1. Masalah Utama

Merancang sebuah sistem berbasis K-Nearest Neighbor untuk estimasi nilai K terbaik menggunakan teknik 5-Fold Cross Validation. Dataset berasal dari Pima India di dalam file Diabetes.csv.

2. Spesifikasi Implementasi

Implementasi dari sistem ini mencakup:

- Pembangunan sistem menggunakan bahasa Python dengan Jupyter Notebook sebagai platform utama.
- Library yang dipakai:
  - Pandas untuk membaca file .csv
  - MatPlotLib untuk memvisualisasikan dalam bentuk scatter plot.
  - Numpy untuk mengatur data berbentuk matriks.

3. Strategi Penyelesaian

3.1. Perhitungan Jarak

Metode yang digunakan untuk menghitung jarak adalah metode Jarak Euclidean dengan rumus,

$$d_1(x1, x2) = \sqrt{\sum_P (x1_P - x2_P)^2}$$

Dimana *d1* adalah jarak, x1 dan x2 merupakan atribut yang akan dilakukan perhitungan, serta P adalah jumlah iterasi untuk tiap-tiap data yang tersedia dalam dataset.

3.2. Pre-processing Data

Dataset yang diberikan di preprocessing menggunakan rumus normalisasi,

$$z = \frac{x - \mu}{\sigma}$$

Dimana *z* adalah *standardized feature*, *x* adalah nilai atribut, μ adalah rerata dari satu kolom atribut, dan σ adalah standar deviasi nilai atribut. Cara ini digunakan untuk memperkecil nilai, serta dapat mewakili nilai data aslinya.

3.3. Klasifikasi KNN

Kolom dalam dataset ini dibedakan menjadi:
Atribut:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigree
- Age

Label:

- Outcome

Atribut yang diekstrak dari dataset, kemudian dilakukan preprocessing menggunakan rumus normalisasi. Semua atribut dan label akan dibagi menjadi 5 lapisan, meliputi data uji dan data latih. Setiap data uji akan dihitung jaraknya menggunakan data latih melalui metode Euclidean. Setelah terkumpul setiap jarak, akan diurutkan dari yang terkecil. Akan diambil data dengan frekuensi label terbesar

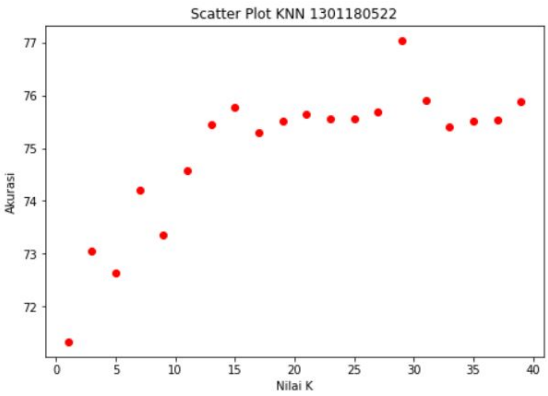
sebanyak *k*-data. Frekuensi label terbesar akan digunakan sebagai data latih untuk diujikan pada data uji.

3.4. 5-Fold Cross Validation

Atribut dan label dibagi menjadi 5 lapisan. Untuk tiap lapisan memegang data yang akan diklasifikasi dan setelah data testing diberikan label data latih, akan dibandingkan dengan label data uji untuk mencari skor akurasinya (antara 0-100). Untuk tiap lapis data terselesaikan, akan muncul skor akurasi lapis tersebut. Setelah akurasi tiap lapis didapatkan, akan dilakukan penghitungan rerata akurasi dari tiap-tiap nilai akurasi lapisan. Proses ini berulang terus-menerus bergantung pada nilai k dari tiap skor akurasi. Setelah skor akurasi rerata dari tiap lapisan didapatkan, program akan memberikan nilai k terbaik berdasarkan skor rerata akurasi.

4. Parameter dan Observasi

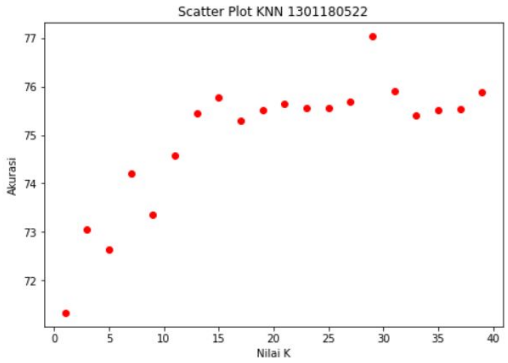
Untuk program ini, nilai maksimal k dibatasi dengan banyak perulangan yang diberikan pada program utama, yang berarti apabila perulangan 20 kali maka sekuens nilai k adalah 1, 3, 5, …, 39. Proses cross validation, dilakukan berdasarkan tiap-tiap nilai k. Grafik skor akurasi untuk tiap nilai K dijelaskan pada scatter plot berikut,



Berdasarkan dari hasil yang didapatkan, parameter terbaik dari skor validasi adalah:

- Nilai k terbaik: 29
- Rerata nilai k: 77.04221 %

5. Output Sistem




```
K terbaik:  29
Rerata akurasi dari K terbaik:  77.04221994366354
```

```
K= 1
Akurasi= 71.33596113385812
----------
K= 3
Akurasi= 73.05463988871581
----------
K= 5
Akurasi= 72.64817022222118
----------
K= 7
Akurasi= 74.21270371047594
----------
K= 9
Akurasi= 73.36675870912191
----------
K= 11
Akurasi= 74.58461695139648
----------
K= 13
Akurasi= 75.45412420407965
----------
K= 15
Akurasi= 75.76791876809699
----------
K= 17
Akurasi= 75.28550828960738
----------
K= 19
Akurasi= 75.52503778524095
----------
K= 21
Akurasi= 75.65574578691671
----------
K= 23
Akurasi= 75.56294569803748
----------
K= 25
Akurasi= 75.54692429295888
----------
K= 27
Akurasi= 75.69365369971321
----------
K= 29
Akurasi= 77.04221994366354
----------
K= 31
Akurasi= 75.89611315435589
----------
K= 33
Akurasi= 75.41454054767189
----------
K= 35
Akurasi= 75.52503778524094
----------
K= 37
Akurasi= 75.54441067754203
----------
K= 39
Akurasi= 75.8750645184436
----------
```