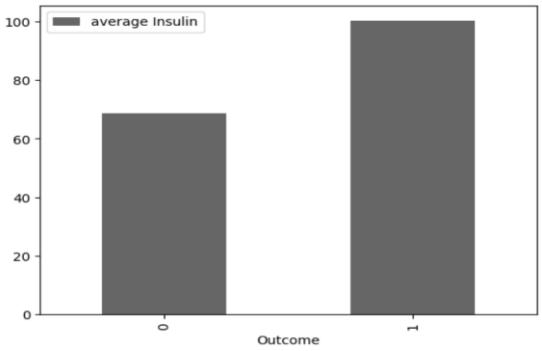
Gambar 6. Grafik histogram

Gambar 6 menunjukkan perbandingan distribusi kadar glukosa antara pasien sehat (Outcome = 0) dan pasien diabetes (Outcome = 1). Pasien sehat (biru) umumnya memiliki kadar glukosa di bawah 120, dengan puncak kepadatan sekitar 100–110. Sebaliknya, distribusi pasien diabetes (kuning) bergeser ke kanan, menandakan kadar glukosa yang lebih tinggi secara umum, dengan banyak kasus melebihi 120 hingga lebih dari 200. Perbedaan ini menguatkan bahwa tingginya kadar glukosa berkaitan erat dengan kondisi diabetes.

4. Hubungan antara Kadar Insulin dan Diabetes

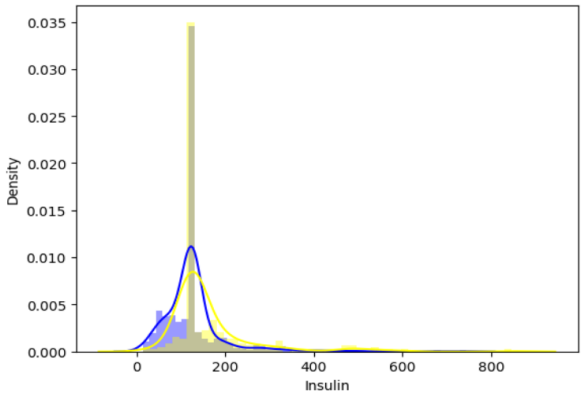
Analisis ini bertujuan melihat perbedaan kadar insulin rata-rata berdasarkan status diabetes. Rata-rata dihitung per kelompok Outcome (0 = sehat, 1 = diabetes) untuk mengetahui apakah pasien diabetes memiliki kadar insulin lebih tinggi dibandingkan yang tidak.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141



Gambar 7. Grafik membandingkan rata-rata insulin pasien diabetes dan non-diabetes.dan non-diabetes.

Sumbu X menunjukkan status pasien (0 = tidak diabetes, 1 = diabetes), sementara sumbu Y menampilkan rata-rata kadar insulin berdasarkan data SQL, yaitu sekitar 68,79 untuk pasien non-diabetes dan 100,34 untuk pasien diabetes. Hasil ini menunjukkan bahwa pasien diabetes memiliki kadar insulin lebih tinggi, mengindikasikan insulin sebagai indikator potensial untuk mendeteksi atau memprediksi risiko diabetes.



Gambar 3. Grafik histogram dan KDE (Kernel Density Estimation) dari data Insulin.

Grafik distribusi insulin menunjukkan sebagian besar nilai di bawah 200 dengan sebaran miring ke kanan, menandakan adanya beberapa pasien dengan kadar insulin sangat tinggi dan banyak data nol. Hal ini mengindikasikan perlunya pembersihan data, termasuk penanganan nilai nol dan outlier sebelum analisis lebih lanjut.

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengidentifikasi pengaruh signifikan beberapa faktor kesehatan, seperti kadar glukosa, insulin, dan BMI, terhadap risiko diabetes pada individu berdasarkan data Pima Indians Diabetes. Visualisasi dan analisis sederhana menunjukkan perbedaan mencolok antara individu yang sehat dan yang mengidap diabetes, memperkuat potensi penggunaan faktor-faktor

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 2, No. 12 Mei 2025
ISSN 3025-0919 (media online)
Hal 2133-2141

tersebut sebagai indikator prediksi. Meskipun demikian, keterbatasan seperti adanya nilai hilang dan terbatasnya variabel kontekstual menjadi catatan penting. Oleh karena itu, pengembangan lebih lanjut disarankan dengan menambahkan variabel seperti pola makan dan riwayat genetik, serta menerapkan algoritma machine learning yang lebih kompleks dan evaluasi model yang lebih mendalam untuk meningkatkan akurasi dan generalisasi prediksi.

REFERENCES

Castleman, Kenneth R., 2004, *Digital Image Processing*, Vol. 1, Ed.2, Prentice Hall, New Jersey.

Gonzales, R., P. 2004, *Digital Image Processing* (*Pemrosesan Citra Digital*), Vol. 1, Ed.2, diterjemahkan oleh Handayani, S., Andri Offset, Yogyakarta.

Wyatt, J. C, danSpiegelhalter, D., 1991, *Field Trials of Medical Decision-Aids: PotentialProblems and Solutions*, Clayton, P. (ed.): *Proc. 15th Symposium on ComputerApplications in Medical Care*, Vol 1, Ed. 2, McGraw Hill Inc, New York.

Yusoff, M, Rahman, S., A., Mutalib, S., and Mohammed, A., 2006, Diagnosing Application Development for Skin Disease Using Backpropagation Neural Network Technique, *Journal of Information Technology*, vol 18, hal 152-159.

Wyatt, J. C, Spiegelhalter, D, 2008, Field Trials of Medical Decision-Aids: PotentialProblems and Solutions, *Proceeding of 15th Symposium on ComputerApplications in Medical Care*, Washington, May 3.

Prasetya, E., 2006, Case Based Reasoning untuk mengidentifikasi kerusakan bangunan, *Tesis*, Program Pasca Sarjana Ilmu Komputer, Univ. Gadjah Mada, Yogyakarta.

Ivan, A.H., 2005, Desain target optimal, *Laporan Penelitian Hibah Bersaing*, Proyek Multitahun, Dikti, Jakarta.

Wallace, V. P., Bamber, J. C. dan Crawford, D. C. 2000. Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminate analysis and artificial neural network. *Journal Physical Medical Biology*, No.45, Vol.3, 2859-2871.

Xavier Pi-Sunyer, F., Becker, C., Bouchard, R.A., Carleton, G. A., Colditz, W., Dietz, J., Foreyt, R. Garrison, S., Grundy, B. C., 1998, Clinical Guidlines on the identification, evaluation, and treatment of overweight and obesity in adults, *Journal of National Institutes of Health*, No.3, Vol.4, 123-130, :http://journals.lww.com/acsm-msse/Abstract/1998/11001/paper_treatment_of_obesity.pdf.

Borglet, C, 2003, Finding Asscociation Rules with Apriori Algorithm,http://www.fuzzy.cs.uniagdeburgde/~borglet/apriori.pdf, diakses tanggal 23 Februari 2007.