Penerapan Metode K-Nearest Neighbors (K-NN) untuk Klasifikasi Risiko Diabetes pada Dataset Kesehatan

Disusun Oleh:

Emelsha Viadra 152022056
Naufal Zaidan 152022168

Prodi Informatika
Fakultas Teknologi Industri
Institut Teknologi Nasional
Bandung

2024

K-Nearest Neighbors (K-NN) adalah algoritma *supervised learning* yang banyak digunakan dalam klasifikasi dan regresi. Algoritma ini bekerja dengan memanfaatkan kedekatan atau kemiripan antara data baru dengan data yang sudah ada untuk menentukan kelas atau nilai prediksinya. Dalam penerapannya, K-NN dimulai dengan menetapkan parameter K, yang merupakan jumlah tetangga terdekat yang akan dipertimbangkan saat mengklasifikasikan data baru. Pemilihan nilai K ini sangat penting, karena K yang terlalu kecil bisa membuat model terlalu peka terhadap data tertentu (*overfitting*), sementara K yang terlalu besar bisa membuat model kehilangan sensitivitas terhadap pola lokal pada data.

Proses utama K-NN melibatkan perhitungan jarak antara data baru dan setiap data dalam dataset yang sudah ada. Jarak ini biasanya dihitung dengan menggunakan metrik seperti *Euclidean distance*, *Manhattan distance*, atau *Minkowski distance*. *Euclidean distance* adalah salah satu metode yang paling umum digunakan, di mana jarak diukur secara lurus antara dua titik dalam ruang fitur. Untuk dataset yang memiliki fitur dengan dimensi yang berbeda-beda, normalisasi data sering kali diperlukan untuk memastikan jarak tersebut relevan.

Setelah semua jarak dihitung, algoritma kemudian memilih K data terdekat, yaitu data yang memiliki jarak terkecil terhadap data baru. Pada tahap akhir, algoritma menentukan kelas data baru berdasarkan mayoritas kelas dari K tetangga terdekat tersebut. Misalnya, jika mayoritas dari K tetangga berada dalam kelas diabetes positif, maka data baru juga akan diklasifikasikan sebagai diabetes positif. Kelebihan utama dari K-NN adalah kesederhanaannya, karena tidak memerlukan proses pelatihan yang rumit. Namun, kelemahannya adalah algoritma ini bisa menjadi lambat jika ukuran dataset sangat besar, karena harus menghitung jarak antara data baru dengan setiap data dalam dataset.

Sejarah algoritma **K-Nearest Neighbors** (K-NN) berakar pada perkembangan awal kecerdasan buatan dan analisis data statistik. Meskipun konsep dasar dari algoritma ini telah dikenal sejak awal abad ke-20, perkembangan signifikan dalam teori klasifikasi dan pengenalan pola baru terjadi pada pertengahan hingga akhir 1900-an.

Berikut adalah beberapa poin penting dalam sejarah algoritma K-NN:

1. **1951: Penemuan oleh Evelyn Fix dan Joseph Hodges**

- Algoritma K-NN pertama kali diperkenalkan oleh **Evelyn Fix** dan **Joseph Hodges** pada tahun 1951 dalam makalah mereka berjudul "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties".
- Mereka bekerja di Universitas California, Berkeley, dan fokus utama penelitian mereka adalah pengembangan metode non-parametrik untuk klasifikasi.
- Makalah ini membahas penggunaan jarak sebagai dasar untuk mengklasifikasikan data baru dengan melihat tetangga terdekat dalam kumpulan data yang telah diberi label.

2. **1967: Penerapan dalam Pengenalan Pola**

- Pada tahun 1967, algoritma K-NN semakin populer ketika digunakan dalam buku berjudul **"The Theory of Pattern Recognition"** oleh Thomas M. Cover dan Peter E. Hart.
- Cover dan Hart mengembangkan teori dasar dari K-NN dan menunjukkan bahwa metode ini bisa sangat efektif untuk pengenalan pola tanpa memerlukan asumsi yang rumit mengenai distribusi data.
- Penelitian mereka menginspirasi banyak ilmuwan lain untuk mengembangkan metode klasifikasi berbasis tetangga terdekat.

3. **1970-an: Perkembangan di Bidang Pengklasifikasian**

- Algoritma K-NN mulai banyak digunakan dalam berbagai aplikasi pengklasifikasian, terutama dalam bidang statistik dan pembelajaran mesin.
- K-NN diadopsi sebagai pendekatan sederhana namun kuat untuk klasifikasi data. Aplikasi utamanya adalah di bidang bioinformatika, pengenalan wajah, dan pengenalan tulisan tangan.

4. **1980-1990-an: Penggunaan dalam Komputasi**

- Pada era ini, K-NN semakin populer dengan munculnya komputasi yang lebih cepat dan lebih terjangkau.
- Penggunaan K-NN meluas ke bidang yang lebih kompleks seperti **komputer vision**, **pengenalan suara**, dan **diagnosis medis**.

5. **2000-an: Kebangkitan Bersama Pembelajaran Mesin**

- Dengan berkembangnya bidang **pembelajaran mesin** (machine learning), K-NN menjadi salah satu algoritma dasar yang sering diajarkan dan digunakan dalam pemodelan awal data.
- K-NN banyak digunakan untuk klasifikasi data di berbagai industri, seperti keuangan, kesehatan, dan perdagangan elektronik.

6. Hingga Saat Ini

- K-NN dianggap sebagai algoritma yang mudah dipahami dan diterapkan, yang tetap relevan hingga kini dalam berbagai aplikasi pembelajaran mesin.
- Meskipun algoritma ini sederhana dan memiliki beberapa keterbatasan, seperti rentan terhadap **overfitting** dan performa yang lambat pada dataset yang besar, K-NN masih populer sebagai metode awal dalam eksplorasi data.

Dalam algoritma K-Nearest Neighbors (K-NN), salah satu metrik yang umum digunakan untuk mengukur jarak antar data adalah Euclidean Distance. Metode ini menghitung jarak lurus antara dua titik dalam ruang fitur, dan sering digunakan karena kesederhanaannya dalam mengukur kedekatan antar titik.

Rumus Euclidean Distance antara dua titik x dan y dengan fitur x1, x2, ..., xn dan y1, y2, ..., yn adalah:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Penjelasan:

- x dan y adalah dua data atau titik yang jaraknya akan dihitung.
- xi dan yi adalah nilai fitur ke-i untuk titik x dan y.
- n adalah jumlah fitur yang ada dalam dataset.

Contoh kasus 1:

Diabetes merupakan salah satu penyakit yang banyak dialami oleh masyarakat, dan deteksi dini sangat penting untuk mencegah komplikasi lebih lanjut. Dalam studi ini, kita akan menggunakan algoritma K-Nearest Neighbors (K-NN) untuk mengklasifikasikan apakah seseorang berisiko terkena diabetes berdasarkan beberapa faktor kesehatan.

Tabel 1.Data Sampel

DATA SAMPEL

No	Gula Darah	Insulin	BMI	Usia	Diabet?
1	89	94	28,1	21	Tidak
2	137	168	43,1	33	Ya
3	78	88	31	26	Ya
4	197	543	30,5	53	Ya
5	189	846	30,1	59	Ya
6	166	175	25,8	51	Ya
7	118	230	45,8	31	Ya
8	103	83	43,3	33	Tidak
9	115	96	34,6	32	Ya
10	126	235	39,3	27	Tidak
11	143	146	36,6	51	Ya
12	125	115	31,1	41	Ya
13	97	140	23,2	22	Tidak
14	145	110	22,2	57	Tidak
15	158	245	31,6	28	Ya
16	88	54	24,8	22	Tidak
17	103	192	24	33	Tidak
18	111	207	37,1	56	Ya
19	180	70	34	26	Tidak
20	171	240	45,4	54	Ya
21	103	82	19,4	22	Tidak
22	101	36	24,2	26	Tidak
23	88	23	24,4	30	Tidak
24	176	300	33,7	58	Ya
25	150	342	34,7	42	Tidak
26	100	71	38,5	26	Tidak
27	110	125	32,4	27	Tidak

Penyelesaian:

Dari data tersebut kita ambil 3 data untuk menjadi acuan pengujian terhadap seluruh data

Tabel 2.Data Uji

DATA UJI

Gula Darah	Insulin	BMI	Usia	Diabet?
187	304	37,7	41	Ya
93	64	28,7	23	Tidak
155	495	34	46	Ya

Tabel 3. Hasil dari pengujian data 1

DATA UJI 1

Jarak Euclidian	Urutan	Diabet?
232,80	17	
145,22	10	
242,50	22	
239,62	21	
542,36	27	
131,62	8	
101,99	6	
236,63	19	
220,31	16	
93,17	5	Tidak
164,32	11	
199,02	14	
188,59	12	
199,74	15	
67,29	3	Ya
269,87	24	
140,90	9	
124,14	7	
234,61	18	
67,68	4	Ya
238,82	20	
282,18	25	
298,43	26	
21,02	1	Ya
53,13	2	Tidak
249,17	23	
195,43	13	

Tabel 3. Hasil dari pengujian data 2

DATA UJI 2

Jarak Euclidian	Urutan	Diabet?
30,34	7	
114,28	16	
28,55	5	Ya
491,08	26	
788,69	27	
135,80	18	
168,93	20	
27,82	4	Tidak
40,30	8	
174,52	21	
100,35	15	
62,89	10	
76,31	12	
77,58	13	
192,40	22	
11,88	1	Tidak
128,86	17	
148,10	19	
87,42	14	
195,70	23	
22,62	3	Tidak
29,62	6	
42,11	9	
252,66	24	
284,48	25	
14,25	2	Tidak
63,56	11	

Tabel 3. Hasil dari pengujian data 3

DATA UJI 3

Jarak Euclidian	Urutan	Diabet?
407,21	19	
327,88	11	
414,71	20	
64,26	1	Ya
352,90	13	
320,33	10	
268,25	7	
415,58	21	
401,24	18	
262,35	6	
349,25	12	
381,23	16	
360,67	14	
385,47	17	
250,68	4	Ya
446,80	25	
307,87	9	
291,53	8	
426,20	23	
255,88	5	Ya
417,21	22	
462,70	26	
477,10	27	
196,49	3	Ya
153,14	2	Tidak
428,04	24	
373,21	15	

Dalam pengujian ini, kita menggunakan algoritma K-Nearest Neighbors (K-NN) dengan K=5 untuk menentukan risiko diabetes pada tiga data uji berdasarkan jarak Euclidean terhadap data sampel.

Data Uji 1:

Lima tetangga terdekat untuk data uji pertama dengan hasil klasifikasi adalah:

- **Jarak 21,02** dari data sampel No. 1 - *"Ya"*
- **Jarak 53,13** dari data sampel No. 2 - *"Tidak"*
- **Jarak 67,29** dari data sampel No. 3 - *"Ya"*
- **Jarak 67,68** dari data sampel No. 4 - *"Ya"*
- **Jarak 93,17** dari data sampel No. 5 - *"Tidak"*

Dari lima tetangga terdekat, tiga di antaranya memiliki hasil *"Ya"* dan dua memiliki hasil *"Tidak"*. Karena mayoritas hasil adalah *"Ya"*, data uji pertama diklasifikasikan sebagai **berisiko diabetes**.

Data Uji 2:

Lima tetangga terdekat untuk data uji kedua dengan hasil klasifikasi adalah:

- **Jarak 11,88** dari data sampel No. 1 - *"Tidak"*
- **Jarak 14,25** dari data sampel No. 2 - *"Tidak"*
- **Jarak 22,62** dari data sampel No. 3 - *"Tidak"*
- **Jarak 27,82** dari data sampel No. 4 - *"Tidak"*
- **Jarak 28,55** dari data sampel No. 5 - *"Ya"*

Dari lima tetangga terdekat, empat di antaranya memiliki hasil *"Tidak"* dan satu memiliki hasil *"Ya"*. Dengan demikian, data uji kedua diklasifikasikan sebagai **tidak berisiko diabetes**.

Data Uji 3:

Lima tetangga terdekat untuk data uji ketiga dengan hasil klasifikasi adalah:

- **Jarak 64,26** dari data sampel No. 1 - *"Ya"*
- **Jarak 153,14** dari data sampel No. 2 - *"Tidak"*
- **Jarak 196,49** dari data sampel No. 3 - *"Ya"*
- **Jarak 250,68** dari data sampel No. 4 - *"Ya"*

- **Jarak 255,88** dari data sampel No. 5 - *"Ya"*

Dari lima tetangga terdekat, empat memiliki hasil *"Ya"* dan satu memiliki hasil *"Tidak"*. Oleh karena itu, data uji ketiga diklasifikasikan sebagai **berisiko diabetes**.

Ringkasan

Dengan menggunakan K=5, hasil klasifikasi menunjukkan bahwa:

- Data uji pertama dan ketiga diklasifikasikan sebagai **berisiko diabetes**.
- Data uji kedua diklasifikasikan sebagai **tidak berisiko diabetes**.