IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR UNTUK KLASIFIKASI JURUSAN PADA PESERTA DIDIK BARU

Nur Aeni Widiastuti
Fakultas Sains dan Teknologi, Program Studi Teknik Informatika
Universitas Islam Nahdlatul Ulama Jepara
Email: nuraeniwidiastuti@unisnu.ac.id

Maulana Azhar
Fakultas Sains dan Teknologi, Program Studi Teknik Informatika
Universitas Islam Nahdlatul Ulama Jepara
Email: 181240000759@unisnu.ac.id

Harminto Mulyo
Fakultas Sains dan Teknologi, Program Studi Teknik Informatika
Universitas Islam Nahdlatul Ulama Jepara
Email: minto@unisnu.ac.id

ABSTRAK

Penjurusan siswa merupakan suatu proses penempatan siswa ke dalam jurusan tertentu sesuai dengan minat dan kemampuan akademiknya sebagai upaya untuk lebih mempermudah siswa-siswi dalam proses pembelajaran. Madrasah Aliyah Darul Hikmah Menganti merupakan sekolahan sederajat dengan SMA, yang memiliki dua jurusan yaitu IPA, dan IPS. Sulitnya dalam mengklasifikasikan jurusan peserta didik baru menjadi kendala bagi pihak sekolah. Karena proses penilaian kriteria yang dilakukan satu per satu. Dari permasalahan tersebut dilakukan penerapan metode *K-Nearest Neighbor* (K-NN) untuk mengklasifikasikan jurusan guna mempermudah dan meminimalisir kesalahan dalam proses penentuan jurusan siswa baru. Data yang awalnya berjumlah 638 record dan 31 atribut, setelah dilakukan preprosessing data yang digunakan berjumlah 635 record dengan 12 atribut yaitu nama, jenis kelamin, minat penjurusan, asal sekolah, anak ke, jumlah saudara, nilai matematika, nilai bahasa Inggris, nilai ipa, nilai bahasa Indonesia, nilai hasil tes, dan rekomendasi jurusan. Setelah dilakukan pengujian menggunakan *K-Fold Cross Validation* dan *Confusion Matrix* untuk evaluasi dan validasi hasil dengan perhitungan jarak *Euclidean Distance* didapatkan nilai k terbaik(optimal) k=3 yang menghasilkan *accuracy*: 97.11%, *precision*: 96.82%, *recall*: 98.33%, dan AUC: 0.951.

Kata kunci: klasifikasi, penjurusan siswa, k-nearest neighbor, Euclidean distance

ABSTRACT

Majoring students is a process of placing students into certain majors in accordance with their interests and academic abilities to make it easier for students in the learning process. Madrasah Aliyah Darul Hikmah Menganti is a school equivalent to SMA, which has two majors, namely science and social studies. The difficulty of classifying the majors of new students is an obstacle for the school. Because the criteria assessment process is carried out one by one. From these problems, the K-Nearest Neighbor (K-NN) method was applied to classify majors to simplify and minimize errors in the process of determining new student majors. The data initially amounted to 638 records and 31 attributes. After preprocessing, the data used amounted to 635 records with 12 attributes, namely name, gender, major interest, school origin, children to, number of siblings, math scores, English grades, science grades, Indonesian language scores, test scores, and major recommendations. After testing using K-Fold Cross Validation and Confusion Matrix for evaluation

and validation of results by calculating the Euclidean Distance distance, the best k value (optimal) k=3 which produces accuracy: 97.11%, precision: 96.82%, recall: 98.33%, and AUC: 0.951.

Keywords: *classification, majoring students, k-nearest neighbor, Euclidean distance*

1. PENDAHULUAN

Pendidikan mempunyai peran penting dalam kehidupan manusia, hal ini berarti bahwa setiap manusia diharapkan untuk dapat selalu berkembang. Oleh sebab itu, pendidikan sering disebut *long life education*. Pendidikan secara umum mempunyai arti suatu proses kehidupan dalam mengembangkan diri setiap individu untuk dapat hidup dan melangsungkan kehidupannya, sehingga menjadi seseorang yang terdidik. Sehingga bisa menjadi orang yang berguna baik bagi Negara, Nusa dan Bangsa.

Madrasah Aliyah (MA) adalah salah satu pendidikan formal berbasis agama islam yang setara dengan Sekolah Menengah Atas (SMA). Madrasah Aliyah Darul Hikmah (MA DH) Menganti merupakan salahsatu sekolah yang berada di kabupaten Jepara. Sekolah ini dikelolah oleh Yayasan Lembaga Pendidikan Islam Darul Hikmah (YLPIDH) yang terdiri dari 2 jurusan yaitu Ilmu Pengetahuan Alam (IPA), dan Ilmu Pengetahuan Sosial (IPS) dengan jumlah rata-rata kelas sepuluh IPA 2 kelas, dan 2 kelas untuk IPS. Penerimaan Peserta Didik Baru (PPDB) dari tahun ke tahun mengalami peningkatan. Meningkatnya Jumlah para pendaftar menjadi kabar baik untuk pihak sekolah.

Penjurusan siswa merupakan suatu teknik penempatan siswa ke dalam jurusan atau peminatan tertentu, sehingga siswa dapat menyerap semua mata pelajaran dengan optimal dan sesuai dengan kemampuan yang dimilikinya. Oleh sebab itu, pihak sekolah akan menyeleksi peserta didik baru terlebih dahulu, untuk mengetahui apakah siswa tersebut layak masuk jurusan yang mereka pilih, dikarenakan saat pendaftaran, masih banyak dari mereka yang milih jurusan bukan berdasarkan minat dari dalam dirinya sendiri, seperti dipaksa oleh orang tua, bahkan banyak juga yang hanya ikut dengan temannya saja. Berdasarkan hasil survey yang dilakukan oleh peneliti melalui wawancara terhadap Peserta didik dapat disimpulkan bahwa meraka kurang nyaman dan tidak cocok masuk jurusan yang mereka pilih.

Sulitnya dalam menentukan jurusan untuk siswa baru menjadi kendala bagi pihak sekolah dalam menghitung setiap kriteria para siswa baru, terkadang dalam proses input data juga mengalami banyak kendala sehingga menjadikan waktu yang kurang efisien karena harus menghitung satu per satu nilai yang akan dijadikan bahan pertimbangan dalam menentukan jurusan siwa baru. Maka dari itu perlu dilakukan penelitian analisis data mining agar dapat memudahkan pihak sekolah dalam pengklasifikasian jurusan siswa baru secara cepat dan tetap.

Klasifikasi adalah sebuah proses pencarian model (atau fungsi) yang menggambarkan dan membedakan kelas dari konsep data atau aturan. Model ini didapatkan dari analisis satu set data pelatihan, model ini digunakan untuk memprediksi kelas dari suatu objek yang label kelasnya belum diketahui (1). Dikarenakan mudah dipahami dan mudah untuk diimplementasi serta menghasilkan tingkat akurasi yang lumayan tinggi(2), maka dalam penelitian ini peneliti memilih menggunakan algoritma KNN.

Beberapa penelitian terkait tentang algoritma klasifikasi diantaranya yang dilakukan oleh (2,3) dalam penelitiannya menerapkan algoritma *Naive Bayes Classifier* Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta menghasilkan akurasi 33,34%. Kemudian pada penelitian (3) dengan menggunakan metode *naive bayes* dari hasil uji coba yang telah dilakukan menunjukkan bahwa tingkat akurasi sistem sebesar 62.86 %, *percision* 52.38% dan *recall* 78.57%.

Penelitian yang dilakukan oleh (4–7) tentang klasifikasi penentuan Jurusan di sekolah menengah pertama dan sekolah menengah kejuruan atau atas dengan menggunakan algoritma C45. menghasilkan akurasi berurut yaitu 79,68% dengan nilai K=7, penelitian dari Sambani dan Nuraeni 92.3 %, kemudian penelitian dari Ikhbal dan kurniawan 68.42%, penelitian dari Prabowo 83.33%.

Penelitian yng dilakukan oleh (8–10)melakukan klasifikasi dengan menggunakan algoritma *K-Nearest Neighbor* (K-NN) dengan akurasi tertinggi 85.42%. dari penelitian tersebut yang relevan

dengan penelitian yang akan peneliti lakukan adalah penelitian dari (8) penentuan jurusan yang dilakukan dikelas X memiliki 2 jurusan yaitu IPA dan IPS dengan mengikuti beberapa serangkaian tes yang dilakukan sekolah. Masalah yang sering dihadapi saat proses penjurusan yaitu pihak sekolahan kesulitan dalam proses menganalisis dan mengevaluasi jurusan siswa baru. Dalam menentukan jurusan, pihak sekolahan perlu melakukan evaluasi data dari Angket Peminatan Jurusan dan hasil Test Psikotest Siswa. Dikarenakan banyaknya data, proses penilaian ini menghabiskan waktu 2 bulan lamanya. Data yang digunakan dalam penelitian ini yaitu data siswa kelas X sebanyak 357 untuk atribut yang digunakan berjumlah 12 atribut terdiri dari Rerata Raport Bahasa Indonesia (RBIN), Rerata Raport Bahasa Inggris (RBI), Rerata Raport Matematika (RM), Rerata Raport IPS (RPIS), Rerata Raport IPA (RIA), USBN Bahasa Indonesia (UBIN), USBN Bahasa Inggris (UBI), USBN Matematika (UM), USBN IPS (UIS), USBN IPA (UIA), Minat dan *Intellectual Quotient* (IQ).Uji coba peratama dilakukan dengan k = 2 yang didaptkan dari pengujian dengan algoritma *K-Means* mengasilkan akurasi sebesar 82.29%. Setelah itu untuk mendapatkan tingkat akurasi yang baik dilakukan proses pencarian nilai k yang terbaik. Dari hasil tersebut didapatkan nilai k = 5 dengan akurasi tertinggi 85,42% termasuk kedalam klasifikasi yang baik.

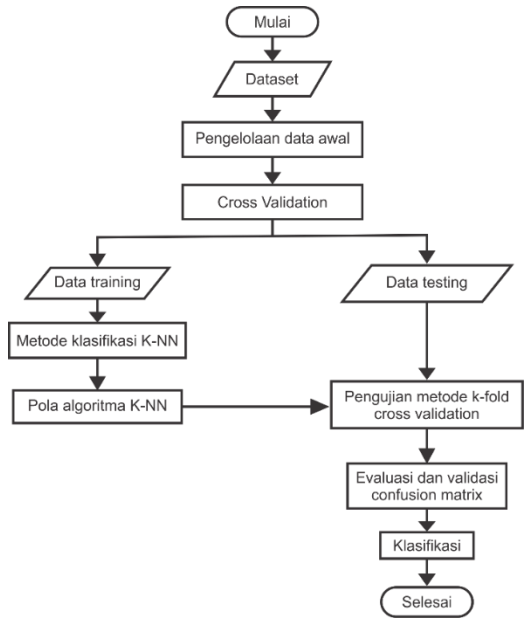
Berdasarkan penelitian diatas pada penelitian ini menggunakan algoritma *K-Nearest Neighbor* (K-NN) dengan data yang berbeda. Penelitian ini bertujuan untuk menentukan jurusan peserta didik baru di Madrasah Aliyah Darul Hikmah (MA DH) Menganti serta mengetahui akurasi yang dihasilkan setelah dicari nilai k yang terbaik. Peneliti memilih Algoritma ini karena mudah diimplementasikan dan sering menghasilkan nilai akurasi yang tinggi dalam pengklasifikasian data [3]. Sehingga dapat meminimalisir kesalahan dalam proses penjurusan dan mendapatkan hasil yang optimal.

Pada peneliti ini, permasalahan yang ada akan diselesaikan dengan menggunakan metode klasifikasi data mining guna mempermudah dan meminimalisir kesalahan dalam proses penentuan jurusan siswa baru dan dapat mengetahui akurasi yang dihasilkan. Algoritma klasifikasi yang digunakan yaitu agoritma *K-Nearest Neighbor* (K-NN). Dengan menggunakan algoritma *K-Nearest Neighbor* (K-NN) bisa dilakukan sebuah klasifikasi berdasarkan jarak terdekat *(Euclidean Distance)* dari data latih *(training)* dengan data uji *(testing)*. Algoritma ini dipilih karena mudah diimplementasikan dan sering menghasilkan nilai akurasi yang tinggi dalam pengklasifikasian data.

2. METODOLOGI PENELITIAN

Dalam penelitian ini, peneliti menggunakan analisis kuantitatif yang merupakan metode penelitian dengan menekankan analisisnya pada pengolahan, penyajian, dan perhitungan data. Data yang digunakan dalam penelitian ini adalah data yang bersifat kategorikal dan nominal yang diperoleh dari MA DH Menganti. Metode pengumpulan data menggunakan teknik wawancara, observasi dan dokumentasi sehingga diperoleh data jumlah peserta didik sejumlah 635 data. Data peserta didik yang ada akan dibagi menjadi data training 508 data dan data testing 127 data. Sedangkan atribut yang digunakan meliputi nama siwa, jenis kelamin, minat, asal sekolah, anak ke-, jumlah saudara, nilai ujian matematika, nilai ujian bahasa Inggris, nilai ujian ilmu pengetahuan alam, nilai ujian bahasa Indonesia, nilai tes.

Tahapan penelitian sesuai Gambar 1 yang menggambarkan dari awal/ dimulainya penelitian sampai akhir untuk mencapai tujuan yang ingin dicapai. Dimulai dari pengumpulan dataset, pengelolaan data awal, pengujian metode K-NN, Evaluasi dan Validasi kemudian Hasil yang ingin dicapai.



Gambar 1. Tahapan Penelitian

Sesuai Gambar 1 tahapan penelitian dapat dijelaskan mulai dari:

2.1. *Teknik Pengumpulan data*

Pada penelitian ini menggunakan Teknik pengumpulan data Wawancara, Observasi dan studi Pustaka. Hasil dari pengumpulan data diperoleh data jumlah peserta didik baru di MA Daru Hikmah Menganti dari Tahun 2016 sampai dengan 2021. Data yang diperoleh ini yang akan dijadikan sebagai dataset pada penelitian ini.

2.2. *Pengelolaan Data Awal*

Data awal yang diperoleh sebanyak 638 record, kemudian dilakukan pengolahan awal data agar data yang diperoleh memiliki kualitas yang baik. Hal ini dilakukan karena tidak smua data dan atribut dapat digunakan pada proses pengolahan. Pada pengolahan awal sebelumnya dilakukan seleksi atribut dan pengkateorian data.

a) Seleksi Atribut
   Pada penelitian ini, variabel independen yang digunakan adalah nama siwa, jenis kelamin, minat, asal sekolah, anak ke-, jumlah saudara, nilai ujian matematika, nilai ujian bahasa Inggris, nilai ujian ilmu pengetahuan alam, nilai ujian bahasa Indonesia, nilai tes. Sedangkan untuk variabel dependen yang digunakan adalah Hasil rekomendasi dengan *output* IPA atau IPS. Dalam dataset ada atribut yang tidak digunakan yaitu Nomor Induk Siswa Nasional (NISN), Nomor Induk Siwa (NIS), tempat lahir, tanggal lahir, alamat, RT, RW, nama ayah, nama ibu, pekerjaan ayah, pekerjaan ibu, pendidikan ayah, pendidikan ibu, Nomor Induk Keluarga (NIK), Nomor Kartu Keluarga (KK). Atribut ini tidak digunakan dikarenakan data tersebut tidak dapat dipakai dalam perhitungan. Dan juga sebagian data banyak yang kosong, karena setiap tahunnya data yang dibutuhkan untuk pendaftaran tidak selalu sama.
b) Pengkategorian Data
   Data awal yang peneliti terima dari MA Darul Hikmah Menganti masih berupa data mentah, oleh karena itu diperlukan penyesuaian. Tipe data yang diterima masih tercampur

antara data kategorikal dan nominal. Tipe data tersebut tidak dapat diolah menggunakan algoritma *K-Nearest Neighbor (K-NN)*. Algoritma *K-Nearest Neighbor (K-NN)* hanya menerima inputan berupa angka atau numerik, hal ini digunakan untuk menghitung jarak dengan rumus *Euclidean Distance*. Data yang diubah terdapat pada atribut jenis kelamin, asal sekolah, minat penjurusan. Berikut merupakan kriteria variabel yang digunakan untuk mengolah data:

Tabel 1. Kriteria Variabel

Atribut	***Keterangan***	***Nilai Variabel***
Jenis kelamin	*Laki-laki*	*1*
	Perempuan	*2*
Minat	*IPA*	*1*
	IPS	*2*
Asal sekolah	*SMP*	*1*
	MTs	*2*

c) Pembersihan Data
Sebelum data diproses menggunakan sebuah algoritma, perlu dilakukan pembersihan pada data untuk mengidentifikasi dan menghilangkan data yang kosong (*value null*) dan data yang bernilai salah (*missing value*). Hal ini dilakukan untuk mencegah terjadinya *error* pada tahapan data mining. Pembersihan data dilakukan dengan cara mengisi dan membuang nilai-nilai yang kosong.

Setelah dilakukan pengelolaan data awal yang berjumlah 368 didapatkan data yang siap digunakan ujicoba sebanyak 365 data, yang terdiri dari 619 data *training* dan 16 data *testing* dengan atribut nama siwa, jenis kelamin, minat, asal sekolah, anak ke-, jumlah saudara, nilai ujian matematika, nilai ujian bahasa Inggris, nilai ujian ilmu pengetahuan alam, nilai ujian bahasa Indonesia, nilai tes penjurusan, dan hasil rekomendasi jurusan dengan input IPA/IPS.

2.3. *Penerapan Algoritma K-NN*

Pada pengujian algoritma, dataset yang sudah dibersihkan kemudian diuji dengan algoritma *K-Nearest Neighbor (K-NN).* Tahapan penerapan algoritmanya sebagai berikut:

Untuk mencari jarak antara dua titik yaitu pada titik data training dan pada titik data testing, maka dapat digunakan rumus persamaan *Euclidean Distance* sebagai berikut:

1. Dataset peserta didik baru akan dibagi menjadi data training dan data testing
2. Menentukan parameter k jumlah tetangga terdekat
3. Menghitung jarak *Euclidean Distance* antara data training dengan data testing

Untuk mencari jarak antara dua titik yaitu pada titik data training dan pada titik data testing, maka dapat digunakan rumus persamaan *Euclidean Distance* sebagai berikut:

$$d(P,Q) = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2} \quad (1)$$

Keterangan:
d(P,Q) = jarak euclidien
P = data ke-
Q = data testing ke-
n = jumlah data training
P = inputan data ke -1 dari data training
Q = inputan data ke -1 dari data testing

4. Mengurutkan jarak hasil perhitungan *Euclidean Distance* no 3 dimulai dari yang mempunyai jarak terkecil.
5. Mengumpulkan klasifikasi berdasarkan nilai k tetangga terdekat
6. Kemudian keluar hasil klasifikasi penjurusan peserta didik baru.

2.4. *Evaluasi dan Validasi*

Data yang diolah menggunakan algoritma *K-Nearest Neighbor* (K-NN) dan menghasilkan model nantinya dilakukan pengujian. Pengujian model berfungsi untuk memperkirakan estimasi akurasi. Pengujian dalam penelitian ini menggunakan *K-fold cross validation*, yang akan dilakukan percobaan sebanyak k. Nilai k yang digunakan yaitu 10 atau disebut dengan 10-*fold cross validation*. Hingga nantinya akan diambil nilai k terbaik dan diaplikasikan ke dalam metode yang diusulkan.

Sedangkan unutk evaluasi dan validasi digunakan lah *confusion matrix* sebagai *performance* yang menghasilkan nilai *Accuracy*, *Precision*, *Recall*, dan *AUC (Area Under Curve)*. Penelitian ini diharapkan mampu memberikan nilai akurasi yang tinggi sehingga dapat meminimalisir kesalahan dalam menentukan hasil. Setelah melalui beberapa tahap barulah didapatkan hasil klasifikasi penjurusan siswa baru.

3. HASIL DAN PEMBAHASAN

Sulitnya dalam mengklasifikasikan jurusan siswa baru menjadi kendala bagi pihak sekolah dalam menghitung setiap kriteria para siswa baru, terkadang dalam proses input data juga mengalami banyak kendala sehingga menjadikan waktu yang kurang efisien karena harus menghitung satu per satu nilai yang akan dijadikan bahan pertimbangan dalam menentukan jurusan siwa baru. Maka dari itu perlu dilakukan penelitian analisis *data mining* agar dapat memudahkan pihak sekolah dalam pengklasifikasian jurusan siswa baru secara cepat dan tepat. Untuk menangani masalah tersebut dapat diselesaikan menggunakan metode klasifikasi *data mining* guna mempermudah dan meminimalisir kesalahan dalam proses penentuan jurusan siswa baru dan dapat mengetahui akurasi yang dihasilkan.

Dari hasil observasi dan wawancara dengan kepala sekolah MA Darul Hikmah Menganti didapatkan data sebanyak 638 record dan memiliki 31 atribut. Dari data tersebut selanjutnya dilakukan *pre-processing* data untuk menghindari *eror* saat dilakukan pengujian. Setelah melalui tahap *pre-processing* didapatkan 635 data, data tersebut akan dibagi menjadi dua yaitu data *training* sebayak 508 dan data *testing* 127 data dengan menggunakan 12 atribut yaitu nama siswa, jenis kelamin, minat, asal sekolah, anak ke, jumlah saudara, nilai matematika, nilai Bahasa Inggris, nilai ilmu pengetahuan alam, nilai Bahasa Indonesia, nilai tes penjurusan, dan hasil rekomendasi penjurusan dengan output jurusan IPA atau IPS. Setelah melalui beberapa tahapan data tersebut diolah menggunakan *software RapidMiner* dengan menggunakan algoritma *K-Nearest Neighbor* (K-NN). Setelah dilakukan uji coba sebanyak 10 kali dengan model *10-fold cross validation* diperoleh akurasi tertinggi yaitu 97.11% dengan nilai K terbaik di K = 3 sesuai Gambar 2.

	K=1	K=3	K=5	K=7	K=9	K=11	K=13	K=15	K=17	K=19
Pengujian 1	92.37%									
Pengujian 2		97.11%								
Pengujian 3			96.29%							
Pengujian 4				95.81%						
Pengujian 5					95.80%					
Pengujian 6						95.32%				
Pengujian 7							95.16%			
Pengujian 8								95.16%		
Pengujian 9									95.16%	
Pengujian 10										95.65%

Gambar 2. Hasil Pengujian *10-fold cross validation*

Berdasarkan hasil penelitian yang diperoleh terdapat beberapa tahapan yang dilakukan dalam penelitian ini yang dimulai dari:

3.1. *Teknik Pengumpulan Data*

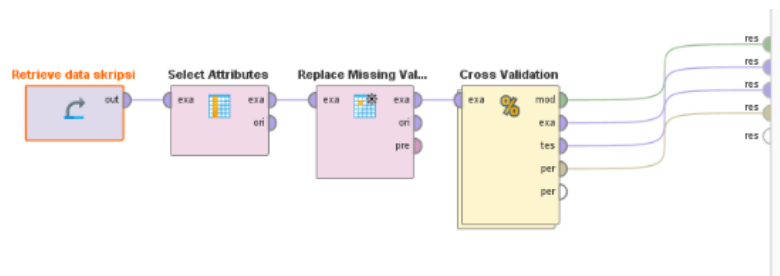
Pada tahapan pengumpulan data menggunakan teknik observasi dan wawancara tentang penjurusan siswa baru di MA Darul Hikmah Menganti. Data yang diperoleh berupa pendaftar peserta didik baru dari tahun 2016-2021 dengan jumlah data sebanyak 638. Dari data yang diperoleh sejumlah 638 data yang terdiri dari 619 data *training* dan 16 data *testing* yang memiliki atribut Nomor Induk Siswa Nasional (NISN), Nomor Induk Siswa (NIS), Nama siswa, Jenis kelamin, Minat Penjurusan, Tempat lahir, Tanggal lahir, anak ke-, jumlah saudara, alamat, nama ayah, nama ibu, pekerjaan ayah, pekerjaan ibu, asal sekolah, Rata-rata penghasilan, No HP orang tua, NIK, No KK, Nilai Ujian Matematika, Bahasa Inggris, Ilmu Pengetahuan Alam, Bahasa Indonesia, Nilai Tes. Sedangkan untuk variabel dependen yang digunakan adalah Hasil Rekomendasi dari sekolah dengan *output* IPA atau IPS.

3.2. *Pengelolaan Data Awal*

Setelah data diperoleh kemudian tahap berikutnya adalah pengelolaan data awal atau *pre-processing* dimulai dari penggabungan data, seleksi atribut, menyusun dan mengkategorikan data, dan pembersihan data. Adapun proses pengolahan data awal menggunakan *software Microsoft Excel* terlebih dahulu baru setelah didapatkan hasil data yang sesuai untuk dilakukan penelitian maka dilanjutkan dengan pengujian menggunakan *software RapidMiner*.

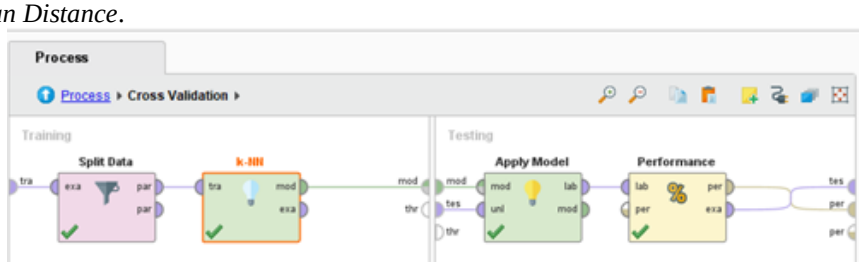
3.3. *Pengujian Model K-Fold Cross Validation*

Pada tahap pengujian model. *Dataset* akan terbagi secara otomatis pada *RapidMiner* menjadi 10 bagian dan akan dilakukan percobaan sebanyak 10 kali atau biasa disebut *10-fold cross validation*. Permodelan *K-Fold Cross Validation* dan *10-fold cross validation* dapat dilihat pada Gambar 3.



Gambar 3. Model *K-Fold Cross Validation*

Pada operator *Cross Validation* dilakukan pemodelan data *training* dan data *testing* menggunakan algoritma *K-Nearest Neighbor*(K-NN) dengan perhitungan jarak menggunakan *Euclidean Distance*.



Gambar 4 Pemodelan Cross Validation

Agar didapatkan nilai k yang terbaik, maka dilakukanlah pengujian model menggunakan *10-fold cross validation* sebanyak 10 kali dengan nilai k pada metode *K-Nearest Neighbor*(K-NN) yang berbeda-beda. Nilai k yang menghasilkan *accuracy* yang terbaiklah yang akan digunakan dalam penelitian ini. Berdasarkan hasil pengujian model *10-fold cross validation*, maka hasil tingkat akurasi tertinggi diperoleh saat pengujian ke 2 dengan nilai k = 3 yang dapat dilihat pada Gambar 2. Berikut penerapan Algoritma *K-Nearest Neighbor*(K-NN):

1. Menetapkan Parameter K
   Berdasarkan hasil pada tahapan pengujian model *10-fold cross validation*, nilai akurasi yang paling tertinggi diperoleh dengan nilai k = 3. Nilai tersebut kemudian diimplementasikan kedalam langkah-langkah algoritma *K-Nearest Neighbor*(K-NN)

2. Menghitung Jarak *Euclidean Distance*
   Dari perbandingan evaluasi algoritma K-NN yang dilakukan pengujian dengan menggunakan RapidMiner didapatkan hasil akurasi tertinggi pada perhitungan jarak *Euclidean Distance* sesuai dengan Tabel 2.

Tabel 2. Perbandingan Metode Hasil Evaluasi Perhitungan Jarak

Metode	***Accuracy***
Cosines Similarity	77.71%
Euclidean Distance	97.11%
Manhattan Distance	96.62%

Dari hasil perbandingan evaluasi jarak algoritma K-NN diperoleh *Euclidean Distance* yang terbaik. Hal ini juga diperoleh perhitungan yang sama ketika dilakukan perhitungan secara manual sesuai dengan rumus *Euclidean Distance* sebagai perhitungan jarak. Berikut adalah perhitungan jarak masing-masing objek yang dicontohkan pada data testing pertama:

1) Perhitungan data *testing* pertama terhadap data *training* pertama didapatkan nilai *Euclidean Distance* 3,118. Untuk perhitunggannya dapat dilihat dibawah ini.

$$d(1{,}620) = \sqrt{\sum_{i=1}^{n} (P_1 - Q_{620})^2}$$

$$d(1{,}620) = \sqrt{\begin{matrix}(A1\ baris\ 1 - A1\ baris\ 620)^2 + (A2\ baris\ 1 - A2\ baris\ 620)^2 + \\ (A3\ baris\ 1 - A3\ baris\ 620)^2 + (A4\ baris\ 1 - A4\ baris\ 620)^2 + \\ (A5\ baris\ 1 - A5\ baris\ 620)^2 + (A6\ baris\ 1 - A6\ baris\ 620)^2 + \\ (A7\ baris\ 1 - A7\ baris\ 620)^2 + (A8\ baris\ 1 - A8\ baris\ 620)^2 + \\ (A9\ baris\ 1 - A9\ baris\ 620)^2 + (A10\ baris\ 1 - A10\ baris\ 620)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(1-1)^2 + (2-2)^2 + (2-2)^2 + (1-2)^2 + (2-1)^2 + \\ (6.8-5.4)^2 + (7.4-7)^2 + (6.8-5.6)^2 + (8.4-6.75)^2 + \\ (7.4-6.2)^2\end{matrix}}$$

$$= \sqrt{0 + 0 + 0 + 1 + 1 + 1.96 + 0.16 + 1.44 + 2.72 + 1.44}$$

$$= \sqrt{9.722}$$

$$= 3.118$$

2) Perhitungan data *testing* pertama terhadap data *training* kedua didapatkan nilai *Euclidean Distance* 3,394. Untuk perhitunggannya dapat dilihat dibawah ini.

$$d(2{,}620) = \sqrt{\sum_{i=1}^{n} (P_2 - Q_{620})^2}$$

$$d(1{,}620) = \sqrt{\begin{matrix}(A1\ baris\ 2 - A1\ baris\ 620)^2 + (A2\ baris\ 2 - A2\ baris\ 620)^2 + \\ (A3\ baris\ 2 - A3\ baris\ 620)^2 + (A4\ baris\ 2 - A4\ baris\ 620)^2 + \\ (A5\ baris\ 2 - A5\ baris\ 620)^2 + (A6\ baris\ 2 - A6\ baris\ 620)^2 + \\ (A7\ baris\ 2 - A7\ baris\ 620)^2 + (A8\ baris\ 2 - A8\ baris\ 620)^2 + \\ (A9\ baris\ 2 - A9\ baris\ 620)^2 + (A10\ baris\ 2 - A10\ baris\ 620)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(2-1)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2 + (3-1)^2 + \\ (7-5.4)^2 + (7.8-7)^2 + (6.4-5.6)^2 + (7.6-6.75)^2 + \\ (7.6-6.2)^2\end{matrix}}$$

$$= \sqrt{1 + 0 + 0 + 0 + 4 + 2.56 + 0.64 + 0.64 + 0.722 + 1.96}$$

$$= \sqrt{11.522}$$

$$= 3.394$$

3) Perhitungan data *testing* pertama terhadap data *training* ketiga didapatkan nilai *Euclidean Distance* 3,924. Untuk perhitunggannya dapat dilihat dibawah ini.

$$d(3{,}620) = \sqrt{\sum_{i=1}^{n} (P_3 - Q_{620})^2}$$

$$d(3{,}620) = \sqrt{\begin{matrix}(A1\ baris\ 3 - A1\ baris\ 620)^2 + (A2\ baris\ 3 - A2\ baris\ 620)^2 + \\ (A3\ baris\ 3 - A3\ baris\ 620)^2 + (A4\ baris\ 3 - A4\ baris\ 620)^2 + \\ (A5\ baris\ 3 - A5\ baris\ 620)^2 + (A6\ baris\ 3 - A6\ baris\ 620)^2 + \\ (A7\ baris\ 3 - A7\ baris\ 620)^2 + (A8\ baris\ 3 - A8\ baris\ 620)^2 + \\ (A9\ baris\ 3 - A9\ baris\ 620)^2 + (A10\ baris\ 3 - A10\ baris\ 620)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(1-1)^2 + (1-2)^2 + (2-2)^2 + (1-2)^2 + (1-1)^2 + \\ (7.2-5.4)^2 + (7.8-7)^2 + (7-5.6)^2 + (8.4-6.75)^2 + \\ (8.4-6.2)^2\end{matrix}}$$

$$= \sqrt{0 + 1 + 0 + 1 + 0 + 3.24 + 0.64 + 1.96 + 2.722 + 4.84}$$

$$= \sqrt{15.402}$$

$$= 3.924$$

Perhitungan diatas juga berlaku terhadap data *training* selanjutnya sampai dengan data ke 619 dan data testing sampai dengan ke 635, sehingga diperoleh hasil perhitungan *Euclidean Distance* sesuai Tabel 3.

Tabel 3. Hasil Perhitungan Euclidean Distance

No	***Nama Siswa***	***Rekomendasi***	***Euclidean Distance***
1	*Achmad Dimas Rindiyanto*	*IPA*	*3.118*
2	*Ade Elya Saputri*	*IPA*	*3.394*
3	*Ahmad Farhan Fian Mubarok*	*IPA*	*3.925*
4	*Ahmad Rofiq*	*IPA*	*3.934*
5	*Alina Febri Nurfiani*	*IPA*	*3.066*
6	*Anisatul Millah*	*IPA*	*3.259*
…	*….*	*…*	*…*
617	*Isna Dwi Savitri*	*IPS*	*2.926*
618	*Linda Puput Kumalasari*	*IPS*	*2.413*
619	*Muhammad Bagus Arif Salafuddin*	*IPS*	*3.392*
620	*M. Illiyyin Yusuf*	*?*	-

3. Menentukan Kelas Mayoritas

Dari perhitungan jarak *Euclidean Distance* maka kelas mayoritas terhadap data testing pertama diranking dari urutan terkecil hingga terbesar dan ditentukan kelas mayoritasnya berdasarkan nilai k = 3. Adapun hasil penentuan kelas mayoritas terhadap data *testing* pertama sesuai Tabel 4.

Tabel 4. Penentuan Kelas Mayoritas

No	*Nama Siswa*	*Euclidean Distance*	*Rank*	*Rekomendasi*
280	Ahmad Ali Ngufron	1.08	1	IPS
596	Muhammad Khoirunna`im	1.30	2	IPS
166	Muhammas Aldi Nababa	1.37	3	IPS
620	M. Illiyyin Yusuf	-	-	IPS

Data diatas telah diurutkan berdasarkan jarak *Euclidean Distance* yang terkecil. Hasil penentuan kelas mayoritas pada data testing pertama yang bernama M. Illiyyin Yusuf termasuk kedalam kelas IPS, dikarenakan semua tetangganya mayoritas masuk rekomendasi IPS. Perhitungan ini juga berlaku pada data testing lainnya, sehingga didapatkan hasil Tabel 5.

Tabel 5. Hasil Klasifikasi Data Testing

No	*Nama*	*A1*	*:*	*REKO-*
620	M. Illiyyin Yusuf	1.0	:	IPS
621	Muhammad Agus Kurniawan	1.0	:	IPA
622	Muhammad Danil Afandi	1.0	:	IPS
623	Muhammad Khoiril Anam	1.0	:	IPS
624	Muhammad Misbahul Huda	1.0	:	IPS
625	Muhammad Nur Hayyun	1.0	:	IPS
626	M. Najib Mustofa	1.0	:	IPS
627	Muhammad Syarif Hidayat	1.0	:	IPA
628	Muhammad Rifqi Riyadi	1.0	:	IPS
629	M. Rizky Ardiansyah	1.0	:	IPS
630	Rina Hasyim	2.0	:	IPS
631	Silviana Juniati	2.0	:	IPA
632	Vera Amelia Putri	2.0	:	IPS
633	Zachrias Novan Andreansyach	1.0	:	IPS
634	Muhammad Diki Pratama	1.0	:	IPS
635	Galang Saputra	1.0	:	IPS

3.1. *Evaluasi dan Validasi*

Evaluasi dan Validasi dalam penelitian ini dilakukan menggunakan *confusion matrix* sebagai performance yang menghasilkan nilai *accuracy*, *precision* dan *recall*. Selain itu, dilakukan pengujian menggunakan kurva ROC (*Receiver Operating Characteristic*) yang mengasilkan nilai AUC (*Area Under Curve*). Untuk hasil confusion matrix dapat dilihat pada Tabel 6.

Tabel 6. Confusion Matrix

	true IPA	*true IPS*	*class precision*
pred. IPA	245	4	98.39%
pred. IPS	14	356	96.22%
class recall	94.59%	98.89%	

Berdasarkan hasil *performance vector* diatas, maka dapat disimpulkan data yang diprediksi dengan benar melalui algoritma *K-Nearest Neighbor (K-NN)* sejumlah 619, sebanyak 245 orang benar diprediksi IPA dan 356 orang benar diprediksi IPS. Kemudian sebanyak 14 orang IPA tetapi diprediksi sebagai IPS dan sebanyak 4 orang IPS tetapi prediksi sebagai IPA. Sehingga akurasi yang dihasilkan 97,09%, *Precision* 94,59%, *Recall* 98,39%. Untuk perhitungan manualnya sebagai berikut:

a) Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; X \; 100\%$$

$$Accuracy = \frac{245 + 356}{242 + 356 + 14 + 4} \; X \; 100\%$$

$$Accuracy = \; 97.09\%$$

b) Precision

$$Precision = \frac{TP}{TP \; + \; FP} \; X \; 100$$

$$Precision = \frac{245}{245 + 14} \; X \; 100\%$$

$$Precision = 94.59\%$$

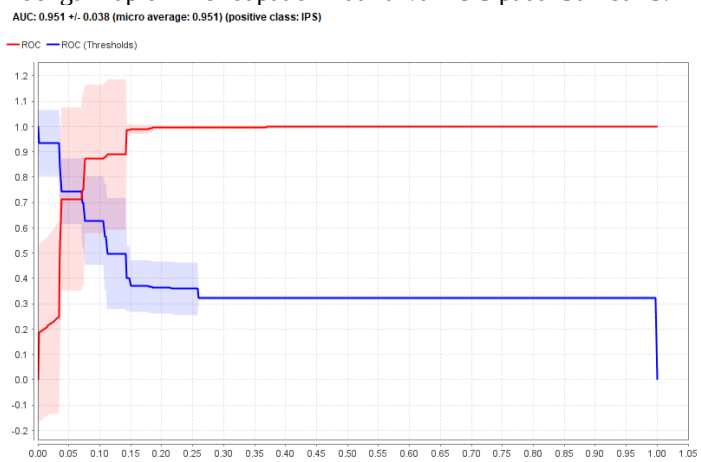
c) *Recall*

$$Recall = \frac{TP}{TP + FN} \; X \; 100\%$$

$$Recall \; = \frac{245}{245 \; + \; 4} \; X \; 100$$

$$Recall \; = 98{,}39\%$$

Untuk uji validitas yang telah dilakukan dapat disimpulkan bahwa nilai AUC (*Area Under Curva*) dari Kurva ROC (*Receiver Operating Characteristic*) terhadap algoritma *K-Nearest Neighbor* (K-NN) adalah sebesar 0.951 sehingga dikategorikan sebagai klasifikasi yang sangat baik. Hasil perhitungan dengan rapidMiner dapat dilihat kurva ROC pada Gambar 5.



Gambar 5. Kurva ROC *(Receiver Operating Characteristic)*

Berdasarkan hasil penelitian yang diperoleh. Jika dibandingkan dengan penelitian (4) (5) (8) dalam penelitian ini nilai akurasi yang dihasilkan lebih tinggi. Sehingga dapat disimpulkan bahwa algoritma *K-Nearest Neighbor* dengan menggunkan jenis data yang sama namun atribut yang berbeda dapat disimpulkan algoritma *K-Nearest Neighbor* menghasilkan akurasi yang lebih baik.

Dikarenakan mudah dipahami dan mudah untuk diimplementasi serta menghasilkan tingkat akurasi yang lumayan tinggi (11).
Dari penelitian data mining diperoleh output sebuah aturan atau pola :
jika Nilai Tes ≤ 7.100 and IPA ≤ 6.450 maka masuk jrusan IPS. Jika MTK > 5.725 maka masuk jurusan IPA. Jika Nilai Tes ≤ 7.300 dan MTK ≤ 5.225 maka masuk jurusan IPS. Jika IPA > 6.900 dan B. Inggris > 6.200 maka masuk jurusan IPA. Jika B. Inggris ≤ 5.500 maka masuk jurusan IPS.

4. KESIMPULAN

Dari hasil penelitian yang telah dilakukan berdasarkan pengujian yang telah dilakukan menggunakan *confusion matrix* dan kurva ROC (*Receiver Operating Characteristic*) bahwa algoritma *K-Nearest Neighbor* (K-NN) memiliki kinerja yang baik dalam mengklasifikasi jurusan peserta didik baru di MA Darul Hikmah Menganti. Didapatkan nilia k optimal dengan angka 3 k =3 dengan akurasi sebesar 97.09%, *precision* 96.82%, *recall* 98.33%, dan nilai AUC *(Area Under Curve)* sebesar 0.951 sehingga dapat dikategorikan sebagai klasfikasi yang sangat baik.

Diharapkan untuk penelitian selanjutnya dapat menerapkan *rule* yang diperoleh ke dalam aplikasi sederhana atau bisa menggunakan algoritma klasifikasi data mining seperti *Decision Tree, ID3, Neural Network* dan lainnya.

DAFTAR PUSTAKA

[1] Nikmatun Ia, Waspada I. "*Implementasi Data Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor"*. Jurnal Simetris. 2019;10(2):421–32.

[2] Mafakhir AZ, Solichin A. "*Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta"*. Fountain of Informatics Journal. 2020 Apr 29;5(1):21.

[3] Riyanah N, Informasi S, Tinggi S, Informatika M, Komputer D, Mandiri N. "*Penerapan Algoritma Naive Bayes Untuk Klasifikasi Penerima Bantuan Surat Keterangan Tidak Mampu*." JTIM: Jurnal Teknologi Informasi dan Multimedia. 2021;2(4):206–13.

[4] Monalisa S, Hadi F. "*Penerapan Algoritma CART Dalam Menentukan Jurusan Siswa di MAN 1 Inhil"*. Jurnal Sisfokom (Sistem Informasi dan Komputer). 2020 Oct 27;9(3):387–94.

[5] Sambani EB, Nuraeni F. *"Penerapan Algoritma C4.5 Untuk Klasifikasi Pola Penjurusan di Sekolah Menengah Kejuruan (SMK) Kota Tasikmalaya"*. CSRID Journal [Internet]. 2017;9(3):149–57. Available from: https://www.doi.org/10.22303/csrid.9.3.2017.149-157.

[6] Fibo M, Ikhbal D, Kurniadi D. "*Menentukan Penjurusan Siswa Dengan Menggunakan Metode Decision Tree Algoritma C4.5 (Studi Kasus: SMA Negeri 2 Padang)"*. Jurnal Vokasi Informatika [Internet]. 2021;1(3):31–7. Available from: http://javit.ppj.unp.ac.id

[7] Prabowo IM, Subiyantto S. "*Sistem Rekomendasi Penjurusan Sekolah Menengah Kejuruan Dengan Algoritma C4.5*". Jurnal Kependidikan. 2017;1(1):139–49.

[8] Mustakim M, Ulya R, Putri SA. "*Pemodelan Modified K-Nearest Neighbor Dalam Klasifikasi Jurusan Siswa di SMAN 6 Pekanbaru"*. In: UIN Sultan Syarif Kasim Riau ISSN. 2021. p. 2579–5406.

[9] Lestari PI, Andriansyah M. "*Analisis K-Nearest NeighborBerdasarkanForward SelectionUntuk Prediksi Status Mahasiswa Non Aktif Pada STMIK Bani Saleh"*. Jurnal Informatika: Jurnal

pengembangan IT (JPIT) [Internet]. 2021;6(3):181–6. Available from: http://pddiktiadmin.kemdikbud.go.id/admin/kemahasiswaan/d

[10] Habibi AM, Santika RR. "*Implementasi Algoritma K-Nearest Neighbor dalam Menentukan Jurusan Menggunakan Metode Euclidean Distance Berbasis Web Pada SMP Setia Gama"*. Jurnal SKANIKA. 2020;3(4):7–14.

[11] Mafakhir AZ, Solichin A. "*Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta"*. Fountain of Informatics Journal. 2020 Apr 29;5(1):21.