JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 7, No. 1, JANUARI, 2022, Pp. 56 – 67
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi

Daru Prasetyawan [(1)*], Rahmadhan Gatra [(2)]
Pusat Teknologi Informasi dan Pangkalan Data (PTIPD), UIN Sunan Kalijaga, Yogyakarta
e-mail : {daru.prasetyawan,rahmadhan.gatra}@uin-suka.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 18 Oktober 2021, direvisi 11 Januari 2022, diterima 12 Januari 2022, dan dipublikasikan 25 Januari 2022.

Abstract

Student academic performance is one measure of success in higher education. Prediction of student academic performance is important because it can help in decision-making. K-Nearest Neighbor (K-NN) algorithm is a method that can be used to predict it. Normalization is needed to scale the attribute value, so the data are in a smaller range than the actual data. Feature selection is used to eliminate irrelevant features. Data cleaning from outliers in the dataset aims to delete data that can affect the classification process. In the classification process, the dataset is divided into a training set by 80% and a validation set by 20% using the cross-validation method. The classification model that is formed is tested using data that is separate from the training data and is evaluated using a confusion matrix. As an evaluation, the K-NN model has 95.85% average accuracy, 95.97% average precision, and 95.84% average recall.

Keywords: Academic Performance, K-NN, Pearson Correlation, Classification, Classification

Abstrak

Prestasi akademik mahasiswa merupakan salah satu ukuran keberhasilan perguruan tinggi. Prediksi prestasi menjadi hal yang penting karena dapat membantu dalam pengambilan keputusan. Algoritma *K-Nearest Neighbor* (K-NN) merupakan algoritma yang dapat digunakan untuk memprediksi prestasi mahasiswa. Normalisasi data diperlukan untuk penguraian nilai atribut sehingga berada dalam kisaran yang lebih kecil dari data sebenarnya. Seleksi fitur digunakan untuk mengeliminasi fitur yang tidak relevan. Pembersihan data dari pencilan di dalam *dataset* bertujuan untuk menghapus data yang dapat mengganggu proses klasifikasi. Proses klasifikasi dilakukan dengan *cross-validation* dengan membagi data menjadi data pelatihan sebesar 80% dan data uji sebesar 20% secara bergantian sebanyak 5 *fold*. Metode *Euclidean*, *Manhattan*, dan *Minkowski* digunakan untuk mengukur jarak di antara dua data. Model klasifikasi yang terbentuk diuji menggunakan data yang terpisah dari data latih dan dievaluasi menggunakan *confusion matrix*. Sebagai hasil evaluasi, diperoleh rata-rata akurasi 95,85%, rata-rata presisi 95,97%, dan rata-rata *recall* 95,84%.

Kata Kunci: Prestasi Akademik, K-NN, Korelasi Pearson, Klasifikasi, Prediksi

1. PENDAHULUAN

Salah satu tujuan perguruan tinggi adalah menciptakan lulusan yang memiliki kemampuan SDM yang unggul. Indeks Prestasi Kumulatif (IPK) merupakan ukuran kemampuan atau prestasi akademik mahasiswa dalam periode tertentu yang dihitung berdasarkan SKS yang telah diambil. Informasi mengenai prediksi prestasi akademik mahasiswa dapat memberi gambaran apakah mahasiswa tersebut akan berhasil memperoleh prestasi akademik yang diharapkan atau tidak. Apabila prestasi akademik seorang mahasiswa dapat diketahui sebelumnya, bahkan pada saat proses seleksi, tentunya dapat membantu perguruan tinggi dalam mengambil keputusan. Pada saat seleksi calon mahasiswa, informasi ini dapat menjadi bahan pertimbangan dalam memutuskan apakah diterima atau tidak. Informasi dan pengetahuan untuk mendukung proses bisnis sebuah perguruan tinggi semakin sangat diperlukan. Ditambah lagi dengan data yang dimiliki sudah sangat besar, tentunya dapat memotivasi untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna dengan melakukan mengekstraksi atau menambang

pengetahuan dari data tesebut. Informasi dan pengetahuan tersebut dapat digali dengan memanfaatkan teknologi informasi, khususnya dibidang kecerdasan buatan.

Kecerdasan buatan (*artificial intelligence*) saat ini telah menjadi perhatian lebih karena sudah sangat berpengaruh dalam kehidupan manusia. Kecerdasan Buatan (AI) adalah suatu mesin, program komputer, dan sistem untuk melakukan fungsi intelektual dan kreatif dari seseorang yang secara mandiri menemukan cara untuk memecahkan masalah, mampu menarik kesimpulan dan mengambil keputusan (Rupesh & Choudaiah, 2019). Tujuan utama kecerdasan buatan adalah untuk membuat mesin menjadi lebih pintar sehingga mesin menjadi lebih bermanfaat.

Pembelajaran mesin merupakan cabang kecerdasan buatan berdasarkan ide bahwa sistem dapat belajar dari data, mengidentifikasi pola, dan membuat keputusan dengan intervensi manusia yang minimal. Tujuan dari pembelajaran mesin adalah untuk belajar dari data (Dey, 2016). Pembelajaran mesin menyediakan kemampuan pada sistem untuk secara otomatis belajar dari pengalaman tanpa diprogram secara eksplisit untuk meningkatkan kemampuannya. Proses pembelajaran dimulai dengan melakukan pengamatan atau sekumpulan data, kemudian melatih mesin (komputer) dengan membangun model pembelajaran mesin menggunakan data dan algoritma tertentu untuk mencari pola dalam data tersebut dan membuat keputusan yang lebih baik di masa depan. Tujuan utamanya adalah untuk memungkinkan sebuah komputer dapat belajar secara otomatis tanpa campur tangan atau bantuan manusia dan menyesuaikan tindakan yang sesuai. Pembelajaran mesin sangat berbeda dengan pemrograman secara tradisional. Pada pemrograman tradisional, program dibuat oleh manusia dengan algoritma tertentu, kemudian diberikan input data dan menghasilkan output berdasarkan aturan-aturan yang ada di dalam algoritma tertentu. Sedangkan pada pembelajaran, data dijalankan pada sebuah mesin untuk melakukan pelatihan, kemudian mesin akan membuat programnya sendiri yang dapat dievaluasi pada saat pengujian. Rekayasa perangkat lunak tradisional menggabungkan data dan aturan yang dibuat manusia untuk menciptakan jawaban atas suatu masalah. Sedangkan pembelajaran mesin menggunakan data dan jawaban untuk menemukan aturan di balik suatu masalah (Chollet, 2017).

Secara umum pembelajaran mesin dibagi menjadi pembelajaran terawasi (*supervised learning*), pembelajaran tak terawasi (*unsupervised learning*), dan pembelajaran penguatan (*reinforcement learning*). Pembelajaran terawasi melibatkan penggunaan model untuk mempelajari pemetaan antara input dan variabel target. Pembelajaran terawasi bertujuan untuk mempelajari pemetaan atau aturan antara serangkaian *input* dan *output*. Pembelajaran terawasi membangun model yang membuat prediksi berdasarkan bukti di hadapan ketidakpastian. Tujuan pembelajaran terawasi adalah untuk membangun model dari distribusi label kelas dalam hal fitur prediktor. Model yang dihasilkan kemudian digunakan untuk menetapkan label kelas dari data yang belum diketahui label kelasnya (Reddy & Babu, 2018).

Klasifikasi merupakan pembalajaran terawasi yang memprediksi label dari suatu kelas. Klasifikasi digunakan untuk menemukan model yang menjelaskan atau membedakan kelas data yang dapat memperkirakan kelas dari suatu data yang kelasnya belum diketahui. Klasifikasi adalah proses menemukan kumpulan pola atau fungsi-fungsi yang mendeskripsikan dan memisahkan kelas data satu dengan lainnya, dapat digunakan untuk memprediksi data yang belum memiliki kelas data tertentu. Klasifikasi adalah proses menggunakan model untuk memprediksi nilai yang tidak diketahui (variabel *output*), menggunakan sejumlah nilai yang diketahui (variabel *input*) (Muhammad & Yan, 2015). Algoritma klasifikasi yang sering digunakan di dalam *data mining* antara lain pohon keputusan (*decision tree*), K-Nearest Neighbors (K-NN), dan jaringan syaraf tiruan (*artificial neural network*).

Algoritma K-Nearest Neighbors (K-NN) merupakan algoritma yang sederhana dan mudah diterapkan dalam pembelajaran mesin yang dapat digunakan untuk memecahkan masalah klasifikasi dan regresi. K-NN merupakan algoritma pembelajaran mesin terawasi (*supervised machine learning*), yaitu algoritma yang mengandalkan data input berlabel untuk mempelajari fungsi yang menghasilkan output yang sesuai ketika diberi data baru tanpa label. K-Nearest

Neighbor (KNN) adalah metode yang diterapkan dalam mengklasifikasikan objek berdasarkan data pembelajaran yang paling dekat dengan objek berdasarkan perbandingan antara data sebelumnya dan saat ini (Lubis et al., 2020).

Salah satu metode *data mining* adalah metode klasifikasi dengan *decission tree*. Metode ini dapat digunakan untuk memprediksi prestasi akademik mahasiswa dengan menggunakan algoritma C4.5. Dengan metode klasifikasi ini, hasil pendidikan masa lampau merupakan variabel yang paling menentukan berhasil atau tidaknya seseorang dalam prestasi (Sabna & Muhardi, 2016). Penelitian tentang prediksi prestasi akademik mahasiswa dengan judul "Prediksi Prestasi Akademik Mahasiswa menggunakan Algoritma Random Forest dan C4.5" menggunakan data jenjang pendidikan, program studi, asal daerah, jenis kelamin, SKS semester sebelumnya, dan IP semester sebelumnya, serta membagi kelas prediksi menjadi 3, yaitu IPK Tinggi, IPK Sedang, dan IPK Rendah. Penelitian tersebut menghasilkan akurasi sebesar 87,1% menggunakan algoritma C4.5 dan 92,4% menggunakan algoritma Random Forest (Linawati et al., 2020). Penelitian dengan judul "*Data Mining* untuk Memprediksi Prestasi Siswa Berdasarkan Sosial ekonomi, Motivasi, Kedisiplinan, dan Prestasi Masa Lalu" bertujuan membuat prediksi prestasi belajar siswa berdasarkan status sosial ekonomi orang tua, motivasi, kedisiplinan siswa, dan prestasi masa lalu menggunakan metode *data mining* dengan algoritma J48 yang menghasilkan akurasi sebesar 95,7% (Susanto dan Sudiyatno, 2014). Algoritma K-NN juga telah digunakan dalam memprediksi tingkat kelulusan siswa dengan menggunakan 104 data siswa kelas VI di sekolah swasta area Bekasi (Purwaningsih & Nurelasari, 2021). Penelitian ini menggunakan Nilai PAS, Nilai US Teori, Nilai US Praktik, Nilai UTS, Nilai Prilaku Siswa sebagai atribut dan menghasilkan akurasi sebesar 96,49%. Penelitian berjudul "Algoritma *K-Nearest Neighbor Classification* Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa" juga memanfaatkan Algoritma K-NN untuk memprediksi prestasi mahasiswa (Mustakim & Oktaviani, 2016). Penelitian tersebut menggunakan data Mahasiswa Program Studi Sistem Informasi UIN Sultan Syarif Kasim Riau dengan atribut Jenis Kelamin, Jenis Tinggal, Umur, Jumlah Satuan Kredit Semester (SKS), dan Jumlah Nilai Mutu (NM). Penelitian tersebut membagi menjadi 4 kelas prediksi (Pujian, Sangat Memuaskan, Memuaskan, dan Kurang Memuaskan) dan menghasilkan akurasi sebesar 82%.

Pada Penelitian ini akan dikembangkan sebuah model *data mining* menggunkan algoritma K-Nearest Neigbor (K-NN). Alasannya karena K-NN memiliki kelebihan lebih efektif didata *training* yang besar dan dapat menghasilkan data yang lebih akurat. Selain itu, penelitian ini akan menggunakan data mengenai latar belakang pendidikan dan ekonomi calon mahasiswa pada saat proses pendaftaran dan seleksi penerimaan mahasiswa baru, sehingga informasi yang dihasilkan juga dapat dimanfaatkan dalam proses seleksi tersebut.

2. METODE PENELITIAN

2.1 Data dan Sumber Data

Penelitian ini menggunakan data profil mahasiswa yang terkait dengan latar belakang calon mahasiswa. Untuk mengumpulkan data, penelitian ini menggunakan metode studi pustaka dan observasi. Metode observasi merupakan metode pengumpulan data dengan melakukan pengamatan langsung pada objek yang akan diteliti. Calon mahasiswa baru diminta untuk mengisikan data ekonomi dan pendidikan pada aplikasi yang disediakan.

Dalam penelitian pada suatu populasi, biasanya menggunakan sampel untuk mewakili populasi tersebut dikarenakan akan memakan waktu yang lama dan biaya yang besar apabila menggunakan populasi keseluruhan. Agar sampel yang digunakan dapat mewakili populasinya, diperlukan suatu standar atau teknik dalam mengambil sampel tersebut. Terdapat banyak teknik yang sering digunakan untuk menentukan jumlah sampel dalam suatu populasi. Salah satu teknik teknik pengambilan sampel yang sering digunakan adalah dengan rumus Slovin. Formula rumus Slovin dapat dilihat pada Pers. (1).

$$s = \frac{N}{1+N.e^2} \quad (1)$$

Di mana s adalah sumlah sampel yang akan dihitung, N adalah jumlah populasi, dan e adalah batas toleransi *error*.

Penelitian ini mengambil sampel dari populasi mahasiswa S1 di UIN Sunan Kalijaga yang berjumlah kurang lebih 18.000 mahasiswa. Dengan menggunakan rumus Slovin, diperoleh jumlah sampel minimal yang harus digunakan dengan tingkat toleransi kesalahan 3% adalah 1.047.

2.2 Analisis Data

Tahap ini juga akan memastikan integritas data sehingga tidak menimbulkan masalah pada proses pelatihan. Data latar belakang mahasiswa yang digunakan sejumlah 1.834 data yang tebagi menjadi 4 kelas, yaitu pujian, sangat memuaskan, memuaskan, dan kurang memuaskan. Pada tahap ini dilakukan seleksi data dengan menyeleksi atribut apa saja yang diperlukan. Dalam *dataset* terdapat 6 atribut, yaitu atribut nilai SMA, nilai akreditasi sekolah, penghasilan keluarga, hutang keluarga, jumlah anggota keluarga, dan Indeks Harga Konsumen (IHK) daerah.

Dari Tabel 1 diketahui bahwa jumlah data keseluruhan yang digunakan adalah 1.050 data. Atribut nilai SMA memiliki nilai minimum 45 dan maksimum 90. Atribut nilai akreditasi memiliki nilai minimum 69 dan maksimum 98,99. Atribut penghasilan keluarga memiliki nilai minimum 500.000 dan maksimum 20.000.000. Atribut hutang memiliki nilai minimum 0 dan maksimum 20.000.000. Atribut anggota keluarga memiliki nilai minimum 2 dan nilai maksimum 10. Atribut IHK memiliki nilai minimum 126,45 dan nilai maksimum 140,66. Nilai minimum dan maksimum pada setiap atribut diperlukan dalam proses normalisasi data.

Tabel 1 Deskripsi Data Latar Belakang Pendidikan dan Ekonomi

	Nilai SMA	Nilai Akreditasi	Penghasilan	Hutang	Anggota Keluarga	IHK
count	1050	1050	1050	1050	1050	1050
mean	70,12	90,49	4278310	4168571	4.93	131,91
std	8,03	5,28	3747454	6091702	1.52	3,61
min	45	69	500000	0	2	126,45
25%	65	88	1500000	0	4	129,13
50%	70	91,01	3000000	0	5	130,76
75%	75	94	5500000	8000000	6	133,27
max	90	98,99	20000000	20000000	10	140,66

2.3 Normalisasi Data

Normalisasi adalah proses menyamakan rentang nilai pada setiap atribut dengan skala tertentu (Nasution et al., 2019). Normalisasi diperlukan ketika ada perbedaan besar di dalam dalam rentang fitur yang berbeda. Normalisasi data di dalam paper ini menggunakan teknik *min-max normalization*, yaitu dengan mentransformasikan data ke dalam *range* antara 0 dan 1. Teknik yang menjaga hubungan antara data asli disebut *min-max normalization*. Normalisasi *min-max* adalah teknik sederhana di mana teknik tersebut dapat sesuai dengan batas data yang telah ditentukan sebelumnya (Patro & Sahu, 2015). *Min-max normalization* sering dikenal dengan penskalaan numerik fitur data. Untuk menghitung nilai normalisasi dari anggota dari himpunan nilai-nilai x yang diamati, digunakan formula seperti pada Pers. (2).

$$y = \frac{x-min}{max-min} \quad (2)$$

Di mana *y* adalah data hasil mormalisasi, *x* adalah data yang akan dinormalisasi, *min* adalah nilai data terkecil, dan *max* adalah nilai data terbesar. Dengan persamaan tersebut berarti nilai terkecil

dari suatu variabel ditransformasikan menjadi 0 dan nilai terbesar ditransformasikan menjadi 1, sehingga akan menghasilkan *range* antara 0 sampai dengan 1.

2.4 Seleksi Fitur

Dataset biasanya memiliki variabel atau fitur yang tidak relevan, sehingga dapat mempengaruhi akurasi klasifikasi. Oleh karena itu fitur-fitur tersebut harus dikeluarkan dari *dataset*. Selain itu, semakin sedikit fitur yang digunakan akan mengakibatkan proses klasifikasi menjadi lebih cepat. Pemilihan fitur, sebagai strategi pra-pemrosesan data, telah terbukti efektif dan efisien dalam menyiapkan data berdimensi tinggi untuk masalah *data mining* dan pembelajaran mesin (Li et al., 2018). Penelitian ini menggunakan *correlation-matrix* dalam proses seleksi fitur. *Correlation-matrix* adalah sebuah tabel yang menunjukan koefisian korelasi antar variabel. Analisis korelasi digunakan untuk mengetahui kekuatan antara hubungan korelasi kedua variabel di mana variabel lainnya dianggap berpengaruh dikendalikan atau dibuat tetap (sebagai variabel kontrol) (Romadloni & Hilman F Pardede, 2019). Korelasi antar variabel dihitung dengan *Pearson Correlation* seperti pada Pers. (3).

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \qquad (3)$$

Di mana r_{xy} adalah korelasi antara *x* dan *y,* x_i adalah nilai variabel *x* ke-i, $\bar{x}$ adalah rata-rata pada variabel *x,* y_i adalah nilai variabel *y* ke-*i,* dan $\bar{y}$ adalah rata-rata pada variabel *y*.

Pers. (3) digunakan untuk menghitung korelasi antar fitur. Fitur-fitur yang akan digunakan dalam klasifikasi akan dihitung dengan variabel target, yaitu predikat kelulusan. Fitur-fitur yang memiliki nilai korelasi dibawah ambang batas yang ditentukan akan dikeluarkan dari *dataset*. Dalam penelitian ini menggunakan nilai ambang batas 0,4 dalam skala 0 sampai 1, sehingga fitur-fitur yang memiliki nilai korelasi dengan variabel target dibawah nilai ambang batas tidak digunakan dalam proses klasifikasi.

2.5 Pembersihan Data

Salah satu hal yang penting dalam tahap pra pemrosesan data adalah memastikan bahwa data tersebut bersih dari data yang menyimpang sangat jauh dari data lainnya atau yang sering disebut dengan pencilan (*outlier*). Dalam statistika, pencilan adalah titik pengamatan yang jauh dari pengamatan lain. Pencilan dalam suatu kumpulan data merupakan tanda bahwa terjadi ketidaknormalan atau kesalahan pengukuran dalam pengambilan data. Hal ini dapat terjadi karena kesalahan dalam pencatatan atau ketidaktelitian dalam pengumpulan data. Pencilan dalam kumpulan data harus dihilangkan karena dapat mempengaruhi hasil penelitian. Penghitungan *z-score* merupakan salah satu cara yang digunakan untuk mendeteksi pencilan dalam kumpulan data. *Z-score* menggambarkan posisi skor mentah dalam hal jarak dari rata-rata, ketika diukur dalam satuan standar deviasi. Metode *z-score* merupakan teknik yang merepresentasikan perilaku sebuah data dalam kaitannya dengan standar deviasi dan rata-rata dari kumpulan argumen/data (Anusha et al., 2019). *Z-score* diformulasikan seperti pada Pers. (4).

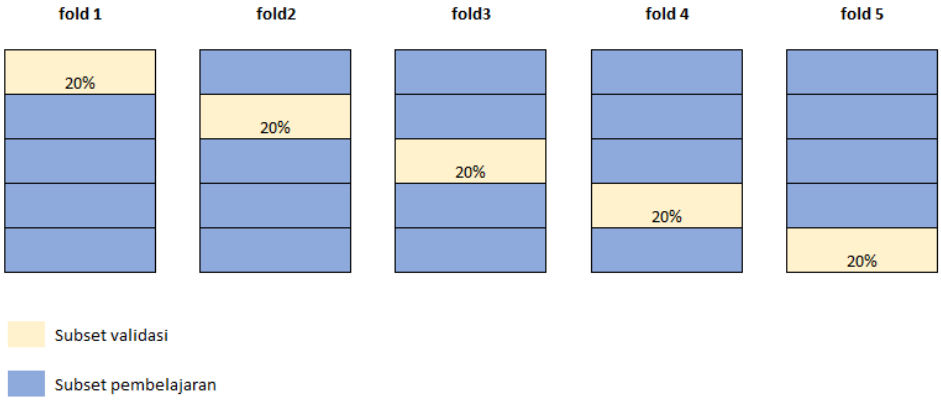
$$z = \frac{x-\bar{x}}{\sigma} \qquad (4)$$

Di mana Z adalah korelasi antara x dan y, x adalah nilai yang akan diukur, $\bar{x}$ adalah rata-rata pada variabel x, dan σ adalah standar deviasi.

Pers. (4) digunakan untuk menghitung nilai z. Data dengan nilai z di atas ambang batas akan dikeluarkan dari *dataset*. Nilai ambang batas yang digunakan dalam penelitian ini adalah 0,4 dari skala 0 sampai 1.

2.6 Klasifikasi

Sebelum melakukan klasifikasi, *dataset* perlu dibagi menjadi data latih dan data uji. Penelitian ini mengambil 20% data sebagai data uji. Dalam menentukan data latih dan data uji, penelitian ini memanfaatkan metode *cross-validation* (CV). *Cross-validation* adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dengan memisahkan data menjadi dua *subset* yaitu *subset* pembelajaran dan *subset* validasi/uji. Model atau algoritma dilatih oleh *subset* pembelajaran dan divalidasi oleh *subset* validasi. *Dataset* akan dibagi menjadi 5 partisi, 4 partisi sebagai subset pembelajaran, dan sisanya sebagai validasi. Proses pembagian antara *subset* pembelajaran dan *subset* validasi akan terus dilakukan sehingga semua data akan berperan sebagai *subset* pembelajaran dan *subset* validasi secara bergantian.



Gambar 1 Pembagian Data Pembelajaran dan Validasi

Penelitian ini akan mencoba menggunakan metode *Euclidean*, *Manhattan*, dan *Minkowski* dalam melakukan perhitungan jarak. Metode dengan akurasi terbaik yang akan digunakan.

2.7 Evaluasi dan Pengujian

Di dalam model klasifikasi, kinerja model yang dihasilkan menggambarkan sejauh mana model tersebut dapat mengklasifikasikan suatu data ke dalam kelas-kelas tertentu. Salah satu metode yang dapat digunakan untuk mengukur kinerja model tersebut adalah *confusion matrix*. Berdasarkan jumlah kelasnya, sistem klasifikasi terbagi menjadi *binary classification* yang hanya mempuanya 2 kelas, dan *multi-class classification* yang memiliki lebih dari 2 kelas). Akurasi merupakan suatu cara yang biasa digunakan untuk mengukur kinerja sistem klasifikasi. Perhitungan akurasi bertujuan untuk memperkirakan seberapa efektif algoritma tersebut dengan menunjukkan probabilitas nilai sebenarnya (*actual*) dan keseluruhan label kelas. dengan kata lain akurasi menilai keefektifan algoritma secara keseluruhan (Sokolova & Lapalme, 2009). Akurasi didefinisikan melalui Pers. (5).

$$Average\ Accuracy\ = \frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fp_i+fn_i+tn_i}}{l} \qquad (5)$$

Di mana *tp* (*true positive*) merupakan jumlah data positif yang diklasifikasikan benar, *tn* (*true negative*) merupakan jumlah data negatif yang diklasifikasikan benar, *fp* (*false positive*) merupakan jumlah data positif yang diklasifikasikan salah, dan *fn* (*false negative*) merupakan jumlah data negatif yang diklasifikasikan salah.

Selain akurasi, presisi dan *recall* juga sering digunakan untuk mengukur kinerja model klasifikasi. *Precision* atau *positive prediction value* merupakan tingkat ketepatan sistem klasifikasi dalam memberikan nilai suatu prediksi. *Precision* mengacu pada persentase hasil klasifikasi yang relevan, sedangkan *recall* mengacu pada persentase total hasil yang relevan yang diklasifikasikan dengan tepat oleh suatu algoritma. *Precision* dihitung dengan membagi jumlah

data positif yang terklasifikasi benar dengan jumlah keseluruhan data yang positif, seperti pada Pers. (6) (Sokolova & Lapalme, 2009).

$$Precision = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp+fp)} \quad (6)$$

Recall dapat didefinisikan sebagai rasio dari jumlah total sampel positif yang terklasifikasi benar dibagi dengan jumlah total sampel positif, sebagaimana dalam Pers. (7).

$$Recall = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp+fp)} \quad (7)$$

3. HASIL DAN PEMBAHASAN

3.1 Normalisasi

Min-max normalization digunakan untuk normalisasi data dengan mentransformasikan data ke dalam *range* antara 0 dan 1. Jika diketahui contoh sampel data x dengan deskripsi:

Nilai SMA (x_1) : 7,33 (*min*: 45; *max*: 90)
Nilai Akreditasi (x_2) : 94 (*min*: 69; *max*: 98,99)
Penghasilan (x_3) : 1000000 (*min*: 500000; *max*: 20000000)
Hutang (x_4) : 0 (*min*: 0; *max*: 20000000)
Anggota keluarga (x_5) : 4 (*min*: 2; *max*: 10)
IHK (x_6) : 130,76 (*min*: 126,45; *max*: 140,66),

dapat dinormalisasikan dengan perhitungan sebagai berikut:

$$y_{(nilai\ SMA)} = \frac{x_{(nilai\ SMA)} - min_{(nilai\ SMA)}}{max_{(nilai\ SMA)} - min_{(nilai\ SMA)}} = \frac{73{,}33-45}{90-45} = 0{,}6295$$

$$y_{(akreditasi)} = \frac{x_{(akreditasi)} - min_{(akreditasi)}}{max_{(akreditasi)} - min_{(akreditasi)}} = \frac{94-69}{98{,}99-69} = 0{,}8336$$

$$y_{(penghasilan)} = \frac{x_{(penghasilan)} - min_{(penghasilan)}}{max_{(penghasilan)} - min_{(penghasilan)}} = \frac{1000000-500000}{20000000-500000} = 0{,}0256$$

$$y_{(hutang)} = \frac{x_{(hutang)} - min_{(hutang)}}{max_{(hutang)} - min_{(hutang)}} = \frac{0-0}{20000000-0} = 0$$

$$y_{(ang,keluarga)} = \frac{x_{(ang,keluarga)} - min_{(ang,keluarga)}}{max_{(ang,keluarga)} - min_{(ang,keluarga)}} = \frac{4-2}{10-2} = 0{,}25$$

$$y_{(ihk)} = \frac{x_{(ihk)} - min_{(ihk)}}{max_{(ihk)} - min_{(ihk)}} = \frac{130{,}76-126{,}45}{140{,}66-126{,}45} = 0{,}3054$$

Hasil dari normalisasi data selanjutnya akan digunakan dalam proses seleksi fitur.

3.2 Seleksi Fitur

Proses seleksi fitur akan mengeluarkan variabel-variabel yang tidak relevan dengan prestasi. Teknik yang digunakan dalam penelitian ini adalah menghitung korelasi linear setiap variabel klasifikasi/fitur dengan variabel target (predikat kelulusan).

$$r_{(nilai\_sma,\ predikat)} = \frac{\sum(nilai\_sma - \overline{nilai\_sma})(predikat_i - \overline{predikat})}{\sqrt{\sum(nilai\_sma_i - \overline{nilai\_sma})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{35{,}0749}{55{,}3118} = 0{,}6341$$

$$r_{(akreditasi,\ predikat)} = \frac{\sum(akreditasi_i - \overline{akreditasi})(predikat_i - \overline{predikat})}{\sqrt{\sum(akreditasi_i - \overline{akreditasi})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{27{,}578640}{54{,}4063} = 0{,}5069$$

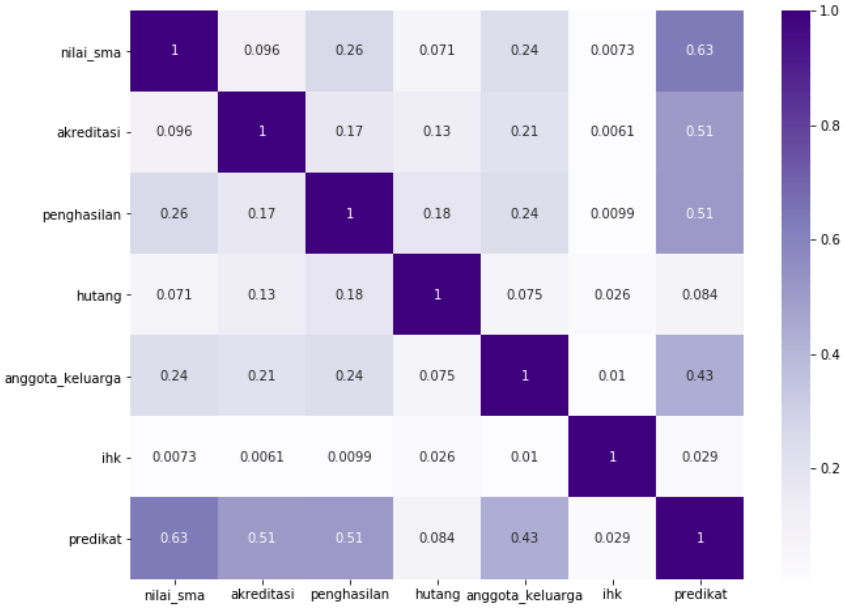
$$r_{(penghasilan,\ predikat)} = \frac{\sum(penghasilan_i - \overline{penghasilan})(predikat_i - \overline{predikat})}{\sqrt{\sum(penghasilan - \overline{penghasilan})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{28{,}5776}{55{,}9087} = 0{,}5111$$

$$r_{(hutang,\ predikat)} = \frac{\sum(hutang_i - \overline{hutang})(predikat_i - \overline{predikat})}{\sqrt{\sum(hutang - \overline{hutang})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{7{,}7944}{93{,}1918} = 0{,}0836$$

$$r_{(ang,keluarga,\ predikat)} = \frac{\sum(ag,keluarga_i - \overline{ag,keluarga})(predikat_i - \overline{predikat})}{\sqrt{\sum(ag,keluarga - \overline{ag,keluarga})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{25{,}2955}{58{,}2462} = 0{,}4343$$

$$r_{(ihk,\ predikat)} = \frac{\sum(ihk_i - \overline{ihk})(predikat_i - \overline{predikat})}{\sqrt{\sum(ihk - \overline{ihk})^2\ \sum(predikat_i - \overline{predikat})^2}} = \frac{2{,}2262}{77{,}7058} = 0{,}0286$$

Variabel target dalam penelitian ini adalah predikat, maka semua variabel dihitung korelasinya dengan variabel predikat. Variabel nilai SMA memiliki nilai korelasi yang paling tinggi, sehingga dapat dikatakan bahwa nilai SMA merupakan variabel yang paling berpengaruh terhadap prestasi mahasiswa. Hasil perhitungan korelasi antar variabel selengkapnya disajikan dalam bentuk matriks korelasi seperti pada Gambar 2.



Gambar 2 Matriks Korelasi Antar Variabel

Dari Gambar 2 diketahui bahwa variabel nilai SMA, akreditasi, penghasilan, dan anggota keluarga berada di atas batas nilai korelasi yang telah ditentukan. Sedangkan variabel hutang dan indeks harga konsumen memiliki nilai korelasi dibawah 0,4. Dengan demikian kedua variabel tersebut tidak digunakan dalam proses pelatihan model, sehingga dalam proses pelatihan model hanya 4 variabel yang akan digunakan, yaitu nilai SMA, akreditasi sekolah, penghasilan keluarga, dan jumlah anggota keluarga.

3.3 Pembersihan Data

Pembersihan data digunakan untuk mengeluarkan data yang menyimpang dari kumpulan data yang akan digunakan dalam proses pelatihan model. Jika diketahui contoh sampel data x (setelah normalisasi) dengan deskripsi:

Nilai SMA (x_1) : 0,63 (*mean*: 0,56; *std*: 0,18)
Nilai Akreditasi (x_2) : 0,83 (*mean*: 0,71; *std*: 0,18)
Penghasilan (x_3) : 0,03 (*mean*: 0,19; *std*:0,18)
Anggota keluarga (x_5) : 0,25 (*mean*: 0,25; *std*: 0,19), perhitungan *z-score* sebagai berikut:

$$Z_{(nilai\ SMA)} = \frac{|x_{(nilai\ SMA)} - mean_{(nilai\ SMA)}|}{std_{(nilai\ SMA)}} = \frac{|0{,}63 - 0{,}56|}{0{,}18} = 0{,}389$$

$$Z_{(akreditasi)} = \frac{|x_{(akreditasi)} - min_{(akreditasi)}|}{std_{(akreditasi)}} = \frac{|0{,}83 - 0{,}71|}{0{,}18} = 0{,}6667$$

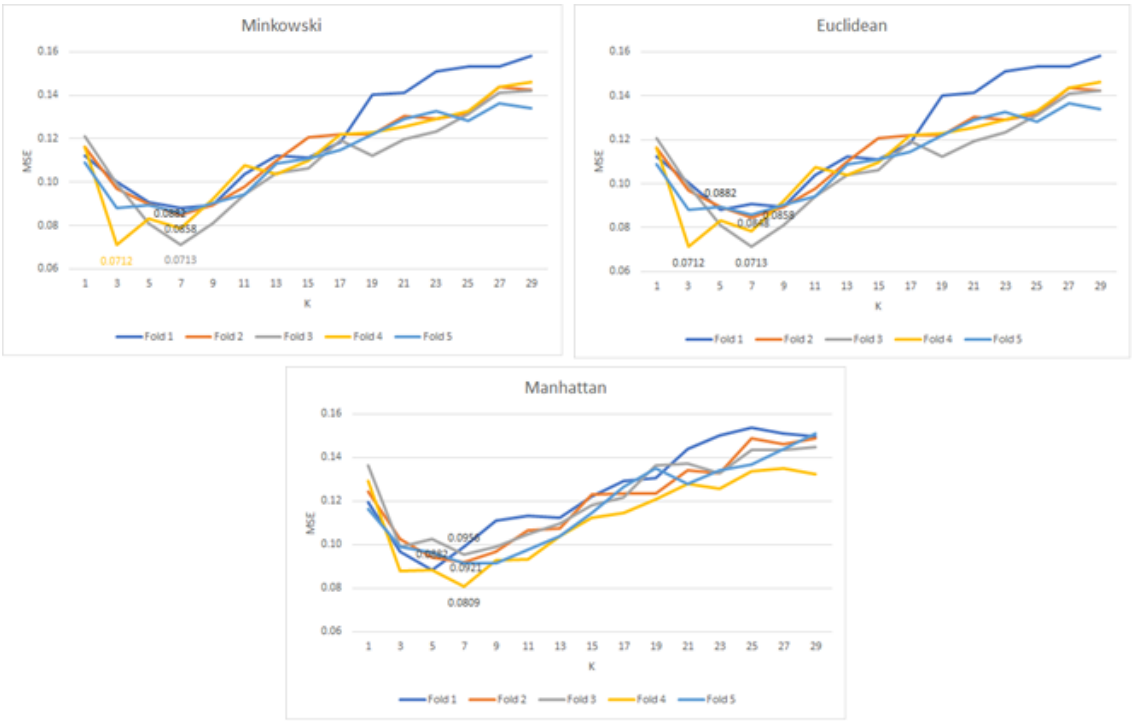
$$Z_{(penghasilan)} = \frac{|x_{(penghasilan)} - min_{(penghasilan)}|}{std_{(penghasilan)}} = \frac{|0{,}03 - 0{,}19|}{0{,}18} = 0{,}8889$$

$$Z_{(ang.keluarga)} = \frac{|x_{(ang.keluarga)} - min_{(ang.keluarga)}|}{std_{(ang.keluarga)}} = \frac{|0{,}25 - 0{,}37|}{0{,}19} = 0{,}6315$$

Dengan menggunakan perhitungan *z-score* diperoleh 14 data dengan nilai *z-score* di luar dari batas yang ditentukan, sehingga data tersebut harus dikeluarkan dari data pelatihan.

3.4 Klasifikasi

Proses klasifikasi dilakukan dengan *cross-validation* dengan membagi data menjadi data pelatihan sebesar 80% dan data uji sebesar 20% secara bergantian sebanyak 5 *fold*. Kemudian dalam proses klasifikasi perhitungan jarak menggunakan metode *Minkowski*, *Euclidean*, dan *Manhattan*. MSE (*Margin Squared Error*) digunakan untuk menentukan nilai k. Jumlah k dengan nilai *error* terendah dipilih untuk melakukan klasifikasi.



Gambar 3 MSE pada Pengukuran Jarak *Minkowski*, *Euclidean*, dan *Manhattan*

Gambar 3 menunjukan grafik nilai MSE untuk setiap *fold* menggunakan metode pengukuran jarak *Minkowski, Euclidean, dan Manhattan*. Pada pengukuran jarak *Minkowski*, nilai MSE terkecil

tarjadi pada *fold-4* dan jumlah k sebanyak 3. Pada pengukuran jarak *Euclidean*, nilai MSE terkecil juga terjadi pada *fold-4* dan jumlah k sebanyak 3. Sedangkan pada pengukuran jarak *Manhattan* nilai MSE terkecil terjadi pada *fold-4* dan jumlah k sebanyak 7.

Tabel 2 Perbandingan Akurasi pada Proses Klasifikasi dengan Perhitungan Jarak *Minkowski, Euclidean, dan Manhattan*

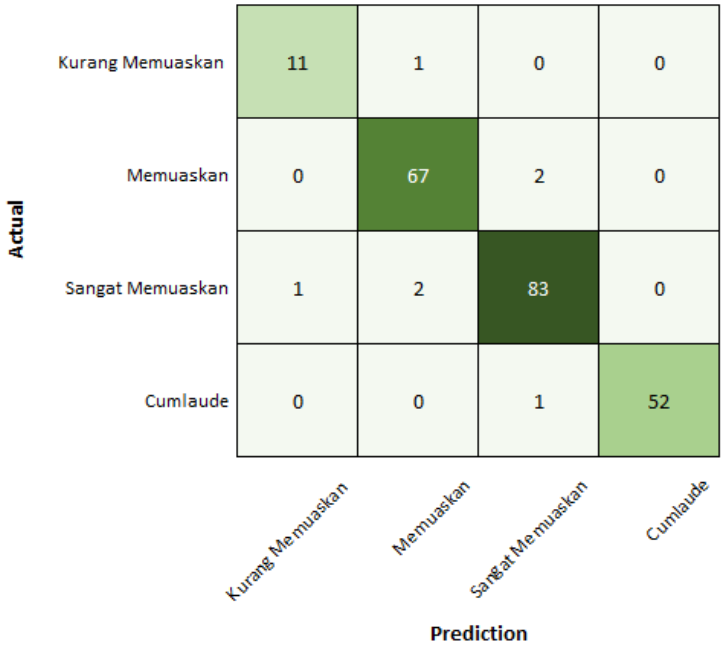
Fold	Minkowski	Euclidean	Manhattan
Fold 1	97,12%	97,12%	94,71%
Fold 2	96,15%	96,15%	96,15%
Fold 3	94,23%	94,23%	95,67%
Fold 4	97,12%	97,12%	96,15%
Fold 5	96,15%	96,15%	95,67%
Rata-rata	**96,15%**	**96,15%**	**95,67%**

Tabel 2 menunjukan perbandingan akurasi saat proses pelatihan menggunakan metode *Minkowski, Euclidean, dan Manhattan.* Akurasi yang sama terjadi pada metode *Minkowski* dan *Euclidean,* yaitu sebesar 96.15%. Sedangkan pada metode *Manhattan,* akurasi yang terjadi lebih rendah dibandingkan *Minkowski* dan *Euclidean*.

3.5 Pengujian dan Evaluasi

Pada tahap pengujian dan evaluasi, sejumlah data uji akan dilakukan klasifikasi dengan menggunakan model klasifikasi yang telah terbentuk. Pengukuran jarak yang digunakan dalam proses klasifikasi adalah metode *Minkowski*, serta jumlah k sebanyak 7. Data uji yang digunakan dalam tahap pengujian pengujian terpisah dari data yang digunakan pada saat pelatihan.

Data uji yang digunakan sebanyak 220 data dengan distribusi data dengan kelas Cumlaude sebanyak 53 data, Sangat Memuaskan sebanyak 86 data, Memuaskan sebanyak 69 data, dan kurang memuaskan sebanyak 12 data. Dalam mengukur kinerja model klasifikasi yang dihasilkan dilakukan dengan *confusion-matrix*. Gambar 6 menunjukan hasil *confusion-matrix* hasil pengujian model klasifikasi.



Gambar 4 *Confusion Matrix* Pengujian Model Klasifikasi

Dari *confusion-matrix* seperti pada Gambar 4 diketahui bahwa pada kelas Cumlaude dengan jumlah sampel sebanyak 53 data terprediksi benar sebanyak 52 data dan hanya 1 data yang terprediksi sebagai kelas Sangat Memuaskan. Pada kelas Sangat Memuaskan dengan jumlah sampel sebanyak 86 terprediksi dengan benar sebanyak 83 data, serta 2 data terprediksi sebagai kelas Memuaskan dan 1 data terprediksi sebagai kelas Kurang memuaskan. Pada kelas Memuaskan dengan jumlah sampel data sebanyak 69 terprediksi secara benar sebanyak 67 data, dan 2 data terprediksi sebagai kelas Sangat memuaskan. Sedangakan pada kelas Kurang Memuaskan dengan jumlah sampel data sebanyak 12 data, 11 data terprediksi dengan benar dan 1 data terprediksi sebagai kelas Memuaskan. Dari *confusion-matrix* tersebut, akurasi model klasifikasi dapat dihitung bersasarkan Pers. (5).

$$Average\ Accuracy = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fp_i + fn_i + tn_i}}{l}$$

$$= \frac{\left(\frac{11}{12}\right)+\left(\frac{67}{69}\right)+\left(\frac{83}{86}\right)+\left(\frac{52}{53}\right)}{4}(100\%) = \frac{0.9167+0.9710+0.9651+0.9811}{4}(100\%) = 95{,}85\%$$

TP dan TN di sini adalah sama, yaitu 213 data karena keduanya adalah jumlah dari semua sampel data yang diklasifikasikan benar, terlepas dari kelasnya. FP dengan jumlah 7 merupakan jumlah sampel data yang seharusnya positif tetapi terklasifikasi salah. FN juga dengan jumlah 7 data merupakan jumlah data yang seharusnya bukan anggota dari suatu kelas tetapi terklasifikasi sebagai anggota kelas tersebut. Sehingga akurasi dari model klasifikasi yang terbentuk adalah 96.82 %. Tabel 3 menunjukan kinerja model klasifikasi (presisi, *recall*, dan *f1-score*).

Tabel 3 Kinerja Model Klasifikasi

Kelas	Precision	Recall	F1-Score	Support
Kurang Memuaskan	0,9167	0,9167	0,92	12
Memuaskan	0,9571	0,9710	0,96	69
Sangat Memuaskan	0,9571	0,9651	0,97	86
Cumlaude	1	0,9808	0,99	53
Total/Avg	**0,9597**	**0,9584**	**0,97**	**220**

4. KESIMPULAN

Metode *data mining* khususnya *K-Nearest Neighbor* dapat digunakan dalam memprediksi prestasi mahasiswa. Selain penambahan jumlah data latih, peningkatan kinerja dapat juga dilakukan dengan melakukan pra pemrosesan seperti normalisasi data, seleksi fitur, dan pembersihan pencilan (*outlier*). Latar belakang pendidikan, khususnya nilai SMA merupakan variabel yang paling berpengaruh terhadap prestasi mahasiswa. Kinerja *K-Nearest Neighbor* mencapai akurasi sebesar 95,85%, presisi sebesar 95,97%, dan recall sebesar 95,84% dalam melakukan prediksi prestasi mahasiswa berdasarkan latar belakang pendidikan dan ekonomi. Prediksi prestasi mahasiswa dapat dipengaruhi oleh banyak variabel, sehingga perlu dikembangkan model klasifikasi untuk melakukan prediksi prestasi mahasiswa dengan mempertimbangkan variabel-variabel lain.

DAFTAR PUSTAKA

Anusha, P. V., Anuradha, C., Murty, P. S. R. C., & Kiran, C. S. (2019). Detecting Outliers in High Dimensional Data Sets using Z-Score Methodology. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 48–53. https://doi.org/10.35940/ijitee.A3910.119119

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.

Dey, A. (2016). Machine Learning Algorithms : A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature

Selection. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Linawati, S., Nurdiani, S., Handayani, K., & Latifah. (2020). *Prediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma Random Forest dan C4.5*. *8*(1), 6–13. https://doi.org/10.31294/jki.v8i1.7827

Lubis, A. R., Lubis, M., & Khowarizmi, A.-. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Muhammad, I., & Yan, Z. (2015). Supervised Machine Learning Approaches: A Survey. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Mustakim, M., & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi, Dan Industri*, *13*(2), 195–202. https://doi.org/10.24014/sitekin.v13i2.1688

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, *4*(1), 78. https://doi.org/10.24114/cess.v4i1.11458

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. *IARJSET*, 20–22. https://doi.org/10.17148/IARJSET.2015.2305

Purwaningsih, E., & Nurelasari, E. (2021). Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa. *Syntax: Jurnal Informatika*, *10*(01), 46–55.

Reddy, R. V. K., & Babu, U. R. (2018). A Review on Classification Techniques in Machine Learning. *International Journal of Advance Research in Science and Engineering*, *7*(3), 40–47.

Romadloni, N. T., & Hilman F Pardede. (2019). Seleksi Fitur Berbasis Pearson Correlation Untuk Optimasi Opinion Mining Review Pelanggan. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(3), 505–510. https://doi.org/10.29207/resti.v3i3.1189

Rupesh, G., & Choudaiah, S. (2019). Artificial Intelligence and its Role in Near Future. *International Journal of Science Research (IJSR)*, *8*(3), 893–898.

Sabna, E., & Muhardi, M. (2016). Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *2*(2), 41. https://doi.org/10.24014/coreit.v2i2.2392

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Susanto, H., & Sudiyatno, S. (2014). Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu. *Jurnal Pendidikan Vokasi*, *4*(2), 222–231. https://doi.org/10.21831/jpv.v4i2.2547