**TUPRO 3 Artificial Intelligence**

**[k-Nearest Neighbour]**

1. **Studi Kasus**

Artificial Intelligence adalah bagaimana mesin/komputer bertingah layaknya manusia dan berfikir layaknya manusia. Dalam AI ada 4 pokok bahasan yaitu searching, reasoning, planning dan learning.

Dalam tugas program 3 ini, pokok bahasan yang ditugaskan adalah learning dengan algoritma k-Nearest Neighbour (k-NN) pada sebuah data set dengan tujuan klasifikasi label class pada data testing.

Untuk melakukan k-Nearest Neighbour terlebih dahulu ditentukan nilai K yang optimal dengan mekanisme k- Fold Cross Validation kemudian k tersebut digunakan untuk melakukan klasifikasi pada data testing dengan kNN. Pada bagian selanjutnya akan dijelaskan bagaimana menemukan K yang tepat dan hasil klasifikasi yang diperoleh oleh k-Nearest Neighbour.

1. **Analisa dan Pembahasan**
2. **k-Nearest Neighbour**

k-Nearest Neighbour adalah salah satu metode klasfikasi pada machine learning dan juga kNN yang paling mudah dimengerti (straightforward) karena proses klasfikasi pada metode ini dengan mengidentifikasi sebuah data testing dengan tetangga terdekat dari data training dan menggunakan tetangga tersebut untuk memberikan data testing tersebut sebuah class berdasarkan class tetangga. Untuk k lebih sering di pakai tidak hanya satu dan berguna untuk k lebih dari satu sehingga metode ini sering disebut dengan k-Nearest Neighbour(k-NN) Classification dimana k nearest neighbours dipakai untuk menetukan label class. k-NN juga memiliki istilah lain yang disebut dengan Memory-Based Classification karena data training ada di memory saat program berjalan kemudian disebut juga Lazy Learning teknik karena tertunda akibat run time dan disebut juga sebagai Example-Based Classification/ Case-Based Classification karena secara langsung mengklasifikasikan dengan data training.

Secara pseudocode k-NN dapat dijelaskan sebagai berikut :

*Untuk setiap point dataTraining:*

*Hitung jarak antara dataTesting dan point sekarang.*

*Urutkan jarak secara ascending*

*Ambil k buah dengan jarak terendah dengan dataTesting*

*Temukan class yang terbanyak di k buah tetangga tersebut*

*Kembalikan class yang terbanyak tersebut sebagai class untuk dataTesting.*

1. **K-fold Cross Validation**

k-Fold cross validation adalah sebuah teknik intensif komputer yang menggunakan keseluruhan data yang ada sebagai data training dan data testing. Seluruh data secara acak dibagi menjadi K buah subset Bk dengan ukuran yang sama dimana Bk merupakan himpunan bagian dari {1…n}. Setelah itu dilakukan iterasi sebanyak K kali. Pada iterasi ke k, subset Bk menjadi data testing, sedangakan subset yang lain menjadi data training. Setelah itu dihitung nilai rata-rata performa hasil dari K buah iterasi.

Kelebihan dari metode ini adalah tidak adanya masalah dalam pembagian data. Setiap data akan menjadi data testing sebanyak satu kali dan akan menjadi data training sebanyak K-1 kali. Kekurangan dari metode ini adalah algoritma pembelajaran harus dilakukan sebanyak K kali yang berarti menggunakan K kali waktu komputasi.

Leave One Out Cross Validation (LOO CV) merupakan n-fold Cross Validation dimana n adalah jumlah data yang tersedia. LOO CV sangat membutuhkan membutuhkan komputasi yang tinggi jika terdapat data dalam jumlah besar.

Biasanya k pada k-Cross validation adalah 5 atau 10.

1. **Hasil**

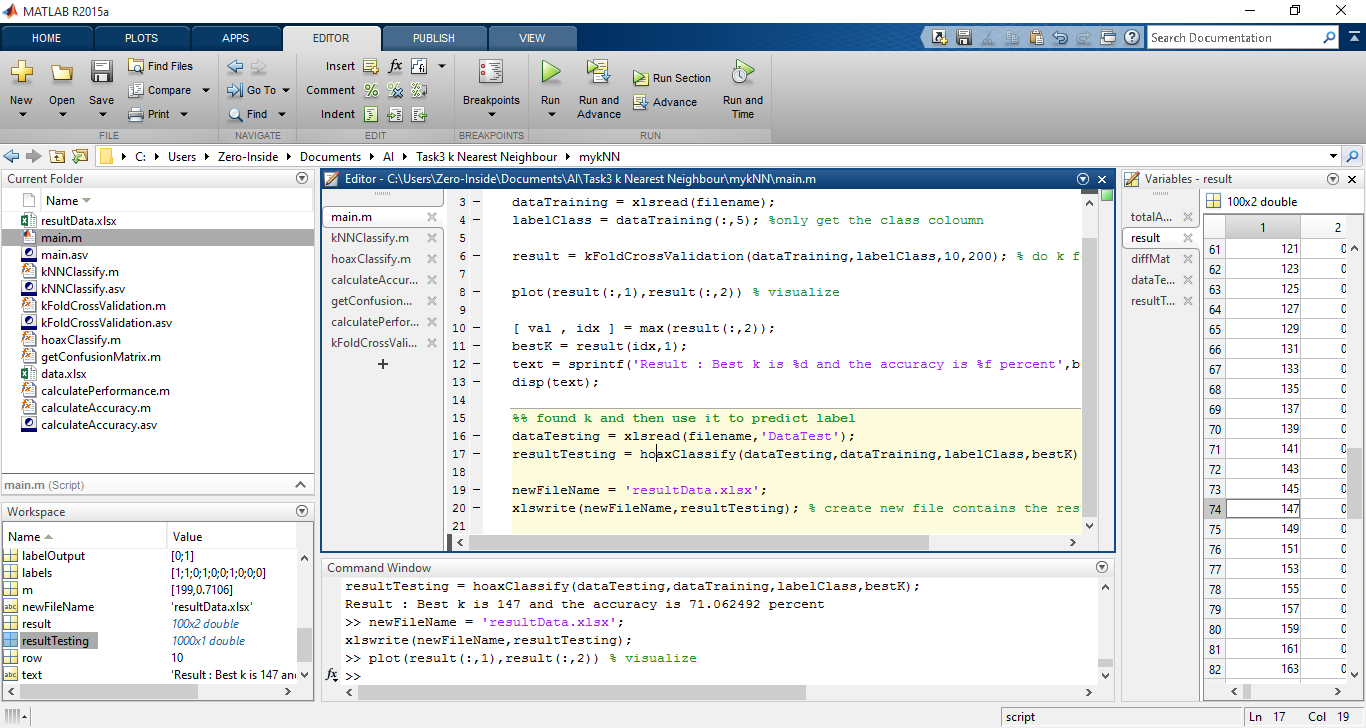
Implementasi dari kedua algoritma diatas adalah pada sebuah data set yang terdiri dari 4000 baris data training, dan 1000 baris data testing. Total adalah sebanyak 5000 data set. Pada setiap data terdapat 6 variable kolom yaitu Berita, Like, Provokasi, Komentar, Emosi dan Hoax. Untuk variable berita menunjukan ID dari setiap baris berita, dalam variable ini berutrut dari B0001 samapi B5000, Like,Provokasi dan Emosi adalah parameter untuk menentukan hoax atau tidaknya suatu berita. Penulis menggunakan Matlab untuk melakukan kedua algoritma tersebut.

Adapun langkah-langkah yang dilakukan adalah sebagai berikut :

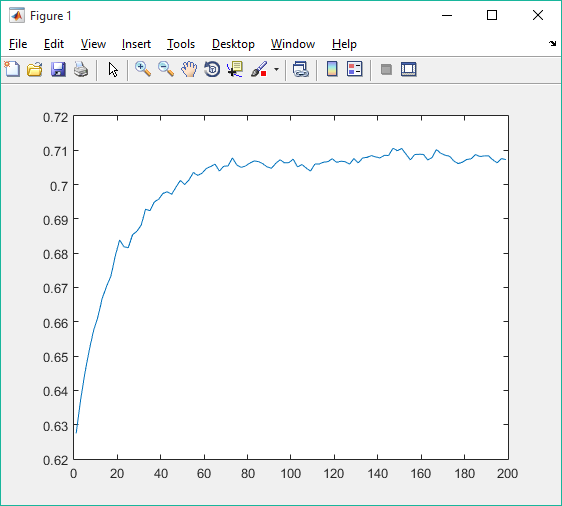
1. Mengimport data set pada sheet 1 ke variable dataTraining
2. Melakukan k-Fold Cross Validation pada dataTraining dengan k adalah 10 dan k maksimal untuk kNN adalah 200 dengan pertambahan k adalah bilangan ganjil, dan hasil dari step ini adalah sebuah k = [1,3,5..199] dan akurasi masing-masing k.
3. Menemukan k maksimal dari hasil k-Fold Cross Validation.
4. Mengimport data set pada sheet 2 ke variable dataTesting
5. Menggunakan k terbaik pada step 3 untuk dilakukan kNN klasifikasi untuk mengklasifikasikan data Testing dan menghasilkan output klasifikasi.
6. Membuat file excel baru dengan hasil dari step 5.
7. **Referensi**

|  |  |
| --- | --- |
| [1] | P. Harrington, Machine Learning in Action, New York: Manning Publications, 2012. |
| [2] | P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifier," UCD-CSI-2007-4, Dublin, 2007. |
| [3] | R. Fitri, "Digilib," [Online]. Available: digilib.itb.ac.id/files/disk1/580/jbptitbpp-gdl-radenfitri-28974-3-2007ta-2.pdf. |

1. **Lampiran**



Gambar 1Output Program



Gambar 2 Visualisasi performa data Training dengan k-Fold Cross Validation untuk mencari k terbaik