Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI

Mardi Yudhi Putra[1,*], Dwi Ismiyana Putri[2]

[1, 2]Fakultas Informatika, Program Studi Rekayasa Perangkat Lunak, Universitas Bina Insani, Bekasi, Indonesia
Email: [1,*]mardi@binainsani.ac.id, [2]dwiismiyana@binainsani.ac.id

Abstrak– Sekolah Menengah Atas atau biasa disebut sebagai SMA dalam proses penjurusan siswa dilakukan setiap awal tahun pelajaran. Ada 2 (dua) jurusan yaitu IPA dan IPS. Dalam melakukan penjurusan IPA maupun IPS siswa kelas XI, saat ini masih dilakukan dengan berdasarkan hasil nilai rerata dari beberapa mata pelajaran tertentu sehingga sering kali menyebabkan ketidaksesuaian minat dan bakat siswa terhadap jurusan. Tujuan penelitian ini adalah untuk membandingkan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* yang memiliki akurasi tertinggi dalam melakukan klasifikasi jurusan IPA dan IPS sehingga membantu pihak sekolah dalam proses klasifikasi jurusan siswa kelas XI. Data siswa yang digunakan pada penelitian ini adalah data nilai semester 2 (dua) dengan jumlah 277 *record* dan menggunakan 4 (empat) atribut dari nilai mata pelajaran antara lain PPKN, Sejarah, Prakarya dan PAI. Salah satu *tools* yang digunakan untuk membantu proses analisis data pada penelitian ini adalah *Rapidminer*. Metode penelitian dilakukan mulai tahap *preprocessing*, *training* data, klasifikasi menggunakan algortima Naïve Bayes dan K-Nearest Neighbor, model data dan mengukur *performance* atau evaluasi. Hasil penelitian menunjukkan algoritma Naïve Bayes memiliki akurasi sebesar 81.82% dengan sampel data sebanyak 55 data dari 277 data. Sedangkan Algoritma K-Nearest Neighbor menunjukkan akurasi sebesar 92.73% dengan sample data yang sama. Hasil kedua algortima menunjukkan algoritma terbaik dengan urutan pertama K-Nearest Neighbor dan urutan kedua Naïve Bayes, tidak ada perbedaan signifikan karena nilai alpha dibawah 0.05. Secara keseluruhan dapat disimpulkan bahwa dengan memanfaatkan algoritma K-Nearest Neighbor memiliki tingkat akurasi yang lebih tinggi dari pada Algortima *Naïve Bayes* pada proses klasifikasi penjurusan IPA maupun IPS pada kelas XI SMA menggunakan algoritma klasifikasi sehingga sesuai dengan minat, bakat dan potensi diri siswa dalam penentuan jurusan sehingga memberikan manfaat dan membantu pihak sekolah menjadi lebih cepat dalam pengklasifikasiannya.

Kata Kunci: Algoritma, Jurusan, K-Nearest Neighbor, Naïve bayes, Siswa

Abstract– Senior High School or commonly referred to as SMA in the process of student majors is carried out at the beginning of each school year. There are 2 (two) majors, namely IPA and IPS. In conducting science and social studies majors for class XI students, currently it is still carried out based on the results of the average scores of certain subjects so that it often causes mismatch of interests and talents of students to majors. The purpose of this study was to compare the Naïve Bayes and K-Nearest Neighbor algorithms which have the highest accuracy in classifying science and social studies majors so as to help the school in the class XI class classification process. The student data used in this study is semester 2 (two) grade data with a total of 277 records and uses 4 (four) attributes of subject scores, including PPKN, History, Craft and PAI. One of the tools used to assist the data analysis process in this research is Rapidminer. The research method is carried out starting from the preprocessing stage, data training, classification using the Naïve Bayes and K-Nearest Neighbor algorithms, data models and measuring performance or evaluation. The results showed that the Naïve Bayes algorithm has an accuracy of 81.82% with a sample of 55 data from 277 data. While the K-Nearest Neighbor Algorithm shows an accuracy of 92.73% with the same sample data. The results of the two algorithms show the best algorithm with the first order K-Nearest Neighbor and the second order Naïve Bayes, there is no significant difference because the alpha value is below 0.05. Overall it can be concluded that by utilizing the K-Nearest Neighbor algorithm has a higher level of accuracy than the Naïve Bayes Algortima in the process of classifying science and social studies majors in class XI SMA using a classification algorithm so that it is in accordance with the interests, talents and potential of students in determining majors. thus providing benefits and helping the school to be faster in classifying.

Keywords: Algorithm, Department, K-Nearest Neighbor, Naïve bayes, Student

1. PENDAHULUAN

Kemajuan perkembangan ilmu pengetahuan dan teknologi telah tersebar dalam seluruh aspek kehidupan masyarakat misalnya aspek bisnis, sosial, maupun aspek dunia pendidikan. Peran teknologi pada dunia tentunya merupakan hal yang sangat penting untuk membantu proses pembelajaran yang ada pada kurikulum. Proses pelaksanaan penjurusan untuk Sekolah Menengah Atas dalam menentukan jurusannya untuk setiap sekolah berbeda-beda. Pada umumnya pelaksanaan penjurusan dilakukan pada saat akan memasuki jenjang SMA dan disisi lain beberapa sekolah dalam proses penjurusan dilakukan ketika kenaikan kelas misalnya dari kelas X ke kelas XII. Penentuan pemilihan jurusan dilakukan berdasarkan kriteria-kriteria dari nilai akademik yang natinya dapat menjadi kriteria dari masing-masing jurusan, berdasarkan hal tersebut maka akan didapatkan peluang siswa yang akan memenuhi semua jurusan atau sebagian dari kriteria jurusan yang ditentukan [1].

Sekolah Menengah Atas (SMA) merupakan salah satu wadah untuk para siswa/siswi agar dapat mewujudkan cita-cita dimasa depan. Agar dapat terwujudnya cita-cita yang diinginkan maka setiap siswa harus memilih jurusan yang tepat. Pada umumnya ada 2 (dua) pilihan jurusan yang dapat dipilih, yaitu jurusan IPA dan IPS. Manfaat dari pemilihan jurusan adalah agar pelajaran yang ditawarkan dapat fokus ke masa depan karena

tidak sedikit siswa ceroboh dalam pemilihan jurusan yang diambil. Pada kenyataannya masih banyak siswa/siswi SMA yang kesulitan mengetahui minat dan bakat dirinya serta pilihan jurusan. Faktor lain juga mempengaruhi dalam melakukan pemilihan jurusan. Tujuan dari penjurusan siswa agar dapat mengarahkan siswa/siswi dapat mengembangkan kemampuan diri dan minat yang dimiki. Jurusan yang kurang tepat dapat merugikan siswa pada masa depan khususnya pada karir. Diharapkan adanya penjurusan dapat meningkatkan potensi dalam diri, minat dan talenta dari siswa sehingga akan berpengaruh positif pada nilai pelajaran jurusannya. sebab akan berpengaruh pula pada siswa siswi yang ingin melakukan studi lanjut.

Memilih jurusan saat memasuki semester ganjil di tahun kedua membuat para siswa kebingungan dalam memilih jurusan yang akan mereka pilih, namun disini kasus yang dijumpai adalah sistem kurikulum tidak sama dengan yang dulu, yang awalnya KTSP jurusan ditentukan saat memasuki semester ganjil di tahun kedua dan berubah menjadi K-13 dimana jurusan ditentukan saat awal pendaftaran masuk sekolah. Banyak juga kasus dijumpai bahwa pemilihan jurusan yang tidak sesuai dengan bakat, kemampuan, minat dan kepribadian dapat mempengaruhi para siswa pada kegiatan belajar mengajar [2].

Salah satu Sekolah SMA memiliki visi terwujudnya siswa siswi yang cerdas, beprestasi yang berlandaskan ilmu pengetahuan dan teknologi serta berwawasan dan berbudaya. Sebagai salah satu penyelenggara pendidikan memiliki peranan penting dalam mewujudkan minat dan bakat serta ketrampilan siswa-siswi. SMA dalam proses penjurusan siswa-siswinya dilakukan setiap awal tahun pelajaran. Penjurusan dilakukan oleh siswa kelas XII, hanya saja dalam proses menjurusan masih dilakukan secara konvensional yakni masih melihat rerata nilai dari mata pelajaran sehingga rentan terhadap kesalahan pemilihan jurusan baik IPA maupun IPS.

Teknologi informasi memiliki peranan yang sangat penting dalam membantu perusahan, organisasi dan institusi pendidikan dalam menjalankan proses bisnisnya. Perkembangan teknologi yang semakin pesat membuat beberapa pekerjaan yang sebelumnya sulit untuk di prediksi dengan teknologi informasi menjadi sangat mudah untuk di prediksi. Dengan memanfaatkan IPTEK dan didukung oleh data maka dalam proses prediksi sesuatu hal bukan lagi merupakan bagian yang sulit terutama dalam bidang pendidikan [3].

Permasalahan pada penelitian ini adalah proses penjurusan IPA atau IPS siswa kelas XI pada Sekolah Menengah Atas (SMA) bahwa pihak sekolah masih mengalami kesulitan karena belum ada klasifikasi pola minat dalam menentukan pilihan jurusan, selain itu juga meminimalisir terjadinya salah penjurusan, sehingga menyebabkan ketidaksesuaian minat bakat siswa dan tentunya akan berpengaruh pada nilai akademik. Pihak sekolah merasa perlu melakukan proses penjurusan IPA dan IPS untuk siswa-siswi kelas XI berdasarkan data nilai dari mata pelajaran sehingga pihak sekolah mengetahui pola penjurusan dan membantu dalam menentukan jurusan mana yang tepat untuk siswa. Penelitian sebelumnya menyatakan dengan pemanfaatan algortima *Decision Tree C4.5*, *Naïve Bayes*, dan *K-Nearest Neighbor* membantu dan mempercepat proses klasifikasi penjurusan minta IPA dan IPS pada SMA.

Penelitian sebelumnya mengenai *data mining* dengan pemanfaatan algoritma klasifikasi dan menggunakan *tools RapidMiner* tentunya sudah banyak dilakukan oleh peneliti lain. Namun, dalam penelitian ini tetap diperlukan referensi-referensi lain dari penelitian sebelumnya sehingga dapat diketahui metode yang akan digunakan serta sebagai bentuk perbandingan kelebihan dan kekurangan dari masing-masing penelitian dalam pemanfaatan algoritma yang digunakan.

Penelitian pertama yang di jadikan acuan adalah Fahdia et al [4] yang melakukan pemanfaatan algoritma *Decision Tree C4.5*, *Naïve Bayes*, dan *K-Nearest Neighbor* untuk orientasi minat mahasiswa dalam penuntasan studi yang memperoleh hasil bahwa metode klasifikasi *Decision Tree C4.5* merupakan yang terbaik dibanding dua metode lainnya dengan nilai uji dengan *Naïve Bayes* sebear 0,226 dan nilai sebesar 0,034 ketika diuji dengan *K-Nearest Neighbor*.

Penelitian yang menjadi acuan berikutnya adalah dilakukan oleh Ikhbal dan Kurniadi [5] memperoleh hasil pengujian dalam pemanfaatan penerapan metode *data mining* menggunakan algoritma C4.5 memiliki tingkat akurasi sebesar 68,42% dalam proses penentuan penjurusan siswa di SMA Negeri 2 Padang tahun ajaran 2020/2021 dengan menggunakan atribut rekomendasi guru BK SMP yang paling berpengaruh dalam menentukan penjurusan siswa sehingga dapat mempercepat proses pengambilan keputusan klasifikasi penjurusan siswa saat proses penerimaan siswa baru.

Penelitian selanjutnya yang dijadikan acuan dalam penelitian ini yaitu penelitian yang dilakukan oleh Nugroho [6] dalam penjurusan siswa SMA Negeri 3 Boyolali menggunakan model klasifikasi dan klastering serta membandingkan tiga buah algoritma yaitu algoritma *C4.5*, *Naïve Bayes*, dan K-Means. Hasil yang didapatkan dari penelitian tersebut pada klasifikasi variabel yang memiliki pengaruh tertinggi terhadap penjurusan siswa adalah nilai rata-rata IPA menggunakan algoritma C4.5 yang dibuktikan dengan IPA menempati simpul akar pada diagram *decision tree*. Pada penelitian ini juga disebutkan bahwa berdasarkan nilai akurasi dan *recall* diperoleh hasil algoritma C4.5 memiliki nilai terbaik dibandingkan metode lain, sedangkan *Naïve Bayes* memiliki nilai lebih baik berdasarkan nilai presisi dibandingkan metode lain.

Tujuan penelitian ini adalah pemanfaatan algoritma dalam proses mengklasifikasi jurusan IPA dan IPS siswa kelas XI menggunakan algoritma klasifikasi sehingga sesuai dengan minat, bakat dan potensi diri siswa dalam penentuan jurusan. Dalam proses pengklasifikasi penelitian ini menggunakan dua algoritma agar dapat

mendapatkan perbandingan dengan klasifikasi terbaik dari algoritma yang digunakan. Algoritma yang digunakan adalah Naïve bayes dan K-Nearest Neighbor (K-NN). Hasil akurasi terbaiklah yang nantinya akan digunakan dan dimanfaatkan untuk diimplementasikan dalam proses penjurusan siswa kelas XII pada SMA.

Pemilihan jurusan yang salah juga akan berpengaruh pada saat siswa akan melanjutkan ke jenjang Pendidikan berikutnya. Silang jurusan antara siswa jurusan IPA dan IPS sering terjadi karena salah dalam memilih jurusan di awal masuk SMA, yang harusnya siswa dengan jurusan IPA menempuh perkuliahan sesuai dengan jurusannya, tetapi berpindah ke jurusan yang seharusnya di ambil oleh siswa jurusan IPS, dan begitu pula sebaliknya. Solusi yang digunakan pada penelitian ini untuk menyelesaikan permasalahan tersebut adalah konsep data mining [7].

Data mining adalah metode yang digunakan dalam pengolahan sejumlah data besar sehingga dapat menemukan sebuah pola, pengetahuan yang tersembunyi sehingga dapat menentukan proses pengambilan keputusan. Proses untuk pengolahan data diperlukan sebuah metode data mining, salah satunya adalah klasifikasi. Algoritma yang dapat digunakan salah satunya naïve bayes [8]. Data mining merupakan sebuah proses dalam menemukan hubungan baru yang menemukan sebuah arti, pola dan kebiasaan dengan memilah-milah sebagian besar data yang disimpan dalam media penyimpanan dengan menggunakan pengenalan pola seperti teknik statistik dan matematika. Data mining merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, *database*, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari *database* [9].

Klasifikasi merupakan salah satu dari metode data mining dimana klasifikasi ini merupakan jenis analisis data yang membantu dalam melakukan prediksi sesuai label kelas sample yang diklasifikasikan. Klasifikasi merupakan proses analisa data yang menghasilkan model-model yang nantinya digambarkan melalui kelas-kelas yang ada didalam data, metode ini memiliki persyaratan yakni attribut datanya harus numerik atau nominal dan label data nominal. Klasifikasi adalah proses dalam menemukan kelas data sehingga dapat memperkirakan kelas dari suatu objek yang labelnya belum diketahui [10].

Algoritma K-NN merupakan algoritma yang termasuk kedalam algoritma *supervised learning*. Supervised learning yaitu proses menemukan pola baru dalam sebuah data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan tujuan algoritma K-NN adalah untuk melakukan klasifikasi objek baru berdasarkan attribut data dan *training samples*. Hasil sample uji yang baru akan diklasifikasikan berdasarkan kategori yang ada pada K-NN. Algoritma ini menggunakan klasifikasi sebagai nilai prediksi dari uji sampel yang baru dan menggunakan jarak *Euclidean Distance.* Yang merupakan jarak yang umum digunakan untuk data numerik [11]. K-NN mempunyai nilai akurasi waktu yang paling baik dalam klasifikasi jika dikomparasi dengan teknik lain dengan hasil pemeriksaan dan aktifitas lain yang dipengaruhi beberapa penetapan aturan [12].

Algoritma Naïve bayes merupakan algoritma yang menggunakan perhitungan probabilitas. Algoritma ini umumnya digunakan untuk menyelesaikan permasalahan prediksi berupa klasifikasi. Algoritma ini juga dikenal sebagai yang memiliki akurasi yang tinggi. Proses klasifikasi pada algoritma ini terdapat 2 fase yakni *fase training* dan *fase testing*. *Fase training* atau bisa disebut sebagai *fase learning* adalah sebagian data telah diketahui kelas datanya untuk model perkiraan. Selanjutnya fase *testing* atau bisa disebut fase *classify* model yang sudah terbentuk diuji dengan sebagian data lainnya agar diketahui akurasi atas model yang sudah terbentuk [13]. Algoritma Naive Bayes salah satu algoritma klasifikasi yang cukup populer digunakan untuk studi kasus pada Data Mining maupun Text Mining, Algortima ini juga termasuk kedalam salah satu metode klasifikasi [14].
Naïve bayes merupakan metode pengklasifikasian yang memanfaatkan probabilitas dan statistik untuk memprediksi peluang di masa depan dengan memanfaatkan pengalaman di masa sebelumnya. Sistem penjurusan di Sekolah Menengah Atas (SMA) merupakan upaya untuk lebih mengarahkan siswa berdasarkan minat dan kemampuan akademiknya [15].

Rapid Miner merupakan sebuah tools yang digunakan untuk pengolahan data mining. Selain itu, digunakan untuk pengujian terhadap data yang diperoleh dari siswa-siswi. RapidMiner digunakan untuk tujuan menguji keakuratan klasifikasi penjurusan siswa-siswi SMA. Data yang dianalisis yakni data *training* nilai mata pelajaran siswa dan data *testing* menggunakan *data sample* [16]. Rapid miner salah satu solusi yang banyak digunakan untuk melakukan analisis data mining hingga prediksi, karena teknik deskriptif dan prediksi dalam memberikan wawasan sehingga menghasilkan keputusan yang paling baik [17].

2. METODE PENELITIAN

Data siswa yang digunakan pada penelitian ini adalah data nilai semester 2 (dua) dengan jumlah 277 record dan menggunakan 4 (empat) attribut dari nilai mata pelajaran antara lain PPKN, Sejarah, Prakarya dan PAI. Metode penelitian yang digunakan diadopsi dari *Model Cross-Standar Industry For Data Mining* (*CRISP-DM*) yang memiliki tahapan mulai dari *business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling* dan *Evaluation*. Berikut ini penjelasan dari masing-masing tahapan [18].

Gambar 1. Proses Data Mining CRISP-DM

2.1 *Businees Understanding*

Salah satu cara yang dilakukan pada penelitian ini dalam memproleh data adalah dengan melakukan proses wawancara kepada pihak yang memiliki tanggung jawab terkait dalam proses penjurusan SMA [18]. Pada tahapan *business understanding* menjelaskan *problems* dan *Objective*. Adapun yang menjadi ***problems*** berdasarkan fakta yang terjadi bahwa Pihak sekolah dalam proses penjurusan IPA dan IPS siswa kelas XI mengalami kesulitan karena belum ada klasifikasi pola minat dalam menentukan pilihan jurusan, selain itu juga meminimalisir terjadinya salah penjurusan, sehingga menyebabkan ketidaksesuaian minat bakat siswa dan tentunya akan berpengaruh pada nilai akademik. Pihak sekolah merasa perlu melakukan proses penjurusan IPA dan IPS untuk siswa-siswi kelas XI berdasarkan data nilai dari mata pelajaran sehingga pihak sekolah mengetahui pola penjurusan dan membantu dalam menentukan jurusan mana yang tepat untuk siswa. ***Objective*** adalah untuk menyelidiki pola penjurusan agar dapat mengklasifikasi jurusan IPA, IPS kelas XI berdasarkan nilai pelajaran PPKN, Sejarah, Prakarya dan PAI.

2.2 *Data Understanding*

Data yang digunakan adalah data nilai mata pelajaran siswa-siswi SMA. Data tersebut akan dibagi menjadi 2 yaitu data *training* sebesar 80% dan data *testing* sebesar 20%. Tahapan ini juga merupakan tahapan dalam pemilihan attribut yang nantinya akan digunakan. Namun untuk pemilihan harus menyesuaikan sesuai dengan variabel yang dibutuhkan. Pihak sekolah meminta bantuan dalam membuat matriks korelasi dengan 8 attribut reguler adalah sebagai berikut.

a. No adalah sebagai jumlah urutan yang nantinya diidentifikasi sebagai id.
b. Nama adalah sejumlah nama siswa-siswi kelas XI yang hendak melakukan penjurusan.
c. Jenis Kelamin adalah jenis kelamin siswa-siswi kelas XI
d. Sejarah adalah nilai dari mata pelajaran sejaran (Nilai semester II)
e. PPKN adalah nilai mata pelajaran Pendidikan Pancasila dan Kewarganegaran (Nilai semester II)
f. Prakarya adalah nilai mata pelajaran prakarya siswa-siswi (Nilai semester II)
g. PAI adalah nilai mata pelajaran Pendidikan Agama Islam (Nilai semester II)
h. Program adalah jurusan IPA dan IPS yang akan diidentifikasi sebagai label.

2.3 *Data Preparation*

Tahap *data preparation* yang dilakukan pada penelitian ini terdiri dari *data cleaning, data integration. Data cleaning* adalah proses untuk membersihkan data seperti mencari dan mengisi nilai yang kosong pada data siswa-siswi (*Missing value*s), mencari dan memecahkan data yang tidak konsisten, menghilangkan *Noisy/Outlier*, mencari data duplikat dan *incomplete data.* Proses pembersihan data dapat berdampak baghkan mempengaruhi proses data mining karena data akan berkurang akibat hasil dari pembersihan. Sedangkan *data integration* adalah

proses untuk menggabungkan data dari tempat penyimpanan yang berbeda dalam satu data atau satu file. Hal ini dapat terjadi karena data bersifat heterogen.

2.4 *Modelling*

Tahap *modelling* pada penelitian ini ada proses melakukan penerapan teknik dan algoritma data mining terhadap data siswa siswa untuk proses penjurusan IPA dan IPS menggunakan bantuan *tools.* Algoritma yang digunakan adalah algoritma *Naïve Bayes dan K-Nearest Neighbor.* Kedua algoritma akan dikomparasi melalui operator *performance* sehingga dihasilkan nilai akurasi yang tertinggi. Selain itu, juga dilakuakn pengujian menggunakan operator t-test sebagai bentuk rangkain akhir komparasi kedua algoritma. Dari hasil komparasi kedua algoritma yang memberikan hasil akurasi yang tertinggi yang akan digunakan untuk proses implementasi klasifikasi penjurusan siswa kelas XI SMA. *Tools* yang digunakan untuk tahapan *modelling* ini adalah menggunakan RapidMiner.

2.5 *Evaluation*

Algoritma yang sudah diimplementasikan pada proses penelitian ini selanjutnya akan diimplementasikan pada klasifikasi penjurusan siswa SMA melalui simulasi model. Evaluasi ini akan dilakukan dengan melihat dan mengamati hasil dari proses klasifikasi dan implementasi algoritma *Naïve Bayes dan K-Nearest Neighbor.* Hasil nilai pengukuran akurasi pada penelitian ini menggunakan *performance* Vektor dengan memperhatikan model evaluasi *confusion matrix* nilai *accuracy, precision* dan AUC. Sehingga dapat diketahui tingkat akurasi dari masing-masing algoritma yang digunakan pada penelitian ini.

3. HASIL DAN PEMBAHASAN

3.1 *Data Preparation Phase*

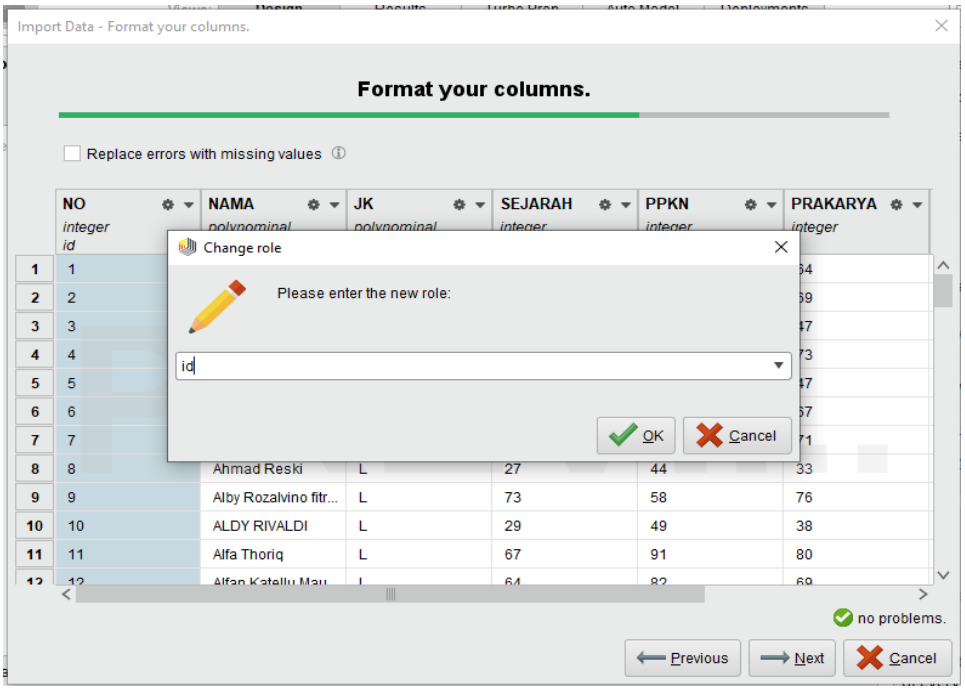
Persiapan awal yang dilakukan adalah mempersiapkan kumpulan data yang akan digunakan untuk fase berikutnya secara keseluruhan. Pada penelitian ini menggunakan sebanyak 277 *record* dengan *Dataset* Nilai Siswa. *Dataset* menggunakan format file excel agar tools Rapid Miner dapat melakukan pembacaan file. Berikut pada gambar 2 ditunjukkan dataset siswa.

	A	B	C	D	E	F	G	H	I
1	NO	NAMA	JK	SEJARAH	PPKN	PRAKARYA	PAI	PROGRAM	
2	1	Adam Fery Sujatmo	L	60	84	64	86	IPA	
3	2	Ade Nazwa Adilla	P	64	69	69	62	IPA	
4	3	Adelia Lestari	P	53	60	47	68	IPA	
5	4	Adi Nova Kurniawan	L	58	80	73	84	IPA	
6	5	Aditya Iswandi	L	67	82	47	49	IPA	
7	6	Afif Fachrozy	L	71	91	67	73	IPA	
8	7	Afra Amaniah	P	62	84	71	78	IPA	
9	8	Ahmad Reski	L	27	44	33	30	IPA	
10	9	Alby Rozalvino fitrajaya	L	73	58	76	81	IPA	
11	10	ALDY RIVALDI	L	29	49	38	51	IPA	
12	11	Alfa Thoriq	L	67	91	80	76	IPA	
13	12	Alfan Katellu Maulana	L	64	82	69	86	IPA	
14	13	Alfian Syarif Hidayatullah	L	71	89	67	73	IPA	
15	14	Ali bintang syari'ati	L	64	76	62	68	IPS	
16	15	Alia apriyanti	P	49	67	58	59	IPA	
17	16	ALIFIAH MATANO DITA	P	76	87	80	92	IPA	
18	17	Alta yulyan insani	L	64	91	76	86	IPA	
19	18	Alya Febrianti	L	49	80	71	65	IPA	
20	19	Amira Alida putri	P	62	87	76	73	IPS	
21	20	Andika	L	64	80	56	62	IPA	
22	21	andini hadifansya	P	53	91	62	73	IPA	
23	22	ANDYOKTAVIADI	L	64	84	71	70	IPS	

Gambar 2. *Dataset* Data Siswa

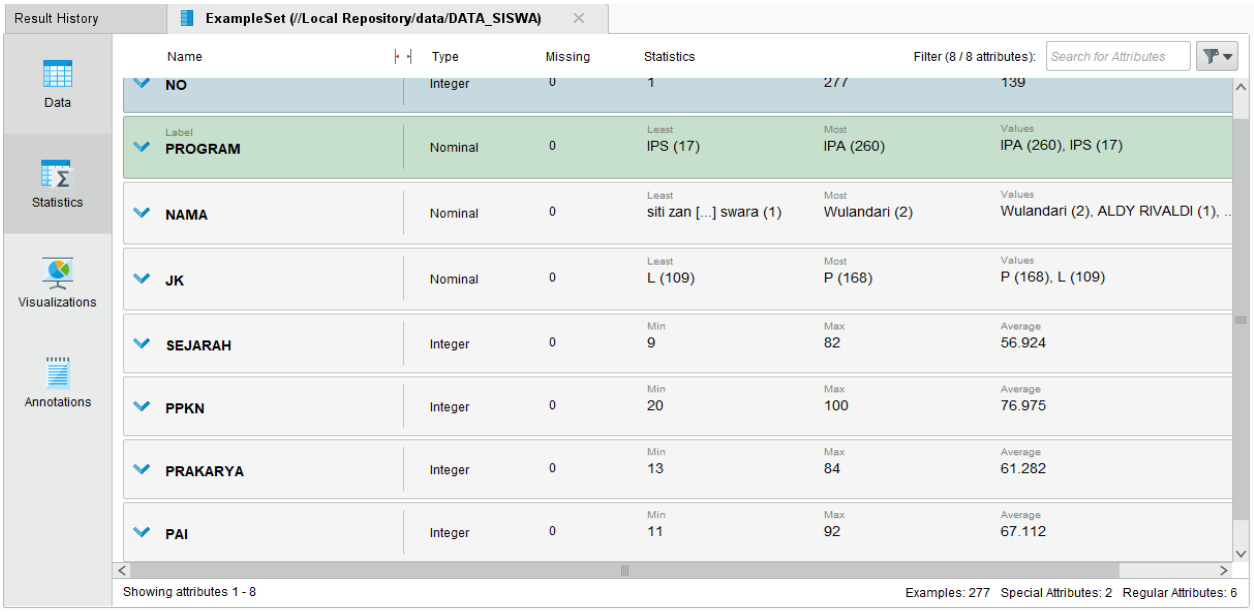
Lalu tahapan selanjutnya adalah melakukan impor data siswa tersebut ke dalam aplikasi Rapid Miner dan menentukan atribut yang akan dijadikan *id* dan *label* atau target dengan melakukan change role. Pada penelitian ini yang dijadikan sebagai *id* adalah nomer dan *label* ada program. Penentuan *id* dan *label* dilakukan karena salah satu prasyarat metode data mining yakni klasifikasi adalah *attribut* bertipe nominal/numerik dan label bertipe nominal. Pada tahap ini juga bisa melakukan *replace errors* dengan *missing values* dengan cara melakukan mengisi checklist diatas tabel. Pengaturan untuk melakukan perubahan *id* dan *label* dilakukan pada bagian *format*

your columns dengan mengklik icon gear pada *attribut* yang akan dijadikan id maupaun label kemudian pilih *change role*. Secara detail dalam penentuan pengaturan *id* dan *label* ditunjukkan pada gambar 3.



Gambar 3. Menentukan id dan label

Lakukan pengecekan pada menu *Statistics* sebagai salah satu bentuk *data cleaning*. Untuk melakukan pengecekan dapat dilakukan secara manual apakah ada data yang *missing*, *error*, *incomplete data*, atau *noisy*. Berdasarkan dari menu *statistics* tidak terlihat *error* ataupun *missing values* pada 277 *ExampleSet* dengan 2 *special attributes* yaitu *id* nomor dan *label* program, serta *regular attributes* sebanyak 6 atribut berupa nama, jk, sejarah, ppkn, prakarya, dan PAI.

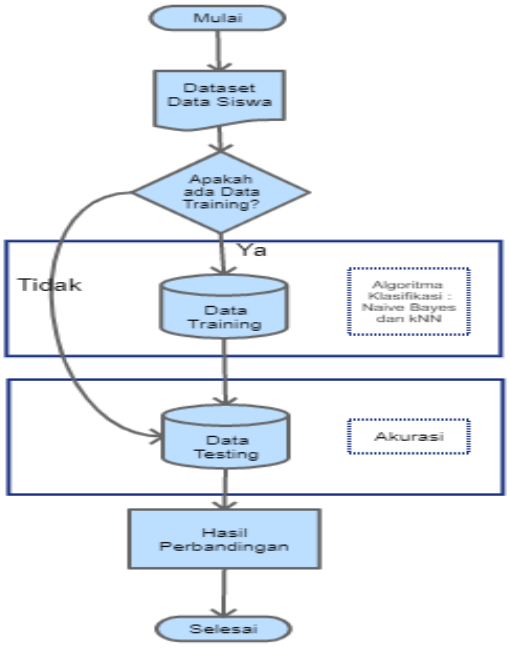


Gambar 4. Pengecekan pada Menu Statistics

3.2 *Modelling Phase*

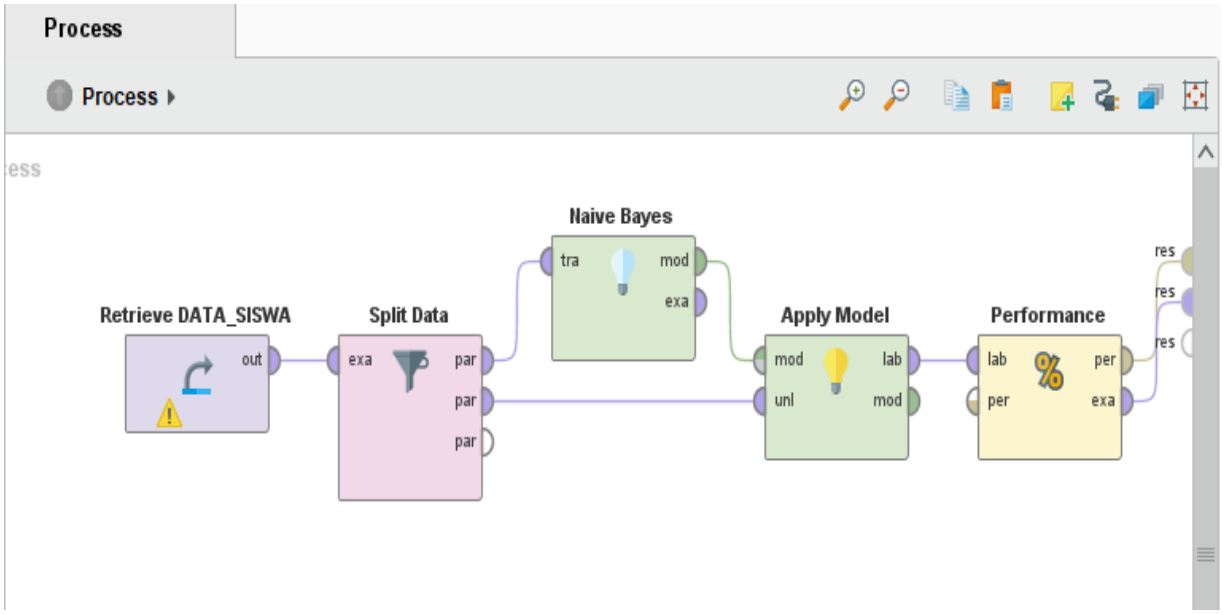
Pada fase ini, peneliti menentukan metode sesuai dengan karakter data yaitu klasifikasi. Tahapan *modelling phase* terlihat pada bagian *data training* dan *data testing*. Untuk *data training* dilakukan menggunakan algoritma klasifikasi yang sudah ditentukan yakni naïve bayes dan K-NN sehingga himpunan data yang berada pada dataset

di *traning*. Selanjutnya dilakukan pengujian data untuk melihat nilai akurasi dari kedua algoritma yang ditentukan. Tahapan pada model yang diusulkan pada fase ini dijabarkan dalam flowchart gambar 5 berikut ini.



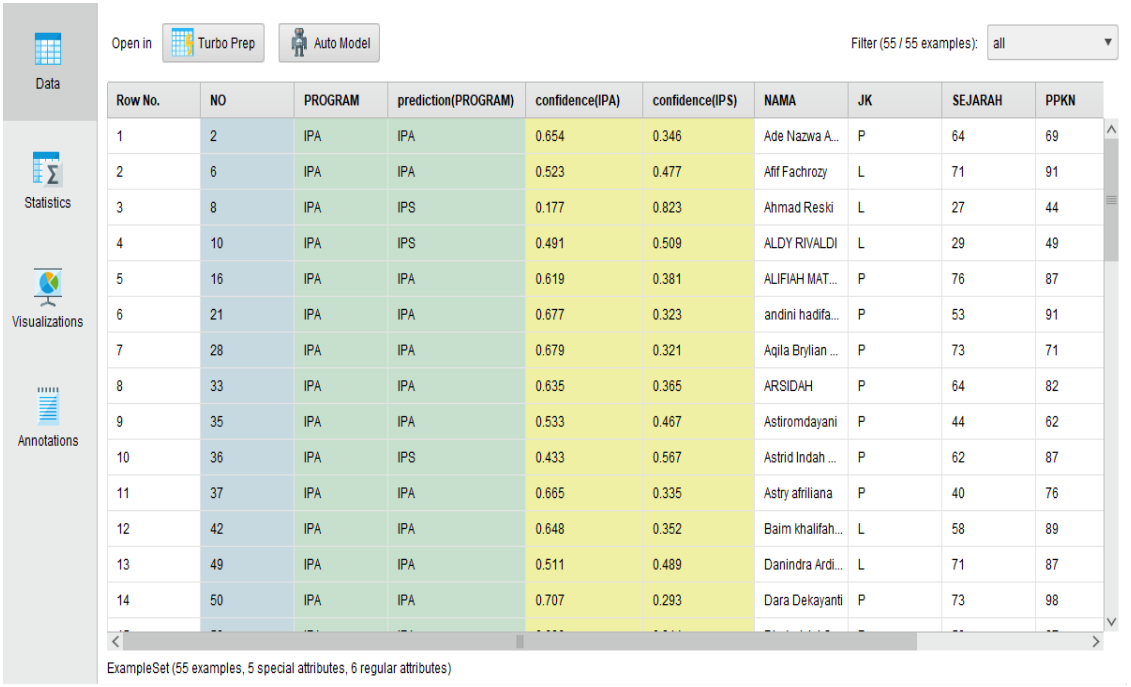
Gambar 5. *Flowchart* model usulan

Tahapan pada RapidMiner untuk *modeling* klasifikasi menggunakan algoritma Naïve Bayes, yaitu dengan menggunakan operator Data (*Split Data*). Pada bagian ini khusus pada penelitian ini membagi menjadi antara *Data Training* sebanyak 80% dan *Data Testing* sebanyak 20%, kemudian menambahkan modeling menggunakan algoritma Naïve Bayes. Lalu untuk mendapatkan hasil dari *dataset* yang ada, dilakukan penambahan operator *Apply Model*. Operator ini digunakan bertujuan untuk mendapatkan prediksi dataset baru hasil implementasi tahapan sebelumnya.



Gambar 6. Proses Model Klasifikasi dengan Algoritma Naive Bayes

Hasil *running* program Rapid Miner setelah implementasi *Apply Model* menghasilkan hasil *prediction* , nilai *confidence* untuk attribut yang sudah dijadikan *label* yakni Program IPA dan IPS. Hasil prediksi dan nilai *confidence* ditunjukkan pada gambar 7 berikut.



Row No.	NO	PROGRAM	prediction(PROGRAM)	confidence(IPA)	confidence(IPS)	NAMA	JK	SEJARAH	PPKN
1	2	IPA	IPA	0.654	0.346	Ade Nazwa A...	P	64	69
2	6	IPA	IPA	0.523	0.477	Afif Fachrozy	L	71	91
3	8	IPA	IPS	0.177	0.823	Ahmad Reski	L	27	44
4	10	IPA	IPS	0.491	0.509	ALDY RIVALDI	L	29	49
5	16	IPA	IPA	0.619	0.381	ALIFIAH MAT...	P	76	87
6	21	IPA	IPA	0.677	0.323	andini hadifa...	P	53	91
7	28	IPA	IPA	0.679	0.321	Aqila Brylian ...	P	73	71
8	33	IPA	IPA	0.635	0.365	ARSIDAH	P	64	82
9	35	IPA	IPA	0.533	0.467	Astiromdayani	P	44	62
10	36	IPA	IPS	0.433	0.567	Astrid Indah ...	P	62	87
11	37	IPA	IPA	0.665	0.335	Astry afriliana	P	40	76
12	42	IPA	IPA	0.648	0.352	Baim khalifah...	L	58	89
13	49	IPA	IPA	0.511	0.489	Danindra Ardi...	L	71	87
14	50	IPA	IPA	0.707	0.293	Dara Dekayanti	P	73	98

ExampleSet (55 examples, 5 special attributes, 6 regular attributes)

Gambar 7. Tahap Apply Model Algoritma Naïve Bayes

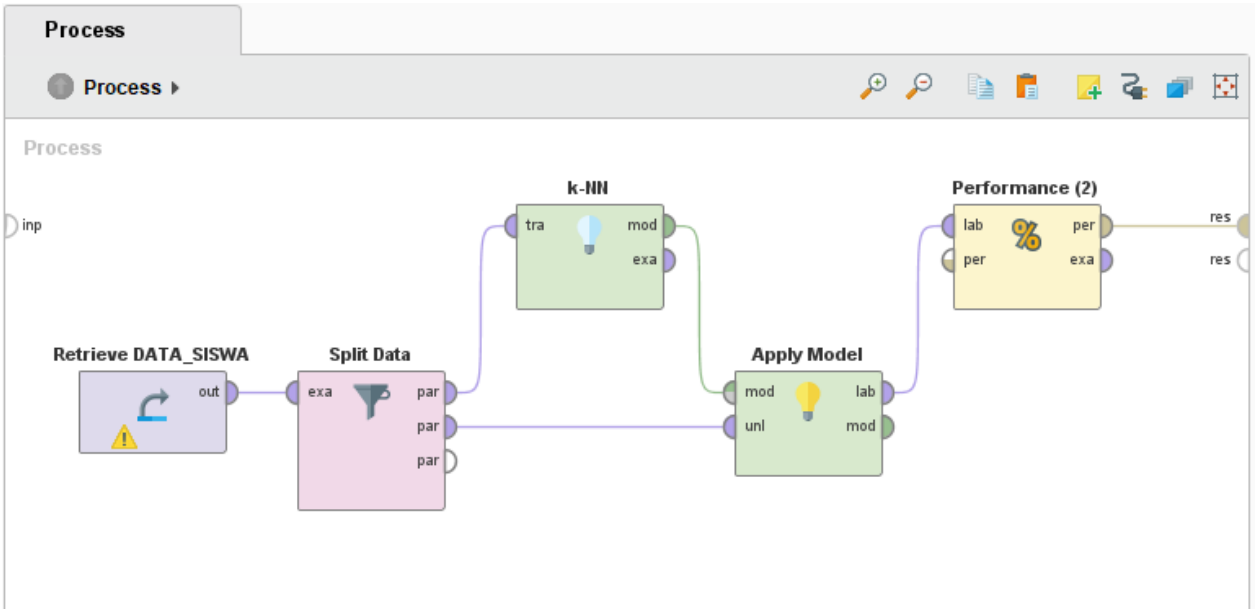
Selanjutnya dilakukan pengecekan *performansi vector* pada Algoritma *Naïve Bayes* sehingga didapatkan hasil akurasi sebesar 81.82%. Hasil prediksi juga ditunjukkan dari *confusion matrix* beserta nilai *AUC* yang disajikan pada gambar 8 berikut.

PerformanceVector

```
PerformanceVector:
accuracy: 81.82%
ConfusionMatrix:
True:   IPA     IPS
IPA:    43      1
IPS:    9       2
precision: 18.18% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    43      1
IPS:    9       2
recall: 66.67% (positive class: IPS)
ConfusionMatrix:
True:   IPA     IPS
IPA:    43      1
IPS:    9       2
AUC (optimistic): 0.865 (positive class: IPS)
AUC: 0.865 (positive class: IPS)
AUC (pessimistic): 0.865 (positive class: IPS)
```

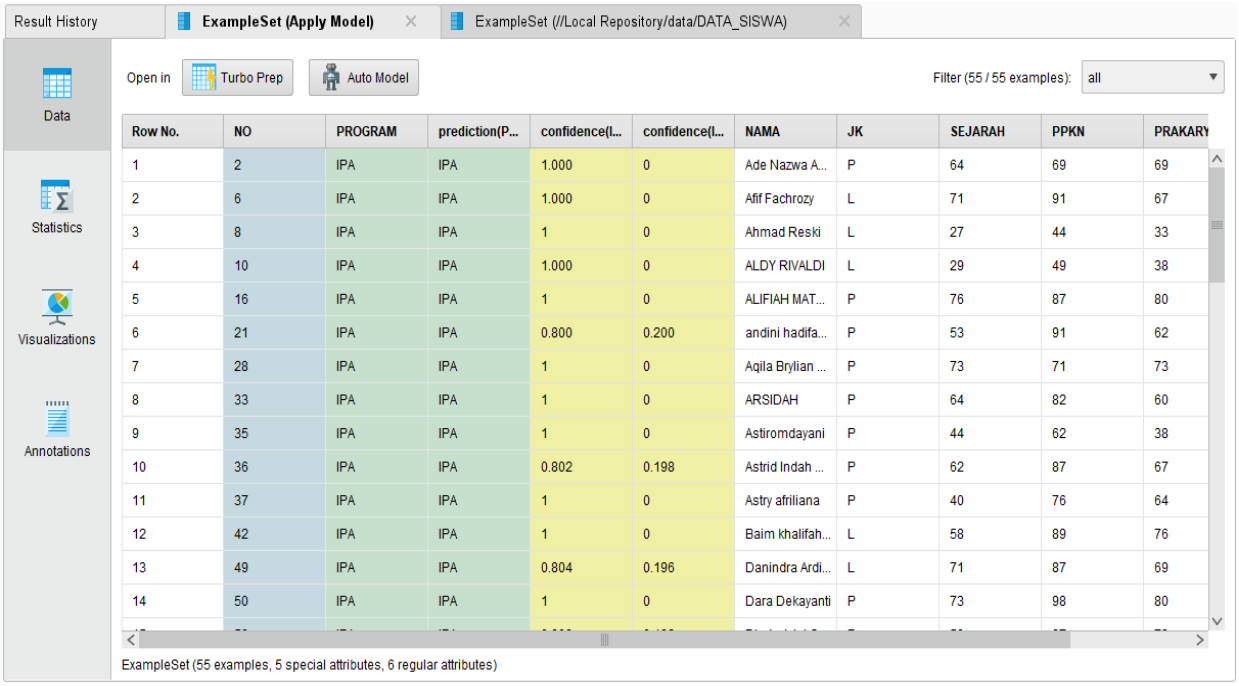
Gambar 8. *Performance* dengan Algoritma Naive Bayes

Tahapan pada Rapid Miner untuk model klasifikasi menggunakan algoritma *K-Nearest Neighbors*, yaitu dengan membagi data (*Split Data*) antara *Data Training* sebanyak 80% dan *Data Testing* sebanyak 20%, kemudian menambahkan *modeling* menggunakan Algoritma K-Nearest Neighbors. Lalu untuk mendapatkan hasil dari *dataset* yang ada, dilakukan penambahan operator *Apply Model.* Untuk implementasi menggunakan algoritma *K-Nearest Neighbors* di tunjukkan pada gambar 9 berikut.



Gambar 9. Proses Model Klasifikasi dengan Algoritma K-Nearest Neighbors

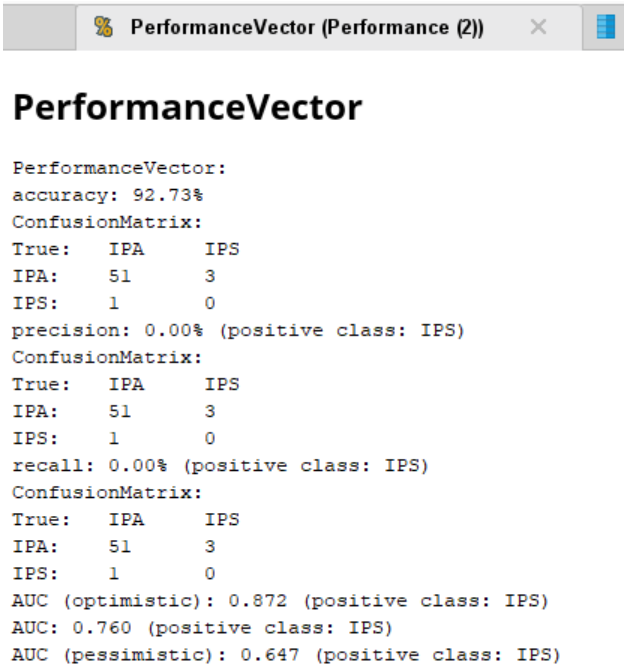
Hasil ekseskusi tools Rapid Miner setelah implementasi *Apply Model* pada algoritma K-NN menghasilkan hasil *prediction*, nilai *confidence* untuk attribut yang sudah dijadikan *label* yakni Program IPA dan IPS. Hasil prediksi dan nilai *confidence* ditunjukkan pada gambar 10 berikut. Nilai confidence disini nilai yang paling mendekati dengan fakta.



Row No.	NO	PROGRAM	prediction(P...	confidence(I...	confidence(I...	NAMA	JK	SEJARAH	PPKN	PRAKARY
1	2	IPA	IPA	1.000	0	Ade Nazwa A...	P	64	69	69
2	6	IPA	IPA	1.000	0	Afif Fachrozy	L	71	91	67
3	8	IPA	IPA	1	0	Ahmad Reski	L	27	44	33
4	10	IPA	IPA	1.000	0	ALDY RIVALDI	L	29	49	38
5	16	IPA	IPA	1	0	ALIFIAH MAT...	P	76	87	80
6	21	IPA	IPA	0.800	0.200	andini hadifa...	P	53	91	62
7	28	IPA	IPA	1	0	Aqila Brylian ...	P	73	71	73
8	33	IPA	IPA	1	0	ARSIDAH	P	64	82	60
9	35	IPA	IPA	1	0	Astiromdayani	P	44	62	38
10	36	IPA	IPA	0.802	0.198	Astrid Indah ...	P	62	87	67
11	37	IPA	IPA	1	0	Astry afriliana	P	40	76	64
12	42	IPA	IPA	1	0	Baim khalifah...	L	58	89	76
13	49	IPA	IPA	0.804	0.196	Danindra Ardi...	L	71	87	69
14	50	IPA	IPA	1	0	Dara Dekayanti	P	73	98	80

Gambar 10. Tahap Apply Model Algoritma K-Nearest Neighbors

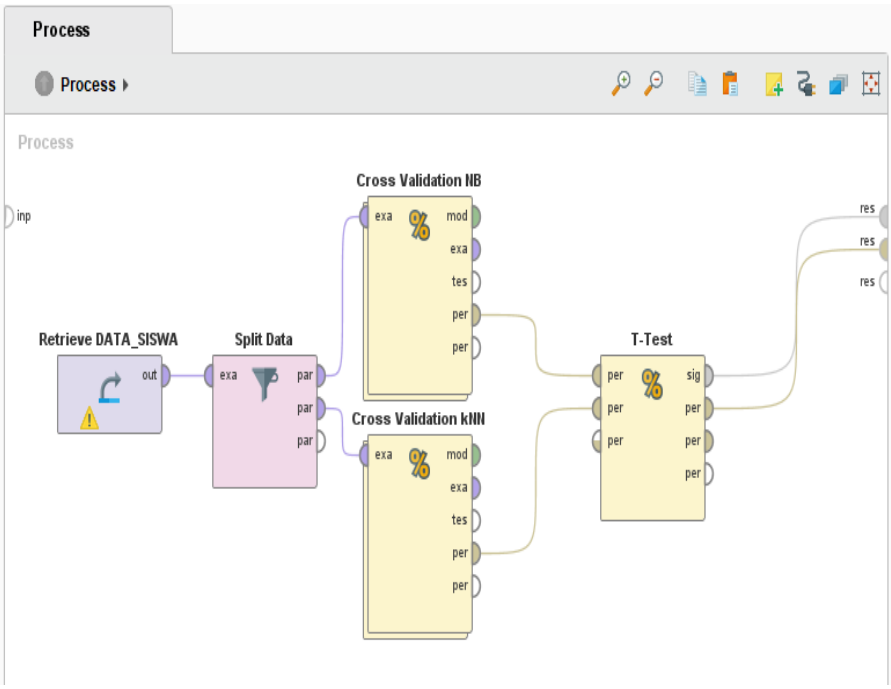
Pada *Apply Model*, setelah didapatkan nilai *prediction* pada label dengan nilai *confidence* atribut IPA dan IPS. Selanjutnya dilakukan pengecekan performansi pada Algoritma *K-Nearest Neighbors* sehingga didapatkan hasil akurasi sebesar 92.73%. selain itu juga telihat nilai hasil prediksi yang disajikan melalui confusion Matrix yang ditunjukkan pada gambar 11 berikut.



Gambar 1. Performance dengan Algoritma K-Nearest Neighbors

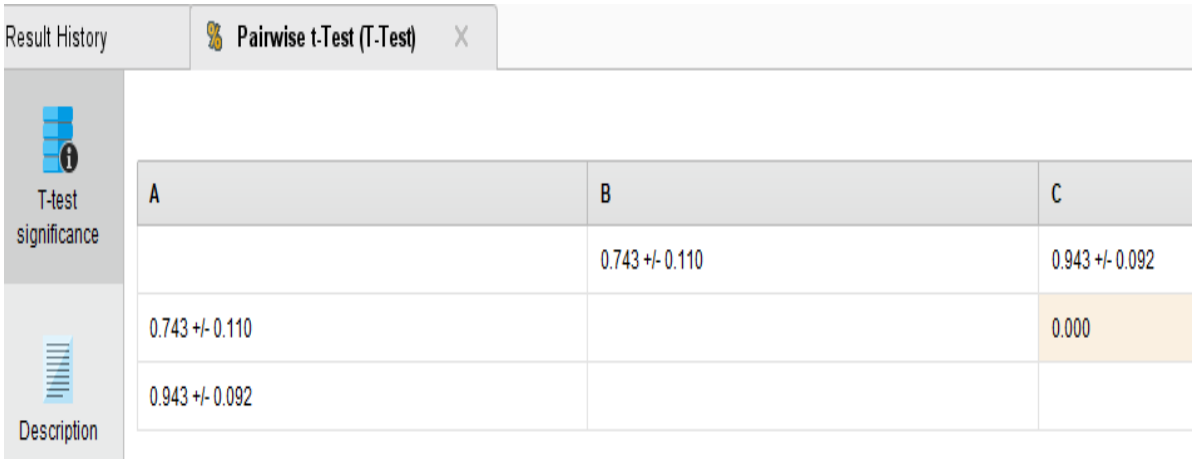
3.3 *Evaluation Phase*

Pada tahapan evaluasi, peneliti melakukan analisis model dan kinerja metode dengan cara melihat nilai akurasi dan membandingkan antara Algoritma *Naive Bayes* dan *K-Nearest Neighbors*, serta menentukan urutan model terbaik dalam pengolahan data siswa. Untuk prosesnya menggunakan operator *corss validation* untuk kedua algortima dan operator T-Test untuk komparasi kedua algoritma. Hasil komparasi ditunjukkan pada gambar berikut.



Gambar 2. Performance dengan Algoritma K-Nearest Neighbors

Dari hasil perbandingan menggunakan T-Test diketahui tidak ada perbedaan signifikan karena nilai alpha dibawah 0.05, dengan nilai perbandingan model terbaik yaitu K-Nearest Neighbors dengan nilai akurasi sebesar 92.73% dan nilai akurasi Naïve Bayes sebesar 81.82%.



A	B	C
	0.743 +/- 0.110	0.943 +/- 0.092
0.743 +/- 0.110		0.000
0.943 +/- 0.092		

Gambar 3. Hasil Perbandingan T-Test Algoritma *Naive Bayes* dan *K-Nearest Neighbors*

4. KESIMPULAN

Berdasarkan hasil pembahasan algoritma *Naïve bayes* dan *K-Nearest Neighbor* untuk proses klasifikasi penjurusan program IPA IPS kelas XI pada Sekolah Menegah Atas, secara keseluruhan memberikan hasil yang cukup baik. Sesuai dengan pembahasan permasalahan Pihak sekolah dalam proses penjurusan IPA dan IPS siswa kelas XI mengalami kesulitan karena belum ada klasifikasi pola minat dalam menentukan pilihan jurusan, selain itu juga meminimalisir terjadinya salah penjurusan, sehingga menyebabkan ketidaksesuaian minat bakat siswa dan tentunya akan berpengaruh pada nilai akademik dibuktikan dengan kedua perbandingan algoritma yang menghasilkan nilai *accuracy* yakni nilai data yang diprediksi mendekati dengan nilai yang sesungguhnya. untuk Algoritma *Naïve bayes* memberikan hasil *accuracy* sebesar 81.82%, dengan hasil *confusion Matrix:*

a. Prediksi IPA – True IPA: jumlah data yang diprediksi IPA dan kenyataannya IPA (True Positif).
b. Prediksi IPS – True IPS: jumlah data yang diprediksi IPS dan kenyataannya IPS (True Negatif).
c. Prediksi IPA – True IPS: jumlah data yang diprediksi IPA tapi kenyataannya IPS (False Positif)
d. Prediksi IPS – True IPA: jumlah data yang diprediksi IPS tapi kenyataannya IPA (False Negatif)

Artinya kesesuaian hasil prediksi dengan kenyataannya menghasilkan nilai true positif. Sedangkan algoritma K-*Nearest Neighbor* memberikan hasil *accuracy* sebesar 92.73%, artinya jumlah data yang diprediksi memilih jurusan IPA dan pada kenyataaannya memilih IPA menghasilkan nilai yang signifikan atau dapat disebut sebagai *True Positif*. Selain itu juga menghasilkan pola klasifikasi sehingga membantu dalam proses pengklasifikasian untuk jurusan yang dipilih.

Hasil perbandingan dari kedua algoritma yang telah dilakukan memalui pengujian dari uji *T-Test* memberikan hasil nilai alpha < 0.05, artinya tidak ada perbedaan yang signifikan antara nilai rata-rata sebenarnya di kedua algoritma tersebut. Sehingga dapat disimpulkan algoritma *K-Nearest Neighbor* yang terbaik yang memiliki tingkat akurasi paling tinggi yang artinya paling mendekati dengan kenyataannya.

REFERENCES

[1] S. W. Nengsih, I. Alfian, D. Aji, and S. Anwar, "Analisis Pengelompokan Penentuan Jurusan Siswa Sma Menggunakan Metode K-Means Clustering," *J. Betrik*, no. 03, pp. 242–248, 2021.

[2] F. Ekawati, "Algoritma Naïve Bayes Untuk Penentuan Jurusan Pada Siswa Madrasah Aliyah," *Technol. J. Ilm.*, vol. 9, no. 1, p. 42, 2018, doi: 10.31602/tji.v9i1.1101.

[3] D. Novianti, "Implementasi Algoritma Naïve Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner," *Paradig. - J. Komput. dan Inform.*, vol. 21, no. 1, pp. 49–54, 2019, doi: 10.31294/p.v21i1.4979.

[4] M. R. Fahdia, D. Riana, F. Amsury, I. Saputra, and N. Ruhyana, "Komparasi Algoritma Klasifikasi untuk Orientasi Minat Mahasiswa dalam Penuntasan Studi," *JIRA J. Inov. dan Ris. Akad.*, vol. 2, no. 7, pp. 970–1007, 2021, doi: 10.47387/jira.v2i7.185.

[5] M. F. D. Ikhbal and D. Kurniadi, "Menentukan Penjurusan Siswa Dengan Menggunakan Metode Decision Tree Algoritma C4.5 (Studi Kasus : SMA Negeri 2 Padang)," *JAVIT(jurnal vokasi Inform.*, vol. 1, no. 3, pp. 31–37, 2021.

[6] Y. S. Nugroho, "Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 1, no. 1, p. 1, 2015, doi: 10.23917/khif.v1i1.1175.

[7] H. Hairani, Muhammad Ridho Hansyah, and Lalu Zazuli Azhar Mardedi, "Integrasi Metode Naive Bayes dengan K-Means dan K-Means-Smote untuk Klasifikasi Jurusan SMAN 3 Mataram," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 8–12, 2020, doi: 10.30864/jsi.v15i1.317.

[8] A. Sabathos Mananta and G. Arther Sandag, "Prediksi Kelulusan Mahasiswa Dalam Memilih Program Magister Menggunakan Algoritma K-NN," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 10, no. 2, pp. 90–96, 2021, doi: 10.30591/smartcomp.v10i2.2488.

[9] A. Z. Mafakhir and A. Solichin, "Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta," *Fountain Informatics J.*, vol. 5, no. 1, p. 21, 2020, doi: 10.21111/fij.v5i1.4007.

[10] F. Ariani, Amir, N. Alam, and K. Rizal, "Klasifikasi Penetapan Status Karyawan Dengan Menggunakan Metode Naïve Bayes," *Paradig. - J. Komput. dan Inform.*, vol. 20, no. 2, pp. 33–38, 2018, doi: 10.31294/p.v.

[11] A. Purwanto *et al.*, "Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma," vol. 2, no. 1, pp. 43–47, 2018.

[12] E. Purwaningsih and E. Nurelasari, "Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa," *Syntax J. Inform.*, vol. 10, no. 01, pp. 46–55, 2021, [Online]. Available: https://journal.unsika.ac.id/index.php/syntax/article/download/5173/2749.

[13] S. Sinaga, R. W. Sembiring, and S. Sumarno, "Penerapan Algoritma Naive Bayes untuk Klasifikasi Prediksi Penerimaan Siswa Baru," vol. 1, no. 1, pp. 55–64, 2022.

[14] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," *J. Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.

[15] A. H. Hailitik, B. S. Djahi, and Y. Y. Nabuasa, "Klasifikasi Jurusan Menggunakan Metode Naïve Bayes Pada Sekolah Menengah Atas Negeri (SMAN) 1 Fatuleu Tengah," *J-Icon*, vol. 5, no. 2, pp. 21–27, 2017, [Online]. Available: http://ejurnal.undana.ac.id/jicon/article/view/361.

[16] S. Marpaung, S. -, and I. -, "Penerapan Metode Naïve Bayes Dalam Memprediksi Prestasi Siswa Di SMA Negeri 1 Panombeian Panei," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 8–13, 2021, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.1522.

[17] I. P. Ninditama, I. P. Ninditama, W. Cholil, M. Akbar, and D. Antoni, "Klasifikasi Keluarga Sejahtera Study Kasus : Kecamatan Kota Palembang," *J. TEKNO KOMPAK*, vol. 15, no. 2, pp. 37–49, 2020, [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknokompak/article/view/1156.

[18] E. B. Sambani and F. Nuraeni, "Penerapan Algoritma C4.5 Untuk Klasifikasi Pola Penjurusan di Sekolah Menengah Kejuruan (SMK) Kota Tasikmalaya," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 9, no. 3, p. 144, 2018, doi: 10.22303/csrid.9.3.2017.144-152.