

KOMPARASI METODE KLASIFIKASI DATA MINING UNTUK PREDIKSI KELULUSAN MAHASISWA

Abdul Rohman¹, Muhammad Rochcham²
email: ^{*1}abdulrohman15@gmail.ac.id, ²muhrochan@gmail.com

Abstrak

Data kelulusan mahasiswa sangat penting, ketika data tersebut di olah dengan menggunakan *data mining*, sehingga menghasilkan pengetahuan dan informasi yang sangat penting, salah satunya dijadikan sebagai alat prediksi.

Dalam penelitian ini dilakukan perbandingan metode *data mining* yaitu algoritma *neural network*, *k-nearest neighbor* dan *Decision Tree* yang diaplikasikan pada data kelulusan mahasiswa. Dari hasil pengujian dengan mengukur kinerja ketiga metode tersebut menggunakan *cross validation*, *confusion matrix* dan kurva *ROC*, diketahui bahwa *neural network* memiliki nilai akurasi yang paling tinggi yakni sebesar 87,32%, diikuti oleh metode algoritma *Decision Tree* dengan nilai akurasi 83,57%, kemudian metode *k-nearest neighbor* dengan akurasi 83,66%. Nilai *AUC* untuk metode *neural network* menunjukkan nilai tertinggi yakni 0,917 dan yang terendah adalah metode *Decision Tree* yaitu 0,844.

Kata Kunci

Data Mining, Mahasiswa, algoritma *neural network*, *k-nearest neighbor*, *Decision Tree*

1. Pendahuluan

Memprediksi kelulusan mahasiswa menggunakan *data mining* dalam perguruan tinggi perlu untuk dilakukan, karena untuk menganalisis mahasiswa dalam proses pembelajaran. Sehingga dengan data tersebut dijadikan sebagai pegangan institusi untuk menentukan sebuah keputusan atau kebijakan.

Banyak penelitian yang mengkaji dan analisis kelulusan mahasiswa menggunakan metode klasifikasi data mining dengan berbagai algoritma, salah satunya dalam penelitian sebelumnya pengujian dengan algoritma K-Nearest Neighbor (K-NN) dengan menggunakan data kelulusan mahasiswa yang tepat dan terlambat. Dan hasil yang di dapat dengan nilai accuracy adalah 82,25% dan nilai AUC adalah 0.500, dengan cluster data k=2 accuracy adalah 79,45% dan nilai AUC adalah 0.826, dengan cluster data k=3 accuracy adalah 83,95% dan nilai AUC adalah 0.853, dengan cluster data k=4 accuracy adalah 82,62% dan nilai AUC adalah 0.874, dengan cluster data k=5 accuracy adalah 85,15% dan nilai AUC adalah 0.888 [1].

Penelitian yang dilakukan oleh Agus Romadhona dkk, prediksi kelulusan mahasiswa tepat waktu berdasarkan usia, jenis kelamin dan indeks prestasi

menggunakan algoritma *Decision Tree* menghasilkan nilai akurasi 90,77 % dengan jumlah data 389 orang [2].

2. Kajian Pustaka

a. Kelulusan Mahasiswa

Output Keberhasilan pembelajaran diperguruan tinggi ditentukan terampilnya mahasiswa dalam keilmuannya [3].

Jumlah lulusan mahasiswa akan berpengaruh terhadap status akreditasi program studi [4]. Faktor-faktor yang dapat mempengaruhi kelulusan mahasiswa antara lain adalah umur, status pernikahan, jenis kelamin, jumlah saudara, nilai akhir sekolah menengah, Indeks Prestasi Semester, gaji orang tua, pekerjaan orang tua [5], [6], [7].

Pada penelitian ini parameter yang digunakan adalah usia, jenis kelamin, indeks prestasi semester satu sampai dengan indeks prestasi semester empat.

b. Data Mining

Data Mining merupakan bidang ilmu untuk proses mendapatkan pengetahuan atau pola dari basis data [8]. Sehingga dengan pola yang diperoleh dapat dijadikan solusi pengambilan keputusan.

Dalam penelitian ini akan memanfaatkan data mining untuk mengklasifikasi data mahasiswa, dengan mengkomparasi 3 algoritma yaitu

Decision Tree, K-NN dan Neural Network.

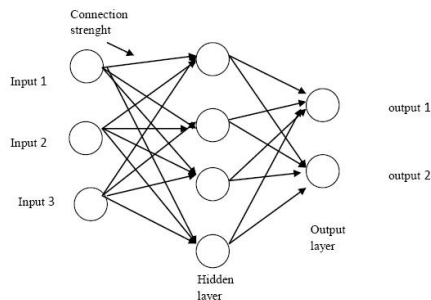
c. Metode Klasifikasi Data Mining

Suatu proses penemuan model yang menggambarkan dan membedakan kelas data yang bertujuan bisa digunakan untuk memprediksi kelas disebut klasifikasi [9].

Proses klasifikasi didasarkan pada empat komponen[10]: (1) kelas, (2) Predictor, (3) Training dataset, dan (4) Testing dataset.

d. Algoritma neural network

Neural Network merupakan algoritma yang meniru fungsi otak manusia yang didalamnya terdapat ribuan dan jutaan unit pengolahan kecil, yang disebut dengan neuron. Dan setiap neuron saling terhubung satu sama lain dan kemudian neuron tersebut memiliki input dan keluaran [11].



Gambar 1. Arsitektur Neural Network

Langkah perhitungan dalam algoritma Neural Network adalah sebagai berikut [12]:

1. Inisialisasi bobot jaringan secara acak (biasanya antara -0.0 sampai 1.0)
2. Untuk setiap data pada *training*, hitung input untuk simpul berdasarkan nilai input dan bobot jaringan saat itu, menggunakan rumus:

$$input_j = \sum_{i=1}^n o_i w_{ij} + \theta_j$$

Keterangan:

o_i = Output simpul i dari layer sebelumnya

w_{ij} = Bobot relasi dari simpul i pada layer sebelumnya ke simpul j

θ_j = bias

3. Berdasarkan input dari langkah dua, selanjutnya membangkitkan output untuk simpul menggunakan fungsi aktifitas sigmoid

$$Output = \frac{1}{1 + e^{-Input}}$$

4. Hitung nilai *Error* antara nilai yang diprediksi dengan nilai sesungguhnya menggunakan rumus:

$$Error_j = Output_j \cdot (1 - Output_j) \cdot (Target_j - Output_j)$$

Keterangan:

$Output_j$ = Output aktual dari simpul j

$Target_j$ = Nilai target yang sudah diketahui pada data *training*

5. Setelah nilai *Error* dihitung, selanjutnya dibalik ke layer sebelumnya. Untuk menghitung nilai *Error* pada *hidden layer*, menggunakan rumus:

$$Error_j = Output_j(1 - Output_j) \sum_{k=1}^n Error_k w_{jk}$$

6. Nilai *Error* yang dihasilkan dari langkah sebelumnya digunakan untuk memperbarui bobot relasi menggunakan rumus:

$$w_{ij} = w_{ij} + l \cdot Error_j \cdot Output_i$$

Keterangan:

w_{ij} = bobot relasi dari unit i pada layer sebelumnya ke unit j

l = *learning rate* (konstanta, nilainya antara 0 sampai dengan 1)

$Error_j$ = *Error* pada output layer simpul j

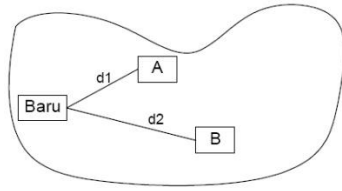
$Output_i$ = Output dari simpul i

e. Algoritma k-Nearest Neighbor (K-NN)

Algoritma *K-Nearest Neighbor* (K-NN) merupakan salah satu metode *data mining* pengklasifikasian data berdasarkan similaritas dengan label data [13]. Similaritas tersebut menggunakan matriks jarak dimana untuk satuan jaraknya menggunakan Euclidean

$$D(a, b) = \sqrt{\sum_{k=1}^d (a_k - b_k)^2}$$

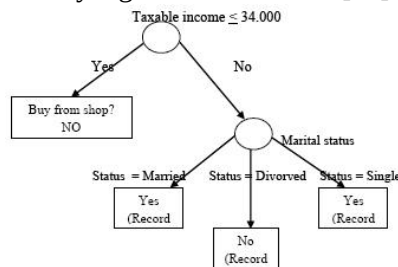
Dimana matriks $D(a, b)$ adalah jarak skalar dari kedua vektor a dan b dari matriks dengan ukuran d dimensi.



Gambar 2. Ilustrasi Tetangga Terdekat

f. Algoritma Decision Tree

Algoritma *Decision Tree* digunakan untuk membangun pohon keputusan yaitu dengan mengubah data menjadi pohon keputusan yang memiliki aturan. [10].



Gambar 3. Contoh Konsep Keputusan Sederhana [10].

Tahapan dalam membuat sebuah pohon keputusan algoritma *Decision Tree* [10], yaitu sebagai berikut:

1. Menyediakan data *training*
2. Menentukan akar dengan rumus *entropy*

$$Entropy(i) = - \sum_{j=1}^m f(i,j) \cdot 2f[(i,j)]$$

3. Hitung nilai *gain* dengan rumus:

$$gain = - \sum_{i=1}^p \frac{n_i}{n} IE(i)$$

4. Untuk menghitung *gain ratio* perlu diketahui suatu term baru yang disebut *Split Inormation* dengan rumus:

$$Split Information = - \sum_{t=1}^c \frac{s_t}{s} \log_2 \frac{s_t}{s}$$

5. Selanjutnya menghitung *ratio*

$$Gainratio(S.A) = \frac{Gain(S.A)}{SplitInformation(S.A)}$$

6. Tahap terakhir Mengulangi langkah ke-2 hingga semua *record* terpartisi

g. Kerangka Pemikiran

Sesuai dengan permasalahan diatas, maka kerangka pemikiran yang dilakukan adalah:



Gambar 4. Kerangka Pemikiran

3. Metode

Dalam penelitian ini menggunakan data kelulusan mahasiswa di Universitas Negeri Semarang.

Data yang diolah sebanyak 410 mahasiswa dari 6 jurusan yaitu; jurusan KTP, PLS, BK, PGSD, PPAUD, dan Psikologi. Dengan jumlah yang tepat waktu sebanyak 289 mahasiswa dan dan terlambat sebanyak 121 orang.

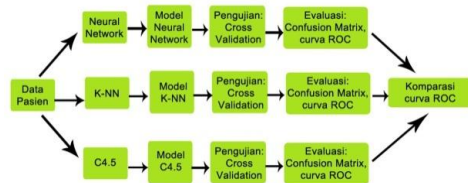
Tabel 2. Dataset Kelulusan Mahasiswa

No	Jurusan	Jenis Kelamin	Umur	IP1	IP2	IP3	IP4	Status
1	KTP	Perempuan	24	3.2	3.36	3.18	3.21	Tepat
2	KTP	Laki-laki	25	3.3	3.45	3.32	3.13	Tepat
3	KTP	Perempuan	25	3.3	3.59	3.42	3.29	Tepat
4	KTP	Perempuan	25	3.55	3.42	3.35	3.55	Tepat
5	KTP	Perempuan	23	3.3	3.45	3.36	3.3	Tepat
6	KTP	Perempuan	24	3.35	3.41	3.18	3.7	Tepat
7	KTP	Perempuan	23	3.1	3.18	3.09	3.07	Tepat
8	KTP	Perempuan	24	3.15	3.23	3.23	3.24	Tepat
9	KTP	Laki-laki	23	3.15	3.18	3.32	3.21	Tepat
10	KTP	Laki-laki	26	3.05	2.6	3.14	3.16	Terlambat
11	KTP	Laki-laki	23	3.15	3.5	3.32	2.74	Tepat
12	KTP	Laki-laki	24	3.2	3.25	3.2	2.88	Tepat
13	KTP	Laki-laki	26	2.9	2.55	3.12	3.23	Terlambat
14	KTP	Laki-laki	24	2.95	3.36	3.18	3.19	Tepat
15	KTP	Perempuan	25	3.4	3.59	3.17	3.26	Tepat
16	KTP	Laki-laki	27	2.88	2.64	3.14	3.32	Terlambat
17	KTP	Laki-laki	25	3.05	3.23	2.95	3.05	Tepat
18	KTP	Laki-laki	23	3.4	3.59	3.25	3.12	Tepat
19	KTP	Laki-laki	25	3	3	3.02	2.44	Tepat
20	KTP	Laki-laki	24	3.2	3.41	3.36	3.38	Tepat
21	KTP	Laki-laki	25	2.75	2.74	3.14	2.52	Terlambat
22	KTP	Perempuan	27	2.73	2.05	3.11	3.07	Terlambat
23	KTP	Laki-laki	23	3.1	2.73	2.68	2.68	Tepat
24	KTP	Laki-laki	25	2.65	3.23	2.82	2.64	Tepat

Model yang diusulkan pada penelitian ini adalah menggunakan tiga metode klasifikasi data mining. Metode yang diusulkan untuk pengolahan dataset kelulusan mahasiswa adalah penggunaan algoritma *Decision Tree*, *neural network*,

dan *k-Nearest Neighbord (K-NN)* dilanjutkan dengan pengujian menggunakan *k-fold cross validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan kurva *ROC*. Tahap selanjutnya adalah memperbandingkan hasil akurasi dan AUC dari tiap model sehingga diperoleh model dari metode klasifikasi yang memperoleh nilai akurasi dan AUC tertinggi.

Dalam tahapannya seperti gambar dibawah ini:

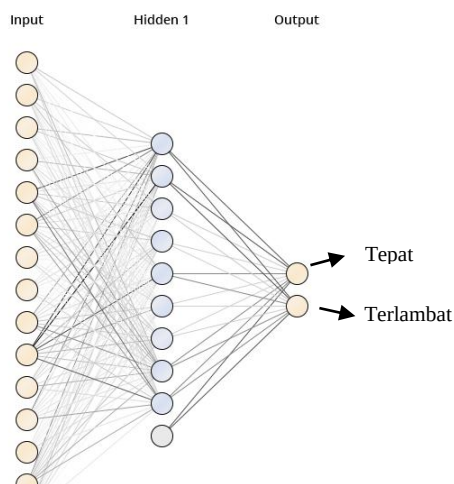


Gambar 5. Metode yang diusulkan

4. Hasil dan Pembahasan

a. Algoritma *Neural Network*

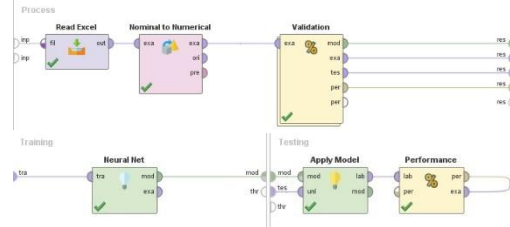
Setelah melakukan Preprocessing data, peneliti melakukan klasifikasi dengan menggunakan algoritma *Neural Network* dengan *training cycles*: 200 dan *learning rate*: 0,3 sehingga menghasilkan model seperti gambar dibawah ini



Gambar 6. Model Algoritma Neural Network

Tahap selanjutnya pengujian model algoritma *neural network* tersebut diatas dengan *Cross Validaton*, yaitu teknik validasi dengan membagi data secara acak kedalam *k* bagian dan masing-masing bagian akan dilakukan proses

klasifikasi[14]. Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak *k*. Data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai *k* dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai *k* yang digunakan berjumlah 10 atau *10-fold Cross Validation*.

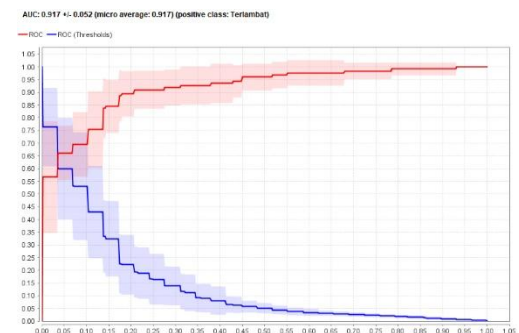


Gambar 7. Pengujian K-Fold Cross Validation Metode Neural Network

Hasil pengujian tersebut diatas diperoleh *accuracy* 87,32% dan *AUC* 0,917.

accuracy: 87.32% +/- 3.21% (micro average: 87.32%)			
	true Tepat	true Terlambat	class precision
pred. Tepat	271	34	88.85%
pred. Terlambat	18	87	82.86%
class recall	93.77%	71.90%	

Gambar 8. Nilai Akurasi *Neural Network*



Gambar 8. Grafik Area Under Curva *Neural Network*

b. Algoritma *K-Nearest Neigbord*

Setelah melakukan Preprocessing dataset kelulusan mahasiswa, peneliti melakukan klasifikasi dengan menggunakan algoritma *K-Nearest Naegboard (K-NN)*, dengan mengklaster *k=2* dari data 410, dengan keadaan sakit sejumlah 257 orang dan keadaan sehat sejumlah 310 orang. Maka model yang

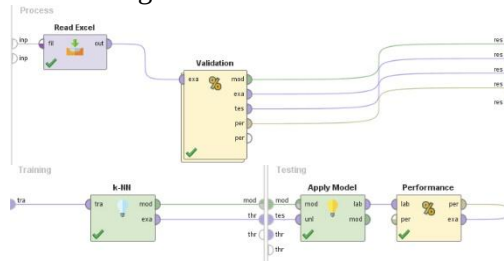
diperoleh seperti hasil eksekusi dibawah ini.

KNNClassification

Weighted 2-Nearest Neighbour model for classification.
The model contains 410 examples with 8 dimensions of the following classes:
Tepat
Terlambat

Gambar 9. Model Algoritma K-Nearest Neighbour

Tahap selanjutnya pengujian model algoritma *K-Nearest Neighbour* tersebut diatas dengan *Cross Validation*.



Gambar 10. Pengujian K-Fold Cross Validation Metode K-Nearest Neighbour

Hasil pengujian tersebut diatas diperoleh *accuracy* 83,66% dan *AUC* 0,842.

accuracy: 83.66% +/- 5.64% (micro average: 83.66%)			
	true Tepat	true Terlambat	class precision
pred. Tepat	257	30	88.01%
pred. Terlambat	32	86	72.89%
class recall	88.95%	71.07%	

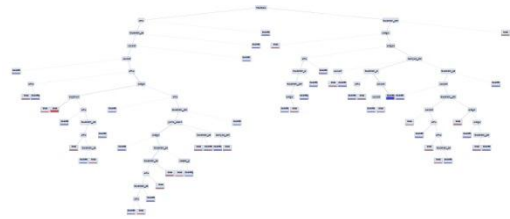
Gambar 11. Nilai Akurasi K-Nearest Neighbour



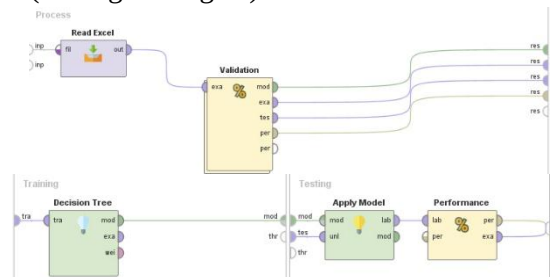
Gambar 12. Grafik Area Under Curva K-Nearest Neighbour

c. Algoritma Decision Tree

Dalam eksperimennya menggunakan algoritma *Decision Tree* terhadap dataset kelulusan mahasiswa menggunakan *criterion* yaitu *gaint_ration* dan *maximal depht* yaitu 20. Maka model yang dihasilkan seperti gambar dibawah ini



Dalam pengujian *K-Fold Cross Validation* Algoritma *Decision Tree* peneliti menggunakan 10 kali percobaan dengan sampling type *Stratified* (bertingkat-tingkat).

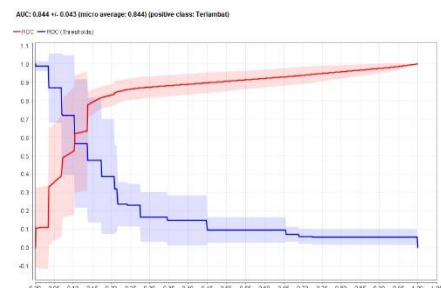


Gambar 14. Pengujian K-Fold Cross Validation Metode Decision Tree

Hasil pengujian tersebut diatas diperoleh *accuracy* 85,37% dan *AUC* 0,844.

accuracy: 85.37% +/- 3.45% (micro average: 85.37%)			
	true Tepat	true Terlambat	class precision
pred. Tepat	253	34	88.55%
pred. Terlambat	26	87	76.99%
class recall	91.05%	71.90%	

Gambar 15. Nilai Akurasi Decision Tree



Gambar 16. Grafik Area Under Curva Decision Tree

Berdasarkan dari analisis pengujian masing-masing metode diatas maka dapat dirangkumkan hasilnya seperti Tabel 3 berikut:

Tabel 3. Perbandingan *Performance* Metode

	NN	K-NN	DT
Accuracy	87,32%	83,66%	85,37%
AUC	0,917	0,842	0,844

Melihat hasil perhitungan yang terangkum pada Tabel 3 diatas dengan menerapkan klasifikasi *performance* keakurasian AUC maka diperoleh hasil penelitian yaitu, terdapat satu metode yang merupakan kategori *Excellent Clasification* yaitu untuk metode *Neural Network* dengan nilai AUC 0,917, dan kategori *Good Clasification* terdapat 2 metode yaitu *K-Nearest Neigboard* dengan AUC 0,842 dan *Decision Tree* dengan nilai AUC 0,844.

Berdasarkan Tabel 3 diatas juga dapat dilihat bahwa nilai akurasi maka dapat diketahui metode yang terbaik adalah *Neural Network* dengan nilai akurasi 87,32%, yang kedua adalah algoritma *Decision Tree* dengan nilai akurasi 85.37%, dan yang terakhir adalah *K-Nearest Neigboard* dengan nilai 83,66%

5. Kesimpulan

Dari penelitian yang telah dilakukan menggunakan algortima *neural network*, *k-nearest neigboard* dan *decision tree* menggunakan data mahasiswa untuk prediksi kelulusan mahasiswa. Model yang dihasilkan dikomparasi untuk dapat diketahui metode terbaik dalam prediksi kelulusan mahasiswa. Untuk mengukur kineja medel digunakan confusion matrix dan kurva ROC, dan diketahui bahwa Metode algoritma *neural network* menghasilkan nilai akurasi yaitu 87,32% dan nilai AUC 0,917, Metode algoritma *Decision Tree* dengan nilai akurasi 85.37% dan nilai AUC 0,844, dan yang terakhir adalah *K-Nearest Neigboard* dengan nilai 83,66% dan nilai AUC 0,842.

Dengan demikian metode *neural network* adalah metode yang terbaik untuk pemecahan masalah prediksi kelulusan mahasiswa.

6. Daftar Pustaka

[1] A. Rohman, "Model Algoritma K-Nearest Neighbor (K-Nn) Untuk

Prediksi Kelulusan Mahasiswa," *Neo Tek.*, vol. 1, no. 1, 2015.

- [2] H. Romadhona, Agus; suprapedi; himawan, "Prediksi Kelulusan Tepat Waktu Mahasiswa Stmik-Ymi," *J. Teknol. Inf.*, vol. 13, no. 1, pp. 69–83, 2017.
- [3] Azwar, *Penyusunan Skala Psikologi*. Yogyakarta: Pustaka Pelajar, 2014.
- [4] S. T. Karamouzis and A. Vrettos, "An Artificial Neural Network for Predicting Student Graduation Outcomes," *Lect. Notes Eng. Comput. Sci.*, vol. 2173, no. 1, pp. 991–994, 2008.
- [5] E. Marselina Silvia Suhartinah, "Graduation Prediction Of Gunadarma University Students Using Algorithm And Naive Bayes C4.5 Algorithm," *Fac. Ind. Technol. Gunadarma Univ.*, 2010.
- [6] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," vol. 7, pp. 292–295, 2010.
- [7] J. Yingkuachat, P. Praneetpolgrang, and B. Kijsirikul, "An Application of the Probabilistic Model to the Prediction of Student Graduation Using Bayesian Belief Networks," *Electr. Eng. Comput. Telecommun. Inf. Technol. Assoc. Thai. (ECTI Thailand)*, pp. 63–71, 2007.
- [8] I. H. Witten, "Data Mining Data Mining Complications: Overfitting Statistical modeling One attribute does all the work?," 2007.
- [9] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," *Canada B. chapter Data Mining Concepts Tech.*, pp. 3–26, 2000.
- [10] F. Gorunescu, *Data Mining: Concepts, Models and Techniques (Intelligent Systems Reference Library)*. 2011.
- [11] R. Kala, H. Vazirani, A. Shukla, and R. Tiwari, "Offline Handwriting Recognition using Genetic Algorithm," *Int. J. Comput. Sci.*, vol. 7, no. 2, pp. 16–25, 2010.
- [12] G. J. Myatt, *Making Sense of Data: A Practical Guide to Exploratory Data*

- Analysis and Data Mining*. 2006.
- [13] D. T. Larose, *Data Mining Methods and Models*. 2006.
- [14] J. Han and M. Kamber, *Mining Stream, Time-Series and Sequence Data*, vol. 54. 2006.