Comparison of C4.5 Algorithm and K-Nearest Neighbors on the Classification of Multiple Intelligence Test Results for Recommended Student Lectures

Yusuf Fadlila Rachman
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
yusufadil064@student.uns.ac.id

Ristu Saptono
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
ristu.saptono@staff.uns.ac.id

Winarno
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
win@staff.uns.ac.id

Abstract

This research uses K-Nearest Neighbors and C4.5 classification algorithm aimed to classify student lecture field based on multiple intelligence test result. There are 2 categories of fields used in the classification, namely Science-Technology (SainsTech) and Social-Humanities (Soshum.) The data used obtained from multiple intelligence test conducted to students of Sebelas Maret University Surakarta. The collected data sets will be transformed and normalized for a classification process. The selection feature is performed to remove any inappropriate attributes.

The test is done by using confusion matrix, by doing 12 experimental scenarios with different datasets and classification techniques. From the experimental results, it is shown that the K-Nearest Neighbors algorithm is better than the C4.5 dataset of normalization result has the highest accuracy of 56.84% to 54.84%.

Keywords: *C4.5, K-Nearest Neighbors, multiple intelligence, Sains, Sosial-Humaniora*

1. Pendahuluan

Kecerdasaan merupakan salah satu faktor yang menentukan masa depan seseorang. Pada dasarnya setiap orang memiliki jenis kecerdasan yang berbeda, kemampuan dan bakat yang bebeda pula. Jenis kecerdasan yang dimiliki dapat diketahui sejak dini. Banyak terdapat tes kepribadian atau psikotest yang dapat membantu untuk mengidentifikasi jenis kecerdasan yang dimiliki siswa. Tes multiple intelligence merupakan salah satu metode yang digunakan untuk mengidentifikasi jenis kecerdasan yang dimiliki siswa. Konsep *Multiple Intelligence* dicetuskan oleh Howard Gardner pada tahun 1983 [**1**]. Howard Gardner menjelaskan bahwa *Multiple Intelligence* atau Kecerdasan Majemuk merupakan bentuk klasifikasi kecerdasan manusia secara spesifik. Dia membagi kecerdasan manusia kedalam 8 jenis kecerdasan yang lebih spesifik [**2**] yaitu

1.Linguistic Intelligence
2.Logical – Mathematical Intelligence
3.Musical Intelligence
4.Bodily – Kinesthetic Intelligece
5.Spacial – Visual Intelligence
6.Interpersonal Intelligence
7.Intrapersonal Intelligence
8.Naturalistic Intelligence

. Klasifikasi merupakan salah satu cabang ilmu dari data mining. Klasifikasi memungkinkan untuk melakukan analisa pada kelompok data untuk mendapatkan sebuah informasi atau pola tertentu. Terdapat beberapa algoritma yang digunakan dalam klasifikasi antara lain C4.5 dan *K-Nearest Neighbor (KNN)*. Algoritma C4.5 merupakan kelompok algoritma *Decision Tree*. Algoritma ini mempunyai input berupa *training samples* dan *samples*. Training samples berupa data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan samples merupakan *field-field data* yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data [**3**]. Algoritma C4.5 dibuat oleh Ross Quinlan yang merupakan pengembangan dari ID3 yang juga dibuat oleh Quinlan [**4**]. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi *missing value*, bisa mengatasi *continue data*, dan *pruning*.

KNN adalah suatu metode yang menggunakan algoritma supervised dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan training sample.

Hasil tes *multiple intelligence* yang dilakukan akan diklasifikasikan menggunakan algoritma KNN dan C4.5 ke dalam dua kelompok yaitu sains-teknologi dan sosial-humaniora.

2. Dasar Teori

2.1. Sistem Rekomendasi

Sistem rekomendasi adalah sebuah peranngkat lunak yang bertujuan untuk membantu pengguna dengan memberikan saran, pilihan atau rekomendasi kepada pengguna ketika dihadapkan pada suatu masalah dengan informasi yang berjumlah besar. Sistem rekomendasi merupakan bagian dari sistem penyaring informasi yang berguna untuk memprediksi nilai atau pilihan pada sebuah masalah yang dialami pengguna [**5**]. Sistem rekomendasi telah berkembang cukup pesat dan digunakan di berbagai bidang. Beberapa contoh penggunaan sistem rekomendasi yang cukup populer antara lain, film, musik, buku, pariwisata, dan lain – lain.

Terdapat dua cara yang dapat digunakan oleh sistem rekomendasi untuk memproduksi sebuah hasil, yaitu dengan *collaborative filtering* dan *content based filtering* [**6**]. Metode *Collaborative Filtering* digunakan dengan cara mengumpulkan dan menganalisa dalam jumlah besar informasi user berupa kebiasaan, kegiatan, dan memprediksi kesukaan user berdasarkan kesamaan dengan user lain. Keuntungan dari penggunaan metode *Collaborative Filtering* adalah tidak begantung pada analisis mesin dan dapat digunakan dalam rekomendasi masalah yang kompleks, seperti rekomendasi film. Contoh algoritma yang termasuk dalam *collaborative filtering* adalah *k-nearest neighbors* [**7**]. Metode *content-based filtering* digunakan berdasarkan dari deskripsi barang dan informasi pengguna [**8**]. Pada metode ini *keyword* digunakan untuk mendeskripsikan barang dan informasi pengguna digunakan untuk menunjukkan tipe barang yang penggunaan suka.

2.2. Data Mining

Data mining merupakan salah satu cabang dari komputer sains. Data mining merupakan sebuah proses menemukan suatu pola dari suatu data set yang besar dan melibatkan metode dalam kecerdasan buatan, pembelajaran mesin, statistik dan sistem database [**9**]. Dalam data mining terdapat dua jenis pembelajaran yaitu *supervised* dan *unsupervised*. *Supevised Learning* digunakan untuk memprediksi suatu nilai sedangkan *Unsupervised Learning* digunakan untuk mencari struktur intrinsik, relasi dalam suatu data yang tidak memerlukan class atau label sebelum dilakukan proses pembelajaran. Contoh dari alogritma *unsupervised* adalah *k-means clusteing* dan *apriori association rules*. Contoh dari supervised adalah *NaiveBayes* dan *C4.5* untuk klasifikasi.

Data mining dapat diklasifikasi berdasarkan fungsi yang dilakukan atau berdasarkan jenis aplikasi yang menggunakannya [**10**], yaitu :

1. *Anomaly detection* : Mengidentifikasi kelompok data acak yang mungkin menarik atau error data yang membutuhkan pengecekan lanjutan.
2. *Association rule learning* : Mencari hubungan antara variabel. Contoh : sebuah supermarket memiliki data kebiasaan belanja pada pengunjung. Memakai association rule learning, supermarket dapat mengidentifikasi produk mana yang sering dibeli dan memakai informasi ini untuk strategi marketing.
3. *Clustering* : Menemukan kelompok atau truktur dalam suatu data yang mungkin mirip dalam suatu hal, tanpa menggunakan struktur data yang telah diketahui.
4. *Classification* : Generalisasi struktur yang dikenal untuk diterapkan pada data baru. Contoh : email program mengelompokkan email sebagai email sampah atau tidak.
5. *Regression* : Mencoba menemukan fungsi yang memodelkan data dengan error yang sedikit.
6. *Summarization* : Menyediakan representasi data set lengkap, termasuk visualisasi dan pembuatan laporan.

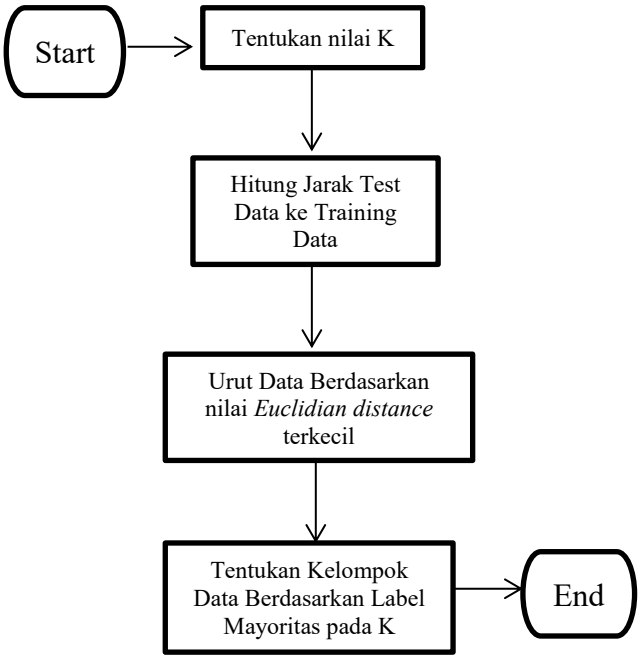
2.3. Klasifikasi

Dalam terminologi pembelajaran mesin, klasifikasi dianggap sebagai salah satu metode pembelajaran, dimana kelompok data pelatihan yang diidentifikasi benar-benar tersedia [**11**]. Klasifikasi termasuk ke dalam *supervised learning* dalam data mining. Klasifikasi terdiri dari prediksi sebuah hasil tertentu berdasarkan input yang diberikan. Untuk memprediksi kemungkinan hasil tersebut, algoritma memproses data training yang berisi sebuah data atribut dan hasil masing-masing, biasanya disebut goal atau atribut prediksi. Algoritma mencoba untuk menemukan hubungan antar atribut yang mungkin digunakan untuk memprediksi hasil. Berikutnya algoritma yang diberikan kelompok data yang tidak terlihat disebut prediction set, yang berisi atribut yang sama, kecuali untuk atribut prediksi yang belum diketahui. Algoritma menganalisis input dan menghasilkan sebuah prediksi. Tingkat keakuratan dari prediksi tersebut menggambarkan seberapa baik algoritma itu. Contoh : pada database medis, *set training* akan memiliki rekaman informasi pasien yang relevan sebelumnya, dimana atribut prediksi adalah apakah pasien memiliki masalah jantung atau tidak.

2.4. K-Nearest Neighbors

Dalam pengenalan pola, algoritma KNN merupakan metode non parametrik yang digunakan untuk klasifikasi dan regresi [**12**]. Algoritma KNN memiliki kelebihan yaitu dapat menghasilkan data yang kuat atau jelas dan efektif jika digunakan pada data yang besar.

Algoritma KNN dapat ditunjukkan pada *flowchart* berikut :



Penentuan nilai k yang baik bergantung pada data yang digunakan. Umumnya, nilai k yang lebih besar mengurangi efek *noise* pada klasifikasi, namun menyebabkan batasan antar kelas sedikit berbeda. Jarak antar data yang digunakan dalam algoritma KNN dapat dihitung menggunakan rumus jarak *Euclidean Distance* dan *Manahattan Distance*. Rumus jarak yang biasa digunakan untuk KNN adalah *Euclidean Distance*,

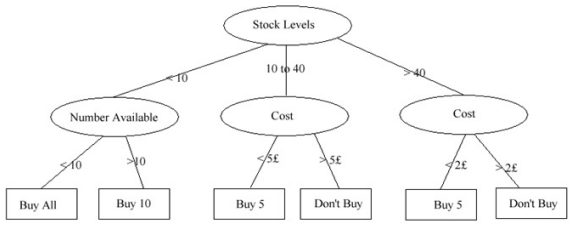
yang merepresentasikan cara berpikir manusia tentang jarak pada kehidupan nyata [**13**]. Berikut adalah rumus *Euclidean Distance* :

$$D_{Euclidean}(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Dimana xk dan yk merupakan atribut ke- k dari x dan y berturut - turut.

2.5. C4.5 Algorithm

Algoritma C4.5 termasuk dalam *supervised learning* dimana membutuhkan sebuah set contoh pelatihan dan tiap contoh dapat dipasangkan, yaitu objek input dan nilai output. Algoritma ini menganalisa *training set* dan membangun *classifier* yang harus dapat mengklasifikasi *training* dan contoh uji dengan benar. Sebuah contoh uji adalah objek input dan algoritma harus memprediksi sebuah nilai output. *Classifier* yang digunakan C4.5 adalah pohon keputusan yang dibangun dari akar hingga daun. Berikut adalah contoh pohon keputusan menggunakan C4.5.



Pohon keputusan dalam algoritma C4.5 dibuat dengan memilih atribut sebagai akar. Kemudian dibuat cabang untuk tiap nilai di dalam akar tersebut. Langkah berikutnya yaitu membagi kasus dalam cabang. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama. Untuk memilih atribut dengan akar, didasarkan pada nilai gain tertinggi dari atribut yang ada. Untuk menghitung gain digunakan rumus berikut [**3**] :

Gain : G(S,A) = E(S) - $\sum_{i=1}^{m} Pr(A_i)E(S_{Ai})$

Keterangan :
G(S,A) = gain dari S setelah pembagian pada atribut A.
E(S) = entropi informasi dari S
m = nilai nr atribut A dalam S
Pr(Ai) = frekuensi dari kasus dimana terdapat Ai dalam S
E(SAi)= subset dari S dengan item yang memiliki nilai Ai

Untuk perhitungan nilai entropi dapat dilihat pada rumus berikut :

Entropy : E(S) = $\sum_{i=1}^{m} - Pr(C_i) * log_2 Pr(C_i)$

Keterangan :
E(S) = entropi informasi dari S
m = nilai nr atribut A dalam S
Pr(Ci) = frekuensi dari kasus dimana terdapat Ai dalam S

2.6. Pruning

Jumlah data yang terlalu banyak dan beragam akan membuat struktur pohon keputusan menjadi terlalu rumit [**11**]. Terdapat dua metode dalam melakukan pemangkasan (*pruning*) dalam pohon keputusan, yaitu :

1. *Prepruning*: menghentikan proses pembuatan cabang pada titik tertentu. Berikut adalah rumus prepruning :

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} - \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

r = nilai perbandingan error rate

n = total sample

z = Φ-1(c)

c = confidence level

2. *Postpruning*: menyederhanakan pohon dengan cara membuang beberapa cabang subtree setelah pohon selesai dibangun. Metode *postpruning* merupakan metode standard untuk algoritma C4.5. Salah satu algoritma yang digunakan dalam postpruning adalah *Reduced Error Pruning* (REP) [**9**].

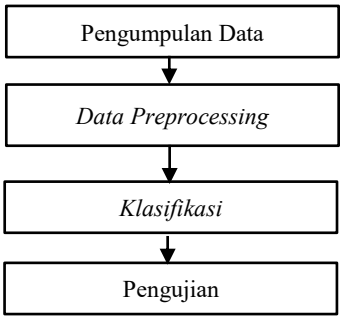
2.7. Multiple Intelligence

Multiple Intelligences merupakan istilah dari kajian teori tentang ilmu kecerdasan yang memiliki makna "kecerdasan majemuk" atau "kecerdasan ganda". Gardner [**14**] pada teorinya menjelaskan bahwa kecerdasan manusia tidak terdiri dari satu tipe kecerdasan saja. Berikut penjelasan 8 tipe kecerdasan menurut Gardner [**14**] :

1. Kecerdasan Linguistik
2. Kecerdasarn Logika-Matematika
3. Kecerdasan Visual-Spasial
4. Kecerdasan Gerak Tubuh (Kinetis)
5. Kecerdasan Musikal
6. Kecerdasan Interpersonal
7. Kecerdasan IntraPersonal
8. Kecerdasan Naturalis

3. Metodologi

Tahapan yang dilakukan oleh penulis dalam menyelesaikan penelitian ini dijelaskan pada Gambar 3.1 berikut ini :



Gambar 3.1 Metodologi Penelitian

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini didapat dengan melakukan survey secara online terhadap mahasiswa Universitas Sebelas Maret. Survey dilakukan dengan menggunakan kuisioner. Data yang diambil kemudian akan diklasifikasi menggunakan algoritma KNN dan C4.5 berdasarkan aturan yang telah ditentukan dan digunakan sebagai data training untuk rekomendasi.

3.2. *Text Preprocessing*

Data yang telah didapat akan dilakukan proses *preprocessing* atau pengolahan data terlebih dahulu dengan tujuan untuk memperoleh pola yang lebih baik. Proses dari *preprocessing* meliputi sebagai berikut :

3.2.1. Data Selection

Seluruh data yang berhasil didapat akan dilakukan proses seleksi. Seleksi dilakukan berdasarkan fakultas tiap mahasiswa. Data yang di gunakan hanya data mahasiswa yang telah menjawab semua pertanyaan survey yang diberikan. Data fakultas kemudian dikelompokkan menjadi 2 bagian yaitu Sains-Teknologi (SainsTek) dan Sosial-Humaniora (Soshum).

3.2.2. Data Cleaning

Data *cleaning* atau pembersihan data dilakukan setelah tahap seleksi. Kelompok data yang telah diseleksi dan disesuaikan dengan kelompoknya, dibersihkan dari data yang hilang, data yang tidak konsisten, dan data yang terjadi duplikasi.

3.2.3. Data Transformation

Data hasil seleksi akan diubah menjadi dataset yang digunakan dalam klasifikasi. Pertama adalah membuat dataset pertama, yaitu dengan melakukan pengisian nilai pada jawaban hasil survey. Sangat Setuju = 4, Setuju = 3, Kurang Setuju = 2, Tidak Setuju = 1. Dataset ini akan disebut dengan data-awal. Dataset kedua dibuat dengan merubah pada nilai numerik yang sudah diisikan dengan abjad A,B,C,D. Nilai 4 diganti dengan A, 3=B, 2=C, 1=D disebut dengan data-transform. Dataset ketiga dibuat dengan mentransformasi dataset kedua kedalam bentuk biner 0 1. Dataset ketiga ini dinamakan sebagai data-transform

3.3. Klasifikasi

Implementasi dan pengembangan aplikasi pada penelitian ini dilakukan dengan mengklasifikasi data training ke dalam kelas-kelas yang telah ditentukan menggunakan algoritma KNN dan algoritma C4.5. Proses penglasifikasian ini bertujuan untuk mendapatkan rule atau aturan yang digunakan dalam rekomendasi. Hasil dari kedua proses klasifikasi akan dibandingkan berdasarkan tingkat akurasi kedua algoritma tersebut.

3.4. Evaluasi

Proses evaluasi pada penelitian ini menggunakan perhitungan akurasi, *precision* dan *recall* dari hasil klasifikasi yang disajikan dengan tabel *confusion matrix* pada Tabel 3.1.

Tabel 3.1 *Confusion Matrix*

Kelas	Klasifikasi Positif (SainsTek)	Klasifikasi Negatif (Soshum
Positif (SainsTek)	TP (True Positif)	FP (False Positif)
Negatif (Soshum)	TN (True Negatif)	FN (False Negatif)

Dengan persamaan perhitungan sebagai berikut :

$$\text{Accuracy} : \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100\%$$

$$\text{Precision} : \frac{TP}{TP+FP} \text{ x } 100\%$$

$$\text{Recall} : \frac{TP}{TP+FN} \text{ x } 100\%$$

4. Hasil Dan Pembahasan

4.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data hasil *survey* yang dilakukan pada mahasiswa Universitas Sebelas Maret (UNS). *Survey* dilakukan secara online menggunakan aplikasi google form kepada 279 *respondent*. Survey yang disajikan berisi data fakultas tiap mahasiswa dan jawaban dari 40 pertanyaan yang merepresentasikan 8 jenis kecerdasan majemuk. Pertanyaan yang digunakan pada *survey* dibagi menjadi 8 kategori kecerdasan majemuk, yaitu Linguistik, Matematika dan Logika, Spasial (Visual), Kinetis (Gerak), Interpersonal, Intrapersonal, dan Spasial. Tiap kategori kecerdasan majemuk berisi 5 pertanyaan dengan total 40 pertanyaan. Terdapat 4 jawaban yang dapat diisikan oleh mahasiswa yaitu sangat setuju (SS), setuju (S), kurang setuju (KS), dan tidak setuju (TS). Keempat jawaban tersebut masing-masing memiliki nilai SS = 4, S = 3, KS = 2, TS = 1.

4.2. *Data Preprocessing*

4.2.1. Data Selection

Seluruh data yang berhasil didapat akan dilakukan proses seleksi. Seleksi dilakukan berdasarkan fakultas tiap mahasiswa. Data yang di gunakan hanya data mahasiswa yang telah menjawab semua pertanyaan *survey* yang diberikan. Data fakultas kemudian dikelompokkan menjadi 2 bagian yaitu Sains dan Sosial-Humaniora (Soshum).

4.2.2. Data Cleaning

Tahap data *cleaning* akan dilakukan pembersihan data. Pembersihan dilakukan dengan menghilangkan *missing values* dan menyeragamkan data-data yang tidak sesuai.

Missing values merupakan suatu kondisi dimana data yang didapat tidak sesuai atau tidak dapat digunakan. Dalam penelitian yang menggunakan survey sebagai sumber data, *missing values* pada data sering terjadi pada data jawaban yang didapat.

4.2.3. Data Transformation

Transformasi data merupakan sebuah langkah untuk merubah kelompok data mentah kedalam kelompok data yang dapat digunakan untuk klasifikasi. Ada tiga tahap transformasi data yang dilakukan dalam penelitian ini. Pertama, data mentah berupa jawaban dari mahasiswa diisikan nilai berdasarkan nilai yang telah ditentukan. Nilai yang ditambahkan seperti berikut : Sangat Setuju = 4, Setuju = 3, Kurang Setuju = 2, Tidak Setuju = 1. Kedua dengan mengganti nilai tadi ke dalam alphabet yang ditentukan, yaitu nilai 1=A, 2=B, 3=C, dan 4=D. Ketiga dengan menggunakan normalisasi terhadap data yang digunakan. Tabel normalisasi data dapat dilihat pada Tabel 4.1:

Tabel 4.1 Dummy Normalisasi Data

Jawaban	A	B	C	D
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

4.3. Klasifikasi C4.5 dan K-Nearest Neighbors

4.3.1. Klasifikasi C4.5

Pelatihan data dilakukan menggunakan metode C4.5 yang diimplementasikan ke dalam bahasa pemrograman Python. Data yang digunakan dalam pelatihan merupakan hasil *preprocessing*. C*onfusion matriks* hasil klasifikasi yang dilakukan dengan algoritma C4.5 dapat dilihat pada Tabel 4.2 :

Tabel 4.2 Confusion Matriks C4.5

Actual Class	SainsTek		Soshum		Total Data
Predicted Class	SainsTek (TP)	Soshum (FN)	SainsTek (FP)	Soshum (TN)	
Data awal-CV	75	75	70	59	279
Data trans-CV	93	57	69	60	279
Data norm-CV	80	70	63	66	279
Data awal-SP	26	29	18	22	95
Data trans-SP	30	25	23	17	95
Data norm-SP	30	25	23	17	95

Keterangan:
Data awal - CV = Klasifikasi data awal dengan metode cross validation
Data awal - SP = Klasifikasi data awal dengan metode percentage split
Data trans – CV = Klasifikasi data hasil transformasi menggunakan cross validation
Data trans – SP = Klasifikasi data hasil transformasi menggunakan percentage split
Data norm - CV = Klasifikasi data hasil normalisasi menggunakan cross validation
Data norm – SP = Klasifikasi data hasil normalisasi menggunakan percentage split

4.3.2. K-Nearest Neighbors

Pelatihan data dilakukan dengan menggunakan metode *k-nearest neighbors* yang diimplementasikan ke dalam bahasa pemrograman Python. Data yang digunakan untuk pelatihan adalah data hasil tahap *preprocessing*. *Confusion matriks* klasifikasi menggunakan algoritma KNN dapat dilihat pada Tabel 4.3 :

Tabel 4.3 Confusion Matriks KNN

Actual Class	SainsTek		Soshum		Total Data
Predicted Class	SainsTek (TP)	Soshum (FN)	SainsTek (FP)	Soshum (TN)	
Data awal-CV	85	65	73	56	279
Data trans-CV	83	67	75	54	279
Data norm-CV	82	68	75	54	279
Data awal-SP	19	36	18	22	95
Data trans-SP	35	20	21	19	95
Data norm-SP	35	20	21	19	95

4.4. Hasil dan Evaluasi

Proses klasifikasi dilakukan menggunakan 2 algoritma, yaitu *C4.5* dan KNN. Data yang digunakan merupakan hasil dari tes *multiple intelligence* yang diberikan kepada 279 responden. Berdasarkan data yang telah didapat, akan dibuat menjadi 3 jenis dataset yang berbeda yaitu data asli, data normalisasi, dan data transform.

Data asli merupakan dataset hasil pengisian nilai pada jawaban hasil survey. Data transform merupakan data set yang dibuat dengan mengubah nilai data asli ke dalam bentuk alphabetik. Data normalisasi merupakan data hasil normalisasi data transform. Ketiga dataset tersebut akan diklasifikasi menggunakan teknik *cross validation* dan *percentage split*, sehingga secara total akan terdapat 12 hasil klasifikasi yang akan dibandingkan.

Nilai *Precision, Recall, Accuracy* Hasil klasifikasi menggunakan *C4.5* dan *KNN* dapat dilihat pada Tabel 4.3 dan 4.4:

Tabel 4.3 Precision Recall Accuracy C4.5

Hasil Klasifikasi	Precision	Recall	Accuracy	FP Rate	TP Rate
Data awal-CV	51.72%	54.26 %	48.03%	54.26 %	50.00 %
Data trans-CV	57.41%	53.49 %	54.84%	53.49 %	62.00 %
Data norm-CV	55.94%	48.84 %	52.33%	48.84 %	53.33 %
Data awal-SP	59.09%	45.00 %	50.53%	45.00 %	47.27 %
Data trans-SP	56.60%	57.50 %	49.47%	57.50 %	54.55 %
Data norm-SP	56.60%	57.50 %	49.47%	57.50 %	54.55 %

Tabel 4.4 Precision Recall Accuracy KNN

Hasil Klasifikasi	Precision	Recall	Accuracy	FP Rate	TP Rate
Data awal-CV	53.80%	56.59 %	50.54%	56.59 %	56.67 %
Data trans-CV	52.53%	58.14 %	49.10%	58.14 %	55.33 %
Data norm-CV	52.23%	58.14 %	48.75%	58.14 %	54.67 %
Data awal-SP	51.35%	45.00 %	43.16%	45.00 %	34.55 %
Data trans-SP	62.50%	52.50 %	56.84%	52.50 %	63.64 %
Data norm-SP	62.50%	52.50 %	56.84%	52.50 %	63.64 %

Berdasarkan Tabel 4.3 dan 4.4 di atas, dapat dilihat bahwa KNN memiliki tingkat akurasi yang lebih tinggi dibanding C4.5 yaitu 56,84% berbanding 54,84%.

Klasifikasi data menggunakan KNN mencapai akurasi terbaik ketika percobaan kelimadan keenam, dimana data yang digunakan adalah data hasil normalisasi menggunakan teknik *split* dan data transformasi. Sedangkan klasifikasi data C4.5 mencapai tingkat akurasi terbaik pada percobaan kedua dengan data transformasi *cross-validation*.

5. Kesimpulan Dan Saran

5.1. Kesimpulan

Hasil dari penelitian yang telah dilakukan dapat disimpulkan bahwa penggunaan metode klasifikasi K-Nearest Neighbors memiliki akurasi yang lebih baik dibanding dengan penggunaan algoritma C4.5 dalam kasus dataset yang sedikit dengan banyak atribut.

Penelitian ini menggunakan beberapa teknik untuk melakukan pengujian pada dataset yang telah didapat. Percobaan pertama dilakukan klasifikiasi data menggunakan C4.5 sebanyak 6 kali menggunakan data awal, data transformasi, dan data normalisasi. Dalam percobaan tersebut, diperoleh nilai akurasi tertinggi ketika klasifikasi data dilakukan menggunakan data transformasi dan data normalisasi yaitu sebesar 54,84%, precision sebesar 57,41% dan recall sebesar 53,49%. Percobaan kedua dilakukan dengan klasifikasi data menggunakan KNN sebanyak 6 kali menggunakan dataset awal, data transformasi, dan data normalisasi.

Diperoleh tingkat akurasi tertinggi ketika klasifikasi dilakukan menggunakan data transformasi dan data normalisasi yaitu 56,84% dengan nilai precision sebesar 62,50% dan recall 52,50%. Nilai akurasi C4.5 yang lebih rendah daripada KNN dapat disebabkan oleh banyaknya atribut yang digunakan untuk membangun pohon keputusan tidak sebanding dengan banyak data yang digunakan, sehingga menyebabkan klasifikasi tidak akurat. Hal ini dapat diatasi dengan melakukan klasifikasi data menggunakan algoritma berjenis supervised yang lebih sederhana, contohnya K-Nearest Neighbors.

5.2. Saran

1. Banyaknya atribut yang digunakan (40 atribut) sebagai acuan untuk klasifikasi tidak sebanding dengan jumlah data (279) yang diperoleh. Hal ini akan menyebabkan klasifikasi data tidak maksimal karena sebaran data yang tidak merata. Salah satu cara yang dapat dilakukan untuk menyelesaikan masalah ini adalah dengan melakukan *selection* pada variabel yang digunakan dan menambah jumlah data yang digunakan.
2. Untuk melakukan klasifikasi data yang memiliki dataset yang kecil, tetapi memiliki atribut yang beragam, sebaiknya tidak menggunakan algoritma *decision tree* karena akan berakibat pada tingkat akurasi yang lebih rendah.

Daftar Pustaka

[1] Lynn Gilman, *The Theory of Multiple Intelligence*. Indiana: Indiana University, 2012.

[2] Robert Slavin, *Educational Psychology*., 2009.

[3] Snati Sunjana, "Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree," in *Seminar Nasional Aplikasi Teknologi Informasi*, 2010, pp. 24-29.

[4] J.R. Quinlan, *Programs for Machine Learning*.: Morgan Kaufmann Publishers, 1993.

[5] Fransesco Ricci, Lior Rokach, and Brarcha Shapira, *Introduction to Recommender System Handbook*.: Springer, 2011.

[6] Si Hosein Jafarkariimi A.T.H. and R. Saadotdoost, "A Naive Recommendation Model for Large Database," *International Journal of Information and Education Technology*, 2011.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in

Recommender System a Case Study," 2000.

[8] Peter Brusilovsky, "The Adaptive Web," 2007.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning : Data Mining, Inference, and Prediction," 2009.

[10] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, *From Data Mining to Knowledge Discovery in Database*., 1996.

[11] Ethem Alpaydin, *Introduction to Machine Learning*.: MIT press, 2010.

[12] N.S. Altman, *An introduction to kernel and nearest-neighbor nonparametric regression*.: The American Statistician, 1992.

[13] D.T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: Wiley, 2005.

[14] Howard Gardner, *Multiple Intelligence*. New York: Basic Books, 1993.