Cyberspace: Jurnal Pendidikan Teknologi Informasi
Volume 4, Nomor 1, Maret 2020, 50-58

IMPLEMENTASI ALGORITMA NAIVE BAYES DAN RANDOM FOREST DALAM MEMPREDIKSI PRESTASI AKADEMIK MAHASISWA UNIVERSITAS ISLAM NEGERI AR-RANIRY BANDA ACEH

Bustami Yusuf[1], Muthmainna Qalbi[2], Basrul[3], Ima Dwitawati[4], Malahayati[5], Mega Ellyadi[6]

[2,3]Program Studi Pendidikan Teknologi Informasi, Fakultas Tarbiyah dan Keguruan
Universitas Islam Negeri Ar-Raniry-Banda Aceh, 23111
[1,4,5]Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Ar-Raniry-Banda Aceh, 23111

Email : bustamiyusoef@ar-raniry.ac.id, muna.muthmainna@gmail.com, basrul.amajid@ar-raniry.ac.id, ima@ar-raniry.ac.id, malahayati_umar@ar-raniry.ac.id, megaellyadi@gmail.com

Abstrak

Prestasi akademik di tentukan oleh dua faktor, yaitu faktor internal yang berasal dari dalam diri individu dalam hal ini mahasiswa dan faktor eksternal yang berasal dari luar diri individu atau hal yang di pengaruhi oleh lingkungan. Ada banyak cara mencari suatu prestasi akademik, salah satunya menggunakan *data mining* yang bertujuan memprediksikan atau mengklasifikasikan data menggunakan algoritma klasifikasi. Penelitian ini bertujuan untuk 1) mengetahui cara menerapkan algoritma Naive Bayes terhadap prestasi mahasiswa, dan 2) melihat keakuratan algoritma *Naive Bayes* terhadap prestasi mahasiswa. Jenis penelitian yang digunakan adalah data sekunder yang berupa data mahasiswa yang di peroleh dari pusat teknologi informasi dan pangkalan data UIN Ar-Raniry. Penelitian ini menggunakan algoritma *naive bayes* dan algoritma *random forest*. Hasil yang di peroleh dari penelitian ini menunjukkan nilai korelasi tertinggi pada variabel IP awal sebesar r=0,783 dan variabel cuti memiliki tingkat korelasi sangat lemah sebesar r=0,054. Nilai keakuratan variabel algoritma naive bayes setelah di cleaning sebesar 78.0% dan variabel algoritma Random Forest sebesar 76,7%.

Kata Kunci : *Naive Bayes, Prestasi Akademik, Random Forest, Prediksi, Motivasi*

Abstract

Academic achievement is determined by two factors, namely internal factors originating from within the individual in this case students and external factors that come from outside the individual or things that are influenced by the environment. There are many ways to find an academic achievement, one of which uses data mining which aims to predict or classify data using a classification algorithm. This study aims to 1) find out how to apply the Naive Bayes algorithm to student achievement, and 2) see the accuracy of the Naive Bayes algorithm to student achievement. This type of research is secondary data in the form of student data obtained from the information technology center and the Ar-Raniry UIN database. This research uses Naive Bayes algorithm and random forest algorithm. The results obtained from this study indicate the highest correlation value in the initial IP variable of r = 0.783 and the leave variable has a very weak correlation level of r = 0.054. The accuracy value of Naive Bayes algorithm after cleaning is 78.0% and Random Forest algorithm variable is 76.7%.

Keywords : *Naive Bayes, Academic Achievement, Random Forest, Prediction, Motivation*

1. Pendahuluan

Naive Bayes merupakan suatu algoritma klasifikasi yang sangat efektif dan juga efisien. Algoritma ini bertujuan untuk melakukan klasifikasi data pada kelas tertentu [1]. Dalam penelitian sebelumnya yang dilakukan oleh Supardi Salmu (2017) tentang prediksi tingkat kelulusan mahasiwa mengatakan bahwa hasil akurasi pada penelitian tersebut sebesar 80,7% dari data training yang berjumlah 1162 data dan data testing berjumlah 587 data[2]. Pada penelitian tersebut ia menggunakan algoritma Naive Bayes.

Selain Algoritma Naive Bayes, terdapat pula algoritma Random Forest yang bertujuan untuk melakukan klasifikasi pada kelas dengan akurat. Pada penelitian yang dilakukan oleh I made Budi Adnyana tentang prediksi lama mahasiswa, mengatakan bahwa algoritma random forest memiliki tingkat keakuratan algoritma sebesar 83,54%, yang berarti tingkat keakuratannya sudah baik[3].

Dalam dunian pendidikan, algoritma sejenis ini dapat di gunakan untuk memprediksi tingkat prestasi siswa atau peseta didi seperti penelitian sebelumnya yang dilakukan oleh Heri Susanto (2014) mengatakan bahwa prestasi siswa atau peserta didik itu berdasarkan status sosial ekonomi orang tua, motivasi, kedisiplinan siswa dan prestasi masa lalu. Variabel motivasi adalah variabel yang menentukan potensi seorang siswa berhasil atau tidak prestasi belajarnya di waktu yang akan datang. Variabel prestasi masa lalu merupakan variabel kedua yang penting dalam keberhasilan siswa menempuh studinya. Hal ini menunjukkan bahwa aspek *knowledge* atau kecerdasan siswa sangat berpengaruh terhadap keberhasilan belajarnya. Sebaliknya, jika kecerdasan siswa tersebut kurang, terdapat kemungkinan siswa tersebut masih tetap berprestasi[4].

Oleh karena itu, pada penelitian ini dilakukan implementasi Algoritma Naive Bayes dan Random Forest dalam memprediksi prestasi akademik mahasiswa Universitas Islam Negeri Ar-Raniry. Penelitian ini di harapkan dapat membantu peneliti untuk memilih algoritma yang tepat untuk memprediksikan tingkat prestasi dalam pendidikan.

2. Metodologi Penelitian

2.1 Alat dan Bahan Penelitian

Penelitian ini membutuhkan alat yang bertujuan untuk mendukung berjalannya implementasi, antara lain:

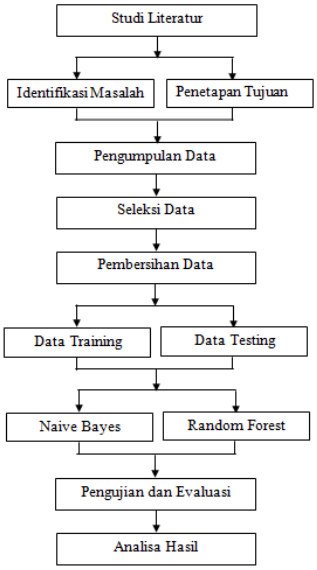
1. Perangkat Keras
   Penelitian ini menggunakan sebuah Laptop Lenovo Ideapad 100 dengan spesifikasi *Processor:* Intel(R) Core(TM) i3-5005U CPU @2.00GHz, RAM 4GB, sistem operasi Windows 10 Pro 64-bit
2. Perangkat lunak
   Penelitian ini menggunakan sebuah perangkat lunak yaitu WEKA.

2.2 Data Training dan Data Testing

Data Training atau data latih adalah data yang sudah ada, sedangkan *Data Testing* adalah data yang sudah berkelas dan sudah berlabeldari target atribut yang digunakan untuk mengklasifikasi suatu data [5].Persentase data training yang digunakan sebesar 60%, 70%, 80% dan 90% dari 1500 sampel, sampel yang tersisa dari besaran persentase pada tiap-tiap data training di gunakan sebagai datang testing.

2.3. Tahapan Penelitian

Tahapan penelitian yang yang dilakukan pada penelitian ini adalah seperti yang terlihat dalam gambar. 1.



Gambar.1. Metode Penelitian

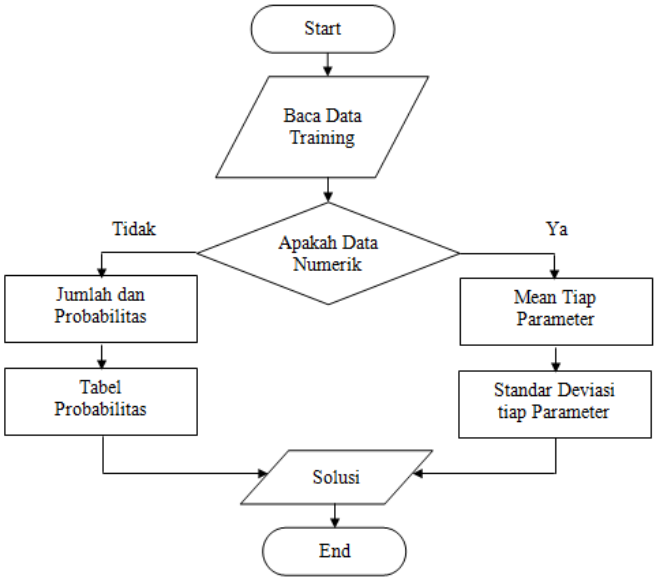
2.4. Sampel

Sampel data dalam penelitian ini berjumlah 1500 data, yang dimana data dibagi menjadi dua bagian,yaitu training set dan Testing set. Pada penelitian ini awalnya terdapat beberapa variabel, diantaranya: jenis kelamin, kategori masuk, tahun masuk, tahun keluar, lamanya kuliah, non aktif/tidak, asal daerah, asal sekolah, keadaan ayah, pendidikan terakhir ayah, pekerjaan ayah, penghasilan ayah, IP semester 1, IP semester 2, IP semester 3, IP semester 4, IP semester 5, IP semester 6, IP semester 7, IP semester 8, IP semester 9, IP semester 10, IPK awal.

2.5. Tahapan Analisis Algoritma Naive Bayes

Berikut tahapan analisis algoritma naive bayes yaitu sebagai berikut[6]:

1. Masukkan data training
2. Melihat apakah data training yang dimasukkan berupa numerik atau tidak
   a. Jika data tersebut numerik, maka yang dihitung mean dan standar deviasi dari tiap parameter yang ada.
   b. Jika data tersebut bukan numerik, maka yang dihitung nilai probabilitasnya, dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut, kemudian buat tabel dari probabilitas yang ada.
3. Setelah itu akan mendapatkan nilai dari tabel mean, standar deviasi, dan probabilitas.

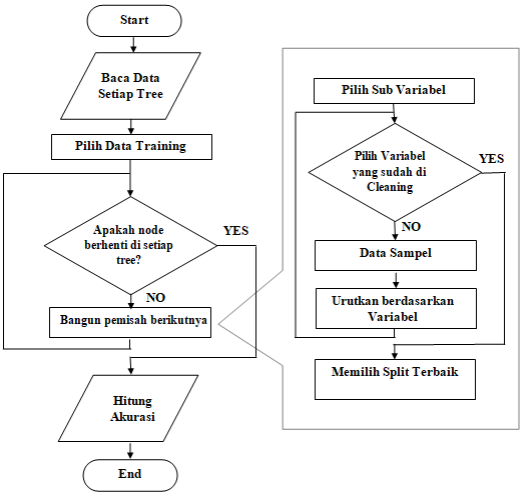


Gambar 2. Alur Algoritma Naive Bayes

2.6.Tahapan Algoritma Random Forest

Berikut tahapan analisis algoritma random forest sebagai berikut[7]:

1. Masukkan data setiap tree
2. Pilih data trainingnya, misalnya 60% yang menjadi data training, maka 40% menjadi data testing .
3. Melihat apakah setiap node (simpul) berhenti disetiap tree atau tidak.
   a. Jika tidak berhenti maka bangun pemisah berikutnya dengan cara Pilih sub variabel lalu pilih variabel yang sudah di cleaning.
   b. Jika sudah maka pilihlah split terbaik
   c. Jika belum maka pilihlah data sampel lalu urutkan berdasarkan variabel, kemudian ulangi proses tersebut sampai mendapatkan split terbaik.
   d. Ulangi proses diatas sampai node berhenti di setiap tree.
   e. Jika berhenti maka hitung nilai akurasi



Gambar 3. Alur Algoritma Random Forest

2.7. Labelling

Labeling adalah proses penentuan label kepada data yang ada. Herlina (2012) mendefinisikan *labelling* sebagai penggambaran sifat yang berhubungan dengan perilaku[8]. Pada penelitian ini IPK menjadi labelnya. Alasannya karena IPK ini merupakan hasil akhir dari prestasi akademik mahasiswa. Kemudian diberikan kelas pada label tersebut dengan melakukan convert data yang dimana variabel IPK awalnya berbentuk numerik, kemudian diubah menjadi kategori.

Tabel 1. Pemberian Nama Kelas Pada Label

IPK	KATEGORI
≥ 3.50	Istimewa
≤ 3.00	Sangat Baik
< 3.00	Baik
< 2.00	Cukup

3. Hasil Dan Pembahasan

3.1 Uji Korelasi

Dalam penelitian ini nilai IPK (Indeks Prestasi Komulatif) menjadi label akhir untuk menentukan prestasi seorang mahasiswa. Dalam menetapkan interval kategori kekuatan korelasi, Sarwono (2012)[9] menetapkan penetapan sebagai berikut :

Tabel 2. Kategori kekuatas Korelasi

0	Tidak ada korelasi
0,00 – 0,25	Korelasi sangat lemah
0,25 – 0,50	Korelasi cukup
0,50 – 0,75	Korelasi kuat
0,75 – 0,99	Korelasi sangat kuat
1	Korelasi sempurna

Pada penelitian ini, jika suatu hubungan tidak sama dengan nol (0), maka dapat dikatakan terjadi hubungan, dimana dihasilkan pada hasil hasil berikut

Tabel 3. Korelasi Variabel

Variabel	Sig	Nilai Pearson Corelation	Keterangan
Jenis Kelamin	0,011	0,065	Berkorelasi Sangat Lemah
Jalur Masuk	0,000	0,324	Berkorelasi Cukup
Lama Kuliah	0,000	0,434	Berkorelasi Cukup
Cuti	0,035	0,054	Berkorelasi Sangat Lemah
Keadaan Ayah	0,860		Tidak berkorelasi
Pendidikan Ayah	0,000	0,166	Berkorelasi Sangat Lemah
Pekerjaan Ayah	0,069		Tidak berkorelasi
Penghasilan Ayah	0,000	0,169	Berkorelasi Sangat Lemah
IP Awal	0,000	0,789	Berkorelasi Sangat Kuat

Setelah dilakukan perhitungan diatas (Tabel 3), variabel yang memiliki hubungan positif antara variabel dengan IPK adalah variabel jenis kelamin, lama kuliah, cuti,

pendidikan ayah, penghasilan ayah, dan IPK awal. Dan tidak mempunyai hubungan yaitu status ayah dan pekerjaan ayah.

Dari Tabel diatas dapat disimpulkan bahwa IPK awal memiliki hubungan paling besar dibandingkan dengan variabel lainnya. Hubungan antara IPK awal dengan IPK akhir memiliki korelasi sangat kuat yaitu (r = 0,783).

3.2 Pengujian Klasifikasi

Pada pengujian ini, peneliti menggunakan aplikasi weka untuk menguji keakuratan dari algoritma naive bayes dan algoritma random forest.

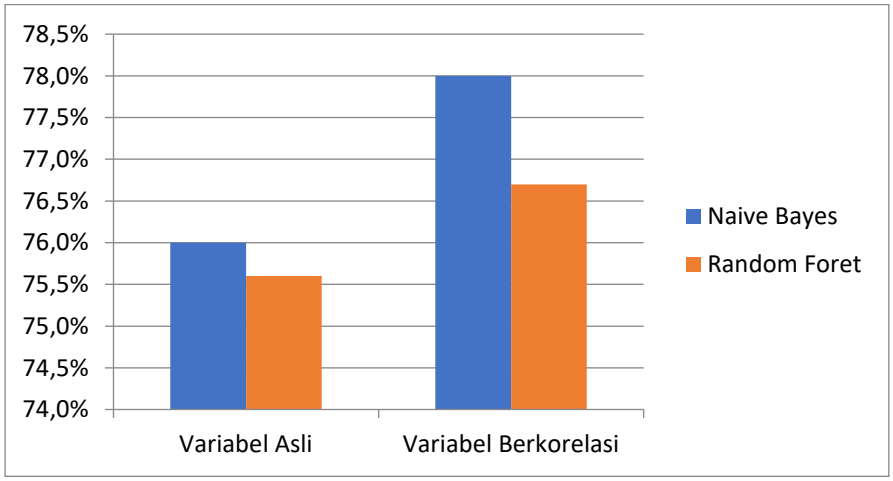
a. Pengujian Korelasi Variabel

Peneliti ingin menguji hasil keakuratan anatara variabel asli dengan variabel yang sudah di cleaning dengan korelasi. Variabel asli yaitu seluruh variabel yang didapat dari PTI-PD dan belum di cleaning. Sedangkan variabel korelasi adalah variabel-variabel yang telah di lakukan proses cleaning (proses dengan membuang variabel yang tidak bisa diolah atau variabel yang tidak diperlukan).

Tabel 4.Pengujian Korelasi Variabel

Pengujian	Variabel	
	Variabel Asli	Variabel Berkorelasi
Naive Bayes	76.0%	78.0%
Random Foret	75.6%	76.6%

Berdasarkan tabel diatas dapat diketahui bahwa nilai ketepatan (keakuratan) tertinggi terletak pada variabel yang telah berkorelasi, Naive bayes sebesar 78,0% dan Random Forest sebesar 76,6%. Berikut ini adalah grafik berdasarkan tabel diatas:



Gambar 4. Pengujian Korelasi Variabel

Dari grafik diatas dapat disimpulkan bahwa nilai semakin berkorelasi tiap variabelnya, maka makin tinggi hasil keakuratannya. Setelah data di cleaning, maka semakin bagus keakuratannya.

3.3 Uji Percentage Split

Terdapat 4 buah *split* yang digunakan dalam penelitian ini, diantaranya 60%, 70%, 80%, 90%. Pada split 60% artinya dari 1500 data maka 60% yang menjadi data training

dan 40% yang menjadi data testing. Pada split 70% artinya dari 1500 data maka 70% yang menjadi data training dan 30% yang menjadi data testing. Pada split 80% artinya dari 1500 data maka 80% yang menjadi data training dan 20% yang menjadi data testing. Pada split 90% artinya dari 1500 data maka 90% yang menjadi data training dan 10% yang menjadi data testing.

Tabel 5. Uji Percentage Split Perbandingan Metode

Percentage Split	Ketepatan (Keakuratan)	
	Naive Bayes	Random Forest
60%	76.5%	76.0%
70%	77.3%	76.9%
80%	77.0%	75.6%
90%	78.0%	76.6%

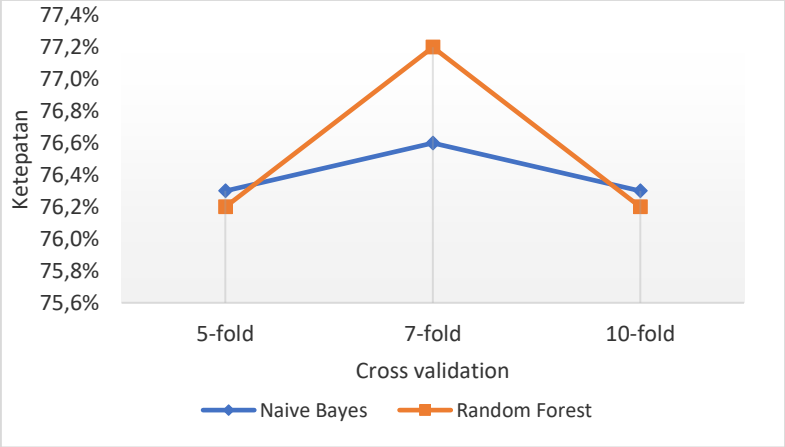
3.4 Uji Cross validation

Pada pengujian ini terdapat banyak pilihan *fold* yang digunakan. Nilai fold yang digunakan yaitu *5-fold*, *7-fold*, dan *10-fold*.

Tabel 6. Uji Cross Validation

Cross-Validation	Ketepatan (Keakuratan)	
	Naive Bayes	Random Forest
5-fold	76.3%	76.2%
7-fold	76.6%	77.2%
10-fold	76.3%	76.2%

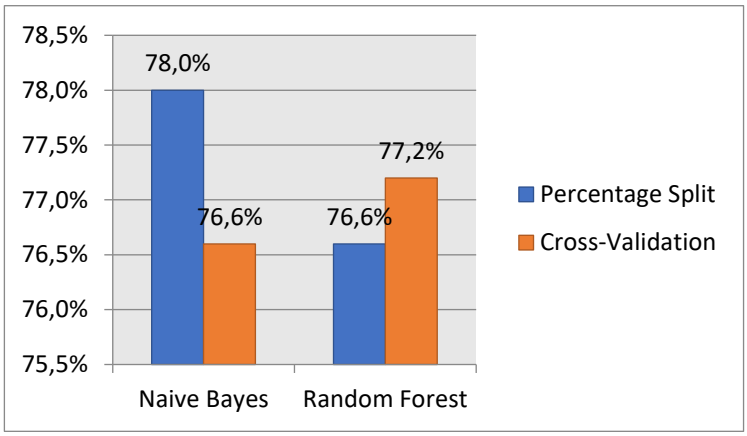
Berdasarkan tabel diatas dapat diketahui bahwa nilai ketepatan (keakuratan) tertinggi terletak pada 7-*fold*, Naive bayes sebesar 76,3% dan Random Forest sebesar 77,2%. Berikut ini adalah grafik berdasarkan tabel diatas:



Gambar 5. Uji Cross Validation

3.5. Perbandingan Percentage Split dan Cross-Validation

Berdasarkan pengujian yang telah dilakukan diatas, telah diperoleh hasil tetinggi dari tiap jenis pengujian. Sehingga peneliti ingin membandingkan hasil dari kedua jenis pengujian tersebut berdasarkan metode Naive Bayes dan Random Forest.

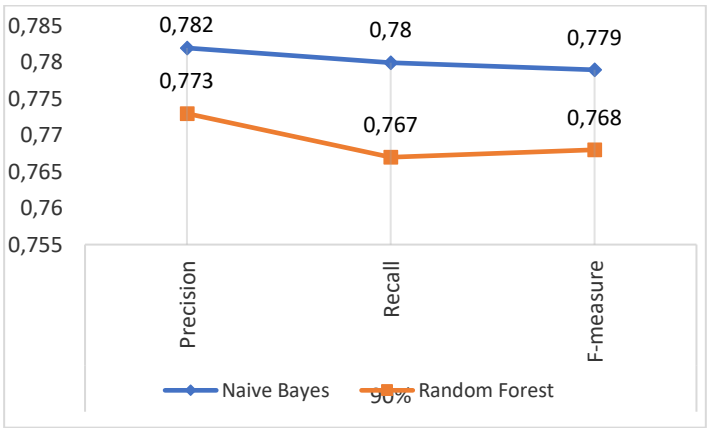


Gambar 6. Perbandingan Percentage Split dan Cross-Validation

Dari grafik diatas dapat disimpulkan bahwa metode Naive Bayes lebih unggul dikarenakan nilai keakuratan Naive Bayes lebih besar dibandingkan metode Random Forest. Nilai tertinggi Naive Bayes sebesar 78,0% sedangkan Random Forest sebesar 77,2%.

3.6. Hasil Evaluasi

Setelah melakukan pengujian diatas, yang akan dilakukan selanjutnya yaitu mengevaluasinya dengan menggunakan metode evaluasi berupa Precision, *Recall*, dan *F-measure*.



Gambar 7. Hasil Evaluasi

4. KESIMPULAN

Setelah di lakukan pengujian pada penelitian ini, dapat di simpulkan bahwa:

1 Setelah melakukan pengujian korelasi, terdapat dua variabel yang tidak berkorelasi yaitu variabel keadaan ayah dan pekerjaan ayah. Variabel yang memiliki korelasi paling besar terhadap label (IPK) adalah variabel IP awal.

2 Pengujian terhadap variabel yang berkorelasi memiliki keakuratan yang lebih tinggi dibandingkan variabel data asli (awal).

3 Dari pengujian cross-validation pada kedua algoritma diperoleh hasil algoritma random forest lebih unggul dibandingkan naïve bayes. Sedangkan pada pengujian percentage split, algoritma naïve lebih unggul dengan tingkat keakuratannya sebesar 78,0% sedangkan random forest sebesar 76,6%. Artinya algoritma naive bayes lebih baik tingkat keakuratannya.

DAFTAR PUSTAKA

[1] Lestari, Diah Indah. 2015. Skripsi. *Analisis Data Siswa Menggunakan Klasifikasi Naive Bayes dalam Data Mining untuk Memprediksi Siswa diterima di PTN*. Yogyakarta: Universitas Negeri Yogyakarta

[2] Salmu, Supardi. *Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes*. Prosiding Seminar Nasional Multidisiplin llmu, 2017, ISSN : 2087-0930.

[3] Adnyana, I Made Budi. 2015. *Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : Stikom Bali)*. CSRID Journal, Vol. 8 No. 3.

[4] Susanto, Heri. 2014. D*ata Mining Untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan Dan Prestasi Masa Lalu*. Jurnal Pendidikan Vokasi Vol 4, No. 2

[5] S.A. Zega. 2014. *Penggunaan Pohon Keputusan untuk Klasifikasi Tingkat Kualitas Mahasisa Berdasarkan Jalur Masuk Kuliah.*Yogyakarta

[6] Imandasari, Tia dkk. 2019. "Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air (Studi Kasus : STIKOM Tunas Bangsa Pematangsiantar)". Prosiding Seminar Nasional Riset Information Science (SENARIS), ISSN: 2686-0260

[7] Maulana Dhawangkara, Skripsi "*Prediksi Intensitas Hujan Kota Surabaya Dengan Matlab Menggunakan Teknik Random Forest Dan Cart (Studi Kasus Kota Surabaya)*" (Surabaya: Institut Teknologi Sepuluh November, 2016) Hal. 26.

[8] Herlina. 2007.*Labeling dan Perkembangan Anak*. FOTA-Salman

[9] Sarwono, Jonathan. 2012. *Mengenal SPSS Satistic 20 : Aplikasi Untuk Riset Ekspreimental*. Elex Media Komputindo. Jakarta.