IJAI (Indonesian Journal of Applied Informatics)
Vol. 5 No. 1 Tahun 2020 pISSN: 2548-3846, eISSN: 2598-5981

Implementasi *Naïve Bayes* untuk Klasifikasi Tunggakan Iuran Sekolah

Rizal Nur Alfi*, Jajam Haerul Jaman, Riza Ibnu Adam
Program Studi S1 Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Email : rizal.16193@student.unsika.ac.id*

Info Artikel

Kata Kunci :
KDD, *naïve bayes*, penunggakan iuran sekolah, *percentage split*, *website*

Keywords :
arrears school tuition, *KDD*, *naïve bayes*, *percentage split*, *website*

Tanggal Artikel
Dikirim : 12 November 2020
Direvisi : 25 November 2020
Diterima : 30 November 2020

Abstrak

Penunggakan iuran sekolah menjadi salah satu permasalahan yang ada pada setiap sekolah ataupun institusi pendidikan lainnya. Salah satu sekolah di Karawang mengalami masalah yang sama dalam penunggakan iuran sekolah yang dilakukan oleh para siswanya. Sekolah tersebut mengalami sejumlah kerugian akibat tunggakan ini. Terhambatnya proses administrasi sekolah akan terjadi dan pihak sekolah harus memiliki strategi untuk menanganinya. Maka dari itu, penelitian ini melakukan klasifikasi terhadap siswa yang akan melakukan tunggakan dengan mengimplementasikan algoritma *Naïve Bayes* dan hasilnya diterapkan ke dalam sebuah sistem berbasis *website*. Penelitian ini menggunakan metodologi *KDD* yang sering digunakan untuk olah *data mining*. Pada perhitungan pengujian diterapkan skenario pembagian *data training* dan *data testing* menggunakan *percentage split* guna mencari pemodelan mana yang menghasilkan kinerja optimal. Hasil akhir pemodelan menghasilkan rasio pembagian 50:50 memiliki hasil yang terbaik dengan nilai Akurasi sebesar 85.461%, nilai Presisi sebesar 0.869, nilai *Recall* sebesar 0.855 dan nilai *F-Measure* sebesar 0.857. Sistem berbasis *website* dibangun sebagai hasil akhir diuji menggunakan *Black Box Testing* dengan metode *Boundary Value Analysis* yang hasilnya menunjukkan tiga skenario uji berhasil dilakukan pada semua elemen *field* sistem yang ditandai dengan keterangan "*Success*" pada kesimpulan.

Abstract

Arrears school tuition are one of the problems that exist in every school or other educational institution. One of schools in Karawang also experienced the same problem in context of arrears school tuition by its students. This school suffered from all students in many of arrears. Obstructed of school administration process will occur and the school must have a strategy to handle it. Therefore, this research classifies students who will do arrears by implementing the Naïve Bayes algorithm and the results are applied into a website for arrears school tuition. This research uses KDD methodology which is often used for data mining process. In the test calculations, scenario for split training data and testing data using a percentage split technique was carried out to find which modeling produces optimal performance. The final result of modeling produces a split ratio of 50:50 has the best results with an accuracy score of 85.461%, a precision score of 0.869, a recall score of 0.855 and a F-Measure score of 0.857. This website is built as the final result and tested using black box testing with the Boundary Value Analysis method, the results show that three test scenarios were successfully carried out on all elements of the system field which are marked with the description "Success" at the conclusion.

1. PENDAHULUAN

Pendidikan adalah salah satu unsur kebutuhan untuk memenuhi kehidupan sebagai manusia yang berintelektual dan berbudi pekerti. Sejatinya, pendidikan bisa kita dapatkan dimanapun kita berada tergantung sudut pandang kita dalam melihat setiap kondisi atau kejadian. Di Indonesia, secara nasional pendidikan terbagi dalam dua jenis, yakni pendidikan luar sekolah dan pendidikan sekolah. Pendidikan luar sekolah merupakan jenis pendidikan yang diperoleh setiap anak dari lingkungan atau dunia luarnya. Pendidikan tersebut mengajarkan tentang wawasan dan arti dalam sebuah kehidupan yang dialaminya sehari-hari. Berbeda dengan pendidikan luar sekolah, pendidikan sekolah merupakan jenis pendidikan yang berjenjang, terstruktur dan berkesinambungan. Pendidikan sekolah memiliki jenjang atau tingkatan dari yang paling dasar hingga tingkatan tinggi dengan struktur yang saling berhubungan dan berkembang. Selain jenjang pendidikan, pendidikan sekolah memiliki beberapa komponen pendidikan yang harus dimiliki oleh setiap instansi pendidikan sekolah di Indonesia [1]. Sudah banyak sekolah yang berdiri di berbagai daerah Indonesia, termasuk di kabupaten Karawang. Salah satu sekolah di Karawang ini merupakan sekolah yang dipilih menjadi objek penelitian karena telah sukses dalam menerapkan komponen-komponen pendidikan seperti yang dikatakan pada buku Amos dan Gracia tentang komponen pendidikan [1].

Namun di samping itu, masih terdapat satu permasalahan yang menjadi hambatan bagi pihak sekolah. Permasalahan ini terletak pada bagian administrasi yang terhambat akibat tunggakan iuran sekolah yang dilakukan oleh para siswa di salah satu sekolah di Karawang tersebut. Permasalahan tunggakan tersebut dapat menjadi krusial jikalau pihak sekolah tidak membuat strategi dalam menanggulangi permasalahan ini. Informasi tunggakan yang diperoleh berdasarkan data informasi dari pihak sekolah selama masa studi tiga tahun akademik terakhir, yakni tahun akademik 2017/2018, tahun akademik 2018/2019 dan tahun akademik 2019/2020. Maka dari itu, diperlukan sebuah solusi untuk menyelesaikan permasalahan tersebut, salah satunya menggunakan teknik klasifikasi dalam bidang *data mining*. Dalam klasifikasi, diketahui memiliki banyak metode dan algoritma yang digunakan, seperti *Naïve Bayes*, *Support Vector Machine*, *Decision Tree*, Jaringan Syaraf Tiruan dan *Fuzzy. Naïve Bayes* merupakan algoritma pada klasifikasi *data mining* yang sering digunakan karena terbukti sebagai algoritma yang unggul dalam memproses tipe data kategori. Selain itu, *Naïve Bayes* memiliki performa yang baik di antara algoritma klasifikasi yang lain [2].

Dasar atas keunggulan dari algoritma *Naïve Bayes* ini ada pada penelitian yang telah dilakukan oleh [3] membahas tentang komparasi *Naïve Bayes*, SVM dan K-NN pada prediksi kelancaran pembayaran TV kabel menghasilkan perbandingan setiap algoritmanya dengan pengujian Akurasi, AUC dan *t-Test*. *Dataset* yang sebanyak 100 *record* dan 7 atribut. Hasilnya menunjukkan bahwa nilai Akurasi *Naïve Bayes* sebesar 96% dengan nilai AUC sebesar 0,99, lalu nilai Akurasi SVM sebesar 66% dengan nilai AUC sebesar 0,786 dan nilai Akurasi K-NN sebesar 92% dengan nilai AUC sebesar 0,971 serta hasil *t-Test* mendapatkan *Naïve Bayes* dan K-NN lebih dominan dibandingkan metode SVM. Dapat disimpulkan bahwa *Naïve Bayes* merupakan algoritma yang akurat dan lebih dominan daripada metode lainnya. Penelitian oleh [4] membahas mengenai analisis kerja algoritma C4.5 dan *Naïve Bayes* dalam memprediksi keberhasilan sekolah menghadapi UN dengan *dataset* yang diperoleh dari rata-rata hasil UN setiap sekolah di Banda Aceh tahun 2012. Pengujian menggunakan *confusion matrix* yang dijalankan pada *tools* RapidMiner. Hasil penelitian ini menghasilkan algoritma C4.5 memiliki nilai Akurasi sebesar 78,50% sedangkan pada *Naïve Bayes* memiliki nilai Akurasi sebesar 95,50% sehingga dapat dijelaskan bahwa *Naïve Bayes* menjadi algoritma yang akurat dan terbaik berdasarkan hasil analisis kerja antara algoritma C4.5 dan *Naïve Bayes*. Penelitian lainnya yang dilakukan oleh [5] membahas tentang perbandingan Akurasi algoritma C4.5 dan *Naïve Bayes* untuk deteksi gangguan autisme pada anak. *Dataset* diperoleh dari observasi dengan mendatangi lembaga yang menangani anak berkebutuhan khusus atau autisme di Bekasi. Hasil menunjukkan algoritma C4.5 memiliki Akurasi sebesar 72% dan *Naïve Bayes* memiliki Akurasi sebesar 73,33%. Walaupun terlihat tidak signifikan, *Naïve Bayes* masih menjadi algoritma yang memiliki Akurasi yang lebih baik dibandingkan dengan algoritma C4.5.

Seperti permasalahan dan hasil dari penelitian yang dijelaskan sebelumnya, maka pada penelitian ini akan melakukan analisis terhadap siswa yang melakukan tunggakan iuran sekolah dengan teknik klasifikasi pada *data mining* menggunakan algoritma *Naïve Bayes* di sekolah tersebut. Penelitian ini akan menerapkan algoritma *Naïve Bayes* dalam membentuk suatu model yang akan menghasilkan sebuah pengetahuan dan informasi terkait kondisi siswa yang dikategorikan tidak akan melakukan tunggakan. Bentuk pengetahuan yang dihasilkan dari penelitian ini akan diterapkan ke dalam sebuah sistem berbasis *website*. Proses pengubahan bentuk pengetahuan yang diterapkan ke dalam bentuk sistem berbasis *website* bertujuan untuk mempermudah pihak sekolah tersebut dalam mengetahui hasil dari penelitian yang dilakukan. Bentuk pengetahuan yang ditampilkan ke dalam sistem berbasis *website* juga berguna karena sistem dapat digunakan oleh pihak staff pengurus sehingga dapat mempermudah pihak sekolah dalam menyelesaikan pemilihan calon siswa terbaik yang dikategorikan ke dalam siswa yang tidak akan menunggak. Selain itu, informasi dapat bermanfaat dan memudahkan pihak sekolah dalam membuat strategi yang tepat bagi para siswa yang akan melakukan tunggakan.

2. METODE PENELITIAN

Metodologi penelitian yang digunakan pada penelitian ini yaitu metodologi *Knowledge Discovery in Database* (*KDD*) [6]. Tahapan-tahapan yang dilakukan akan dijelaskan sebagai berikut:

2.1 *Database*

Tahap ini sebagai awal dari proses *KDD*. Tahap ini dimulai dengan persiapan data dengan cara pengumpulan data siswa pada salah satu sekolah di Karawang dan memahami permasalahan tunggakan iuran sekolah yang dilakukan oleh siswa sekolah tersebut. Pengumpulan data pun dilakukan di sekolah tersebut dengan kriteria data seperti data informasi siswa dan data tunggakan siswa selama periode tahun akademik 2017/2018, 2018/2019 dan 2019/2020.

2.2 *Data Cleaning*

Tahap ini akan dilakukan pembersihan kepada semua data yang telah diperoleh dari tahap sebelumnya. Pembersihan ini dilakukan dengan cara membersihkan/menghilangkan data-data yang dianggap ambigu, seperti *missing value*, data yang inkonsisten dan duplikasi data. Pembersihan dilakukan pada data informasi siswa dan data tunggakan siswa di salah satu sekolah di Karawang selama periode tahun akademik 2017/2018, 2018/2019 dan 2019/2020.

2.3 *Data Integration*

Tahap *data integration* melakukan penggabungan data yang awalnya terpisah dalam beberapa *resource* atau beberapa bentuk tabel menjadi satu tabel saja. Tahap ini akan dilakukan penggabungan data pada data yang sudah diperoleh sebelumnya, yakni data informasi siswa dan data tunggakan siswa yang bersumber dari pihak salah sekolah di Karawang. Selain itu, penggabungan data pun dilakukan dengan pencocokan kembali setiap atribut data yang disesuaikan antara keseluruhan informasi siswa dengan jumlah tunggakan, sehingga terciptanya *dataset* yang baru.

2.4 *Data Selection*

Pada tahap ini akan dilakukan pemilihan terhadap atribut apa saja yang akan digunakan dan dipilih sesuai dengan kebutuhan selama proses penelitian. Tahap ini dilakukan pada *dataset* baru yang berasal dari salah satu sekolah di Karawang tentang data informasi siswa dan data tunggakan siswa dengan kondisi yang sudah digabungkan dan sudah disinkronisasi. Tahap ini nantinya akan memilih atau menyeleksi atribut apa saja pada data yang sudah diperoleh dari sekolah tersebut dan yang sudah digabungkan menjadi *dataset* yang baru. Teknik penyeleksian atribut pada tahap ini berdasarkan kebutuhan yang disesuaikan dan referensi-referensi yang digunakan, tanpa ada teknik khusus ataupun perhitungan lainnya.

2.5 *Data Transformation*

Selanjutnya, untuk tahap *data transformation* akan dilakukan perubahan pada data yang disesuaikan dengan kebutuhan teknik atau metode. Tahap ini dilakukan pada *dataset* baru yang berasal dari salah satu sekolah di Karawang tentang data informasi siswa dan data tunggakan siswa dengan kondisi sudah digabung dan sudah diseleksi atributnya. Data perlu diubah sebab pada penelitian ini algoritma yang digunakan yakni *Naïve Bayes* yang memerlukan data dalam bentuk kategori. Penentuan kategori disesuaikan dengan rentang nilai pada *dataset* dan beberapa referensi yang digunakan sebagai acuan bentuk kategori. Selain perubahan pada data, tahap ini pun melakukan perubahan pada jumlah data karena *dataset* yang digunakan mengalami kondisi yang tidak seimbang (*imbalanced data*). Sehingga, untuk menangani masalah tersebut dilakukan penerapan teknik SMOTE yang berfungsi untuk menambahkan jumlah data dalam bentuk data sintetis dengan besaran yang ditentukan sesuai dengan kebutuhan.

2.6 *Data Mining*

Tahap *data mining* akan melakukan pengolahan dan perhitungan terhadap data sesuai dengan alur algoritma yang diterapkan. Tahap ini akan mengolah dan menerapkan metode *Naïve Bayes* dengan cara menghitung nilai probabilitas di setiap atribut dan untuk hasil akhirnya akan dihitung secara keseluruhan. Selain itu, penelitian ini akan melakukan skenario menggunakan teknik *percentage split* dalam pembagian data latih dan data uji dengan komposisi 90:10, 80:20, 70:30, 60:40 dan 50:50. Pemodelan *Naïve Bayes* ini dibantu menggunakan *tools* WEKA versi 3.9.4.

2.7 *Pattern Evaluation*

Tahap mengevaluasi dan menguji model tersebut untuk mengetahui seberapa efektif dan akurat model yang sudah dihasilkan. Tahap evaluasi yang digunakan yaitu menggunakan *confusion matrix* dengan parameter yang dipilih yakni nilai Akurasi (*accuracy*), Presisi (*precision*), *Recall* dan *F-Measure* di setiap skenario yang telah dilakukan. Hasil akhir di antara semua

skenario akan dibandingkan untuk melihat komposisi mana yang memiliki nilai-nilai parameter yang paling tinggi. Selain itu, hasilnya ditampilkan dalam bentuk tabel dan grafik secara rinci agar memudahkan pembaca dalam memahami hasil dari pemodelan. Pada tahap ini pun dilakukan perbandingan hasil antara perhitungan model terbaik menggunakan *tools* dengan perhitungan secara manual.

2.7 *Knowledge Presentation*

Tahap terakhir pada metodologi *KDD* ini yaitu penyajian pengetahuan dan informasi yang sudah didapatkan pada tahap *data mining*. Penelitian ini akan menyajikan seluruh proses dan pengetahuan akhir dalam bentuk laporan penelitian tertulis dan terstruktur agar dapat dipahami oleh para pembaca sehingga dapat digunakan sebagai referensi pada penelitian selanjutnya. Selain itu, bentuk pengetahuan dan informasi yang didapat dari hasil pemodelan akan diterapkan ke dalam sistem berbasis *website*. Sistem akan dibangun tanpa ada penjelasan tentang perancangan desain yang rinci dan detail, sehingga sistem berbasis *website* ini hanya sebagai produk hasil akhir yang telah diterapkan bentuk pengetahuan yang diperoleh dari proses *data mining* dan dapat digunakan oleh staff pengurus sekolah tersebut.

3. HASIL DAN PEMBAHASAN

3.1 *Database*

Data yang sudah dikumpulkan dari pihak salah satu sekolah di Karawang dengan proses perizinan yang sesuai dengan prosedur penelitian akan diproses pada tahap pertama ini. *Dataset* yang diperoleh sebanyak tiga *dataset* yang isinya dibedakan berdasarkan tahun akademik, yakni *dataset* tahun akademik 2017/2018, *dataset* tahun akademik 2018/2019 dan *dataset* tahun akademik 2019/2020. Ketiga *dataset* ini memiliki komponen atribut yang sama seperti 'Nama', 'Alamat', 'Jenis Kelamin', 'Kelas', Identitas Orang Tua yang terdiri dari 'Tahun Lahir', 'Pendidikan Terakhir', 'Pekerjaan' dan 'Penghasilan', serta 'Jumlah Tunggakan'. Jumlah *record data* dari ketiga *dataset* sebanyak 1,274 siswa dengan rincian jumlah siswa tahun akademik 2017/2018 sebanyak 431 siswa, jumlah siswa tahun akademik 2018/2019 sebanyak 421 siswa dan jumlah siswa tahun akademik 2019/2020 sebanyak 422 siswa.

3.2 *Data Cleaning*

3.2.1 *Penanganan Data Duplikasi*

Pada penelitian ini, *dataset* diperoleh berdasarkan tahun akademik sekolah tersebut. Ada tiga masa tahun akademik, yaitu tahun akademik 2017/2018, tahun akademik 2018/2019 dan tahun akademik 2019/2020. Dikarenakan berdasarkan tahun akademik dan waktunya dalam tiga tahun akademik terakhir, maka ada beberapa siswa yang sama dalam tiga tahun akademik (*dataset*) ini. Semisal, X merupakan siswa kelas tahun pertamanya di sekolah tersebut yang baru saja masuk mengikuti kegiatan sekolah pada tahun akademik 2017/2018. Lalu, memasuki tahun akademik 2018/2019, X naik kelas di sekolah tersebut dan seterusnya hingga tahun akademik 2019/2020, X sudah berada di tahun akhirnya di sekolah tersebut. Dari analogi tersebut, maka ada beberapa siswa yang sama di setiap *dataset* sehingga kondisi ini dinamakan dengan duplikasi data. Cara penanganan duplikasi ini yakni dengan menggunakan kondisi data siswa yang terbaru pada setiap tahun akademiknya. Semisal X yang sudah berada di tahun akhir (tahun akademik 2019/2020) akan digunakan datanya karena kondisi X merupakan data yang terbaru, sedangkan data X pada tahun akademik 2017/2018 dan 2018/2019 akan dibuang/dihapus. Setelah dilakukan pembuangan terhadap data duplikasi, jumlah *record data* dari ketiga *dataset* telah berkurang menjadi 152 data untuk tahun akademik 2017/2018, 142 data untuk tahun akademik 2018/2019 dan 422 data untuk tahun akademik 2019/2020.

3.2.2 *Penanganan Missing Value*

Pada *dataset* yang digunakan pada penelitian ini, diketahui terdapat *missing value* pada beberapa nilai di atribut *dataset* ketika pengecekan terhadap semua baris data.

Tabel 1. Jumlah Data *Missing Value*

Dataset	*Valid*	*Missing*
Dataset TA 2017/2018	49	103
Dataset TA 2018/2019	45	97
Dataset TA 2019/2020	290	132
Jumlah	384	332

Pada penelitian ini, penanganan *missing value* menggunakan *Listwise Deletion* atau menghapus baris data yang memiliki *missing value* seperti yang dilakukan oleh [7]`. Untuk data yang valid/lengkap, total datanya menjadi 384 data yang terbagi

menjadi 49 data untuk *dataset* tahun akademik 2017/2018, 45 data untuk *dataset* tahun akademik 2018/2019 dan 290 data untuk *dataset* tahun akademik 2019/2020.

3.2.3 *Penanganan Inkonsistensi Data*

Diketahui, *dataset* yang digunakan pada penelitian ini memiliki data yang inkonsisten. Data tersebut terletak pada nilai yang ada pada atribut 'Penghasilan Bapak' dan 'Penghasilan Ibu'.

Tabel 2. Data Yang Inkonsisten

Range Penghasilan	Frekuensi
Kurang dari Rp. 1,000,000	1
Rp. 500,000 - Rp. 999,999	3
Rp. 500,000 - Rp. 999,999	2
Rp. 1,000,000 - Rp. 1,999,999	12

Diketahui bahwa terdapat nilai yang berbeda namun memiliki makna yang sama, seperti pada nilai *range* 'Penghasilan' senilai 'Kurang dari Rp. 1,0000,000' yang memiliki makna serupa pada nilai *range* 'Rp. 500,000-Rp. 999,999' dan makna serupa pada nilai *range* 'Rp. 1,000,000-Rp. 1,999,999' dengan *range* yang bernilai 'Rp. 1,000,000-Rp. 2,000,000'. Masing-masing nilai *range* yang digunakan yakni nilai *range* 'Rp. 500,000-Rp. 999,999' dan nilai *range* 'Rp. 1,000,000-Rp. 1,999,999'.

3.3 *Data Integration*

Pada tahap data *integration* ini akan dilakukan teknis penggabungan semua *dataset* yang digunakan pada penelitian ini. Penggabungan dilakukan dari *dataset* tahun akademik 2017/2018, *dataset* tahun akademik 2018/2019 dan *dataset* tahun akademik 2019/2020 sehingga total *record* data pada *dataset* yang baru ini berjumlah sebanyak 384 data.

3.4 *Data Selection*

Tahap ini merupakan seleksi atau pemilihan atribut yang akan digunakan untuk melakukan proses pengolahan pada tahap selanjutnya. *Dataset* ini memiliki 13 atribut yaitu 'Nama', 'Jenis Kelamin', 'Kelas', 'Alamat', 'Tahun Lahir Bapak', 'Tahun Lahir Ibu', 'Pendidikan Bapak', 'Pendidikan Ibu', 'Pekerjaan Bapak', 'Pekerjaan Ibu', 'Penghasilan Bapak', 'Penghasilan Ibu' dan 'Jumlah Tunggakan'. Selanjutnya, dari semua atribut ini akan dipilih hanya beberapa saja yaitu 'Alamat', 'Pendidikan Bapak', 'Pendidikan Ibu', 'Pekerjaan Bapak', 'Pekerjaan Ibu', 'Penghasilan Bapak', 'Penghasilan Ibu' dan 'Jumlah Tunggakan'.

3.5 *Data Transformation*

Pada penelitian ini, jenis atribut pada *dataset* yang digunakan sudah dalam bentuk kategorik semua, kecuali pada atribut 'Jumlah Tunggakan' sebagai *class target*. Maka dari itu, atribut 'Jumlah Tunggakan' akan ditransformasikan ke dalam bentuk kategorik.

Tabel 3. Transformasi pada *Class Target*

Jumlah Tunggakan	→	Keterangan
Tunggakan = 0 (nol)	→	Tidak
Tunggakan > 0 (nol)	→	Ya

Sebelumnya, *class target* atau atribut ini bernama 'Jumlah Tunggakan' dengan berisikan nilai jumlah tunggakan yang dilakukan oleh siswa dari salah satu sekolah di Karawang tersebut. Karena penelitian ini berfokus kepada siswa yang melakukan tunggakan, maka besaran jumlah tunggakan tidak akan berpengaruh. Selanjutnya, diketahui *dataset* ini mengalami kondisi *imbalaced data* antara nilai 'Ya' dan nilai 'Tidak' pada *class target*. Untuk menangani kendala tersebut, maka akan dilakukan sebuah teknik SMOTE (*Synthetic Minority Over-sampling Technique*) pada *dataset* ini [8]. Dengan penerapan teknik tersebut, jumlah awal dari *dataset* adalah 384 data yang terdiri dari 364 data 'Tidak' dan 20 data 'Ya' menjadi sebanyak 564 data yang terdiri dari 364 data 'Tidak' dan 200 data 'Ya'. Data dibangkitkan sebesar 900% yang dihitung menggunakan WEKA.

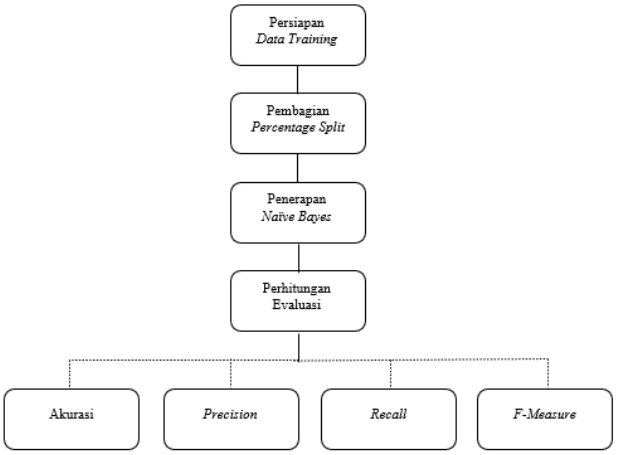
3.6 *Data Mining*

Pada tahap ini, seluruh data akan diolah untuk menentukan sebuah pola/pengetahuan model dengan menerapkan algoritma klasifikasi yang sudah ditentukan, yaitu menggunakan algoritma *Naïve Bayes*. Klasifikasi *Naïve Bayes* adalah teknik klasifikasi yang menggunakan Teorema *Bayes* dengan tujuan untuk menghitung nilai probabilitas tiap atribut dengan asumsi

bahwa antar tiap atribut dengan atribut lain tidak saling tergantung (independen) [9]. Klasifikasi *Naïve Bayes* menggunakan persamaan sebagai berikut:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (1)$$

Dengan,
X = Sampel data yang memiliki kelas (label)
H = Hipotesis X pada data kelas (label)
P(H) = Peluang hipotesis H (*prior*)
P(X) = Peluang sampel data yang memiliki kelas (label)
P(X|H) = Peluang data sampel X akan memengaruhi hipotesis H (*likelihood*)
P(H|X) = Suatu hipotesis H dengan peluang akhir bersyarat (*conditional probability*)



Gambar 1. Skenario Tahapan *Data Mining*

Penerapan *Naïve Bayes* dilakukan menggunakan *tools* WEKA dengan pembagian data menggunakan *Percentage Split*. Rasio pembagian data tersebut sebesar 90:10, 80:20, 70:30, 60:40 dan 50:50 [10]. *Dataset* yang digunakan pada penelitian ini berjumlah 564 data dan sudah melewati beberapa tahapan *data mining* sebelumnya. *Dataset* akan diklasifikasikan menggunakan algoritma *Naïve Bayes* untuk menghasilkan pemodelan.

3.7 *Pattern Evaluation*

Setelah melewati tahap *data mining*, selanjutnya yakni mengumpulkan semua hasil pemodelan dari seluruh skenario yang sudah dilakukan. Skenario yang dimaksud adalah skenario dalam membagi *data training* dan *data testing* menggunakan teknik *percentage split* pada WEKA. Parameter evaluasi yang dilakukan pada penelitian ini menggunakan *Confusion Matrix* [6] dengan nilai Akurasi, Presisi, *Recall* dan *F-Measure* seperti yang dilakukab oleh [11]. Berikut adalah hasil pemodelan dari beberapa skenario yang dilakukan:

Tabel 4. Hasil Pemodelan *Naïve Bayes* Menggunakan *Percentage Split*

Percentage Split	*Confusion Matrix*			
	Akurasi	**Presisi**	***Recall***	***F-Measure***
90:10	80.537%	0.827	0.804	0.802
80:20	81.416%	0.840	0.814	0.816
70:30	83.432%	0.853	0.834	0.836
60:40	84.513%	0.862	0.845	0.847
50:50	85.461%	0.869	0.855	0.857

Selain itu, pada penelitian ini pun dilakukan perbandingan antara perhitungan secara manual dengan perhitungan pada WEKA yang terpilih sebagai pemodelan terbaik, yakni pemodelan pada pembagian rasio sebesar 50:50. Maka dari itu, hasil dari kedua perbandingan tersebut adalah sebagai berikut:

Tabel 5. Perbandingan Hasil Perhitungan dari Model Terbaik

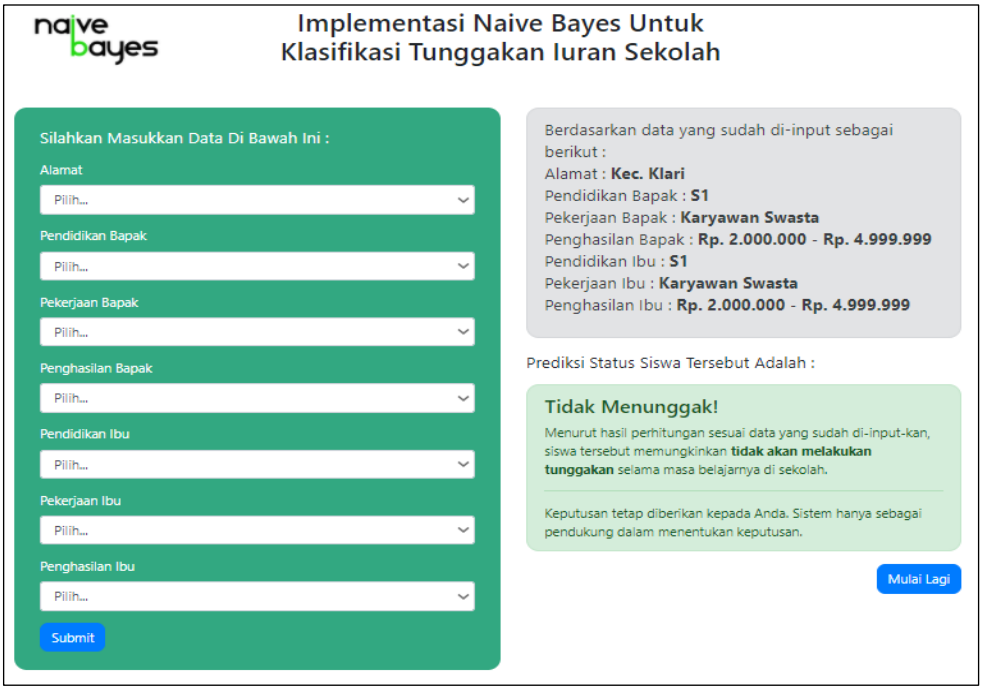
Hasil Perhitungan	*Confusion Matrix*			
	Akurasi	**Presisi**	***Recall***	***F-Measure***
Menggunakan WEKA	85.461%	0.869	0.855	0.857
Perhitungan Manual	81.560%	0.836	0.816	0.819

Hasil pemodelan terbaik masih didapatkan pada perhitungan menggunakan WEKA karena banyak faktor yang memengaruhi dalam proses selama perhitungan berlangsung. Selain itu, faktor lain yang terdapat antara kedua perhitungan tersebut yaitu terletak pada penentuan *data testing*. Pada perhitungan menggunakan WEKA, baik *data training* maupun *data testing* ditentukan dan dipilih secara otomatis oleh WEKA. Sedangkan pada perhitungan secara manual, penentuan *data testing* ditentukan secara acak tanpa menggunakan teknik apapun, sehingga,hasil dari kedua pengujian tersebut tentu tidak akan sama.

3.6 *Knowledge Presentation*

Berdasarkan pada tahap sebelumnya, pemodelan menggunakan algoritma *Naïve Bayes* ini menghasilkan model yang optimal ketika pembagian *data training* dan *data testing* dibagi sebesar 50:50. Hasil tersebut diperoleh dari perhitungan menggunakan *tools* WEKA dengan penggunaan opsi algoritma *Naïve Bayes* secara *default*. Hasil dari pemodelan yang sudah ditetapkan akan diimplementasikan ke dalam bentuk sistem berbasis *website* dalam menentukan siswa yang diidentifikasi akan menunggak di salah satu sekolah di Karawang sebagai hasil dan produk akhir dari penelitian *data mining* ini.

Sistem yang dibangun berbasis *website* dengan tampilan cukup sederhana dan mudah digunakan. Akses kepada *website* ini tidak menggunakan proses *login*, sehingga seluruh staff internal sekolah tersebut dapat menggunakannya tanpa perlu sebuah hak akses. Dasar logika sistem berasal dari perhitungan algoritma *Naïve Bayes* pada tahap *data mining* sebelumnya, yang diperoleh dari WEKA yang menghasilkan nilai probabilitas model yang sudah diprosesnya. Nilai probabilitas tersebut dijadikan sebagai logika perhitungan pada sistem ini. Nilai probabilitas yang digunakan pada sistem ini berasal dari data-data yang sudah dipilih secara otomatis oleh WEKA dan telah dihitung sesuai dengan algoritma *Naïve Bayes*. Sehingga, sistem ini menerapkan hasil perhitungan algoritma *Naïve Bayes* dengan menggunakan nilai probabilitasnya.



Gambar 2. Tampilan Hasil dari *Website*

3.6.1 *Evaluasi Sistem menggunakan Black Box Testing*

Pada penelitian ini dilakukan pengujian pada sistem menggunakan pengujian *Black Box* dengan fokus pengujian pada fungsionalitas sistem berbasis *website* [12]. Pengujian akan dilakukan pada halaman pengisian *form* data siswa untuk klasifikasi tunggakan iuran sekolah di salah satu sekolah di Karawang ini menggunakan metode *BVA* (*Boundary Value Analysis*) dengan skenario yang sudah dibuat sesuai dengan kebutuhan pengujian [13]. Skenario pengujian akan dilakukan seperti di bawah ini:

Tabel 6. Hasil Uji Validasi menggunakan *BVA*

Field	Skenario Uji	Hasil Yang Diharapkan	Hasil Sesungguhnya	Kesimpulan
Alamat	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"Kec. Klari"*)	*True*	*True*	*Success*
Pendidikan Bapak	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"S1"*)	*True*	*True*	*Success*
Pekerjaan Bapak	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"Karyawan Swasta"*)	*True*	*True*	*Success*
Penghasilan Bapak	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"Rp. 2.000.000-Rp. 4.999.999"*)	*True*	*True*	*Success*
Pendidikan Ibu	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"S1"*)	*True*	*True*	*Success*
Pekerjaan Ibu	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"Karyawan Swasta"*)	*True*	*True*	*Success*
Penghasilan Ibu	Menampilkan semua isi nilai	*True*	*True*	*Success*
	Field diisi kosong	*False*	*True*	*Success*
	Field diisi dengan pilihan yang sudah tersedia (*"Rp. 2.000.000-Rp. 4.999.999"*)	*True*	*True*	*Success*

Dari hasil uji validasi pada tabel diatas, maka dapat disimpulkan bahwa fungsionalitas pada sistem yang sudah dibuat telah memenuhi persyaratan fungsional dengan hasil kesimpulan pada setiap skenario uji ditandai dengan keterangan "*Success*". Masing-masing dari setiap *field* diuji dengan tiga skenario uji yang sama karena semua elemen dari setiap *field* berbentuk *dropdown* dan nilai yang dicantumkan sudah tersedia dalam *database* sistem.

4. KESIMPULAN

Klasifikasi terhadap siswa yang melakukan tunggakan iuran sekolah pada salah satu sekolah di Karawang dengan menerapkan algoritma *Naïve Bayes* dilakukan dengan mengolahnya menggunakan metode *KDD* (*Knowledge Discovery in Database*). *Dataset* yang digunakan diperoleh dari pihak sekolah dengan total data sebanyak 564 data berdasarkan tiga tahun akademik terakhir yaitu tahun akademik 2017/2018, tahun akademik 2018/2019 dan tahun akademik 2019/2020. *Dataset* tersebut telah melalui proses 8 tahap pada *data mining* sehingga diperoleh hasil pemodelan yang terbaik. Pemodelan terbaik didapatkan pada pemodelan dengan rasio pembagian 50:50 dengan nilai Akurasi sebesar 85.461%, Presisi sebesar 0,869, *Recall* sebesar 0.855 dan *F-Measure* sebesar 0.857. Jumlah data yang berhasil diklasifikasikan sebanyak 241 data dan jumlah data yang tidak berhasil diklasifikasikan sebanyak 41 data dari total *data testing* sebanyak 282 data. Selanjutnya untuk hasil evaluasi sistem berbasis *website* yang dibangun, diuji menggunakan *Black Box Testing* metode *Boundary Value Analysis* (*BVA*) dengan menerapkan tiga skenario uji pada setiap elemen *field form* sistem. Hasil dari pengujian tersebut menghasilkan bahwa semua elemen *field* pada *form* sistem bekerja dengan baik sesuai dengan persyaratan fungsional yang diharapkan, ditandai dengan keterangan "*Success*" pada setiap kesimpulan skenario uji yang telah dilakukan.

DAFTAR PUSTAKA

[1] A. Neolaka and G. A. A. Neolaka, *Landasan Pendidikan: Dasar Pengenalan Diri Sendiri Menuju Perubahan Hidup*. Depok: Kencana, 2017.

[2] A. P. Wibawa, M. G. A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," in *Prosiding Seminar Ilmu Komputer dan Teknologi Informasi*, 2018, pp. 134–138.

[3] M. E. Lasulika, "Komparasi Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran TV Kabel," *Ilk. J. Ilm.*, vol. 11, no. 1, pp. 11–16, 2019.

[4] Y. Angraini, S. Fauziah, and J. L. Putra, "Analisis Kinerja Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Keberhasilan Sekolah Menghadapi UN," *J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 5, no. 2, pp. 285–290, 2020.

[5] B. Sugara, D. Adidarma, and S. Budilaksono, "Perbandingan Akurasi Algoritma C4.5 dan Naïve Bayes Untuk Deteksi Dini Gangguan Autisme Pada Anak," *J. IKRA-ITH Inform.*, vol. 3, no. 1, pp. 119–128, 2019.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann, 2012.

[7] A. S. Suweleh, D. Susilowati, and Hairani, "Aplikasi Penentuan Penerima Beasiswa Menggunakan Algoritma C4.5," *J. BITe J. Bumigora Inf. Technol.*, vol. 2, no. 1, pp. 12–21, 2020.

[8] N. Sulistyowati and M. Jajuli, "Integrasi Naive Bayes Dengan Teknik Sampling SMOTE Untuk Menangani Data Tidak Seimbang," *J. Nuansa Inform.*, vol. 14, no. 01, pp. 34–37, 2020.

[9] A. S. Putra, "Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes Classification (Studi Kasus Posyandu Ngudi Luhur)," Skripsi. Universitas Sanata Dharma, Yogyakarta, 2018.

[10] J. E. Sembodo, E. B. Setiawan, and Z. K. A. Baizal, "A Framework for Classifying Indonesian News Curator in Twitter," *TELKOMNIKA*, vol. 15, no. 1, pp. 357–364, 2017.

[11] J. H. Jaman, A. R. Sanjaya, and Carudin, "Klasifikasi Jenis Mobil Paling Diminati Di Indonesia Menggunakan Algoritma Naive Bayes," *Fakt. Exacta*, vol. 13, no. 1, pp. 18–25, 2020.

[12] G. W. Setiawan, "Pengujian Perangkat Lunak Menggunakan Metode Black Box Studi Kasus Exelsa Universitas Sanata Dharma," Skripsi. Universitas Sanata Dharma Yogyakarta, 2011.

[13] D. Andriansyah, "Pengujian Kotak Hitam Boundary Value Analysis Pada Sistem Informasi Manajemen Konseling Tugas Akhir," *Indones. J. Netw. Secur.*, vol. 7, no. 1, pp. 13–18, 2018.