Optimasi Metode *K-Nearest Neighbours* dengan *Backward Elimination* Menggunakan Dataset *Software Effort Estimation*

Wawan Nugroho
STMIK Nusa Mandiri
Email :14002382@nusamandiri.ac.id

Abstrak - Klasifikasi sering digunakan untuk menentukan suatu keputusan sesuai pengetahuan baru yang didapat dari pengolahan data lampau menggunakan perhitungan suatu algoritma. Metode penelitian yang digunakan dalam penelitian ini metode eksperimental adapun tahapan metode tersebut pesiapan dataset, preprosesing, seleksi fitur kemudian evauasi model dengan RMSE dan AE. *K-Nearest Neighbor* merupakan salah satu algoritma yang digunakan untuk klasifiksi dan juga prediksi yang menggunakan metode *supervised learning*. Algoritma K-Nearest Neighbor memiliki keunggulan pelatihan yang sangat cepat, sederhana dan mudah dipahami, *K-Nearest Neighbor* juga memiliki kekurangan dalam menentukan nilai *K* dan pemilihan atribut terbaik. Untuk megoptimalkan algorittma *K-Nearest Neighbor* digunakan seleksi fitur *Backward Elimination,* memiliki fungsi untuk mengoptimalkan kinerja suatu model dengan sistem kinerja mundur, digunakan untuk memilih atribut yang paling relevan. Hasil penelitian menunjukan bahwa *K-Nearest Neighbor* dengan *Backward Elimination* memiliki *Root Mean Square Erorr* (RMSE) dan *Absolute Erorr* (AE) pada dateset *Software Effot Estimation* menujukan hasil yang lebih optimal dibandingkan hanya menggunakan algoritma *K-Nearest Neighbor*.
Kata Kunci : Optimasi, *K-Neares Neighbor, Backward Elimination*

Abstract - *Classification is widely used to determine decisions according to new knowledge obtained from past data processing using the calculation of an algorithm. The research method used in this research is experimental method, while the steps of the method are dataset preparation, preprocessing, feature selection then evaluation of the model with RMSE and AE. K-Nearest Neighbor is one of the algorithms used for classification and prediction using the supervised learning method. The K-Nearest Neighbor algorithm has the advantage of training that is very fast, simple and easy to understand, K-Nearest Neighbor also has deficiencies in determining the K value and selecting the best attributes. To optimize the K-Nearest Neighbor algorithm, the Backward Elimination feature selection is used, which has a function to optimize the performance of a model with a backward performance system, used to select the most relevant attributes. The results showed that K-Nearest Neighbor with Backward Elimination has Root Mean Square Error (RMSE) and Absolute Error (AE) on the Effort Estimation software dataset, which shows more optimal results than using only the K-Nearest Neighbor algorithm.*
Keyword : Optimization, K-Neares Neighbor, Backward Elimination

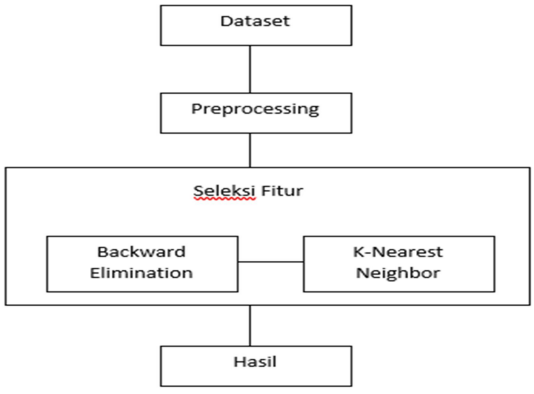
I. PENDAHULUAN

Data mining merupakan ilmu yang memanfaatkan data yang sebelumnya kurang terpakai Untuk mendapatkan suatu informasi atau pengetahuan baru. Data Mining merupakan bagian bidang ilmu yang menyatukan teknik pembelajaran, pengenalan pola, statistik, database, serta visualisasi dalam mengatasi masalah ekstraksi suatu informasi dari basis data yang besar (Nengsih, 2017) Salah satu fungsi utama dari data mining adalah klasifikasi. Klasifikasi sering digunakan dalam menentukan suatu keputusan sesuai pengetahuan baru yang didapat dari pengolahan data lampau dengan menggunakan perhitungan suatu algoritma. Algoritma *K-Nearest Neighbor* (k-NN) adalah salah satu metode digunakan untuk klasifikasi dan prediksi (Noviana et al., 2019) yang menggunakan metode *supervised* (Yuita Arum Sari, 2018) K-NN juga memiliki keunggulan pelatihan yang sangat cepat dan sederhana sehingga mudah dipelajari (Iriantoro et al., 2018). Namun ada kekurangan dalam K-NN perlu penentuan nilai *K* dan untuk pemilihan atribut terbaik (Bode, 2017). Berbagai penelitian telah dilakukan untuk mendapatkan nilai akurasi dengan menerapkan fitur seleksi. Dalam penelitian (Prasetio et al., 2020). Seleksi Fitur Dan Optimasi Parameter K-NN Berbasis Algoritma Genetika Pada Dataset Medis Hasil eksperimen menunjukkan bahwa metode yang diusulka dapat mencapai kinerja yang baik, dibandingkan dengan hasil pengklasifikasi lain. Sedangkan dalam (Indriyanti et al., 2017) Peningkatan akurasi algoritma K-NN dengan seleksi fitur *gain ratio* untuk klasifikasi penyakit diabetes mellitus, hasil penelitian yang telah dilakukan dapat disimpulkan bahwa

penggunaan algoritma seleksi fitur *gain ratio* efektif meningkatkan akurasi dari klasifikasi penyakit diabetus mellitus dengan menerapkan algoritma K-NN. Adapun kenaikan tertinggi didapatkan pada nilai treshold 0,152 dengan hanya mempertahankan 4 atribut dari keseluruhan 8 atribut data. Pada penelitian (Bode, 2017) untuk meningkatkan hasil akurasi metode K-NN perlu ditambahkan seleksi fitur *Backward Elimination.* Dengan seleksi fitur tingkat akurasi yang dihasilkan meningkat secara signifikan dilihat dari penurunan nilai RMSE (Indra Kurniawan, 2019). Penelitian yang dilakukan (Achmad Saiful Rizal & Moch. Lutfi, 2020) penggunaan metode *K-Nearest Neighbor* yang digambungkan dengan metode *Backward Eimination* pada dataset pemilu menghasilkan akurasi sebesar 96.03%. Penelitian (Drajana, 2018) Algoritma *K-Nearest Neighbor* menggunakan *Backward Elimination* menghasikan model terbaik yang dilihat berdasarkan nilai *error* terkecil yaitu 0.109 dari sebelumnya yaitu 0.111. Dari beberapa penelitian tersebut ada kekurangan pada metode K-NN, sehingga pada penelitian ini menggunakan optimasi seleksi fitur dengan *Backward Elimination. Backward Elimination* dapat digunakan untuk menguji semua variabel yang dianggap tidak relevan (Ary & Rismiati, 2019) Algoritma *Backward Elimination* ini memungkinkan untuk mendapatkan bebrapa atribut yang memiliki kemampuan klasifikasi rendah secara individu apabila digabungkan dengan atribut lainnya akan mendapatkan akurasi yang cukup tinggi (Gamadarenda & Waspada, 2020) sehingga hasil yang diharapkan akan lebih optimal.

II. METODE PENELITIAN

Metode penelitian yang diterapkan peneliti dalam penelitian ini adalah metode eksperimental. Tahapan metode eksperimental yang dilakukan adalah sebagai berikut:



Sumber : Penelitian Mandiri (2020)

Gambar 1. Metodologi Penelitian

1. Dataset

Langkah pertama yang dilakukan peneliti adalah menyiapkan dataset yang akan diteliti, dataset diunduh dari laman UCI Repository.

2. Data Preprocessing

Tahapan selanjutnya *Preprocessing dataset* dengan mengapus bagian yang tidak perlu sepeti nomer urut atau karakter yang tidak diinginkan.

3. Seleksi Fitur

Selajutnya mengoptimalkan seleksi fitur menggunakan *Backward Elimination* pada algoritma *K-Nearest Neighbor.*

4. Hasil

Hasil evaluasi merupakan tahapan terakhir dari penelitian ini, dimana hasil metode *K-Nearest Neigbhor* dengan *Backward Elmination* dievaluasi menggukan *Root Mean Square Erorr* (RMSE) dan *Absolute Erorr* (AE).

2.1. Metode Pengumpulan data

Tahapan pertama yang dilakukan dalam penelitian ini adalah pengumpulan data. Data yang digunakan dalam peneitian ini adalah data *public* yaitu dataset *Software Effort Estimation* dengan jumlah 2 (dua) dataset yang berbeda. Berikut spesifikasi dari masing-masing dataset :

Tabel 1. Deskripsi Dataset

Dataset	Jumlah Sampel	Jumlah Fitur
Albrecht	24	8
Cemerer	15	8

Sumber : Hasil Penelitian (2020)

2.2. *K-Nearest Neighbour* (K-NN)

Algoritma *K-Nearest Neighbour* (k-NN) merupakan salah satu metode *non-parametrik* yang dapat digunakan untuk (Drajana, 2018), salah satu agoritma klasifikasi (Safriandono, 2017) sering digunakan untuk klasifikasi, meskipun dapat juga digunakan untuk estimasi dan prediksi. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasarkan atribut dan data latih (Wahyuni & Suparman, 2020).

Algoritma KNN pada dasarnya dilakukan dengan langkah berikut (Yuita Arum Sari, 2018).

1. Menentukan nilai k
2. Mengitung jarak antar data dengan semua data latih.
3. Sejumlah k data dipilih yang paling dekat Dengan data masukan. Data Akan diklasifikasikan kepada kelas yang memiliki jumlah kelas yang sama dengan nilai terbanyak.

2.3. Seleksi Fitur

Backward Elimination adalah metode yang memiliki fungsi untuk mengoptimalkan kinerja suatu model dengan sistem kinerja

mundur. Pemilihan dilakukan dengan cara memilih variabel kedepan yaitu dengan menguji semua variabel kemudian menghapus variabel-variabel yang dianggap tidak relevan. Variabel yang diproses satu per satu, jika variabel dianggap tidak berpengaruh atau tidak signifikan dalam model maka akan dihapus dari model tersebut (Bode, 2017). Berikut Langkah-langkah metode *Backward Elimination* (Ghani et al., 2019):

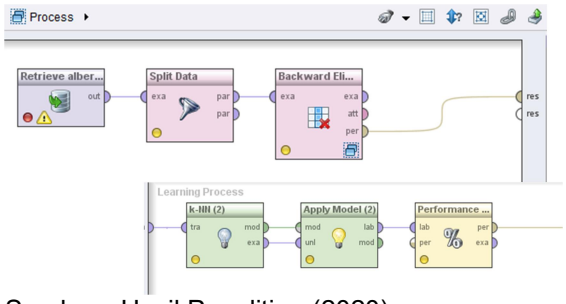
a. Membuat model degan meregredikan variabel reson Y dengan semua variabel prekdiktor.
b. Mengeluarkan satu-saru vaiabel predictor dengan melakukan pengujian terhadap parameternya dengan *partial Ftest*. Niali Fparsial terkecil dibandingkan dengan Ftabel :
   1. Jika Fparsial < Ftabel, maka X yang bersangkutan dikeluakan dari model dan dilanjutkan dengan pembuatan model baru tanpa variable tersebut.
   2. Jika Fparsial >Ftabel, maka proses dihentikan artinya tidak ada variabel yang perlu dikeluarkan dan persamman terakhir yang dipilih.

III. HASIL DAN PEMBAHASAN

Hasi dalam penelitian dilakukan dalam dua eksperimen yaitu terhadap algoritma *K-Nearest Neighbor* tanpa seleksi fitur serta eksperimen terhadap algoritma *K-Nearest Neighbor* dengan *Backward Elimination*.

3.1. Hasil Eksperimen

Dalam percobaan pertama peneliti menerapkan algoritma *K-Nearest Neighbor* terlebih dahulu tanpa menggunakan *Backward Elimination* untuk seleksi fitur. Percobaan ini dilakukan terhadap dataset yang telah divalidasi berdasarkan hasil yang telah dilakukan dengan *tools* RapidMider dengan tahapan sebagai berikut :



Sumber : Hasil Penelitian (2020)

Gambar 2. Proses pada *Tools* RapidMiner

Pada gambar 2 menampilkan data yang telah di*import*, selanjutnya diproses dengan membagi data *testing* dan *training* dengan split data, selanjutnya data dioptimasi menggunakan *Backward Elimination*, data diproses ulang menggunakan K-NN selanjutnya model diaplikasikan dan terahir evaluasi performansi dengan RMSE dan AE.

Dalam Tabel berikut merupakan hasil dari pemodelan data berdasarkan nilai K dengan ekperimen nilai K 3, 5, 7, 9 kemudian menghasilkan nilai *Root Mean Square Error* (RMSE) dan *Absolute Error* (AE) dengan menggunakan algoritma *K-Nearest Neighbor*.

Tabel 2. Dataset Albrecht

	K-NN	
Nilai *K*	RMSE	AE
3	12.437	10.137
5	20.925	7.589
7	24.916	8.240
9	27.233	9.645

Sumber : Hasil Penelitian (2020)

Hasil percobaan dengan jumlah k yang berbeda, terlihat bahwa *error* terendah adalah k=3 sebesar 12.437.

Tabel 3. Dataset Cemerer

	K-NN	
Nilai *K*	RMSE	AE
3	8.265	8.077
5	8.760	8.646
7	8.206	8.144
9	7.613	7.557

Sumber : Hasil Penelitian (2020)

Hasil percobaan dengan jumlah k yang berbeda, terlihat bahwa *error* terendah adalah k=9 sebesar 7.613.

Hasil eksperimen pertama menunjukan hasil *error* pada kedua dataset masih terlalu tinggi dengan nilai 12.437 pada dataset Albrecht dan 7.613 pada dataset Cemerer.

Percobaan selanjutnya menggunakan algoritma *K-Nearest Neighbor* dengan penambahan *Backward Elimination*, berikut adalah hasi *Root Mean Square Error* (RMSE) dan *Absolute Error* (AE). :

Tabel 4. Dataset Albrecht

	K-NN + BE	
Nilai *K*	RMSE	AE
3	3.710	2.957
5	15.818	8.623
7	20.128	9.879
9	23.296	11.277

Sumber : Hasil Penelitian (2020)

Hasil percobaan dengan jumlah k yang berbeda, terlihat bahwa *error* terendah adalah k=3 sebesar 3.710.

Tabel 5. Dataset Cemerer

	K-NN + BE	
Nilai *K*	RMSE	AE
3	1.871	1.530
5	2.582	2.171
7	3.061	2.708
9	3.202	2.717

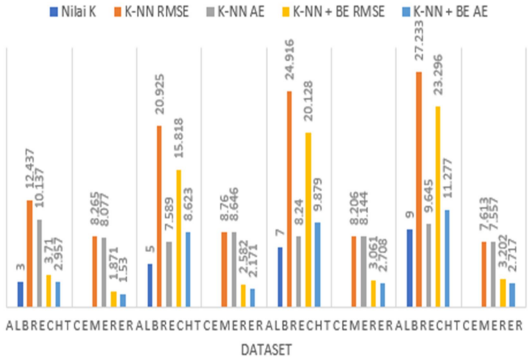
Sumber : Hasil Penelitian (2020)

Hasil percobaan dengan jumlah k yang berbeda, terlihat bahwa *error* terendah adalah k=3 sebesar 1.871.

Hasil eksperimen kedua menunjukan hasil yang cukup baik dengan error terkecil pada dataset Albrecht 3.710 dan pada dataset Cemerer 1.871.

Percobaan tersebut dilakukan beberapa kali karena setiap percobaan menggunakan nilai *k* yang berbeda sehingga mendapatkan nilai *Root Mean Square Erorr* (RMSE) dan *Absolute Erorr* (AE) yang berbeda.

Berikut merupakan hasil perbandingan algoritma *K-Neares Neighbor* sebelum dan sedudah dioptimasi menggunakan *Backward Elimination.*



Sumber : Hasil Penelitian (2020)

Gambar 3 Hasil Evaluasi RMSE dan AE

Eksperimen pada penelitian ini menghasilkan perbandingan hasil optimasi aloritma *K - Nearest Neighbor* dengan Backward Elimination ditunjukan pada warna kuning dan tanpa optimasi menggunakan Backward Elimination ditunjukan pada warna merah.

IV. KESIMPULAN

Simpulan dari penelitian ini adalah tingkat akurasi yang dihasilkan oleh algoritma *K - Nearest Neighbor* dalam dataset *Software Effort Estimation* menghasilkan RMSE dan AE yang cukup tinggi Sedangkan untuk hasil klasifikasi algoritma *K-Nearest Neighbor* setelah dilakukan seleksi fitur menggunakan *Backward Elimination* menunjukan peningkatan RMSE dan AE yang cukup optimal. Untuk peneitian selanjutnya agar dapat melekukan uji coba menggunakan seleksi fitur selain *Backward Elimination* dengan tujuan mendapatkan hasil *error* yang lebih kecil.

V. REFERENSI

Achmad Saiful Rizal, & Moch. Lutfi. (2020). Prediksi Hasil Pemilu Legislatif Menggunakan Algoritma K-Nearest Neighbor Berbasis Backward Elimination. *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, *3*(1), 27–41. https://doi.org/10.31598/jurnalresistor.v3i1.517

Ary, M., & Rismiati, D. A. F. (2019). Ukuran Akurasi Klasifikasi Penyakit Mesothelioma Menggunakan Algoritma K-Nearest Neighbor dan Backward Elimination. *SATIN - Sains Dan Teknologi Informasi*, *5*(1), 11–18. https://doi.org/10.33372/stn.v5i1.444

Bode, A. (2017). K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika. *ILKOM Jurnal Ilmiah*, *9*(2), 188–195. https://doi.org/10.33096/ilkom.v9i2.139.188-195

Drajana, I. C. R. (2018). Prediksi Jumlah Produksi Coconut Oil Menggunakan k-Nearest Neighbor dan Backward Elimination bagian dari pohon digunakan manusia , sehingga tumbuhan ini dianggap. *Tecnoscienza*, *13*(1), 51–64.

Gamadarenda, I. W., & Waspada, I. (2020). Implementasi Data Mining untuk Deteksi Penyakit Ginjal Kronis (PGK) menggunakan K-Nearest Neighbor (KNN) dengan Backward Elimination. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(2), 417. https://doi.org/10.25126/jtiik.2020721896

Ghani, A. D., Salman, N., & Mustikasari. (2019). Algoritma k-Nearest Neighbor Berbasis Backward Elimination Pada Client Telemarketing. *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *8*(2), 141–150.

Indra Kurniawan, A. F. A. (2019). KOMPARASI METODE KOMBINASI SELEKSI FITUR DAN MACHINE LEARNING K-NEAREST NEIGHBOR PADA DATASET LABEL HOURS SOFTWARE EFFORT ESTIMATION. *Jurnal Sistem Informasi Dan Telematika*, *10*.

Indriyanti, Sugianti, D., & Karomi, M. A. Al. (2017). Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur G ain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus. *IC-Tech*, *7*(2), 1–6. https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/3

Iriantoro, D. N. D., Dewi, C., & Fitriani, D. (2018). Klasifikasi pada Penyakit Dental Caries Menggunakan Gabungan K-Nearest Neighbor dan Algoritme Genetika. *Klasifikasi Pada Penyakit Dental Caries Menggunakan Gabungan K-Nearest Neighbor Dan Algoritme Genetika*, *2*(8), 2926–2933.

Nengsih, W. (2017). Analisa Akurasi Permodelan Supervised Dan Unsupervised. *Sebatik 1410-3737*, *23*(2), 285–291. https://jurnal.wicida.ac.id/index.php/sebatik/article/view/771

Noviana, D., Susanti, Y., & Susanto, I. (2019). Analisis Rekomendasi Penerima Beasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Algoritma C4.5. *Seminar Nasional Penelitian Pendidikan Matematika (SNP2M) 2019 UMT*, 79–87.

Prasetio, R. T., Adhirajasa, U., & Sanjaya, R. (2020). *SELEKSI FITUR DAN OPTIMASI PARAMETER k-NN BERBASIS ALGORITMA GENETIKA PADA*. *2*(2), 213–221.

Safriandono, A. N. (2017). *ALGORITMA K-NEAREST NEIGHBOR BERBASIS FORWARD SELECTION UNTUK MENDIAGNOSIS PENYAKIT JANTUNG*. *3*(1).

Wahyuni, E., & Suparman, S. (2020). A Comparison of Outlier Detection Techniques in Data Mining. *Science, Technology, Engineering, Economics, Education, and Mathematics*, *1*(1), 139–147.

Yuita Arum Sari, A. A. (2018). Optimasi K-Nearest Neighbour Menggunakan Particle Swarm Optimization Optimasi K-Nearest Neighbour Menggunakan Particle Swarm Optimization pada Sistem Pakar untuk Monitoring Pengendalian Hama pada Tanaman Jeruk. *Jurnal Teknologi*, *2*(July), 13.