Klasifikasi Penyakit Diabetes Menggunakan Algoritma *K-Nearest Neighbor* Optimasi *K-Fold Cross Validation*

Diah Siti Fatimah Azzahrah[1], Alamsyah[2]

[1,2]Jurusan Ilmu Komputer, FMIPA, Universitas Negeri Semarang
Email: [1]diah.azzahrah@gmail.com, [2]alamsyah@mail.unnes.ac.id

Abstrak

Diabetes merupakan penyakit yang terjadi karena adanya peningkatan kadar gula darah yang melebihi batas normal. Penyakit ini juga berisiko pada komplikasi. *World Health Organization* (WHO) menyatakan jumlah penderita diabetes terus mengalami peningkatan dari tahun ke tahun dan berpotensi meningkat lebih tinggi dibandingkan dengan tahun-tahun sebelumnya. Melihat tingginya angka masyarakat yang terkena diabetes maka perlu dilakukan pendeteksian penyakit diabetes secara dini sebagai upaya untuk meminimalisir munculnya komplikasi dan kematian. Penelitian ini menggunakan data mining dengan metode klasifikasi. Tujuan dari penelitian ini untuk mendapatkan hasil akurasi dan melihat kinerja algoritma. Algoritma klasifikasi yang digunakan adalah algoritma *K-Nearest Neighbor* (KNN). *Dataset* yang digunakan adalah PIDD. Dalam penelitian ini data dibagi menjadi 80:20 untuk data *training* dan data *testing*. Penelitian ini melakukan *k-fold cross validation*. Berdasarkan hasil penelitian menunjukkan bahwa metode yang digunakan berhasil memperoleh akurasi yang tinggi dalam memprediksi penyakit diabetes sebesar 78,10%.
Kata Kunci: Diabetes Mellitus, *Data Mining*, Klasifikasi, KNN, *k-fold cross validation*

Abstract

Diabetes is a disease that occurs due to an increase in blood sugar levels that exceed normal limits. This disease is also at risk for complications. The World Health Organization (WHO) states that the number of people with diabetes continues to increase from year to year and has the potential to increase higher than in previous years. Seeing the high number of people affected by diabetes, the detection of diabetes is an effort to detect it early, minimize complications and death. In this research implement data mining with classification methods. This study aims to obtain accuracy results and see the performance of the algorithm. The classification algorithm used is the K-Nearest Neighbor (KNN) algorithm. The dataset used is PIDD. The dataset is divided into 80:20 for training data and testing data. This study performs k-fold cross validation. Based on the results of the study showed that the method used succeeded in obtaining high accuracy in predicting diabetes by 78.10%..
***Keyword**: Diabetes Mellitus, Data Mining, Classification, KNN, k-fold cross validation*

1. PENDAHULUAN

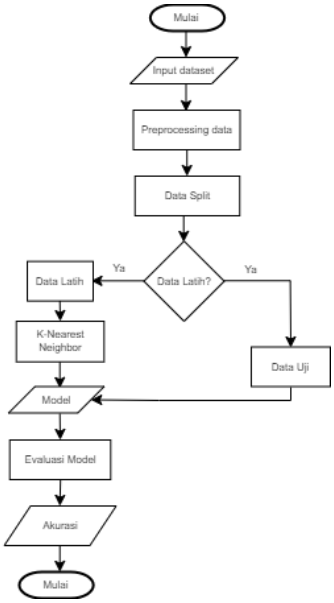
Diabetes merupakan penyakit yang disebabkan karena adanya peningkatan kadar gula darah (glukosa darah) dalam tubuh yang melampaui batas normal [1]. Penyakit diabetes dapat terjadinya komplikasi yang sangat berbahaya terhadap penderita diabetes. Menurut *World Health Organization* (WHO) bahwa jumlah penderita diabetes terus mengalami peningkatan dari tahun ke tahun dan berpotensi meningkat lebih tinggi dibandingkan dengan tahun-tahun sebelumnya [2]. Penyakit diabetes masuk ke dalam penyebab kematian terbanyak di dunia. Jumlah pasien yang

meninggal akibat diabetes di dunia telah mencapai 6,7 juta jiwa pada tahun 2021 [3]. Oleh karena itu, diperlukan suatu teknologi yang dapat mendeteksi suatu penyakit sehingga penderita dapat ditangani lebih awal dengan pengobatan, serta mampu meminimalisir risiko untuk terkena komplikasi hingga kematian. Penyakit diabetes dapat dideteksi dengan menggunakan data historis pasien yang berisikan informasi mengenai gejala-gejala atau kondisi suatu pasien. Adapun bentuk nyata dari pengimplementasian proses teknologi tersebut adalah *data mining*. *Data mining* merupakan proses yang mempelajari alur kerja data atribut yang saling berkaitan untuk menemukan sebuah pola dari data yang berukuran besar [4], [5]. Salah satu teknik yang paling sering digunakan adalah teknik klasifikasi. Klasifikasi adalah suatu proses untuk menemukan pola dengan membangun model klasifikasi dari algoritma yang digunakan berdasarkan variabel dependen dan variabel independen [6]. Berbagai penelitian terkait deteksi penyakit diabetes telah dilakukan dengan menggunakan algoritma *k-nearest neighbor*, *support vector machine*, *naive bayes*, dan algoritma lainnya.

Pada penelitian [7] metode klasifikasi KNN diaplikasikan pada *dataset* penderita penyakit diabetes. Akan tetapi, hasil yang didapatkan tidak cukup baik karena data yang digunakan jumlahnya cukup kecil. Adapun akurasi yang di dapatkan sebesar 39% pada nilai k=3. Selain itu, pada penelitian tersebut juga belum menerapkan *cross validation*. Penelitian yang turut membahas diabetes yaitu [8] menggunakan algoritma *support vector machine*, *naive bayes*, dan *decision tree*. Pada penelitian tersebut nilai akurasi yang didapatkan sebesar 76,30% untuk algoritma *naive bayes*. Akan tetapi, penelitian ini juga belum mengimplementasikan *cross validation*. Sedangkan *cross validation* merupakan teknik tambahan dari *data mining* untuk memperoleh hasil akurasi yang lebih maksimal. Metode *k-nearest neighbor* memiliki fungsi untuk prediksi maupun klasifikasi terhadap suatu objek berdasarkan nilai k tetangga terdekat [9], [10], [11]. Oleh karena itu, pada penelitian yang diusulkan akan melakukan deteksi dan klasifikasi penyakit diabetes menggunakan algoritma *K-Nearest Neighbor* dengan diimplementasikan *k-fold cross validation* untuk meningkatkan hasil akurasi.

2. METODE

Pada metode penelitian ini akan menjelaskan serangkaian dari tahapan yang akan dilakukan dalam menyelesaikan permasalahan. Tahapan tersebut dapat dilihat dalam bentuk *flowchart* pada Gambar 1.



Gambar 1. Tahapan penelitian

Adapun penjelasan lebih lanjut mengenai tahapan penelitian

1. *Dataset*
   Data yang digunakan pada penelitian ini merupakan *Pima Indians Diabetes Database* (PIDD) yang bersifat publik didapatkan dari *link* *https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*. *Dataset* ini berisikan 768 *records* dengan 9 variabel [12].
2. *Preprocessing Data*
   Tahap *preprocessing data* yaitu proses untuk mempersiapkan data sebelum masuk ke tahap pemodelan menggunakan algoritma. Proses tersebut terdiri dari *data cleaning*, *handling outliers*, *feature selection,* dan *feature scaling*.
3. *Data Split*
   Data split merupakan pembagian *dataset* yang terdiri dari *data training* dan *data testing*. Pada penelitian ini akan dilakukan pembagian dengan besaran *data training* sebesar 80% dan *data testing* sebesar 20%.
4. Model
   Model yang akan dibuat pada tahap ini dengan menggunakan algoritma *k-nearest neighbor* (KNN). Model ini akan dibuat setelah dilakukannya proses *preprocessing data* dan *data split*.
5. Evaluasi Model
   Tahap dari evaluasi model ini dilakukan untuk mendapatkan nilai akurasi algoritma yang digunakan pada penelitian ini disertai dengan implementasi *k-fold cross validation*.

3. HASIL DAN PEMBAHASAN

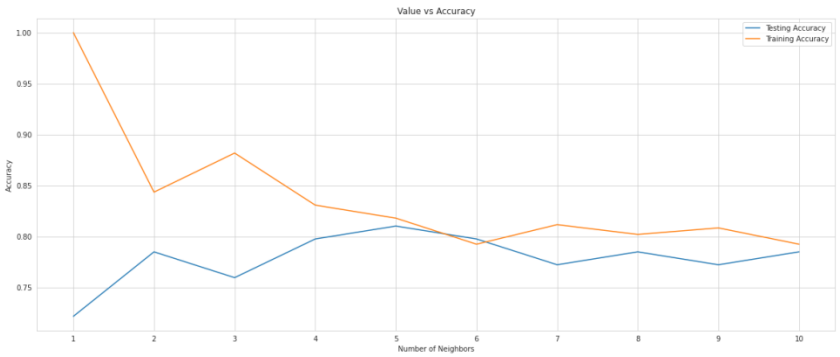
Penelitian ini menggunakan *dataset Pima Indians Diabetes Database*. Tahapan awal yang perlu dilakukan adalah melakukan proses *data cleaning* yang merupakan bagian dari tahapan *prepocessing data.* Hasil dari proses *data cleaning* dapat dilihat pada Gambar 2.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
3	1	89.0	66	23	94.0	28.1	0.167	21	0
4	0	137.0	40	35	168.0	43.1	2.288	33	1
6	3	78.0	50	32	88.0	31.0	0.248	26	1
8	2	197.0	70	45	543.0	30.5	0.158	53	1
13	1	189.0	60	23	846.0	30.1	0.398	59	1
...	...	...	...	...	...	...	...	...	...
753	0	181.0	88	44	510.0	43.3	0.222	26	1
755	1	128.0	88	39	110.0	36.5	1.057	37	1
760	2	88.0	58	26	16.0	28.4	0.766	22	0
763	10	101.0	76	48	180.0	32.9	0.171	63	0
765	5	121.0	72	23	112.0	26.2	0.245	30	0

392 rows × 9 columns

Gambar 2. *Data cleaning*

Langkah selanjutnya adalah menentukan variabel dependen dan independen. Pada penelitian ini variabel independen yang digunakan adalah *Pregnancies*, *Glucose*, BMI, dan *Age*. Sedangkan *Outcome* merupakan variabel dependen. Setelah dilakukannya *preprocessing data*, *data split*, dan pembuatan model. Maka dapat dilihat hasil dari grafik model *testing accuracy* dan *training accuracy* pada Gambar 3.



Gambar 3. Grafik evaluasi model

Setelah mendapatkan ilustrasi dari model *k-nearest neighbor*. Maka dilakukan implementasi dari *k-fold cross validation* untuk mendapatkan hasil yang lebih akurat. Pada Gambar 4 merupakan hasil dari eksperimen pada penelitian ini.

```
k = 1 -> Test Accuracy 0.7215189873417721  & Accuracy Cross Validation: 0.7352564102564102
k = 2 -> Test Accuracy 0.7848101265822784  & Accuracy Cross Validation: 0.7170512820512821
k = 3 -> Test Accuracy 0.759493670886076  & Accuracy Cross Validation: 0.7401923076923077
k = 4 -> Test Accuracy 0.7974683544303798  & Accuracy Cross Validation: 0.7428205128205129
k = 5 -> Test Accuracy 0.810126582278481  & Accuracy Cross Validation: 0.7810897435897436
k = 6 -> Test Accuracy 0.7974683544303798  & Accuracy Cross Validation: 0.7657051282051281
k = 7 -> Test Accuracy 0.7721518987341772  & Accuracy Cross Validation: 0.7580769230769231
k = 8 -> Test Accuracy 0.7848101265822784  & Accuracy Cross Validation: 0.7555128205128205
k = 9 -> Test Accuracy 0.7721518987341772  & Accuracy Cross Validation: 0.7656410256410255
k = 10 -> Test Accuracy 0.7848101265822784  & Accuracy Cross Validation: 0.7631410256410256
```

Gambar 4. Hasil akurasi dan *k-fold cross validation*

4. SIMPULAN

Berdasarkan hasil pembahasan dan penelitian yang telah dilakukan, algoritma *k-nearest neighbor* (KNN) menghasilkan akurasi tertinggi dengan mengimplementasi *k-fold cross validation* sebesar 78,10% pada nilai $k = 5$.

5. REFERENSI

[1] Suryanegara, G. A. B., Adiwijaya, dan Purbolaksono, M. D. 2021. Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*. Vol. 5(1): 114-112.

[2] Najib, A., Nurcahyono, D., dan Setiawan, R. P. P. 2019. Klasifikasi Diagnosa Penyakit Diabetes Mellitus (Dm) Menggunakan Algoritma C4.4. *Just TI (Jurnal Sains Terapan Teknologi Informasi)*. Vol. 11(2): 47-58.

[3] Putry, N. M. 2022. Komparasi Algoritma KNN dan Naïve Bayes untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus. *EVOLUSI: Jurnal Sains dan Manajemen.* Vol. 10(1): 45-57.

[4] Ridwan, A. 2020. Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal Sistem Komputer dan Kecerdasan Buatan.* Vol. 4,(1): 15–21.

[5] Afif, A. 2020. Penerapan Algoritma Naïve Bayes untuk Klasifikasi Penyakit Diabetes Mellitus di Rumah Sakit Aisyiah. Jurnal Ilmu Komputer dan Matematika. Vol. 1(1): 40-46.

[6] Abdurrahman, G. 2022. Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier. Jurnal Sistem dan Teknologi Informasi. Vol. 7(1): 59-66.

[7] Argina, A. M. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. Indonesian Journal of Data and Science. Vol. 1(2), 29-33.

[8] Sisodia, D., dan Sisodia, D. S. 2018. Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science.* Vol. 132: 1578-1585.

[9] Lutfi, M. 2019. Implementasi Metode K-Nearest Neighbor dan Bagging untuk Klasifikasi Mutu Produksi Jagung. *AGROMIX.* Vol. 10(2): 130-137.

[10] Nurida, R., Sugiharti, E., dan Alamsyah, A. (2019). Implementation of Fuzzy K-Nearest Neighbor Method in Decision Support System for Identification of Under-five Children Nutritional Status Based on Anthropometry Index. *Journal of Advances in Information Systems and Technology.* Vol. 1(1): 83-89.

[11] Dinariyah, I. 2021. Accuracy Enhancement in Face Recognition using 1D-PCA & 2D-PCA based on Multilevel Reverse-biorthogonal Wavelet Transform with KNN Classifier. *Journal of Physics: Conference Series.* Vol. 1918(4).

[12] Learning, U. M. 2016. Pima Indians Diabetes Database. *Kaggle.*