LAPORAN OBSERVASI LEARNING KNN K-FOLD CROSS VALIDATION

PENGANTAR KECERDASAN BUATAN IF-42-03

Disusun oleh :

SYA RAIHAN HEGGI (1301184219)

S1 INFORMATIKA

FAKULTAS INFORMATIKA

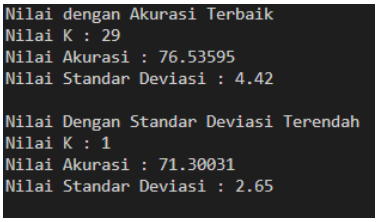
UNIVERSITAS TELKOM

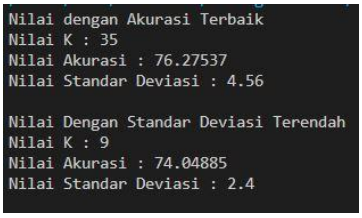
BANDUNG

2020

- **Pemilihan ukuran jarak yang digunakan**
  Dalam program yang dibuat jarak dataset awal merupakan sebuah data berbentuk csv yang memiliki panjang 768 data yang nantinya akan dibagi menjadi 5 bagian dengan masing-masing isi data berisi 20% atau sekitar **154 data akan dijadikan data test dan 614 data akan menjadi data training** , dimana data ini akan berisi 9 kolom yang berisi data berikut ini Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome, dimana **Atribut** berisi *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age* dan **Label** *Outcome.*
- **Teknik Data Prapemrosesan**
  Data sebelum diproses akan diload terlebih dahulu dari file .csv kedalam program dengan menggunakan library csv dan kemudian data akan di clipping sesuai dengan foldnya dengan aturan yang sudah diberikan pada penugasan yaitu sebagai berikut ini,
  1. **Data training set 1-614 dan Data testing set 615:768**
  2. **Data training set 1-461 + 615-768 dan Data testing set 462-614**
  3. **Data training set 1-307 + 462-768 dan Data testing set 308:461**
  4. **Data training set 1-154 + 308-768 dan Data testing set 155:307**
  5. **Data training set 155-768 dan Data testing set 1:154**

  Kemudian menentukan tipe data mana yang cocok untuk data ini, berdasarkan hasil observasi tipe data ini cocok untuk diproses menggunakan rumus perhitungan jarak numerik, kemudian data akan dilakukan scaling dengan metode Standarisasi dengan menggunakan tumus berikut ini $\boldsymbol{Xscal} = \frac{\boldsymbol{Xi} - \mu}{\boldsymbol{\sigma}}$ dimana X (nilai variable kolom), μ (nilai rataan) dan **σ** (Nilai Standar Deviasi) sehingga nilai yang dihasilkan tidak memiliki jarak yang sangat luas dan mengurangi kemungkinan bias yang terjadi, berikut ini hasil jika menggunakan normalisasi (kanan) dan standarisasi (kiri) keduanya dapat digunakan dan normalisasi juga dapat digunakan pada algoritma seperti KNN.





- **Teknik Rekayasa Pemrosesan.**
  Data yang sudah dipersiapkan terlebih dahulu tadi akan diproses dengan cara pertama dilakukan memilih nilai k yang diinginkan dan lakukan looping sebanyak yang diinginkan, untuk observasi ini dilakukan sebanyak 40 kali loop mencari nilai k atau sama dengan 20 akurasi data per dataset di fold, kemudian melakukan prediksi pada proses ini akan dilakukan perhitungan distance, mengelompokkan berada di tetangga mana, dan menghitung akurasinya, dari hasil observasi didapati.

  Metode perhitungan distance :
  1. **Euclidean :** dengan menggunakan Euclidean didapati nilai maksimum K=29 dengan nilai akurasi 76.53595 (77%) dan nilai yang memiliki standar deviasi terendah akan didapati pada K=1 dengan nilai akurasi 71.30031 (71%)
  2. **Manhattan :** dengan menggunakan perhitungan Manhattan nilai maksimum K=21 dengan nilai akurasi 76. 67183 (77%) dan bila melihat pada yang memiliki standar deviasi terendah akan didapati pada K= 17 dengan nilai akurasi 75.22704 (75%)

3. **Minkowski :** dengan menggunakan perhitungan Minkowski nilai maksimum didapati pada K=21 dimana nilai akurasinya 75.22532 dan bila melihat yang memiliki standar deviasi terendah berada pada K=1 dengan akurasi 70.90127

```
Nilai dengan Akurasi Terbaik
Nilai K : 29
Nilai Akurasi : 76.53595
Nilai Standar Deviasi : 4.42

Nilai Dengan Standar Deviasi Terendah
Nilai K : 1
Nilai Akurasi : 71.30031
Nilai Standar Deviasi : 2.65
```

```
Nilai dengan Akurasi Terbaik
Nilai K : 21
Nilai Akurasi : 76.67183
Nilai Standar Deviasi : 4.68

Nilai Dengan Standar Deviasi Terendah
Nilai K : 17
Nilai Akurasi : 75.22704
Nilai Standar Deviasi : 3.45
```

```
Nilai dengan Akurasi Terbaik
Nilai K : 13
Nilai Akurasi : 75.74733
Nilai Standar Deviasi : 5.85

Nilai Dengan Standar Deviasi Terendah
Nilai K : 7
Nilai Akurasi : 72.73736
Nilai Standar Deviasi : 3.92
```

Dari hasil pengujian penggunaan metode penghitungan jarak didapati dengan menggunakan Manhattan hasil yang didapati memiliki akurasi tertinggi serta bila diperhatikan juga standar deviasi dari nilai akurasi yang dihasilkan setiap fold dari setiap distance pun Manhattan masih lebih baik dari menggunakan Euclidean dan Minkowski. Kemudian setelah menentukan untuk menggunakan Manhattan maka yang akan dilakukan adalah menghimpun setiap akurasi dari setiap nilai K dan kemudian nilai itu akan digunakan untuk mendapatkan nilai rataannya. Setelah itu nanti terakhir dipilih nilai K yang memiliki akurasi rataan terbaik itulah yang akan ditampilkan pada program, dan selain itu juga akan dihitung standar deviasi dari nilai rataan yang dihasilkan untuk memastikan data tidak tersebar jauh dan memiliki hasil yang stabil, kemudian akan ditampilkan hasilnya.

- **Strategi Penggunaan KNN.**

  Penggunaan KNN disini menggunakan bantuan 5 Fold Cross Validation semakin banyak ini maka semakin banyak validasi yang dilakukan, yang perlu diperhatikan disini adalah metode perhitungan jarak apa yang digunakan kemudian pencocokan nilai mirip dengan tetangga yang mana, berapa k yang digunakan dan berapa hasil akurasi yang dihasilkan dari percobaan, dari hasil observasi didapati kNN akan digunakan dengan parameter berikut.
  - K-Fold Cross Validation : 5
  - Data Scaling : Standarization
  - N Percobaan = 40
  - K = 21 (Akurasi Terbaik)
  - Metode Distance : Manhattan
  - Label : [0,1] (Positif, Negatif) Diabetes
- **Pemilihan nilai k terbaik untuk proses seleksi dan estimasi model kNN**

  Dari hasil yang dilakukan didapati nilai k terbaik terdapat pada K=21 dengan nilai akurasi 76.67183 dengan standar deviasi akurasi dari cross validation adalah 4.68 dan dapat dilihat dari running program dengan parameter terbaik berikut ini.

```
Nilai dengan Akurasi Terbaik
Nilai K : 21
Nilai Akurasi : 76.67183
Nilai Standar Deviasi : 4.68

Nilai Dengan Standar Deviasi Terendah
Nilai K : 17
Nilai Akurasi : 75.22704
Nilai Standar Deviasi : 3.45
```