# KLASIFIKASI DIABETES TIPE 2 MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOUR

**Ivandari[1], Wahyu Setianto[2], M. Adib Al Karomi**
[1,2]STMIK Widya Pratama Pekalongan
ivandarialkaromi@gmail.com, kian@stmik-wp.ac.id, adib.comp@gmail.com

**RINGKASAN**

Penyakit diabetes adalah penyakit yang banyak menimbulkan kematian. Menurut data dari WHO sepanjang tahun 2019 tercatat ada 2juta kematian yang diakibatkan penyakit diabetes. Pencatatan kondisi pasien banyak dilakukan untuk keperluan medis. Banyaknya pencatatan atau data yang tidak digunakan hanya menjadi sampah digital. Data mining hadir dengan klasifikasi untuk mengolah data menjadi pengetahuan baru. Pengenalan pola dari data dicari dengan model perhitungan algoritmik sebagaimana statistic. Salah satu algoritma klasifikasi terbaik dan banyak digunakan untuk dataset berdimensi tinggi adalah KNN. Penelitian ini menggunakan dataset diabetes dari uci repository yang dirilis pada 2020. Hasil penelitian menunjukkan bahwa tingkat akurasi algoritma KNN untuk klasifikasi data diabetes adalah 92,50%. Hasil ini menunjukkan performa algoritma KNN baik dan layak digunakan.

Kata Kunci : Data mining, KNN, diabetes

## 1. PENDAHULUAN

### 1.1 Latar Belakang

Diabetes adalah salah satu penyakit paling mematikan di dunia. Pada kurun waktu Januari sampai dengan Desember 2019 tercatat ada 2 juta kematian yang diakibatkan oleh diabetes (WHO 2023). Beberapa kasus mematikan akibat diabetes adalah kesalahan pola makan pasien yang kurang sehat dan mengakibatkan kadar gula darah (glukosa) berlebih sehingga tubuh tidak mampu mengontrol glukosa dalam darah (Ejiyi et al. 2023). Penyakit diabetes tipe 2 bukanlah penyakit keturunan dan tidak dapat diturunkan orang tua kepada anaknya. Penanganan dini terhadap penyakit ini dapat mengurangi tingkat keparahan serta mengurangi resiko kamatian dari pasien.

Perkembangan ilmu computer membuat terobosan yang sangat baik dalam segala bidang, termasik salah satunya di bidang kesehatan. Data mining merupakan bidang ilmu computer yang dapat menemukan pola tertentu dari sebuah kumpulan dataset (Maimoon and Rokach 2010). Pola tersebut terbentuk dari sebuah hasil perhitungan algoritmik yang merupakan pengembangan dari beberapa perhitungan statistik (Han and Kamber 2006). Perhitungan algoritmik tersebut nantinya dapat dibandingkan satu sama lain, lalu dilakukan evaluasi guna mendapatkan model algoritma terbaik untuk sebuah dataset (Ian H Witten, Frank, and Hall 2011).

Selain prediksi, estimasi, klustering dan asosiasi, klasifikasi merupakan salah satu model yang dikerjakan atau tujuan data mining. Proses klasifikasi terbukti dapat menangani dataset yang bervariasi dan dengan jumlah yang besar. Di bidang kesehatan, klasifikasi pernah dilakukan untuk melakukan deteksi dini penyakit ginjal (Gamadarenda and Waspada 2018). Selain itu proses klasifikasi juga dilakukan untuk deteksi kanker payudara (Kurniawan and Ivandari 2017) serta deteksi dini diabetes tipe 2 (Aguilera-Venegas et al. 2023). Utnuk klasifikasi penyakit diabetes pada 2021 dilakukan komparasi algoritma menggunakan dataset PIMA. Hasilnya algoritma dengan performa terbaik adalah *soft voting classifier* dengan tingkat akurasi 79,08% (Kumari, Kumar, and Mittal 2021). Kemudian pada awal 2023 dilakukan komparasi serupa dengan menggunakan *Gradient Boost Machine* (GBM). Hasilnya GBM memperoleh tingkat akurasi terbaik yaitu 80,4% (Carpinteiro et al. 2023).

Proses klasifikasi tidak terlepas dari adanya data training dan data testing. Dengan pembagian data yang tepat dapat menghasilkan performa algoritma yang lebih baik. Model dan tipe data sangat mempengaruhi performa sebuah algoritma. Salah satu algoritma terbaik untuk klasifikasi adalah KNN (Wu et al. 2007)

Penelitian ini menggunakan dataset diabetes dari *uci repository*. Uci repository adalah web penyedia dataset public yang banyak digunakan untuk pengujian algoritma. Penelitian ini menggunakan algoritma KNN untuk klasifikasi dataset diabetes.

Proses validasi data dilakukan dengan menggunakan *10 folds cross validation.* Sedangkan proses evaluasi perhitungan dilakukan dengan menggunakan *confusion matrix.* Hasilnya algoritma KNN memiliki tingkat akurasi sebesar 92,50%. Performa algoritma KNN dalam penelitian ini termasuk baik dan memuaskan.

## 2 METODE PENELITIAN

Metode penelitian terbaik yang dapat digunakan dalam penelitian jenis ini adalah eksperimental. Proses perhitungan dilakukan dengan menggunakan aplikasi bantu rapid miner. Secara lebih terperinci metode penelitian yang dilakukan terbagi menjadi beberapa tahapan sebagaimana berikut:

### 2.1 Pengumpulan Data

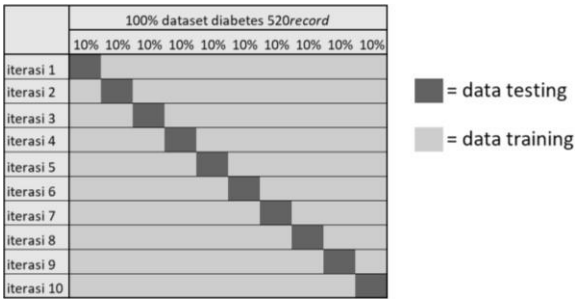
Tahapan awal dalam penelitian ini adalah pengumpulan data. Data yang digunakan adalah dataset public yang dapat diakses di https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset. Dataset ini memiliki 17 atribut dan 520 *record* data.

### 2.2 Analisa Atribut Data

Dalam tahapan analisa atribut data dipaparkan keseluruhan atribut dengan tipe datanya. Selain itu, dalam sebuah proses klasifikasi salah satu atribut wajib digunakan sebagai atribut label. Tipe atribut dalam proses data mining sederhana misalnya atribut dengan tipe numeric dan nominal. Tipe numeric adalah atribut yang didalamnya memiliki nilai yang dapat dihitung secara matematis. Sedangkan atribut nominal adalah atribut yang pada umumnya hanya memiliki nilai yang tidak dapat dihitung secara matematis antar satu dengan yang lain. Banyaknya atribut yang relevan dapat mempermudah proses perhitungan KNN (Alkaromi 2014).

### 2.3 Validasi

Validasi merupakan proses yang wajib ada dalam sebuah klasifikasi. Beberapa cara untuk melakukan validasi menyesuaikan dengan permasalahan dan proses klasifikasi yang dilakukan. *Cross validation* adalah proses validasi yang banyak digunakan dan terbukti memiliki kemampuan yang baik untuk proses klasifikasi (Ian H Witten. Eibe Frank. Mark A Hall 2011). Proses dalam *cross validation* ini adalah membagi data menjadi beberapa bagian lalu satu bagian dijadikan sebagai *data testing*, sisanya digunakan sebagai *data training.* Pembagian data ini disesuaikan dengan kebutuhan yang ada. Adapun yang proses yang paling banyak digunakan dalam penelitian klasifikasi adalah *10 folds cross validation* (Ivandari and Al Karomi 2021b). Gambar 1 merupakan gambaran proses dari *10 folds cross validation.*




**Gambar 1.** Gambaran *10 folds cross validation*

Dari gambar 1 diperlihatkan ada 10 kali iterasi atau perulangan. Iterasi 1 digunakan data 10% pertama atau 52 *recors* pertama dari keseluruhan 520 *records* untuk *testing* dan sisanya 468 *records* digunakan sebagai data *training.* Iterasi berikutnya dilakukan sampai keseluruhan data memperoleh porsi 1x digunakan sebagai data *testing.*

### 2.4 Perhitungan K-Nearest Neighbor

KNN merupakan algoritma yang menggunakan pembobotan seluruh atribut untuk mencari nilai *similarity.* Nilai *similarity* atau kedekatan ini digunakan untuk mencari label dari data *testing* yang dibandingkan sebelumnya dengan keseluruhan data *training.* Pada umumnya nilai yang digunakan antara 0 sampai dengan 1. Perkembangan algoritma KNN ini banyak dilakukan dan beberapa yang popular adalah mencocokkan kasus baru dan kasus yang lama dengan menggunakan pembobotan tertentu (Kusrini and Taufiq 2009). Pengaruh atribut dalam data sangatlah besar dalam keberhasilan algoritma KNN. Banyaknya atribut yang relevan dapat meningkatkan performa dan tingkat akurasi algoritma KNN (Ivandari 2014).

### 2.5 Evaluasi Algoritma

Evaluasi merupakan proses untuk menilai performa sebuah algoritma. Proses evaluasi banyak dilakukan dengan berbagai metode. Salah satu metode evaluasi terbaik dan banyak digunakan dalam penelitian klasifikasi adalah *confusion matrix. Confusion matrix* atau matrix kebingungan sebenarnya adalah prosentase dari proses klasifikasi yang sesuai dengan kondisi atau label sebenarnya (Ivandari and Al Karomi 2021a). Hasil akhir dari *confusion matrix* berupa prosentase yang juga digunakan sebagai nilai atau tingkat akurasi sebuah algoritma. Gambar 4 merupakan representasi dari *confusion matrix* (Gorunescu 2011).Proses evaluasi *confusion matrix* dapat dihitung dengan menggunakan rumus (1) berikut.

| CLASSIFICATION | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class = YES | Class = NO |
| OBSERVED CLASS | Class = YES | *a* (*true positive*-TP) | *b* (*false negative* -FN) |
| | Class = NO | *c* (*false positive*-FP) | *d* (*true negative*-TN) |

**Gambar 2.** Representasi *confusion matrix* (Gorunescu 2011)

$$\boldsymbol{accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

**a** = (*true positive*) = jumlah prediksi YES dengan label YES = klasifikasi benar

**b** = (*false negative*) = jumlah prediksi NO dengan label YES = klasifikasi salah

**c** = (*false positive*) = jumlah prediksi YES dengan label NO = klasifikasi salah

**d** = (*true negative*) = jumlah prediksi NO dengan label NO = klasifikasi benar

## 3. HASIL DAN PEMBAHASAN

Dalam penelitian ini dataset yang didapatkan dari uci repository sudah dalam bentuk csv. Data ini didapatkan dari 520 pasien Rumah Sakit Diabetes Sylhet, Bangladesh. Setelah adanya persetujuan dari Dokter terkait data ini baru didonasikan pada 7 November 2020 (Diabetes and Hospital in Sylhet 2020). Link download dataset dapat diakses dan diunduh di https://archive.ics.uci.edu/static/public/529/early+stage+diabetes+risk+prediction+dataset.zip. Tabel 1 merupakan metadata dari *early stage diabetes risk prediction dataset.*

**Tabel 1.** Metadata *early stage diabetes risk prediction dataset*

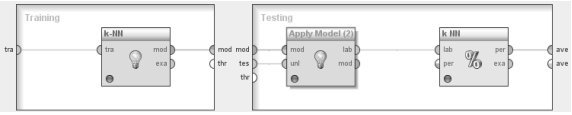
| Role | Name | Type | Statistics | Range | Missings |
|---|---|---|---|---|---|
| label | class | binominal | mode = Positive (320), least = Negative (200) | Positive (320), Negative (200) | 0.0 |
| regular | Age | integer | avg = 48.029 +/- 12.151 | [16.000 ; 90.000] | 0.0 |
| regular | Gender | binominal | mode = Male (328), least = Female (192) | Male (328), Female (192) | 0.0 |
| regular | Polyuria | binominal | mode = No (262), least = Yes (258) | No (262), Yes (258) | 0.0 |
| regular | Polydipsia | binominal | mode = No (287), least = Yes (233) | Yes (233), No (287) | 0.0 |
| regular | sudden weight loss | binominal | mode = No (303), least = Yes (217) | No (303), Yes (217) | 0.0 |
| regular | weakness | binominal | mode = Yes (305), least = No (215) | Yes (305), No (215) | 0.0 |
| regular | Polyphagia | binominal | mode = No (283), least = Yes (237) | No (283), Yes (237) | 0.0 |

| | | | | | |
|---|---|---|---|---|---|
| regular | Genital thrush | binominal | mode = No (404), least = Yes (116) | No (404), Yes (116) | 0.0 |
| regular | visual blurring | binominal | mode = No (287), least = Yes (233) | No (287), Yes (233) | 0.0 |
| regular | Itching | binominal | mode = No (267), least = Yes (253) | Yes (253), No (267) | 0.0 |
| regular | Irritability | binominal | mode = No (394), least = Yes (126) | No (394), Yes (126) | 0.0 |
| regular | delayed healing | binominal | mode = No (281), least = Yes (239) | Yes (239), No (281) | 0.0 |
| regular | partial paresis | binominal | mode = No (296), least = Yes (224) | No (296), Yes (224) | 0.0 |
| regular | muscle stiffness | binominal | mode = No (325), least = Yes (195) | Yes (195), No (325) | 0.0 |
| regular | Alopecia | binominal | mode = No (341), least = Yes (179) | Yes (179), No (341) | 0.0 |
| regular | Obesity | binominal | mode = No (432), least = Yes (88) | Yes (88), No (432) | 0.0 |

Dari metadata dataset diabetes pada tabel 1 terlihat ada 17 atribut dengan 1 atribut label yaitu atribut class. Atribut class memiliki tipe binominal dengan varian positif sebanyak 320 dan negative sebanyak 200 *record*. Atribut usia merupakan satu satunya atribut dengan tipe integer. Range usia dalam dataset ini adalah 16 sampai dengan 90 tahun. Selanjutnya 15 atribut regular lainnya memiliki tipe binominal sebagaimana ada di tabel 1. Data diabetes yang ada termasuk *balance* dengan rata rata variasi data tidak kurang dari 30%.

### 3.1 Hasil Perhitungan

Proses perhitungan klasifikasi menggunakan rapid miner. Dalam aplikasi rapid miner dataset, model algoritma, validasi dan evaluasi dapat di *drag and drop* sesuai kebutuhan user. Dataset yang ada sudah dalam bentuk csv sehingga tidak perlu dilakukan konversi di aplikasi rapid miner. Gambar 2 merupakan proses perhitungan dalam aplikasi rapid miner.




**Gambar 3.** Proses di rapid miner

Hasil dari proses perhitungan tersebut adalah sebuah *confusion matrix* sebagaimana terlihat pada tabel 2 berikut. Matrix tersebut menunjukkan tingkat akurasi dari algoritma KNN untuk klasifikasi diabetes adalah 92,50%.

**Tabel 2.** Hasil c*onfusion matrix*

| | **True positive** | **True negative** | **Class precission** |
|---|---|---|---|
| **Pred. positive** | 296 | 15 | 95,18% |
| **Pred. negative** | 24 | 185 | 88,52% |
| **Class recall** | 92,5% | 92,50% | |

Tabel 2 menunjukkan sebanyak 296 diprediksi positif dengan hasil sebenarnya positif, 185 diprediksi negative dengan hasil sebenarnya negative. Dengan kata lain hasil prediksi dari perhitungan algoritma KNN yang sesuai dengan label sebenarnya adalah sebanyak 481. Dengan total *record* sebanyak 520 tingkat akurasi dapat dihitung dengan 481/520 (481 dibagi 520). Hasilnya adalah 92,5%.

### 3.2 Pembahasan

KNN merupakan model algoritma yang dapat menangani data berdimensi tinggi. Selain itu KNN juga dapat melakukan perhitungan data numeric ataupun nominal. Ini terbukti dengan tingkat akurasi yang tinggi dalam klasifikasi data diabetes. Data ini memiliki 17 atribut dengan 520 *record*. Variasi data didalamnya ada numeric atau integer serta ada beberapa variasi nominal. Ini terbukti dengan performa yang tetap baik dengan mencatatakan tingkat akurasi sebesar 92,5%.

## 4. KESIMPULAN DAN SARAN

### 4.1 Kesimpulan

Klasifikasi data diabetes menggunakan KNN menghasilkan tingkat akurasi sebesar 92,50%. Performa KNN dalam penelitian ini tergolong baik. Data yang digunakan adalah dataset dari *uci repository* yang dirilis pada akhir 2020. Selanjutnya untuk meningkatkan performa dapat dilakukan pemilihan fitur atau *feature selection* untuk menghilangkan altribut yang kurang berpengaruh dalam klasifikasi.

## 5. DAFTAR PUSTAKA


Aguilera-Venegas, Gabriel, Amador López-Molina, Gemma Rojo-Martínez, and José Luis Galán-García. 2023. "Comparing and Tuning Machine Learning Algorithms to Predict Type 2 Diabetes Mellitus." *Journal of Computational and Applied Mathematics* 427: 115115. https://doi.org/10.1016/j.cam.2023.115115.

Alkaromi, M Adib. 2014. "Information Gain Untuk Pemilihan Fitur Pada Klasifikasi Heregistrasi Calon Mahasiswa Dengan Menggunakan K-NN."

Carpinteiro, César, João Lopes, António Abelha, and Manuel Filipe Santos. 2023. "A Comparative Study of Classification Algorithms for Early Detection of Diabetes." *Procedia Computer Science* 220: 868–73. https://doi.org/10.1016/j.procs.2023.03.117.

Diabetes, Sylhet, and Bangladesh Hospital in Sylhet. 2020. "Early Stage Diabetes Risk Prediction Dataset." 2020. https://doi.org/10.24432/C5VG8H.

Ejiyi, Chukwuebuka Joseph, Zhen Qin, Joan Amos, Makuachukwu Bennedith Ejiyi, Ann Nnani, Thomas Ugochukwu Ejiyi, Victor Kwaku Agbesi, Chidimma Diokpo, and Chidinma Okpara. 2023. "A Robust Predictive Diagnosis Model for Diabetes Mellitus Using Shapley-Incorporated Machine Learning Algorithms." *Healthcare Analytics* 3 (December 2022): 100166. https://doi.org/10.1016/j.health.2023.100166.

Gamadarenda, ikhsan wisnuadji, and Indra Waspada. 2018. "Implementasi Data Mining Untuk Deteksi Penyakit Ginjal Kronis (Pgk) Menggunakan K-Nearest Neighbor (Knn) Dengan Backward Elimination" 7 (2): 417–26. https://doi.org/10.25126/jtiik.202071896.

Gorunescu, Florin. 2011. *Data Mining: Concepts; Models and Techniques*. Springer.

Han, Jiawei, and Micheline Kamber. 2006. "Data Mining: Concepts and Techniques Second Edition" 40 (6): 9823. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Ian H Witten. Eibe Frank. Mark A Hall. 2011. *Data Mining 3rd*.

Ivandari. 2014. "Improved Performance Algorithm K-Nearest Neighbor Classification in High Dimension Data." *IC Tech* IX-April 2: 5–9.

Ivandari, and M. Adib Al Karomi. 2021a. "Algoritma K-NN Untuk Klasifikasi Dataset Covid-19 Survillance." *IC Tech* 16 (1): 12–15. https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137.

———. 2021b. "Classification of Covid-19 Survillance Datasets Using the Decision Tree Algorithm." *Jaict* 6 (1): 44–49. https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.

Kumari, Saloni, Deepika Kumar, and Mamta Mittal. 2021. "An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Using Soft Voting Classifier." *International Journal of Cognitive Computing in Engineering* 2 (January): 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001.

Kurniawan, M. Faisal, and Ivandari. 2017. "Komparasi Algoritma Data Mining Untuk Klasifikasi Kanker Payudara." *IC Tech* I April 20: 1–8.

Kusrini, and Luthfi Emha Taufiq. 2009. *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Maimoon, Oded, and Lior Rokach. 2010. *Data Mining and Knowledge Discovery Handbook*. Vol. 40. Springer. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14*, 106–13. https://doi.org/10.1145/2643604.2643631.

WHO. 2023. "Diabetes." 2023. https://www.who.int/news-room/fact-sheets/detail/diabetes.

Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Vol. 40. Elsevier. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining*. *Knowledge and Information Systems*. Vol. 14. https://doi.org/10.1007/s10115-007-0114-2.