Journal of Information System, Informatics and Computing

Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

PENERAPAN ALGORITMA K-NEAREST NEIGHBOR (K-NN) DENGAN PENCARIAN OPTIMAL UNTUK PREDIKSI PRESTASI SISWA

Yuyun Umaidah[1], Purwantoro[2]
Teknik Informatika[1], Teknik Informatika[2]
Fakultas Ilmu Komputer[1], Fakultas Ilmu Komputer[2]
Universitas Singaperbangsa Karawang[1], Universitas Singaperbangsa Karawang[2]
yuyun.umaidah@staff.unsika.ac.id[1], purwantoro.masbro@staff.unsika.ac.id[2]

Abstrak

Pendidikan merupakan hal yang penting untuk meningkatkan kualitas siswa. Dengan pendidikan siswa dapat mencapai hasil-hasil yang diperoleh yaitu prestasi. Prestasi merupakan wujud nyata kualitas yang diperoleh siswa atas usaha dan kerja keras dalam belajar. Penelitian ini memanfaatkan teknik data mining menggunakan algoritma K-Nearest Neighbor (K-NN) dengan pencarian K-Optimal menggunakan metode k-fold cross validation untuk memprediksi prestasi siswa. Kriteria yang digunakan adalah: Les Tambahan, Jurusan, Nilai rata-rata rapor mata pelajaran pokok, Nilai rata-rata rapor mata pelajaran penjurusan, Nilai kedisiplinan, Jarak Tempuh, Ekstrakurikuler, Organisasi, dan Prestasi. Metodologi yang digunakan adalah CRISP-DM dan Performa Algoritma dilihat dari nilai accuracy, precision, recall, dan AUC dengan melakukan pemilihan k-fold cross validation (k=2, k=3, k=4, k=5, k=6, k=7, k=8, k=9, k=10). Setelah diperoleh hasil terbaik dari pemilihan k-fold cross validation akan dilakukan pengujian dengan pemilihan klaster k-NN (klaster1, klaster 2, klaster 3, klaster 4 dan klaster 5). Dari penelitian diperoleh hasil terbaik terdapat pada k=5 (5-fold cross validation) pada klaster 2 dengan hasil accuracy = 93.63%, precision=95.77%, recall=96.58% dan AUC=0.782.
Kata Kunci: K-nearest neighbor, k-fold cross validation, k-optimal, CRISP-DM

I. PENDAHULUAN

Pendidikan merupakan salah satu faktor kemajuan dan kemandirian bangsa. Semakin maju pendidikan suatu bangsa, maka akan semakin maju dan mandiri bangsa tersebut. Melalui pendidikan para generasi penerus bangsa dibentuk kualitasnya. Pendidikan nasional berfungsi mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermatabat dalam rangka mencerdaskan kehidupan bangsa, bertujuan untuk mengembangkan potensi peserta didik agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga Negara yang demokratis serta bertanggung jawab erisi tentang latar belakang penelitian, permasalahan, batasan penelitian, tujuan dan manfaat penelitian [1].

SMA Negeri 4 Karawang merupakan salah satu sekolah menengah atas berakreditasi A yang ingin mewujudkan tujuan Pendidikan, salah satunya dengan memiliki kualitas dan manajemen pembelajaran yang baik. Salah satu indikator kualitas dan menejemen sekolah atau Lembaga Pendidikan yang baik dapat dilihat dari prestasi belajar siswa. Pada SMA Negeri 4 masih banyak siswa yang merasakan permasalahan dalam proses pembelajaran, diantaranya siswa yang mendapat kesempatan berlajar yang baik, kemampuan yang cukup baik, namun hasil yang dicapai justru menunjukkan bahwa
anak tersebut tidak memiliki prestasi. Hal ini menunjukkan bahwa terdapat hambatan dan masalah dalam proses pencapaian prestasi siswa itu sendiri, baik dalam proses berprestasi di sekolah maupun di luar Pihak sekolah harus mengetahui prestasi siswa tersebut, sehingga pihak sekolah dapat melakukan langkah langkah antisipasi sejak dini terhadap siswanya dan kualitas dari sekolah tersebut akan terlihat lebih baik. Penting bagi siswa untuk mengenal prestasinya, karena dengan mengetahui hasil yang sudah dicapai maka siswa mendapat dorongan untuk meningkatkan prestasi yang telah didapatkan sebelumnya. Adapun untuk pencapaian prestasi dapat dinilai dari proses pencapaian siswa di dalam sekolah maupun luar sekolah. Bagi siswa yang memiliki prestasi dapat dilihat pencapaian prestasi siswa dalam hal akademik maupun non akademik. Adapun faktor yang mempengaruhi prestasi siswa ada empat yaitu, Sosial Ekonomi, Motivasi, Kedisiplinan, dan Prestasi Masa Lalu [2].

Dalam membantu pihak sekolah untuk mengetahui prestasi siswa, akan dilakukan penelian untuk memprediksi prestasi siswa.

Penelitian sebelumnya untuk prediksi sudah pernah dilakukan oleh Novianti dan Prasetyo dengan judul Penerapan Algoritma K-Nearest Neighbor untuk prediksi waktu kelulusan mahasiswa, dengan menggunakan 7

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

atribut sebagai kriteria kelulusan diperoleh akurasi untuk program studi Teknik Informatika sebesar 84% dan program studi Sistem Informasi sebesar 87% [3]. Sedangkan penelitian yang dilakukan oleh Banjarsari dkk dengan judul Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4 dengan metode *k-fold cross validation* memperoleh hasil k=5 dengan tingkat akurasi 80% [4]. Berdasarkan penelitian diatas, dalam penelitian ini akan dilakukan pencarian nilai k-optimal dengan metode *k-fold cross validation* dengan menggunakan algoritma K-Nearest Neighbor untuk menentukan akurasi, pressision, recall dan AUC terbaik dalam memprediksi prestasi siswa.

II. LITERATUR DAN METODE

2.1 Data Mining

Menurut Gartner Group data mining adalah suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunaan teknik pengenaan pola seperti teknik statistika dan matematika [5].

Kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa factor [5], antara lain:

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam data warehouse sehingga seluruh perusahaan memiliki akses ke dalam database yang baik.
3. Adanya peningkatan akses data melalui navigasi web dan intranet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknolgi perangkat lunak data mining (ketersediaan teknologi).
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

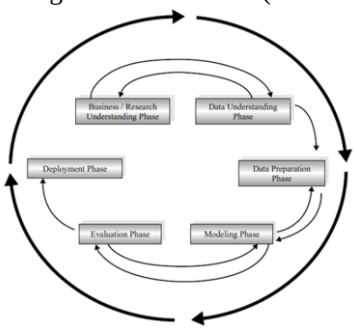
2.2 Klasifikasi

Dalam klasifikasi, ada target variabel kategoris, seperti pada contoh penggolongan pendapatan yang dibagi menjadi tiga kelas atau kategori: berpenghasilan tinggi, pendapatan menengah, dan berpenghasilan rendah. Model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi tentang variabel target serta satu set input atau variabel prediktor [5]. Contoh lain klasifikasi dalam bisnis dan penelitian adalah :

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit.

2.3 CRISP-DM

Sebuah Cross- Industry diperlukan sebagai suatu standar industry yang netral, tool yang netral dan aplikasi yang netral untuk pendekatan data mining. Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan suatu standar untuk pengembangan model data mining. Standar ini dikembangkan tahun 1996 oleh analis yang mewakili DaimlerChrysler, SPSS dan NCR. Didalam CRISP-DM menyediakan kepemilikan dan tersedia proses standar yang bebas untuk data mining yang sesuai sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Dalam CRISP-DM, Sebuah proyek data mining memiliki siklus hidup yang terbagi dalam enam fase (Gambar 1).



Gambar 1. Fase CRISP-DM
(Sumber: Larose, 2005)

Enam fase CRSIP-DM [5]:

1. Bussines Understanding

Pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan, kebutuhan dari perspektif bisnis. Kegiatannya antara lain menetukan sasaran atau tujuan bisnis, memahami situasi bisnis, menentukan tujuan data

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

mining dan membuat perencanaan strategi serta jadwal penelitian.

2. Data Understanding

Fase mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai, mengidentifikasikan masalah yang berkaitan dengan kualitas data, mendeteksi subset yang menarik dari data untuk membuat hipotesa awal.

3. Data Preparation

Sering disebut sebagai fase yang padat karya. Aktifitas yang dilakukan antara lain memilih table dan field yang akan di transformasikan ke dalam database baru untuk bahan data mining (set data mentah).

4. Modelling

Fase menentukan teknik data mining yang digunakan, menentukan tools data mining teknik data mining, algoritma data mining, menentukan parameter dengan nilai yang optimal.

5. Evaluation

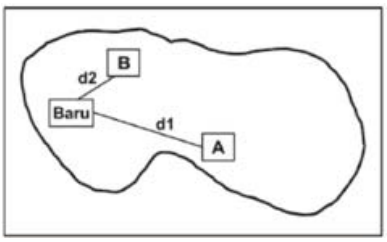
Fase interpretasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan pada fase sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama.

6. Deployment

Fase penyusunan laporan atau presentasi pengetahuan yang didapat dari evaluasi pada proses data mining.

2.4 K-Nearest neighbour

Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Misalkan diinginkan untuk mencari solusi terhadap seorang pasien baru dengan menggunakan solusi dari pasien terdahulu. Untuk mencari kasus pasien mana yang akan digunakan, maka dihitung kedekatan kasus pasien baru dengan semua kasus pasien lama kasus pasien lama dengan kedekatan terbesarlah yang akan diambil solusinya untuk digunakan pada kasus pasien baru [6].



Gambar 2. Ilustrasi Kedekatan Kasus
(Sumber: Kusrini dan Emha, 2009)

Pada Gambar 2 terdapat dua pasien lama: A dan B, ketika ada pasien baru, maka solusi yang akan diambil adalah solusi dari pasien terdekat dari pasien baru. Seandainya d1 adalah kedekatan antara 8 pasien baru dan pasien A, sedangkan d2 adalah kedekatan antara pasien baru dan pasien B, karena d2 lebih dekat dari d1, maka solusi dari pasien B-lah yang akan digunakan untuk memberikan solusi pasien baru. Untuk melakukan perhitungan kedekatan antara dua kasus, terdapat beberapa rumus diantaranya adalah sebagai berikut [3]:

$$Similarity\ (T,S) = \frac{\sum_{i=1}^{n} f(T_i,S_i)*w_i}{w_i} \quad (1)$$

Keterangan:
T: kasus baru
S: kasus yang ada dalam penyimpanan
N: jumlah atribut dalam setiap kasus
I: atribut individu antara 1 s.d n
f: fungsi similarity i antara kasus T dan S
wi: bobot yang diberikan pada atribut ke-i

Kemiripan (similarity) adalah ukuran derajat numerik dimana dua objeknya mirip, nilai 0 jika tidak mirip dan 1 jika mirip penuh. Formula kemiripan dua data dengan satu atribut adalah sebagai berikut:

$$S = \begin{cases} 1 & \text{jika } x = y \\ 0 & \text{jika } x \neq y \end{cases} \quad (2)$$

Adapun rumus untuk pemberian bobot pada setiap atribut adalah sebagai berikut:
1. Input nilai kriteria masing-masing model
2. Input bobot masing-masing kriteria
3. Hitung normalisasi dari bobot

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

$$NK = \frac{\sum_{i=1}^{n}(SBK)xBBT\%}{n} \quad (3)$$

$$Nilai\ Akhir = \frac{\sum NK}{n} \quad (4)$$

Dimana: SBK : Kriteria
BBT : Bobot Kriteria
NK : Nilai Kriteria

Pengukuran kinerja algoritma dilakukan dengan cara membandingkan antara hasil prediksi algoritma klasifikasi dengan nilai target variabel data testing sebagai data sebenarnya. Maka secara logika sederhana dapat disimpulkan kinerja algoritma adalah sebagai berikut:

$$Kinerja = \frac{jumlah\ instance\ yang\ diprediksi\ benar}{jumlah\ instance} x100\% \quad (5)$$

2.5 Metode Evaluasi dan Validasi

Terdapat beberapa teknik pengujian untuk evaluasi dan validasi kinerja suatu algoritma, seperti:

a. Confusion Matrix

Evaluasi Model klasifikasi didasarkan pada pengujian untuk memprediksi objek yang benar dan salah, urutan pengujian ditabulasikan dalam Confusion matriks, dimana kelas yang diprediksi ditampilkan di bagian atas matriks dan kelas yang diamati disisi kiri matriks. Setiap sel berisi angka menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi.

Tabel 1 Confusion Matrix untuk 2 kelas

Classification	**Predicted Class**	
	Class = Yes	**Class = No**
Class = Yes	TP	FN
Class = No	FP	TN

TP adalah jumlah record positif yang diklasifikasikan sebagai positif, FP adalah jumlah record negative yang diklasifikasikan sebagai positif, FN adalah jumlah record positif yang diklasifijasikan sebagai negative, TN adalah jumlah record negative yang diklasifikasikan sebagai negative. Evaluasi dengan confusion matrix menghasilkan akurasi dan laju error. Akurasi adalah presentasi dari total data yang diprediksi secara benar. Laju error adalah presentase dari total data yang diprediksi secara salah.

$$Akurasi = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Total jumlah prediksi yang dilakukan}} = \frac{a+d}{a+b+c+d} \times 100\%$$

Semua algoritma klasifikasi berusaha membentuk model yang memiliki akurasi tinggi (laju error yang rendah), tetapi umumnya model yang dibangun dapat memprediksi dengan benar pada semua data latih, ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan [8].

Dari Confusion Matrix dapat juga menghitung Presicion, Recall, F-Measure, TP Rate, dan FP Rate [9].

1. Precision (Positive predict Value) adalah tingkat ketepatan hasil klasifikasi terhadap suatu kejadian.

Prescision=TP/(TP+FP)×100%

2. Recall (Sensitivity) adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relavan dengan query.

Recall=TP/(TP+FN)×100%

Keterangan:
TP : True Positive
FP : False Positive
FN : False Negative

b. Curva Receiver Operating (ROC)

Kurva ROC adalah salah satu tenik yang dapat memvisualisasikan, mengorganisasikan dan memilih classifier berdasarkan performanya. Receiver Operating Characteristic (ROC) merupakan hasil dari pengukuran klasifikasi dalam bentuk 2 dimensi dimana garis horizontal menggambarkan nilai false positive dan gars vertical sebagai true positive. Visualisasi perhitungan digambarkan dengan ROC Curve (AUC). AUC sering

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

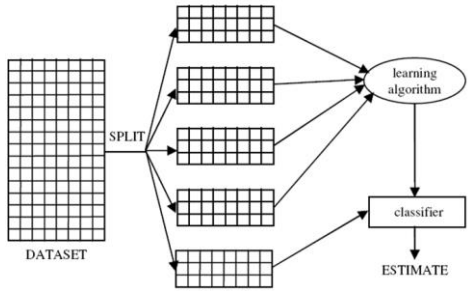
digunakan untuk mengukur kualitas classifier probabilistic. Level pengukuran kualitas classifier menggunakan ROC dilihat berdasarkan akurasi dengan rentang yang diperlihatkan dalam table berikut:

Tabel 2. Nilai AUC dan Keterangan

Rentang Akurasi	Kualitas *Classifier*
0.90-1.00	*Exellent Classification*
0.80-0.90	*Good Classification*
0.70-0.80	*Fair Classification*
0.60-0.70	*Poor Classification*
< 0.60	*Failure*

c. K-Fold Cross Validation

Salah satu pendekatan alternatif untuk "train dan test" yang sering di adopsi dalam beberapa kasus (dan beberapa lainnya terlepas dari ukurannya) yang disebut dengan k-fold cross validation [9], dengan cara menguji besarnya error pada data test [10]. Kita gunakan k-1 sampel untuk training dan 1 sampel sisanya untuk testing. Misalnya ada 10 subset data, kita menggunakan 9 subset untuk training dan 1 subset sisanya untuk testing. Ada 10 kali training dimana pada masing-masing training ada 9 subset data untuk training dan 1 subset digunakan untuk testing. Dari situ lalu di hitung rata-rata error dan standar deviasi error [11]. Setiap bagian k pada gilirannya digunakan sebagai ujian menetapkan dan k lainnya - 1 bagian digunakan sebagai training set [9].

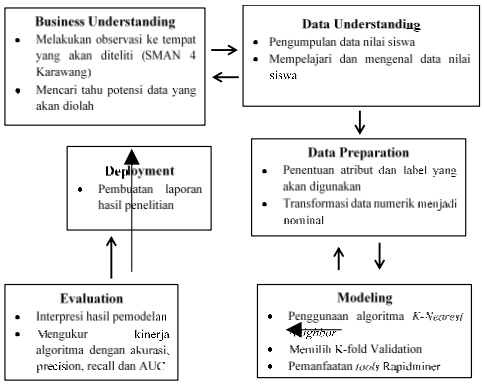


Gambar 3. K-Fold Cross Validation
(Sumber: Bramer, 2007)

III. METODE

Dalam penelitian ini penulis akan menggunakan metodologi penelitian kuantitatif dengan memanfaatkan metode yang ada dalam data mining yaitu metode CRISP-DM (Cross Industry Stadard Process For Data Mining). Adapun tahapan yang terdapat didalam CRISP-DM diantaranya :

1. Business Understanding
2. Data Unsertanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



Gambar 4. Tahapan Penelitian

Pada Gambar 4 menjelaskan tentang tahapan metodologi penelitian yang penjelasannya sebagai berikut:

1. Busines Understanding

Dengan melakukan observasi ke tempat yang akan dijadikan objek penelitian yaitu SMA Negeri 4 Karawang maka akan mendapatkan data yang dibutuhkan untuk menyelesaikan penelitian ini. Data yang didapat nantinya diharapkan dapat memberi informasi untuk mencapai pemahaman bisnis yang mengacu pada prestasi siswa yang ada di SMA Negeri 4 tersebut.

2. Data Understanding

Dalam tahap ini akan dilakukan pengumpulan data yang telah didapat dari data sekolah siswa SMA Negeri 4 Karawang kelas 12 tahun ajaran 2017/2018 dengan jumlah data 423 *records*, yang kemudian data tersebut akan dipahami lebih lanjut. Tahap memahami dengan data-data yang sudah dikumpulkan dan berusaha menemukan wawasan awal mengenai informasi apa saja yang bisa didapatkan serta tahap untuk mengevaluasi kualitas dan kelengkapan data.

3. Data Preparation

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

Tahap ini akan dilakukan pemilihan kriteria data yang akan digunakan dengan menentukan atribut dan label. Dan juga akan merubah data yang sebelumnya berupa data numerik menjadi data nominal.

4. Modeling

Tahap ini adalah pekerjaan berat yang perlu dilaksanakan secara intensif. Pada tahap ini akan dilakukan proses klasifikasi data dengan algoritma K-Nearest Neighbor dengan melakukan pemilihan k-fold cross validation sehingga nanti akan diperoleh nilai yang terbaik dengan menggunakan alat tool pengolahan data RapidMiner.

5. Evaluation

Evaluasi merupakan fase interpretasi terhadap hasil pemodelan data mining. Berdasarkan tujuan yang telah dijabarkan pada fase pemahaman bisnis, maka pada fase evaluasi ini akan dilakukan pengukuran kinerja algoritma yang digunakan dengan melihat nilai accuracy, precision, recall dan AUC.

6. Deployment

Hasil yang akan didapat dari penelitian ini adalah berupa laporan hasil analisa data prediksi prestasi siswa, dari hasil yang telah dilakukan dan dievaluasi maka akan didapatkan hasil nilai terbaik dari parameter accuracy, precision, recall dan juga tingkat efesiensi AUC yang diukur dengan ROC.

IV. HASIL DAN PEMBAHASAN

Berdasarkan metodologi yang telah dijelaskan sebelumnya, terdapat beberapa tahapan yang harus dilakukan, diantaranya:

1. Business Understanding

Tujuan bisnis dari penelitian ini adalah mengukur tingkat keberhasilan belajar dalam menyelesaikan pendidikan siswa. Hal ini penting dilakukan agar pihak sekolah dapat melakukan langkah-langkah antisipasi terhadap siswanya, sehingga kualitas siswa di sekolah tersebut dapat terlihat tetap baik atau lebih baik.

a. Menilai Situasi

Adapun kondisi data saat ini pada SMA Negeri 4 Karawang yaitu sebagai berikut:

1. Pada SMA Negeri 4 terdapat data yang bisa dijadikan untuk informasi ataupun pengetahuan. Banyaknya data tersebut namun belum diolah sehingga menghasilkan kumpulan data yang berukuran besar, tetapi tidak mempunyai nilai guna yang lebih.
2. Kumpulan data siswa belum diolah secara maksimal untuk untuk menyusun langkah-langkah antisipasi sejak dini terhadap prestasi siswa.

b. Menentukan Tujuan Data Mining

Tujuan data mining atau tujuan penelitian ini adalah menggali untuk mendapatkan pengetahuan baru dari dataset prestasi siswa. Lalu mengetahui hasil kinerja dan membandingkan hasil klasifikasi dengan menggunakan algoritma K-Nearest Neighbor dengan pemilihan K-Optimal untuk memprediksi prestasi siswa.

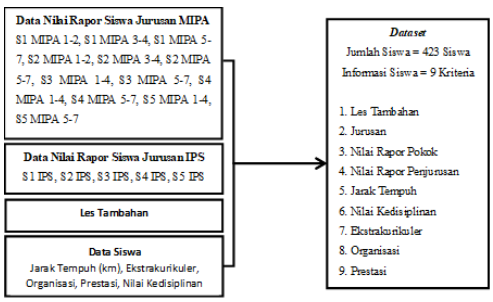
2. Data Understanding

Tahap mengumpulkan data yang didapat dari SMA Negeri 4 Karawang mengenai data siswa yang mempengaruhi prestasi pada siswa. Data yang digunakan pada penelitian ini berjumlah 423 dari data siswa kelas XII tahun ajaran 2017-2018 dan mengambil data:

1. Rata-rata nilai mata pelajaran pokok dari rapor
2. Rata-rata nilai mata pelajaran penjurusan dari rapor
3. Jurusan siswa
4. Jarak tempat tinggal siswa
5. Nilai kedisiplinan
6. Ekstrakurikuler
7. Organisasi
8. Prestasi
9. Data les tambahan yang didapatkan dari hasil angket yang disebar oleh peneliti.

3. Data Preparation

Mementukan atribut dan label yang digunakan selanjutnya mengintegrasi data serta mentranformasi data yang sebelumnya data numerik menjadi nominal.



Gambar 5. Integrasi Data

Journal of Information System, Informatics and Computing
Vol.3 No.2 Desember 2019
e-ISSN : 2597-3673 *(Online)*
p-ISSN : 2579-5201 *(Printed)*

Dari beberapa data yang diperoleh digabungkan menjadi 1 file untuk data diolah ke tahapan selanjtnya.

lesT	Jurusan	Nilai Rapor J	Nilai Rapor P	Jarak Tempuh	....	PRESTASI
Tidak	MIPA	Amat Baik	Baik	Jauh		BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Dekat		TIDAK BERPRESTASI
Ya	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		BERPRESTASI
Ya	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Ya	MIPA	Baik	Baik	Dekat		TIDAK BERPRESTASI
Tidak	MIPA	Amat Baik	Baik	Sedang		BERPRESTASI
Ya	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Ya	MIPA	Amat Baik	Baik	Dekat		BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		BERPRESTASI
Tidak	MIPA	Baik	Amat Baik	Sedang		TIDAK BERPRESTASI
Ya	MIPA	Amat Baik	Baik	Jauh		BERPRESTASI
Tidak	MIPA	Baik	Baik	Dekat		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Ya	MIPA	Baik	Baik	Sedang		BERPRESTASI
Ya	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Dekat		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Dekat		BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Amat Baik	Jauh		BERPRESTASI
Tidak	MIPA	Baik	Baik	Sedang		TIDAK BERPRESTASI
Ya	MIPA	Baik	Baik	Dekat		TIDAK BERPRESTASI
Tidak	MIPA	Baik	Baik	Jauh		TIDAK BERPRESTASI
Tidak	MIPA	Amat Baik	Baik	Dekat		BERPRESTASI

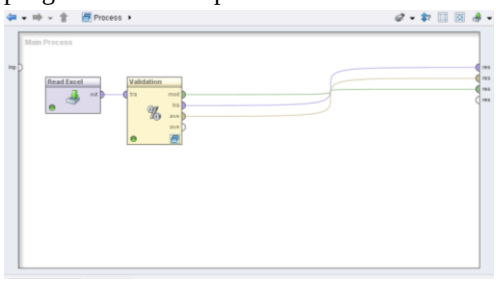
Gambar 6. Dataset Prestasi Siswa

Setelah dilakukan integrasi data selanjutnya akan dilakukan tranformasi data dengan mengubah data yang numerik menjadi nominal. Data yang ditranformasi yaitu: nilai raport Jurusan, nilai raport pokok dan jarak tempuh.

4. Modeling

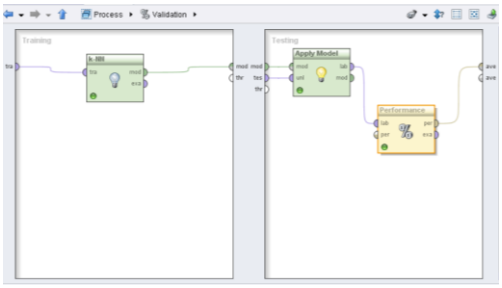
Pada penelitian ini akan dilakukan prediksi prestasi siswa dengan Teknik data mining,

Menggunakan algoritma K-Nearest Neighbor dengan melakukan pemilihan nilai *k-fold cross validation*. Dimana hasil terbaik dari nilai tersebut akan digunakan untuk menentukan klister pada algoritma K-Nearest Neighbor (k-NN). Pengujian ini menggunakan tool pengolahan data RapidMiner 5.



Gambar 7. Penggunaan Operator Fold-Cross Validation

Pada Gambar 7 akan dilakukan pemilihan nilai k-fold cross validation (k=2, k=3, k=4, k=5, k=6, k=7, k=8, k=9, k=10). Sampai ditemukan nilai k berapa yang menghasilkan performa terbaik.



Gambar 8. Penggunaan Operator k-NN

Selanjutnya pada Gambar 8 dilakukan pemilihan klaster pada algoritma k-NN sehingga dengan eksperimen yang dilakukan akan memperoleh kinerja terbaik dari algoritma *K-Nearest Neighbor*.

5. Evaluation

Dari hasil pengujian yang dilakukan pada tahapan modeling akan diperoleh model yang selanjutnya model tersebut akan dievaluasi untuk mengukur kinerja dari algoritma k-NN dengan melihat tingkat akurasi, *precision*, *recall* dan AUC, dimana hasilnya ada pada table berikut:

Tabel 3. Pemilihan k-Fold Cross Validation

Validation	Accuracy	Precision	Recall	AUC
K=2	91.02%	91.75%	98.01%	0.500
K=3	89.83%	91.92%	96.30%	0.500
K=4	89.83%	90.97%	97.43%	0.500
K=5	**91.73%**	**92.29%**	**98.28%**	**0.500**
K=6	89.37%	91.41%	96.30%	0.500
K=7	90.07%	91.26%	97.43%	0.500
K=8	90.07%	91.06%	97.72%	0.500
K=9	90.07%	91.77%	96.87%	0.500
K=10	91.02%	91.61%	98.29%	0.500

Pada Tabel 3 diperoleh hasil, bahwa pada k=5 memperoleh hasil lebih baik dengan nilai *accuracy*=91.73%, *precision*=92.29%, *recall*=98.28% dan AUC=0.500.

Table 4. Klaster k-NN

5-fold cross Validation	Accuracy	Precision	Recall	AUC
Klaster 1	91.73%	92.29%		

			98.28 %	0.500
Klaster 2	**93.63%**	**95.77%**	**96.58 %**	**0.782**
Klaster 3	90.06%	90.34%	98.57 %	0.834
Klaster 4	91.01%	91.73%	98.00 %	0.856
Klaster 5	89.11%	89.22%	98.86 %	0.867

Pada Tabel 4 diperoleh hasil terbaik, terdapat pada klaster 2 (k-2) pada k-NN dengan nilai a*ccuracy*= 93.63%, *precision*=95.77%, *recall*=96.58% dan AUC=0.782.

6. Deployment

Pada tahapan ini akan dibuat laporan penelitian yang nantinya akan memberikan informasi kepada pihak sekolah SMA N 4 terkait prediksi prestasi siswa, dimana dari penelitian diharapkan dapat membantu pihak sekolah dalam pengambilan keputusan.

V. KESIMPULAN

Dari hasil penelitian dapat disimpulkan:

1. Algoritma *K-Nearest Neighbor* pada Klaster 2 mampu memberikan hasil yang baik dalam memprediksi prestasi siswa dengan nilai a*ccuracy*= 93.63%, *precision*=95.77%, *recall*=96.58% dan AUC=0.782.
2. Pemilihan nilai *k-fold cross validation* mampu menghasilkan nilai K-Optimal dalam mengukur kinerja algoritma yang digunakan.

REFERENASI

[1] Depdiknas. (2003). "Sistem Pendidikan Nasional" Undang-Undang Republik Indonesia No. 20 Tahun 2013.

[2] H. Susanto, Sudiyatno. (2014). "Data Mining Untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplian dan Prestasi". Jurnal Pendidikan Vokasi, Universitas Negeri Yogyakarta.

[3] Andi, G.T, Dian, P. (2017). "Penerapan Algoritma K-Nearest Neighbor (k-NN) Untuk Prediksi Waktu Kelulusan Mahasiswa". Seminar Nasional APTIKOM, Jayapura.

[4] Mutiara, A. B, Irwan, B, Andi, F. (2015). "Penerapan K-Optimal Pada Algoritma KNN Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Kompute Fmipa Unlam Berdasarkan IP sampai dengan Semester 4". Kumpulan Jurnal Ilmu Komputer, Volume:2, No.2 Sepetember 2015, ISSN: 2406-7857

[5] Larose, D. T. (2005). "Discovering Knowledge in Data". Canada: Wiley Interscience.

[6] Kusrini., Luthfi E. T. (2009) "Algoritma Data Mining". Andi, Yogyakarta.

[7] Sumarlin. (2015). "Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerimaan Beasiswa PPA dan BBM". Jurnal Sistem Informasi Bisnis, 01(2015), On-line: http://ejournal.undip.ac.id/index.php/jsinbis

[8] Jiawei Han, Micheline Kamber, and Jian Pei. (2011). "Data Mining Concepts and Techniques Third Edition", Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA, ISBN 978-0- 12-381479-1

[9] Bramer, M. (2007). "Principles of Data Mining". London: Springer.

[10] Santosa, B. (2007). "Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis". Yogyakarta: Graha Ilmu.