Laporan Tugas Pemrograman 3 (*K-Nearest Neighbor*)

Mata Kuliah Kecerdasan Buatan

dibuat oleh Dimas Bayu Nugraha (1301181096)

1. Permasalahan

K-Nearest Neighbor adalah salah satu algoritma *learning* dalam ranah kecerdasan buatan. Algoritma ini bekerja dengan cara mencari data mana saja yang mirip dengan data yang dicari menggunakan rumus perhitungan jarak. Dalam algoritma ini k yang dimaksud adalah banyak data yang berdekatan dengan data yang dicari untuk dievaluasi kemudian. *K-Nearest Neighbor* dapat digunakan baik dalam kasus klasifikasi maupun regresi.

Dalam laporan ini digunakan dataset *Pima Indians Diabetes*. Dataset ini berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases*. Dataset ini berisi data pasien wanita yang berumur minimal 21 tahun dimana terdapat dua kelas yaitu pasien tersebut mengidap diabetes atau tidak mengidap diabetes. Dataset ini memiliki 9 kolom yang secara rinci yaitu sebagai berikut:

- Pregnancies : Berapa kali orang tersebut pernah hamil.
- Glucose : Kandungan glukosa dalam darah.
- BloodPressure : Tekanan darah diastolik dalam mm Hg.
- SkinThickness : Ketebalan lipatan kulit trisep dalam mm.
- Insulin : Kandungan insulin dalam mu U/ml.
- BMI : Indeks masa tubuh dalam kg/m^2.
- DiabetesPedigreeFunction : Fungsi keturunan diabetes.
- Age : Umur dalam satuan tahun.
- Outcome : Variabel kelas apakah seseorang diabetes atau tidak.

2. Analisis

Beberapa hal yang diobservasi dalam laporan ini adalah sebagai berikut:

2.1. Prapemrosesan Data

Dalam laporan ini digunakan teknik prapemrosesan data dengan menggantikan nilai data yang dianggap hilang. Pada dataset ini terdapat fitur yang dianggap beberapa nilai data-nya hilang yaitu pada fitur Glucose, BloodPressure, BMI, dan SkinThickness. Alasan nilai data dari fitur tersebut dianggap hilang adalah karena pada nilai data tersebut nol sedangkan tidak mungkin kadar gula darah, tekanan darah, berat badan, maupun ketebalan kulit seseorang bernilai nol. Cara untuk mengatasi kasus ini adalah dengan menggunakan metode *knn regressor imputer* yaitu mengisi nilai data yang dianggap hilang tersebut berdasarkan rata-rata dari nilai data lain yang mirip dengan data tersebut.

Selain nilai data yang dianggap hilang nilai data yang termasuk pencilan juga akan digantikan dengan nilai baru menggunakan teknik yang sama yaitu *knn regressor imputer*. Setelah itu fitur Pregnancies akan dihilangkan karena dapat meningkatkan akurasi di akhir (cara mengetahuinya yaitu dengan *trial and error*). Setelah menghilangkan fitur tersebut dataset akan dilakukan normalisasi untuk mencegah variasi ukuran yang dimiliki tiap fitur. Teknik ini sangat berguna dalam *K-Nearest Neighbor* yang mengandalkan jarak antar data untuk mencari data yang mirip. Akibat dari tidak menggunakan normalisasi adalah jarak akan menjadi tidak berarti karena mungkin suatu fitur akan lebih mendominasi dibanding fitur lainnya.

2.2. *K-Fold Cross Validation*

Dataset yang sudah di praproses kemudian di-*fold* menjadi 5-*fold* yang setiap *fold*-nya dibagi menjadi *data train* dan *data validation*.

2.3. Rumus Jarak

Dalam laporan ini akan dilakukan observasi dengan menggunakan beberapa rumus jarak yaitu Euclidean, Manhattan, Chebyshev, Chi-square, dan Cosine Similarity.

2.4. Pemilihan Nilai K

Dalam laporan ini akan dilakukan observasi k dengan jarak [1, 100].

3. Hasil Observasi

Dengan menggunakan observasi yang sudah direncanakan sebelumnya, didapatkan hasil sebagai berikut:

Table 1 : Hasil observasi akurasi terbaik menggunakan 5 rumus jarak dan 100 k

	Rumus Jarak				
	Euclidean	Manhattan	Chebyshev	Chi-square	Cosine Sim.
k	23	45	16	35	11
Akurasi *Fold*	78.64 %	78.39 %	78.12 %	75.39 %	72.66 %

Berdasarkan hasil observasi didapat akurasi fold terbaik yaitu 78.64 % dengan data yang sudah di praproses dan di-*fold* dengan parameter rumus jarak menggunakan Euclidean dan nilai k sama dengan 23.

4. Lampiran

```
Accuracy = 0.7864103217044394 with k = 23 and using Euclidean distance
```

Figure 1 : Hasil output akhir program menggunakan k = 23 dan rumus jarak euclidean

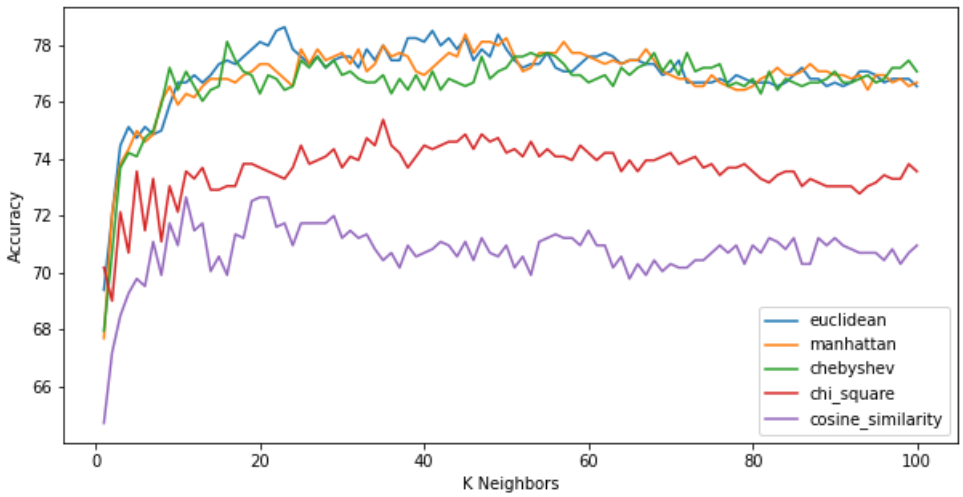

Figure 2 : Grafik perubahan akurasi terhadap k