p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING UNTUK PENELUSURAN MINAT CALON MAHASISWA BARU

Budiman[1], Zatin Niqotaini[2]
[1,2]Sistem Informasi, Fakultas Teknologi dan Informatika
Universitas Informatika dan Bisnis Indonesia
E-mail: *[1]budiman1982@gmail.com, [2]zatinniqotaini@unibi.ac.id

Abstrak

Masa pandemi AMIK HASS mengalami kesulitan dalam menentukan calon mahasiswa baru. Sehingga dalam upaya menarik minat masyarakat Bagian Marketing telah menerapkan beberapa strategi dalam menarik minat calon mahasiswa untuk bergabung menjadi mahasiswa baru. Teknik data mining yang digunakan dalam memprediksi adalah klasifikasi di antaranya *Naïve Bayes*, Decision Tree J48, dan *K-Nearest Neighbor*. Tujuan dari penelitian ini adalah melakukan analisis perbandingan terhadap algoritma klasifikasi data mining dengan menggunakan tools WEKA. Metode yang yang digunakan dalam penelitian ini, menggunakan CRISP-DM. Data set yang digunakan ketiga klasifikasi sebanyak 5934 record dengan *mode test percentage split* yaitu 70% sebanyak 4154 sebagai *data training* dan 30% sebanyak 1780 data sebagai *data testing*. Berdasarkan hasil pengujian terhadap ketiga model klasifikasi maka nilai *accuracy* tertinggi diperoleh pada klasifikasi *Decision Tree J48* yang memperoleh nilai 90,3%. Sedangkan klasifikasi *K-Nearest Neighbor* memiliki *accuracy* lebih rendah yaitu 87,52% dan klasifikasi *Naïve Bayes* memiliki *accuracy* yaitu 87,24%. Perbandingan hasil pengujian AUROC *Decision Tree J48* memiliki nilai tertinggi 0,9654 sedangkan hasil *Naïve Bayes* yaitu 0,9461 dan hasil *K-Nearest Neighbor* yaitu 0,9343. Ketiga klasifikasi memiliki nilai AUC diatas 0,90 masuk ke dalam kategori Excellent Classification.

Kata kunci: Data Mining, Klasifikasi, *Naïve Bayes, Decision Tree J48, K-Nearest Neighbor*

Abstract

During the pandemic period, AMIK HASS faces difficulties to determine new student candidates. In order to attract public interest, the Marketing Department has implemented several strategies to attract prospective students to become new students. The data mining technique used in predicting is a classification that includes Naïve Bayes, J48 Decision Tree, and K-Nearest Neighbor. This study aims to perform a comparative analysis of data mining classification algorithms using WEKA tools. The method used in this study is CRISP-DM. The dataset used by the three classifications is 5.934 records with split mode, the percentage of testing is 70% as much as 4154 as training data and 30% as much as 1780 data as test data. Based on the test results on the three classification models, the highest accuracy value is obtained in the J48 Decision Tree classification, which has a value of 90.3%. While the K-Nearest Neighbor classification has a lower accuracy of 87.52% and the Naïve Bayes classification has an accuracy of 87.24%. The comparison of the AUROC J48 Decision Tree test results has the highest value of 0.9654 while the Naïve Bayes results are 0.9461 and the K-Nearest Neighbor results are 0.9343. The three classifications with ABK scores above 0.90 are included in the excellent classification category.

Keywords: *Data Mining, Classification, Naïve Bayes, J48 Decision Tree, K-Nearest Neighbor*

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

1 PENDAHULUAN

Perkembangan teknologi informasi saat ini semakin pesat, seiiring dengan perkembangan tersebut maka transaksi data meningkat dalam hitungan detik sehingga volume data setiap tahun tumbuh dengan berbagai jenis data yang tersebar. Komputerisasi masyarakat saat ini secara substansial telah meningkatkan kemampuan untuk menghasilkan dan mengumpulkan data dari berbagai sumber untuk dikonsumsi oleh semua kalangan. Bahkan Big Data telah membanjiri hampir setiap aspek kehidupan masyarakat.

Pertumbuhan data yang eksplosif telah tersimpan, sementara data telah menghasilkan kebutuhan yang mendesak untuk teknik baru dan tools otomatis yang dapat membantu dengan cerdas dalam mentransformasikan data berjumlah besar menjadi informasi yang bermanfaat dan pengetahuan. Hal ini menyebabkan berkembangnya ilmu komputer yang disebut data mining dengan berbagai aplikasinya.

Perkembangan data mining saat ini merupakan perangkat teknologi informasi yang sangat kuat digunakan dalam dunia bisnis yang kompetitif, terutama ketika masyarakat memasuki era Big Data. Dalam beberapa tahun terakhir, aplikasi data mining telah menjadi strategi bisnis yang penting bagi kebanyakan perusahaan yang ingin menarik pelanggan baru dan mempertahankan pelanggan yang sudah ada. Sedangkan data mining dalam dunia pendidikan telah banyak diterapkan perguruan tinggi dalam melakukan prediksi terhadap mahasiswa diantaranya mengukur kinerja mahasiswa dalam pembelajaran, kelulusan tepat waktu, drop out, penerimaan mahasiswa baru dan lain-lain. Dampak Covid-19 membawa dampak perubahan yang sangat besar terhadap kehidupan masyarakat, penggunaan teknologi informasi dapat dijadikan solusi untuk mengatasi masalah kehidupan seperti Work From Home dan School From Home.

AMIK HASS merupakan perguruan tinggi swasta di kota, dalam melaksanakan tridharma perguruan tinggi mahasiswa merupakan faktor utama dalam menentukan kelangsungan dan keberadaan sebuah perguruan tinggi, pada masa pandemi AMIK HASS mengalami kesulitan dalam menentukan calon mahasiswa baru. Sehingga dalam upaya menarik minat masyarakat Bagian Marketing telah menerapkan beberapa strategi dalam menarik minat calon mahasiswa untuk bergabung menjadi mahasiswa baru. Penelusuran potensi dan minat merupakan salah satu strategi yang digunakan oleh Bagian Marketing dalam menentukan target siswa SMA/SMK/MA yang akan menjadi calon mahasiswa baru. Melalui data penelusuran potensi dan minat, Bagian Marketing harus memiliki pola atau aturan dalam menentukan potensi calon mahasiswa baru sehingga membutuhkan waktu yang lama.

Penelitian dengan kajian data mining telah banyak dilakukan oleh para peneliti, di bawah ini merupakan penelitian terdahulu yang paling relevan telah dilakukan oleh peneliti dengan kajian klasifikasi data mining.

Penelitian pertama dilakukan oleh Rony Setiawan. Peneliti melakukan penelitian penerapan data mining menggunakan algoritma *K-Means Clustering* untuk menentukan strategi promosi mahasiswa baru. Penelitian ini membahas tentang penerapan data mining, menggunakan algoritma *K-Means Clustering* untuk menghasilkan profil yang memiliki kemiripan atribut yang sama. Metode yang digunakan adalah CRISP- DM dengan melalui proses *business understanding, data understanding, data preparation, modeling, evaluation dan deployment*. Atribut yang digunakan usia, agama, status ujian, kelulusan, registrasi, kelamin, gelombang pendaftaran, gelombang registrasi, Nilai Tes, Jurusan, program studi, sumber informasi, asal kota, lokasi kampus, pekerjaan orang tua, jenis sekolah, pembayaran uang registrasi. Menghasilkan jumlah kluster 4 (k=4) dengan cluster pertama 17007 calon mahasiswa, *cluster* kedua sebanyak 83 calon mahasiswa, *cluster* ketiga sebanyak 12919 calon mahasiswa dan

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

cluster keempat sebanyak 356 calon mahasiswa.[1]

Penelitian kedua dilakukan Norzam Yahya dan Arief Jananto, peneliti melakukan penelitian komparasi kinerja algoritma C4.5 dan *Naïve Bayes* untuk prediksi kegiatan penerimaan mahasiswa baru, didapatkan hasil tertinggi pada pengujian dengan *data training* sebanyak 2133 dan *data testing* sebanyak 533 yaitu algoritma C45 memiliki nilai *accuracy* lebih tinggi sebesar 88.74% dibandingkan algoritma *Naïve Bayes* yang memiliki *accuracy* sebesar 87.24%.[2]

Penelitian ketiga dilakukan Vivek Raj S. N. dan S. K. Manivannan, peneliti melakukan penelitian memprediksi dan mengidentifikasi mahasiswa yang cenderung gagal dalam menyelesaikan ujian. Algoritma Machine Learning yang digunakan adalah *Rep Tree, Jrip, Random Forest, Random Tree, Naïve Bayes*. Hasil yang diperoleh dari semua algoritma klasifikasi hanya *Rep Tree* yang memiliki nilai *precision* dan *f score* yang signifikan. Algoritma *Random Forest* keseluruhan memiliki nilai *accuracy* 66,14% dan *f score* sebesar 0,482 lebih rendah dari 0,5. Algoritma *Naives Bayes* keseluruhan memiliki nilai *accuracy* 62,20% tapi memiliki *f score* hanya 0,429.[3]

Penelitian keempat dilakukan Er. AmlanJyoti Baruah, dkk. Peneliti melakukan penelitian analisis komparasi perbedaan algoritma klasifikasi berdasarkan kinerja akademik mahasiswa menggunakan tools WEKA. Algoritma klasifikasi yang digunakan diantaranya *J48, Random Forest, Rap Tree*, LMT, *Naïve Bayes*, *BayesNet* dan *PART* dengan kuesionar pada mahasiswa tingkat akhir. Berdasarkan hasil percobaan *J48* memiliki *accuracy* 85,81%, *Random Forest* memiliki *accuracy* 100%, *Rap Tree* memiliki *accuracy* 68%, *LMT* memiliki *accuracy* 80%, *Naïve Bayes* memiliki *accuracy* 78%, *BayesNet* memiliki *accuracy* 70%, *PART* memiliki *accuracy* 90%.[4]

Penelitian kelima dilakukan Nidhi, Mukesh Kumar, Nandini Nayar, dan Gaurav Mehta. Peneliti melakukan penelitian prediksi kinerja akademik mahasiswa dengan menggunakan Teknik data mining. Algoritma klasifikasi yang digunakan yaitu *Naïve Bayes, Decision Tree, Random-Forest, JRip* dan *ZeroR*. Hasil yang diperoleh dari penelitian tersebut bahwa algoritma klasifikasi seperti *One Rule, Joint Reserve Intelligence Program* dan *Decision Tree* memiliki tingkat *accuracy* lebih dari 80% untuk memprediksi mahasiswa.[5]

Penelitian keenam dilakukan Sri Widaningsih. Peneliti melakukan penelitian perbandingan metode data mining untuk prediksi nilai dan waktu kelulusan mahasiswa prodi teknik informatika dengan algoritma C4.5, *Naïve Bayes*, *K-Nearest Neighbor* dan *Support Vector Machine*. Variabel-variabel prediktor terdiri dari jenis kelamin dan indeks prestasi dari semester 3 hingga 6. Hasil akhir dari keempat algoritma tersebut diperoleh bahwa algoritma *Naïve Bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu dan IPK $\geq$ 3 dengan nilai *accuracy* (76,79%), *error* (23,17%), dan *Area Under Curve* (0,850).[6]

Penelitian ketujuh dilakukan Azahari, Yulindawati, Dewi Rosita dan Syamsuddin Mallala. Peneliti melakukan penelitian komparasi data mining *Naïve Bayes* dan *Neural Network* memprediksi masa studi mahasiswa S1. Atribut yang digunakan yaitu, umur saat masuk kuliah, klasifikasi kota asal Sekolah Menengah Atas, pekerjaan ayah, program studi, kelas, jumlah saudara, dan Indeks Prestasi Kumulatif (IPK). Hasil penelitian menunjukkan banyaknya kelemahan dari hasil prediksi *Naive Bayes* dikarenakan tingkat *accuracy* kevalidannya tergolong tidak terlalu tinggi. Sedangkan *accuracy* prediksi *Neural Network* adalah 72,58%, sehingga metode alternatif inilah yang lebih baik.[7]

Penelitian kedelapan dilakukan Safitri linawati, Rizky Ade Safitri, Ahmad Rifqy Alfiyan, Witriana Endah Pangesti dan Monikka Nur Winnarto. Peneliti melakukan penelitian perbandingan algoritma klasifikasi *Naive Bayes*

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

dan *Support Vector Machine* pada studi kasus pemberian penerima beasiswa PPA. Berdasarkan hasil komparasi antara algoritma *Naive Bayes* dan *Support Vector Machine* yang dilakukan untuk mengklasifikasikan nilai *accuracy* tertinggi dengan 5 variabel dan jumlah data sebesar 122 dapat disimpulkan bahwa algoritma *Naive Bayes* memiliki tingkat *accuracy* lebih tinggi yaitu 90,90% dibandingkan dengan metode *Support Vector Machine* yaitu 89.25%.[8]

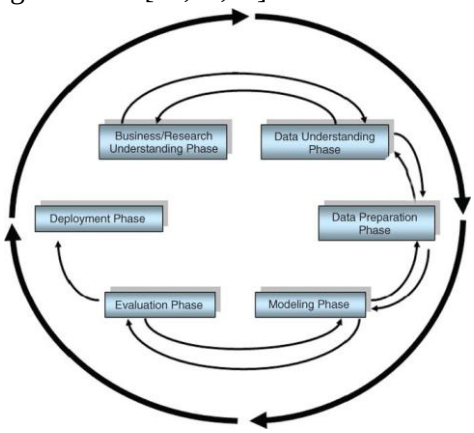
Penelitian kesembilan dilakukan Hermanto, Ali Mustopa dan Antonius Yadi Kuntoro. Peneliti melakukan penelitian algoritma klasifikasi *Naïve Bayes* dan *Support Vector Machine* dalam layanan komplain mahasiswa. Hasil Penelitian diperoleh algoritma *Support Vector Machine* memiliki nilai *accuracy* tertinggi dibanding *Naïve Bayes*. Nilai AUC = 0,922. untuk metode *Support vector Machine* dengan menggunakan *data set* sisfo akademik mahasiswa (students.bsi.ac.id) memiliki 84,45%, dari algoritma *Naïve Bayes* memiliki tingkat *accuracy* sekitar 69.75% dan nilai AUC=0,679.[9]

Penelitian kesepuluh dilakukan Budiman, dkk. Peneliti melakukan penelitian implementasi datamining menggunakan algoritma *Naïve Bayes* dan *Decision Tree J48* dalam menentukan pemilihan konsentrasi. Hasil yang diperoleh terdapat 4 pola atau pattern yang terbentuk untuk menentukan pemilihan konsentrasi, sehingga bagian akademik dapat membantu mahasiswa dalam menentukan pemilihan konsentrasi. Nilai *accuracy* tertinggi diperoleh pada klasifikasi *Naïve Bayes* yang memperoleh nilai 71,4%. Sedangkan klasifikasi *Decision Tree J48* memiliki *accuracy* lebih rendah yaitu 64,3%.[10]

Berdasarkan permasalahan di atas dan penelitian terdahulu maka diperlukan aturan atau pola dalam memprediksi calon mahasiswa baru melalui penelusuran potensi dan minat. Peneliti tertarik untuk melakukan penelitian data mining, adapun teknik data mining yang digunakan dalam memprediksi adalah klasifikasi di antaranya *Naïve Bayes*, *Decision Tree J48*, dan *K-Nearest Neighbor*. Tujuan dari penelitian ini adalah melakukan analisis perbandingan terhadap algoritma klasifikasi data mining dengan menggunakan tools WEKA, sehingga peneliti dapat mengetahui efektifitas algoritma klasifikasi data mining yang terbaik terhadap *data set* penelusuran potensi dan minat calon mahasiswa baru.

2 METODE PENELITIAN

Metode yang dilakukan dalam penelitian merujuk pada metode yang dikemukakan Larose (2015) terlihat pada gambar 1 yaitu data mining memiliki enam fase CRISP-DM (*Cross Industry Standard Process for Data Mining*) sebagai berikut:[11,12,13]



Gambar 1 Metode *Cross Industry Standard Process for Data Mining* atau CRISP-DM

a. Fase Pemahaman Bisnis (*Business Understanding Phase*) akan dilakukan pemahaman tujuan bisnis, menilai situasi, dan menentukan tujuan data mining. Berdasarkan pemahaman tersebut, kemudian menterjemahkan pengetahuan ke dalam pendefinisian masalah data mining. Selanjutnya akan ditentukan strategi untuk mencapai tujuan tersebut.
b. Fase Pemahaman Data (*Data Understanding Phase*) akan dilakukan pengumpulan data awal, mendeskripsikan data, mengeksplorasi data dan memverfikasi kualitas data. Sehingga akan

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

terdeteksi bagian yang menarik dari data yang dapat digunakan untuk menggali informasi yang tersembunyi.

c. Fase Pengolahan Data (*Data Preparation Phase*) akan dilakukan pendeskripsian data set, memilih data, mengintegrasikan data dan membangun data. Proses ini juga mencakup pemilihan tabel, record dan atribut data termasuk pembersihan dan transformasi data. Sehingga akan terbentuk data set akhir sebagai data mentah yang akan diproses pada fase pemodelan.
d. Fase Pemodelan (*Modeling Phase*) akan dilakukan pemilihan teknik pemodelan, membangun model dan menilai model. Saat membangun model dari data set yang digunakan, penelitian ini menggunakan *mode test percentage split* yaitu membagi dua bagian dari data set sebagai *data training* dan *data testing* yang akan diujikan terhadap model yang telah dibentuk.
e. Fase Evaluasi (*Evaluation Phase*) akan dilakukan jika model sudah terbentuk dengan harapan memiliki kualitas yang baik dilihat dari sudut pandang analisa data. Adapun tahapan yang akan dilakukan yaitu evaluasi hasil, meninjau proses dan menentukan langkah selanjutnya. Evaluasi yang dilakukan terhadap keefektifan dan kualitas data model sebelum digunakan dan menentukan apakah model dapat mencapai tujuan berdasarkan fase awal *Business Understanding* yang ditetapkan. Sehingga dapat ditentukan apakah ada masalah bisnis yang belum ditentukan.
f. Fase Penyebaran (*Deployment Phase*), tahap ini merupakan pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna.

3 HASIL DAN PEMBAHASAN

3.1 Fase Pemahaman Bisnis (Business Understanding Phase)

Penelitian ini bertujuan untuk melakukan analisis perbandingan kinerja algoritma klasifikasi diantaranya *Naïve Bayes*, Decision Tree dan *K-Nearest Neighbor* terhadap data set penelusuran potensi dan minat calon mahasiswa baru serta mengidentifikasi pola-pola yang terbentuk dari data masa lampau. Sehingga dalam upaya menarik minat masyarakat Bagian Marketing dapat menerapkan strategi pengambilan keputusan dalam menarik minat calon mahasiswa untuk bergabung menjadi mahasiswa baru pada masa pandemi Covid-19.

Tujuan Bisnis yaitu meningkat jumlah pendaftar calon mahasiswa baru. Penilaian Situasi, AMIK HASS Bandung merupakan perguruan tinggi di Kota Bandung yang memiliki Program Studi Manajemen Informatika terbagi menjadi empat konsentrasi yaitu Komputer Administrasi, Komputerisasi Akuntansi, Multimedia dan E-Commerce. Tujuan Data Mining yaitu menentukan strategi marketing sehingga dapat menenukan profil yang bertujuan untuk mengetahui profil calon mahasiswa baru yang melakukan pendaftaran di AMIK HASS Bandung dan mengetahui strategi marketing berdasarkan pada profil calon mahasiswa baru.

3.2 Fase Pemahaman Data (Data Understanding Phase)

Pada fase pemahaman data terdiri dari beberapa proses diantaranya pengumpulan data awal, deskripsi data, eksplorasi data dan verifikasi kualitas data. Proses pengumpulan data bertujuan untuk mempelajari struktur data pada Bagian Marketing melalui penelusuran potensi dan minat calon mahasiswa baru terlihat pada Tabel 1, selanjutnya menentukan atribut yang akan digunakan dalam proses data mining. Proses deskripsi data bertujuan untuk menganalisis sumber data awal yang diperoleh dari database Bagian Marketing. Sehingga dapat

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

menentukan atribut yang akan diproses dalam data mining.

Tabel 1 Atribut Penelusuran Potensi dan Minat Calon Mahasiswa Baru

No	Atribut	Keterangan
1.	Tanggal Pendaftaran	Tanggal pendaftaran calon mahasiswa baru di AMIK HASS Bandung
2.	Nama	Nama calon mahasiswa baru
3.	Jenis Kelamin	Jenis kelamin calon mahasiswa baru
4.	Tempat Lahir	Tempat tanggal lahir calon mahasiswa baru
5.	Tanggal Lahir	Tanggal lahir calon mahasiswa baru
6.	Alamat	Alamat calon mahasiswa baru
7.	Telepon	Telepon calon mahasiswa baru
8.	Asal Sekolah	Asal sekolah calon mahasiswa baru
9.	Jurusan	Jurusan yang diambil calon mahasiswa baru pada saat sekolah terakhir
10.	Nama Orang Tua	Nama orang tua calon mahasiswa baru
11.	Pekerjaan Orang Tua	Pekerjaan orang tua calon mahasiswa baru
12.	Peminatan	Jurusan yang akan dipilih calon mahasiswa baru setelah lulus dari sekolah
13.	Minat Kuliah	Minat calon mahasiswa baru untuk kuliah di AMIK HASS Bandung

Proses eksplorasi data bertujuan untuk menentukan Sumber data yang digunakan dalam penelitian ini adalah data penelusuran potensi dan minat calon mahasiswa pada tahun 2013-2018 pada Bagian Marketing AMIK HASS Bandung

Proses verifikasi kualitas data bertujuan untuk memastikan data yang digunakan pada proses data mining tidak ada kesalahan, yaitu melalui langkah sebagai berikut: memastikan sumber record terisi dengan data dan tidak ada anomaly data. Jika terjadi data kosong maka sesuai dengan ketentuan record tersebut akan dihapus. Kemudian memastikan tidak ada duplikasi data, jika terdapat duplikasi maka salah satu data akan dihapus.

3.3 Fase Pengolahan Data (Data Preparation Phase)

Pada fase pengolahan data dilakukan pada *data set* penelusuran potensi dan minat calon mahasiswa baru sebanyak 6866 record dengan 13 atribut. Tahap yang dilakukan pada proses meliputi pemilihan data, pembersihan data dan transformasi data. Proses pemilihan data ini dilakukan untuk memilih atribut yang akan digunakan pada pembentukan model algoritma klasifikasi. Proses pembersihan data dilakukan untuk proses cleaning dari data awal sebanyak 6866 record semuanya tidak utuh atau lengkap, dimana data yang tidak lengkap dibuang atau tidak digunakan. Sehingga yang tersisa sebanyak 5934 record. Sedangkan pada proses transformasi data terdapat beberapa atribut yang dapat disederhanakan seperti atribut Alamat menjadi atribut Domisili dan atribut Tanggal Pendaftaran menjadi atribut Status. Kemudian dilakukan konversi nilai dari atribut Alamat terhadap atribut Domisili dengan menentukan wilayah tempat tinggal berdasarkan Kota atau Kabupaten. Konversi nilai atribut tanggal pendaftaran terhadap atribut Status pun dilakukan menjadi Daftar atau Tidak Daftar. Nilai atribut Peminatan awalnya berisi data jurusan yang diminati calon mahasiswa baru dikonversi menjadi beberapa kategori diantaranya Komputer, Non Komputer dan Kerja dalam hal ini tidak melanjutkan untuk kuliah.

Tabel 2 dibawah ini merupakan atribut terpilih beserta potongan record yang akan digunakan sebagai *data set* pengujian terhadap algoritma klasifikasi diantaranya *Naïve Bayes*, Decision Tree dan *K-Nearest Neighbor*.

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

Tabel 2 *Data set* Pengujian

Asal Sekolah	Jenis Kelamin	Domisili	Jurusan	Pekerjaan Orang Tua	Peminatan	Minat Kuliah	Status
MA Hifaz	P	Garut	IPS	Lain-Lain	Komputer	Berminat	Daftar
MAN Ciparay	P	Kabupaten Bandung	IPS	Wiraswasta	Komputer	Berminat	Daftar
MAN Cililin	L	Bandung Barat	IPA	Wiraswasta	Komputer	Berminat	Daftar
MAN Subang	L	Subang	IPA	Wiraswasta	Non Komputer	Berminat	Daftar
SMAN 1 Ciparay	L	Kabupaten Bandung	IPA	PNS	Kerja	Berminat	Daftar
SMAN 1 Serang Panjang	L	Subang	IPA	Karyawan Swasta	Komputer	Berminat	Daftar
SMAN 2 Subang	P	Subang	IPA	Wiraswasta	Komputer	Tidak Berminat	Daftar
SMAN 1 Subang	L	Subang	IPA	Wiraswasta	Komputer	Berminat	Daftar
SMA PGRI Subang	P	Subang	IPS	Karyawan Swasta	Non Komputer	Tidak Berminat	Daftar
SMA PGRI Ciranjang	P	Cianjur	IPA	IRT	Komputer	Berminat	Daftar
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
SMA PGRI Ciranjang	P	Cianjur	IPS	Wiraswasta	Komputer	Berminat	Daftar

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

3.4 Fase Pemodelan (Modeling Phase)

Penelitian berbentuk pengujian terhadap data penelusuran potensi dan minat calon mahasiswa baru memilih yang diambil dari Bagian Marketing yang dituangkan dalam bentuk tabel. Data tersebut akan dilakukan tiga kali percobaan menggunakan *Naïve Bayes*[14,10], *Decision Tree J48*[10,15,16] dan *K-Nearest Neighbor*[17,18] dengan tools pembelajaran mesin "WEKA" dengan menggunakan *mode test percentage split* sebesar 70% sebagai *data training* dan 30% sebagai *data testing*.

Prediksi yang dilakukan dalam penelitian ini adalah menentukan pendaftaran berdasarkan penelusuran potensi dan minat calon mahasiswa baru berikut ini atribut prediksi yang digunakan dalam penelitian ini diantaranya Asal Sekolah, Jenis Kelamin, Jurusan, Pekerjaan Orang Tua, Peminatan, Minat Kuliah dan Status.

3.5 Fase Evaluasi (Evaluation Phase)

Pengujian prediksi dilakukan dengan menggunakan tiga teknik klasifikasi yaitu *Naïve Bayes* dan *Decision Tree J48* dan *K-Nearest Neighbor*. Data set yang digunakan dalam penelitian ini sebanyak 5934 record dengan *mode test percentage split* yaitu 70% atau 4154 data digunakan sebagai *data training* sebagai pembentukan model dan 30% atau 1780 data sebagai *data testing* yang akan diujikan terhadap model yang telah dibentuk.

a. *Naïve Bayes*

Gambar 2 merupakan hasil pengujian klasifikasi *Naïve Bayes* terhadap *data training*, pengujian dilakukan dengan metode yang sama pada *Decision Tree J48* dan *K-Nearest Neighbor*.

```
=== Summary ===

Correctly Classified Instances        1553               87.2472 %
Incorrectly Classified Instances       227               12.7528 %
Kappa statistic                          0.7451
Mean absolute error                      0.1665
Root mean squared error                  0.2953
Relative absolute error                 33.6102 %
Root relative squared error             59.339  %
Total Number of Instances             1780

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.836    0.083    0.924      0.836    0.878      0.749    0.946     0.961     DAFTAR
                 0.917    0.164    0.821      0.917    0.866      0.749    0.946     0.923     TIDAK DAFTAR
Weighted Avg.    0.872    0.120    0.878      0.872    0.873      0.749    0.946     0.944

=== Confusion Matrix ===

   a   b   <-- classified as
 817 160 |   a = DAFTAR
  67 736 |   b = TIDAK DAFTAR
```

Gambar 2 Hasil Pengujian *Naïve Bayes*

Pada Gambar 2 menunjukkan hasil pengujian menggunakan klasifikasi *Naïve Bayes* memiliki tingkat *accuracy* sebesar 87,24% atau 1553 *instance* yang menyatakan rasio prediksi benar dengan keseluruhan *data testing* yang diujikan, sedangkan *mean absolute error* sebesar 0,1665. *Precision* pada kelas Daftar 92,4% menunjukkan bahwa persentase calon mahasiswa baru yang benar melakukan pendaftaran dari keseluruhan calon mahasiswa baru yang diprediksi melakukan pendaftaran. Kelas Tidak Daftar memiliki *Precision* sebesar 82,1% menunjukkan bahwa persentase calon mahasiswa baru yang benar tidak melakukan pendaftaran dari keseluruhan calon mahasiswa baru mahasiswa yang diprediksi tidak melakukan pendaftaran.

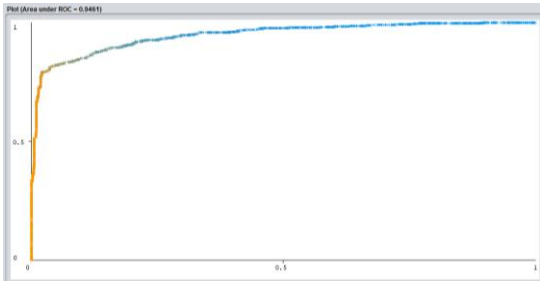
Recall untuk kelas Daftar sebesar 83,6% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya melakukan pendaftaran. Sedangkan *Recall* untuk kelas Tidak daftar sebesar 91,7% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi tidak melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya tidak melakukan pendaftaran.

Confusion Matrix Naïve Bayes, baris pertama terdapat "817 160" menunjukkan bahwa ada *instance class* Daftar dalam *data testing* diantaranya 817 benar di prediksi sebagai Daftar dan 160 salah diklasifikasikan sebagai Tidak Daftar. Pada

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

Baris kedua terdapat "67 736" menunjukkan bahwa ada *instance class* Tidak Daftar dalam *data testing* diantaranya 67 salah diklasifikasi sebagai Daftar dan 736 benar diklasifikasikan Tidak Daftar.

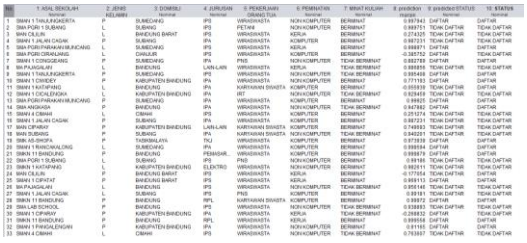


Gambar 3 Area Under ROC

Gambar 3 menunjukkan *Area Under ROC* dihitung untuk mengukur perbedaan performansi. Kurva ROC menunjukkan *accuracy* dan membandingkan klasifikasi secara visual dengan false positive sebagai garis horizontal dan true negative sebagai garis vertical. Berdasarkan hasil pengujian menggunakan pengukuran *Naïve Bayes* didapatkan hasil AUROC 0,9461.

Gambar 4 menunjukkan hasil prediksi terhadap *data testing* sebanyak 1780 record setelah model dibentuk menggunakan klasifikasi *Naïve Bayes*. Sedangkan Waktu yang dibutuhkan dalam test model selama 0,05 detik.

Gambar 4 Hasil Prediksi Klasifikasi *Naïve Bayes*.



No.	1. ASAL SEKOLAH Nominal	2. JENIS KELAMIN Nominal	3. DOMISILI Nominal	4. JURUSAN Nominal	5. PENDIDIKAN ORANG TUA Nominal	6. PEKERJAAN Nominal	7. MINAT KULIAH Nominal	8. prediction Nominal	9. predicted STATUS Nominal	10. STATUS Nominal
1	SMAN 1 TAHULIKAGERTA	P	SUMEDANG	IPS	WIRASWASTA	NON KOMPUTER	BERMINAT	0.997943	DAFTAR	DAFTAR
2	SMA PGRI 1 SUBANG	L	SUBANG	IPS	PETANI	NON KOMPUTER	BERMINAT	0.999771	TIDAK DAFTAR	TIDAK DAFTAR
3	MAN CILEUN	L	BANDUNG BARAT	IPS	WIRASWASTA	KERJA	BERMINAT	0.274325	TIDAK DAFTAR	TIDAK DAFTAR
4	SMAN 1 JALAN CAGAK	P	SUBANG	IPA	WIRASWASTA	KOMPUTER	BERMINAT	0.997311	TIDAK DAFTAR	TIDAK DAFTAR
5	SMA PGRI PARAKAN MUNCANG	L	SUMEDANG	IPS	WIRASWASTA	KERJA	BERMINAT	0.998971	DAFTAR	DAFTAR
6	SMA PGRI CIBRALANG	L	CIANJUR	IPS	WIRASWASTA	KOMPUTER	BERMINAT	-0.387752	DAFTAR	TIDAK DAFTAR
7	SMAN 1 CONGGEANG	P	SUMEDANG	IPA	PNS	NON KOMPUTER	TIDAK BERMINAT	0.992788	DAFTAR	DAFTAR
8	MA PAJAGALAH	L	BANDUNG	LAIN-LAIN	WIRASWASTA	KERJA	TIDAK BERMINAT	0.999816	TIDAK DAFTAR	TIDAK DAFTAR
9	SMAN 1 TAHULIKAGERTA	P	SUMEDANG	IPS	WIRASWASTA	NON KOMPUTER	TIDAK BERMINAT	0.995488	DAFTAR	DAFTAR
10	SMAN 1 CIMEDEK	P	KABUPATEN BANDUNG	IPA	WIRASWASTA	NON KOMPUTER	BERMINAT	0.771183	DAFTAR	DAFTAR
11	SMAN 1 KATAPANG	L	BANDUNG	IPA	KARYAWAN SWASTA	KOMPUTER	BERMINAT	-0.055839	TIDAK DAFTAR	DAFTAR
12	SMAN 1 CIDALEGAWA	L	KABUPATEN BANDUNG	IPA	IRT	NON KOMPUTER	TIDAK BERMINAT	0.929459	TIDAK DAFTAR	TIDAK DAFTAR
13	SMA PGRI PARAKAN MUNCANG	P	SUMEDANG	IPA	WIRASWASTA	KOMPUTER	BERMINAT	0.999825	DAFTAR	DAFTAR
14	SMA MADANA	P	BANDUNG	IPA	WIRASWASTA	NON KOMPUTER	TIDAK BERMINAT	0.997291	DAFTAR	DAFTAR
15	SMAN 1 CIBIRUNG	P	SUBANG	IPA	WIRASWASTA	KOMPUTER	BERMINAT	0.997291	TIDAK DAFTAR	TIDAK DAFTAR
16	SMAN 1 JALAN CAGAK	P	SUBANG	IPA	WIRASWASTA	KOMPUTER	BERMINAT	0.997211	TIDAK DAFTAR	TIDAK DAFTAR
17	MAN CIPARAY	P	KABUPATEN BANDUNG	LAIN-LAIN	KARYAWAN SWASTA	KOMPUTER	BERMINAT	0.749863	TIDAK DAFTAR	TIDAK DAFTAR
18	MAN DUBAIRG	P	SUBANG	IPA	KARYAWAN SWASTA	NON KOMPUTER	TIDAK BERMINAT	0.840201	TIDAK DAFTAR	TIDAK DAFTAR
19	SMK AS-SHOFA	P	TASIKMALAYA	TKJ	WIRASWASTA	KOMPUTER	BERMINAT	0.973818	DAFTAR	DAFTAR
20	SMAN 1 RANCAWALONG	L	SUMEDANG	IPA	WIRASWASTA	KOMPUTER	BERMINAT	0.999594	DAFTAR	DAFTAR
21	SMAN 11 BANDUNG	P	BANDUNG	PEMASAR	WIRASWASTA	KOMPUTER	BERMINAT	0.999879	DAFTAR	DAFTAR
22	SMA PGRI 1 SUBANG	L	SUBANG	IPS	PNS	NON KOMPUTER	BERMINAT	0.99186	TIDAK DAFTAR	TIDAK DAFTAR
23	SMAN 1 KATAPANG	L	KABUPATEN BANDUNG	ELEKTRO	WIRASWASTA	NON KOMPUTER	BERMINAT	0.992911	TIDAK DAFTAR	TIDAK DAFTAR
24	MAN CILEUN	P	BANDUNG BARAT	IPS	WIRASWASTA	KERJA	BERMINAT	-0.177054	TIDAK DAFTAR	DAFTAR
25	SMAN 1 CIPATAT	P	BANDUNG BARAT	IPS	WIRASWASTA	KERJA	BERMINAT	0.999111	DAFTAR	DAFTAR
26	MA PAJAGALAH	L	BANDUNG	IPS	WIRASWASTA	NON KOMPUTER	TIDAK BERMINAT	0.956146	TIDAK DAFTAR	TIDAK DAFTAR
27	SMAN 1 JALAN CAGAK	L	SUBANG	IPS	PNS	KOMPUTER	BERMINAT	0.99181	TIDAK DAFTAR	TIDAK DAFTAR
28	SMKN 11 BANDUNG	P	BANDUNG	IPA	KARYAWAN SWASTA	KOMPUTER	BERMINAT	0.99972	DAFTAR	DAFTAR
29	SMA LAB SCHOOL	P	BANDUNG	IPS	WIRASWASTA	NON KOMPUTER	TIDAK BERMINAT	0.938843	TIDAK DAFTAR	TIDAK DAFTAR
30	SMAN 1 CIPARAY	P	KABUPATEN BANDUNG	IPA	WIRASWASTA	KERJA	TIDAK BERMINAT	-0.268832	DAFTAR	TIDAK DAFTAR
31	SMKN 11 BANDUNG	P	BANDUNG	IPA	WIRASWASTA	KERJA	BERMINAT	0.999594	DAFTAR	DAFTAR
32	SMAN 1 PANGALENGAN	P	KABUPATEN BANDUNG	IPA	WIRASWASTA	NON KOMPUTER	BERMINAT	0.91165	DAFTAR	DAFTAR
33	SMAN 4 CIMAHI	L	CIMAHI	IPS	WIRASWASTA	NON KOMPUTER	TIDAK BERMINAT	0.793357	TIDAK DAFTAR	TIDAK DAFTAR

b. *Decision Tree J48*

Gambar 5 merupakan hasil pengujian klasifikasi *Decision Tree J48* terhadap *data training*, pengujian dilakukan dengan metode yang sama pada *Naïve Bayes* dan K-Nearet Neighbor.

```
=== Summary ===

Correctly Classified Instances        1608               90.3371 %
Incorrectly Classified Instances       172                9.6629 %
Kappa statistic                          0.8056
Mean absolute error                      0.1297
Root mean squared error                  0.2631
Relative absolute error                 26.1894 %
Root relative squared error             52.879  %
Total Number of Instances             1780

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.895    0.086    0.927      0.895    0.910      0.806    0.965     0.974     DAFTAR
                 0.914    0.105    0.877      0.914    0.895      0.806    0.965     0.946     TIDAK DAFTAR
Weighted Avg.    0.903    0.095    0.904      0.903    0.904      0.806    0.965     0.961

=== Confusion Matrix ===

   a   b   <-- classified as
 874 103 |   a = DAFTAR
  69 734 |   b = TIDAK DAFTAR
```

Gambar 5 Hasil Pengujian *Decision Tree J48*

Pada Gambar 5 hasil pengujian menggunakan klasifikasi *Decision Tree J48* memiliki tingkat *accuracy* sebesar 90,33% atau 1608 *instance* yang menyatakan rasio prediksi benar dengan keseluruhan *data testing* yang diujikan, sedangkan *mean absolute error* sebesar 0,1297. *Precision* pada kelas Daftar 92,7% menunjukkan bahwa persentase calon mahasiswa baru yang benar melakukan pendaftaran dari keseluruhan calon mahasiswa baru yang diprediksi melakukan pendaftaran. Kelas Tidak Daftar memiliki *Precision* sebesar 87,7% menunjukkan bahwa persentase calon mahasiswa baru yang benar tidak melakukan pendaftaran dari keseluruhan calon mahasiswa baru mahasiswa yang diprediksi tidak melakukan pendaftaran.

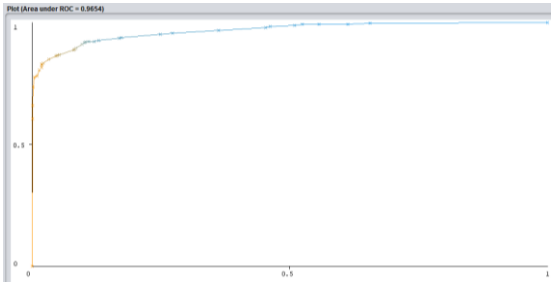
Recall untuk kelas Daftar sebesar 89,5% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya melakukan pendaftaran. Sedangkan *Recall* untuk kelas Tidak daftar sebesar 91,4% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi tidak melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya tidak melakukan pendaftaran.

Confusion Matrix Decision Tree J48, baris pertama terdapat "874 103" menunjukkan bahwa ada *instance class* Daftar dalam *data testing* diantaranya 874 benar di prediksi sebagai Daftar dan 103 salah diklasifikasikan sebagai Tidak Daftar.

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

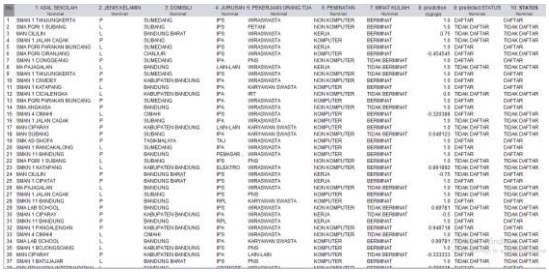
Pada Baris kedua terdapat "69 734" menunjukkan bahwa ada *instance class* Tidak Daftar dalam *data testing* diantaranya 69 salah diklasifikasi sebagai Daftar dan 734 benar diklasifikasikan Tidak Daftar.



Gambar 6 *Area Under ROC Decision Tree J48*

Gambar 6 menunjukkan *Area Under ROC* dihitung untuk mengukur perbedaan performansi. Berdasarkan hasil pengujian menggunakan pengukuran *Decision Tree J48* didapatkan hasil AUROC 0,9654.

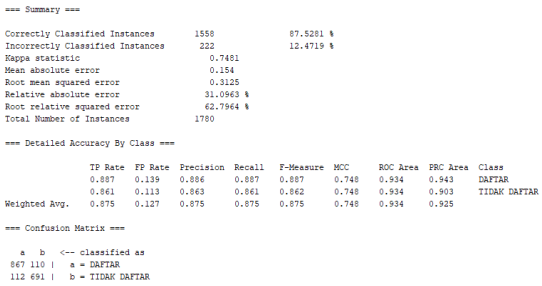
Gambar 7 menunjukkan hasil prediksi terhadap *data testing* sebanyak 1780 record setelah model dibentuk menggunakan klasifikasi *Decision Tree J48*. Sedangkan Waktu yang dibutuhkan dalam test model selama 0,05 detik.



Gambar 7 Hasil Prediksi Klasifikasi *Decision Tree J48*.

c. *K-Nearest Neighbor*

Gambar 8 merupakan hasil pengujian klasifikasi *K-Nearest Neighbor* terhadap *data training*, pengujian dilakukan dengan metode yang sama pada *Naïve Bayes* dan *Decision Tree J48*.



```
=== Summary ===

Correctly Classified Instances        1558               87.5281 %
Incorrectly Classified Instances       222               12.4719 %
Kappa statistic                          0.7481
Mean absolute error                      0.154
Root mean squared error                  0.3125
Relative absolute error                 31.0963 %
Root relative squared error             62.7964 %
Total Number of Instances             1780

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.887    0.139    0.886      0.887    0.887      0.748    0.934     0.943     DAFTAR
                 0.861    0.113    0.863      0.861    0.862      0.748    0.934     0.903     TIDAK DAFTAR
Weighted Avg.    0.875    0.127    0.875      0.875    0.875      0.748    0.934     0.925

=== Confusion Matrix ===

   a   b   <-- classified as
 867 110 |   a = DAFTAR
 112 691 |   b = TIDAK DAFTAR
```

Gambar 8 Hasil Pengujian *K-Nearest Neighbor*

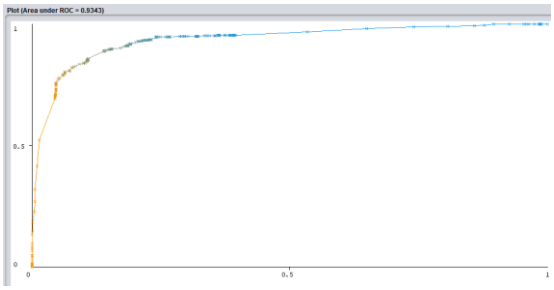
Pada Gambar 8 hasil pengujian menggunakan klasifikasi *K-Nearest Neighbor* memiliki tingkat *accuracy* sebesar 87,52% atau 1558 *instance* yang menyatakan rasio prediksi benar dengan keseluruhan *data testing* yang diujikan, sedangkan *mean absolute error* sebesar 0,154. *Precision* pada kelas Daftar 88,6% menunjukkan bahwa persentase calon mahasiswa baru yang benar melakukan pendaftaran dari keseluruhan calon mahasiswa baru yang diprediksi melakukan pendaftaran. Kelas Tidak Daftar memiliki *Precision* sebesar 86,3% menunjukkan bahwa persentase calon mahasiswa baru yang benar tidak melakukan pendaftaran dari keseluruhan calon mahasiswa baru mahasiswa yang diprediksi tidak melakukan pendaftaran.

Recall untuk kelas Daftar sebesar 88,7% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya melakukan pendaftaran. Sedangkan *Recall* untuk kelas Tidak daftar sebesar 86,1% menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi tidak melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya tidak melakukan pendaftaran.

Confusion Matrix K-Nearest Neighbor, baris pertama terdapat "867 110" menunjukkan bahwa ada *instance class* Daftar dalam *data testing* diantaranya 867 benar di prediksi sebagai Daftar dan 110

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

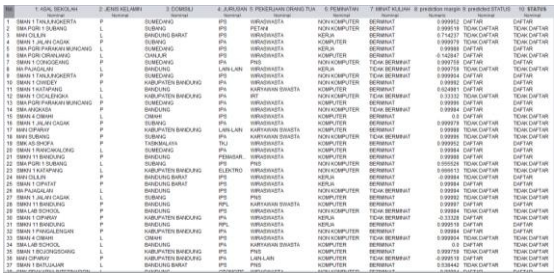
salah diklasifikasikan sebagai Tidak Daftar. Pada Baris kedua terdapat "112 691" menunjukkan bahwa ada *instance class* Tidak Daftar dalam *data testing* diantaranya 112 salah diklasifikasi sebagai Daftar dan 691 benar diklasifikasikan Tidak Daftar.



Gambar 9 *Area Under ROC K-Nearest Neighbor*

Gambar 9 menunjukkan *Area Under ROC* dihitung untuk mengukur perbedaan performansi. Berdasarkan hasil pengujian menggunakan pengukuran *K-Nearest Neighbor* didapatkan hasil AUROC 0,9343.

Gambar 10 menunjukkan hasil prediksi terhadap *data testing* sebanyak 1780 record setelah model dibentuk menggunakan klasifikasi *K-Nearest Neighbor*. Sedangkan Waktu yang dibutuhkan dalam test model selama 1.2 detik.



Gambar 10 Hasil Prediksi Klasifikasi *K-Nearest Neighbor*

3.6 Fase Penyebaran (Deployment Phase)

Perbandingan algoritma klasifikasi pada penelitian ini berdasarkan hasil output pada ketiga teknik klasifikasi yaitu *Naïve Bayes*, *Decision Tree J48* dan *K-Nearest Neighbor* menggunakan tools WEKA dengan mode split test terdiri dari 70% atau 4154 data sebagai *data training* dan 30% atau 1780 data sebagai *data testing* diantaranya *Correctly Classified Instances, Incorrectly Classified Instances, Mean absolute error, Average Precision, Average Recall, F-Measure, Time taken to build model* dan *Time taken to test model on test split*. Tabel 3 menunjukkan perbandingan kinerja pada masing-masing algoritma klasifikasi pada penelusuran potensi dan minat calon mahasiswa baru.

Tabel 3 Perbandingan Kinerja Algoritma Klasifikasi

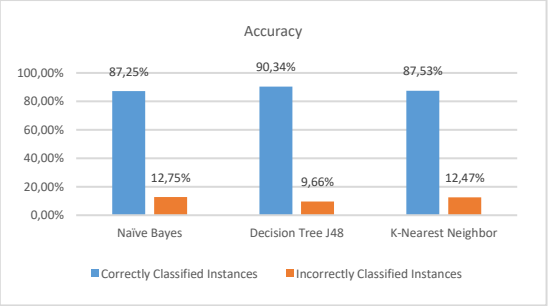
Hasil Pengujian	NB	J48	KNN	Baik
Correctly Classified Instances	87,24 %	90,33 %	87,52 %	J48
Incorrectly Classified Instances	12,76 %	9,67 %	12,48 %	J48
Mean absolute error	0,1665	0,1297	0,154	J48
Average Precision	0,878	0,904	0,875	J48
Average Recall	0,872	0,903	0,875	J48
F-Measure	0,873	0,904	0,875	J48
Time taken to build model	0,02	0,09	0	KNN
Time taken to test	0,03	0,05	1,2	NB

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

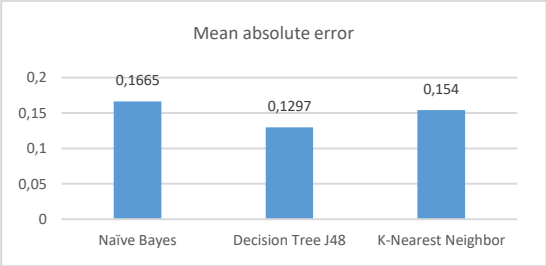
model on test split				
Area Under ROC	0,9461	0,9654	0,9343	J48

Accuracy klasifikasi secara umum dapat ditentukan berdasarkan *Correctly Classified Instances*. Pada gambar 11 menunjukkan perbandingan *accuracy* prediksi pada ketiga algoritma. Berdasarkan gambar 11 *accuracy* prediksi tertinggi adalah *Decision Tree J48* dengan nilai *accuracy* sebesar 90,34% sedangkan *K-Nearest Neighbor* nilai *accuracy* sebesar 87,53 dan *Naïve Bayes* nilai *accuracy* sebesar 87,25%. Hal ini menunjukkan bahwa *Decision Tree J48* merupakan algoritma yang lebih baik dari sisi *accuracy* prediksi dibandingkan *Naïve Bayes* dan *K-Nearest Neighbor* terhadap data set. Semakin tinggi nilai *accuracy* prediksi maka model klasifikasi semakin baik.



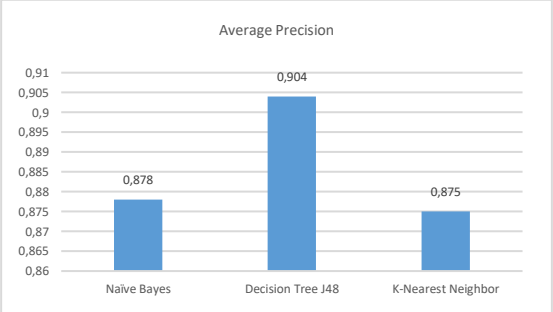
Gambar 11 Grafik Perbandingan *Correctly Classified Instances*.

Pada gambar 12 merupakan grafik perbandingan *mean absolute error* untuk ketiga algoritma. Berdasarkan gambar 12 algoritma *Decision Tree J48* memiliki nilai *mean absolute error* paling rendah yaitu sebesar 0,1297 dibandingkan *K-Nearest Neighbor* yang memiliki nilai sebesar 0,154 dan *Naïve Bayes* yang memiliki nilai sebesar 0,1665. Semakin kecil nilai *mean absolute error* maka model klasifikasi semakin baik.



Gambar 12 Grafik Perbandingan Mean Absolute Error

Pada gambar 13 merupakan grafik perbandingan *Average Precision* pada kelas Daftar dan Tidak Daftar, hal ini menunjukkan bahwa rata-rata persentase calon mahasiswa baru yang benar melakukan pendaftaran dari keseluruhan calon mahasiswa baru yang diprediksi melakukan pendaftaran dan persentase calon mahasiswa baru yang benar tidak melakukan pendaftaran dari keseluruhan calon mahasiswa baru mahasiswa yang diprediksi tidak melakukan pendaftaran. Berdasarkan gambar 13 terlihat bahwa perbandingan rata-rata *Precision* tertinggi pada *Decision Tree J48* sebesar 0,904 sedangkan *Naïve Bayes* sebesar 0,878 dan *K-Nearest Neighbor* sebesar 0,875. Semakin tinggi nilai *Precision* maka model klasifikasi semakin baik.



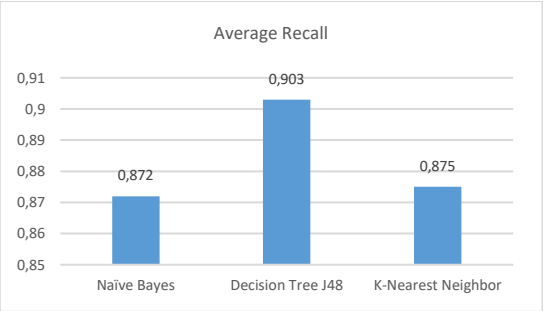
Gambar 13 Grafik Perbandingan *Average Precision*

Pada gambar 15 merupakan grafik perbandingan *Average Recall* untuk kelas Daftar yang menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang

Terakreditasi SINTA 5

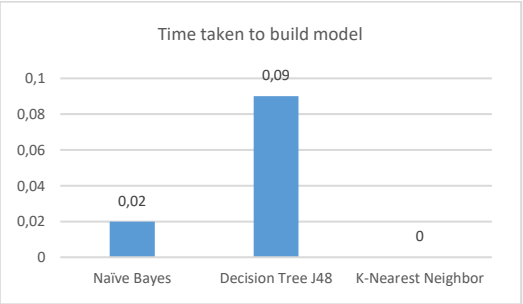
p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

sebenarnya melakukan pendaftaran dan *Recall* untuk kelas Tidak Daftar yang menunjukkan bahwa persentase calon mahasiswa baru yang diprediksi tidak melakukan pendaftaran dibandingkan keseluruhan calon mahasiswa baru yang sebenarnya tidak melakukan pendaftaran. Berdasarkan gambar 15 terlihat perbandingan *Average Recall* nilai tertinggi pada *Decision Tree J48* sebesar 0,903 sedangkan *K-Nearest Neighbor* sebesar 0,875 dan *Naïve Bayes* sebesar 0,872. Semakin tinggi *Average Recall* nilai maka model klasifikasi semakin baik.



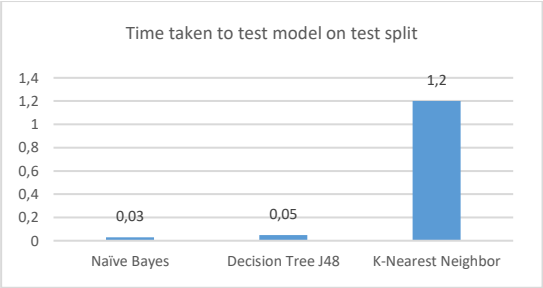
Gambar 14 Grafik Perbandingan *Average Recall*

*P*ada gambar 15 merupakan grafik perbandingan waktu yang dibutuhkan dalam membentuk model pada *data training* dalam satuan detik. Berdasarkan gambar 15 terlihat perbandingan waktu yang dibutuhkan *K-Nearest Neighbor* paling rendah yaitu 0 detik sedangkan *Naïve Bayes* membutuhkan waktu 0,02 dan Decision Tree J48 membutuhkan waktu 0,09 detik. Semakin rendah waktu yang dibutuhkan maka model semakin baik.



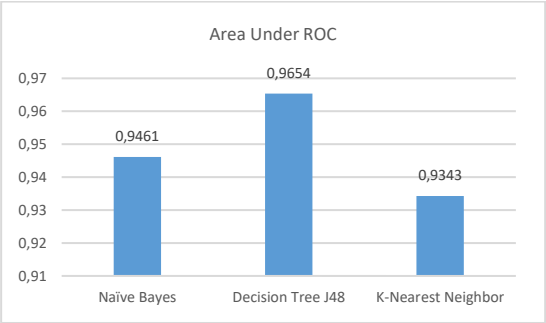
Gambar 15 Grafik Perbandingan *Time taken to build model*

Pada gambar 16 merupakan grafik perbandingan waktu yang dibutuhkan dalam melakukan pengujian *data testing* pada model yang telah terbentuk dalam satuan detik. Berdasarkan gambar 16 terlihat perbandingan waktu yang dibutuhkan dalam pengujian *Naïve Bayes* paling rendah yaitu 0,03 detik sedangkan *Decision Tree J48* membutuhkan waktu 0,05 dan *K-Nearest Neighbor* membutuhkan waktu 1,2 detik. Semakin rendah waktu yang dibutuhkan maka model semakin baik.



Gambar 16 Grafik Perbandingan *Time taken to test model on test split*

Pada gambar 17 merupakan grafik perbandingan *Area Under ROC* dihitung untuk mengukur perbedaan performansi. Berdasarkan gambar 17 terlihat perbandingan hasil pengujian menggunakan pengukuran *Decision Tree J48* tertinggi yaitu 0,9654 sedangkan hasil pengujian menggunakan pengukuran *Naïve Bayes* yaitu 0,9461 sedangkan dan hasil pengujian menggunakan pengukuran *K-Nearest Neighbor* yaitu 0,9343.



Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

Gambar 17 Grafik Perbandingan *Area Under ROC*

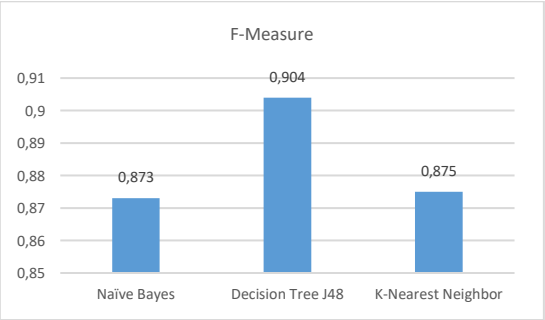
Menurut Gorenescu untuk mengklasifikasikan keakuratan tes diagnosa pengujian menggunakan *Area Under Curve* dapat menggunakan tabel 4 dibawah ini:[19,20]

Tabel 4 Kriteria nilai AUC Nilai

Nilai AUC	Interpretasi
0,90 - 1,00	*excellent classification*
0,80 - 0,90	*good classification*
0,70 - 0,80	*fair classification*
0,60 - 0,70	*poor classification*
0,50 - 0,60	*failure*

Berdasarkan hasil pengujian dan panduan diatas, maka ketiga model masuk ke dalam kategori *Excellent Classification*. Semakin tinggi nilai *Area Under ROC* maka performansi model klasifikasi semakin baik.

Pada Gambar 18 merupakan grafik perbandingan hasil pengukuran *accuracy F-Measure* pada ketiga model klasifikasi. Berdasarkan gambar 18, perbandingan *accuracy F-Measure* tertinggi berada pada *Decision Tree J48* dengan nilai 0,904 sedangkan *K-Nearest Neighbor* sebesar 0,875 dan *Naïve Bayes* sebesar 0,873. Semakin tinggi hasil pengukurun *accuracy F-Measure* maka model klasifikasi semakin baik.



Gambar 18 Grafik Perbandingan *F-Measure*

4 KESIMPULAN

Penelitian ini dilakukan untuk menganalisa perbandingan algoritma klasifikasi data mining untuk penelusuran potensi dan minat calon mahasiswa baru. Berdasarkan hasil pengujian terhadap ketiga model algoritma klasifikasi diantaranya *Naïve Bayes*, *Decision Tree J48* dan *K-Nearest Neighbor* dengan menggunakan data set yang sama maka nilai *accuracy* tertinggi diperoleh pada klasifikasi *Decision Tree J48* yang memperoleh nilai 90,3% terdiri dari 1608 instance yang terklarifikasi benar dari 1780 *data testing*. Sedangkan klasifikasi *K-Nearest Neighbor* memiliki *accuracy* lebih rendah yaitu 87.52% terdiri dari 1558 *instance* yang terklarifikasi benar dari 1780 *data testing* dan klasifikasi *Naïve Bayes* memiliki *accuracy* lebih rendah dari kedua klasifikasi yaitu 87.24% terdiri dari 1553 *instance* yang terklarifikasi benar dari 1780 *data testing*.

Perbandingan hasil pengujian *Area Under Curve ROC*, pengukuran *Decision Tree J48* tertinggi yaitu 0,9654 sedangkan hasil pengujian menggunakan pengukuran *Naïve Bayes* yaitu 0,9461 sedangkan dan hasil pengujian menggunakan pengukuran *K-Nearest Neighbor* yaitu 0,9343. Ketiga klasifikasi memiliki nilai *Area Under Curve* diatas 0,90 masuk ke dalam kategori Excellent Classification. Semakin tinggi nilai *Area Under ROC* maka performansi model klasifikasi semakin baik.

Nilai *mean absolute error* pada klasifikasi *Decision Tree J48* memiliki nilai lebih rendah dibandingkan klasifikasi *Naïve Bayes* dan *K-Nearest Neighbor* yaitu 0,1297. Semakin kecil nilai *mean absolute error* maka model klasifikasi semakin baik. Hasil penelitian yang dilakukan dapat disimpulkan bahwa klasifikasi *Decision Tree J48* dapat digunakan sebagai analisa oleh Bagian Marketing pada masa pandemi Covid-19. Hal ini diperkuat dengan hasil analisis perbandingan pada penelitian ini bahwa *Decision Tree J48* mampu menganalisa penelusuran potensi dan minat calon

Terakreditasi SINTA 5

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

mahasiswa baru yang melakukan pendaftaran dan tidak melakukan pendaftaran sebanyak 90,3%.

5. SARAN

Untuk memperoleh hasil akurasi dan pola yang baik terhadap data set maka diperlukan penelitian lebih lanjut dengan penambahan algoritma klasifikasi yang lainnya sebagai perbandingan seperti *Support Vector Machine, Random Forest* dan lain-lain.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi yang telah memberikan kesempatan untuk melaksanakan penelitian pada skema Penelitian Dasar Pemula tahun pelaksanaan 2021 serta Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Informatika dan Bisnis Indonesia yang telah mendorong dan memfasilitasi setiap pelaksanaan penelitian..

DAFTAR PUSTAKA

[1] R. Setiawan, Aug. 2016, Penerapan Data Mining Menggunakan Algoritma K-Means Clustering untuk Menentukan Strategi Promosi Mahasiswa Baru (Studi Kasus : Politeknik Lp3i Jakarta), *J. LENTERA ICT*, vol. 3, no. 1, pp. 76–92, Accessed: Feb. 25, 2021. [Online]. Available: https://plj.ac.id/ojs/index.php/jrict/article/view/27.

[2] N. Yahya and A. Jananto, 2019, Komparasi Kinerja Algoritma C.45 dan Naive Bayes untuk Prediksi Kegiatan Penerimaan Mahasiswa Baru (Studi Kasus : Universitas STIKUBANK Semarang), in *Prosiding SENDI_U*, pp. 221–228, Accessed: Mar. 08, 2021. [Online]. Available: https://www.unisbank.ac.id/ojs/index.php/sendi_u/article/view/7389.

[3] V. Raj and S. K. Manivannan, 2020, Predicting Student Failure in University Examination using Machine Learning Algorithms, *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 956–959, doi: 10.35940/ijitee.E2643.039520.

[4] E. A. Baruah, S. Baruah, and J. Goswami, 2020, A Comparative Analysis of Different Classification Algorithms based on Students' Academic Performance Using WEKA, *IOSR J. Comput. Eng.*, vol. 22, no. 1, pp. 49–56, doi: 10.9790/0661-2201024956.

[5] Nidhi, M. Kumar, N. Nayar, and G. Mehta, 2020, Student's Academic Performance Prediction in Academic using Data Mining Techniques, in *1st International Conference on Intelligent Communication and Computational Research (ICICCR-2020)*, pp. 1–5, Accessed: Mar. 08, 2021. [Online]. Available: https://ssrn.com/abstract=3565874.

[6] S. Widaningsih, Apr. 2019, Komparasi Kinerja Algoritma C.45 dan Naive Bayes untuk Prediksi Kegiatan Penerimaan Mahasiswa Baru (Studi Kasus : Universitas STIKUBANK Semarang), *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, doi: 10.36787/jti.v13i1.78.

[7] A. Azahari, Y. Yulindawati, D. Rosita, and S. Mallala, May 2020, Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1, *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, pp. 443–452, doi: 10.25126/jtiik.2020732093.

[8] S. Linawati, R. A. Safitri, A. R. Alfiyan, W. E. Pangesti, and M. N. Winnarto, Mar. 2020, Perbandingan Algoritma Klasifikasi Naive Bayes dan SVM Pada Studi Kasus Pemberian Penerima Beasiswa PPA, *Swabumi*, vol. 8, no. 1, pp. 71–75, doi:

p-ISSN : 1858-3911, e-ISSN : 2614-5405
https://journal.uniku.ac.id/index.php/ilkom

10.31294/swabumi.v8i1.7708.

[9] Hermanto, A. Mustopa, and A. Y. Kuntoro, Feb. 2020, Algoritma Klasifikasi Naive Bayes dan Support Vector Machine dalam Layanan Komplain Mahasiswa, *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 211–220, doi: 10.33480/jitk.v5i2.1181.

[10] B. Budiman, R. Nursyanti, R. Y. R. Alamsyah, and I. Akbar, Sep. 2020, Data Mining Implementation Using Naïve Bayes Algorithm and Decision Tree J48 In Determining Concentration Selection, *Int. J. Quant. Res. Model.*, vol. 1, no. 3, pp. 123–134, doi: 10.46336/ijqrm.v1i3.72.

[11] D. T. Larose and C. D. Larose, 2015, *Data Mining and Predictive Analytics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

[12] E. Sabna and M. Muhardi, 2016, Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar, *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 2, p. 41, doi: 10.24014/coreit.v2i2.2392.

[13] D. Astuti, A. R. Iskandar, and A. Febrianti, May 2019, Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering, *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 60–72, doi: 10.20895/inista.v1i2.71.

[14] P. Hari Santoso, May 2020, Application of Data Mining Classification for Covid-19 Infected Status Using Algortima Naïve Method, *J. Mantik*, vol. 4, no. 36, pp. 267–275, Accessed: Apr. 11, 2021. [Online]. Available: https://iocscience.org/ejournal/index.php/mantik/index.

[15] M. F. Maulana and M. Defriani, Mar. 2020, Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period, *PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log.*, vol. 8, no. 1, pp. 39–48, doi: 10.33558/piksel.v8i1.2018.

[16] S. Khruahong and P. Tadkerd, Oct. 2020, Analysis of Scholarship Consideration Using J48 Decision Tree Algorithm for Data Mining, in *International Conference on Cooperative Design, Visualization and Engineering*, vol. 12341, pp. 230–238, doi: 10.1007/978-3-030-60816-3_26.

[17] M. Kück and M. Freitag, Jan. 2021, Forecasting of Customer Demands for Production Planning by Local K-Nearest Neighbor Models, *Int. J. Prod. Econ.*, vol. 231, p. 107837, doi: 10.1016/j.ijpe.2020.107837.

[18] H. Tang *et al.*, 2020, Predicting Green Consumption Behaviors of Students Using Efficient Firefly Grey Wolf-Assisted K-Nearest Neighbor Classifiers, *IEEE Access*, vol. 8, pp. 35546–35562, doi: 10.1109/ACCESS.2020.2973763.

[19] F. Gorunescu, 2011, *Data Mining: Concepts, Models and Techniques*. Berlin: Springer-Verlag Berlin Heidelberg.

[20] Suwarno and A. Abdillah, 2016, Penerapan Algoritma Bayesian Regularization Backpropagation Untuk Memprediksi Penyakit Diabetes, *J. MIPA*, vol. 39, no. 2, pp. 150–158, Accessed: Mar. 16, 2021. [Online]. Available: http://journal.unnes.ac.id/nju/index.php/JM.