KLASIFIKASI NILAI MAHASISWA UNIVERSITAS LAMPUNG MENGGUNAKAN ALGORITMA *NAÏVE BAYES* DAN *RANDOM FOREST*

(Skripsi)

Oleh

REKTI NURUL HIDAYAH
1817031004

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG

2023

ABSTRACT

CLASSIFICATION OF STUDENT SCORE OF LAMPUNG UNIVERSITY USING NAÏVE BAYES AND RANDOM FOREST ALGORITHM

By

REKTI NURUL HIDAYAH

At the tertiary level, at the end of each semester, students receive learning outcomes for each course taken. The end of student assessment in a course is based on the total score obtained from each learning achievement and the Final Semester Examination (UAS), which has a predetermined weight. Classification is a supervised learning method that seeks to identify correlations between input features and target features. Two classification methods were used in this study, namely the naïve bayes and random forest methods and compared the two methods so that they could be categorized as good classification methods. The stages in this research are data selection, data preprocessing, building and testing the model, and model evaluation. As for the results of this study, it was found that the naïve bayes and random forest methods had the highest level of accuracy when using the splitting data compared to using the k-fold cross validation. Based on the evaluation results of the model with the confusion matrix, the random forest method is the best method for classifying Lampung University student grades when compared to the naïve bayes method with an accuracy rate of 99.38%. Furthermore, the performance of the two methods is better if after boosting with Gradient Boosting (GB), namely for the naïve bayes method it is 99.89% and the random forest method is 99.45%.

Keywords: Student, Final Grade, Classification, Naïve Bayes, Random Forest.

ABSTRAK

KLASIFIKASI NILAI MAHASISWA UNIVERSITAS LAMPUNG MENGGUNAKAN ALGORITMA *NAÏVE BAYES* DAN *RANDOM FOREST*

Oleh

REKTI NURUL HIDAYAH

Ditingkat perguruan tinggi, setiap akhir semester, mahasiswa mendapatkan hasil belajar untuk setiap mata kuliah yang diambil. Akhir dari penilaian mahasiswa pada suatu mata kuliah didasarkan pada total nilai yang diperoleh dari setiap pencapaian pembelajaran dan Ujian Akhir Semester (UAS), yang memiliki bobot yang telah ditetapkan. Klasifikasi merupakan metode *supervised learning* yang berusaha untuk mengidentifikasi korelasi antara fitur input dan fitur target. Dua metode klasifikasi digunakan dalam penelitian ini yaitu metode *naïve bayes* dan *random forest* serta membandingkan kedua metode tersebut agar dapat dikategorikan sebagai metode klasifikasi yang baik. Tahapan dalam penelitian ini yaitu seleksi data, *preprocessing* data, membangun dan menguji model, serta evaluasi model. Adapun hasil penelitian ini diperoleh bahwa metode *naïve bayes* dan *random forest* memiliki tingkat akurasi tertinggi jika menggunakan *splitting* data dibandingkan dengan menggunakan *k-fold cross validation*. Berdasarkan hasil evaluasi model dengan *confusion matrix*, metode *random forest* merupakan metode terbaik untuk mengklasifikasi nilai mahasiswa Universitas Lampung jika dibandingkan dengan metode *naïve bayes* dengan tingkat akurasi sebesar 99,38%. Selanjutnya untuk performa dari kedua metode lebih baik jika sesudah dilakukan *boosting* dengan *Gradient Boosting* (GB), yaitu untuk metode *naïve bayes* sebesar 99,89% dan metode *random forest* sebesar 99,45%.

Kata Kunci: Mahasiswa, Nilai Akhir, Klasifikasi, *Naïve Bayes*, *Random Forest*.

KLASIFIKASI NILAI MAHASISWA UNIVERSITAS LAMPUNG MENGGUNAKAN ALGORITMA *NAÏVE BAYES* DAN *RANDOM FOREST*

Oleh

REKTI NURUL HIDAYAH

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	:	**KLASIFIKASI NILAI MAHASISWA UNIVERSITAS LAMPUNG MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN RANDOM FOREST**
Nama Mahasiswa	:	Rekti Nurul Hidayah
Nomor Pokok Mahasiswa	:	1817031004
Program Studi	:	Matematika
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Dian Kurniasari, S.Si., M.Sc.	**Dr. Notiragayu, S.Si., M.Si.**
NIP. 19690305 199603 2 001	**NIP. 19731109 200012 2 001**

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 19740316 200501 1 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dian Kurniasari, S.Si., M.Sc.** .....

Sekretaris : **Dr. Notiragayu, S.Si., M.Si.** .....

Penguji
Buka Pembimbing : **Ir. Warsono, M.S., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi : 18 Juli 2023

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	: Rekti Nurul Hidayah
Nomor Pokok Mahasiswa	: 1817031004
Jurusan	: Matematika
Judul Skripsi	: **KLASIFIKASI NILAI MAHASISWA UNIVERSITAS LAMPUNG MENGGUNAKAN ALGORITMA *NAÏVE BAYES* DAN *RANDOM FOREST***

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 18 Juli 2023

Penulis,

Rekti Nurul Hidayah

RIWAYAT HIDUP

Penulis bernama Rekti Nurul Hidayah lahir di Rajabasa Lama pada 25 Desember 1999. Penulis merupakan anak pertama dari dua bersaudara dari pasangan Bapak Rekadin dan Ibu Katini.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) Aisyiyah Bustanul Athfal pada tahun 2005-2006. Kemudian menempuh pendidikan Sekolah Dasar (SD) di SDN 3 Rajabasa Lama pada tahun 2006-2012. Kemudian melanjutkan ke Sekolah Menengah Pertama di SMPN 1 Labuhan Ratu pada tahun 2012-2015, Sekolah Menengah Atas di SMAN 1 Way Jepara pada tahun 2015-2018. Pada tahun 2018 penulis terdaftar sebagai mahasiswa Program Studi S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN.

Pada Tahun 2020 penulis melakukan Kuliah Kerja Praktik (KP) di Badan Pusat Statistik Sukadana serta mengikuti kegiatan Kerja Kuliah Nyata di Desa Rajabasa Lama, Lampung Timur.

KATA INSPIRASI

“Sesungguhnya sesudah kesulitan itu ada kemudahan.”

(Q.S. Al-Insyirah : 6)

“Allah tidak akan membebani seseorang melainkan dengan kesanggupannya.”

(Q.S. Al-Baqarah : 286)

“Yakinlah ada sesuatu yang menantimu setelah banyak kesabaran (yang kau jalani), yang akan membuatmu terpana hingga kau lupa betapa pedihnya rasa sakit.”

(Ali bin Abi Thalib)

“Jangan menyakiti dirimu dengan selalu berfikir kenapa mereka berkata demikian, kenapa mereka berbuat demikian? Percayalah pada Tuhanmu, percaya pada dirimu, selama mereka hanya manusia, maka yang mereka mampu hanyalah berkata-kata.”

SIMPLE RULES IN LIFE

“Jika kamu tidak mengejar apa yang kamu inginkan, kamu tidak akan mendapatkannya."
“Jika kamu tidak bertanya, maka jawabannya selalu tidak.”
“Jika kamu tidak melangkah, maka kamu akan selalu ditempat yang sama.”

PERSEMBAHAN

Alhamdulillah, puji dan syukur kepada Allah SWT atas nikmat serta hidayahNya sehingga skripsi ini dapat terselesaikan dengan baik dan tepat pada waktunya.

Oleh karena itu, dengan rasa syukur dan bahagia saya persembahkan rasa terimakasih saya kepada:

Bapak Rekadin dan Ibu Katini

Terima kasih kepada kedua orang tuaku atas segala pengorbanan, motivasi, doa dan ridho kalian serta dukungannya selama ini. Terimakasih telah memberikan pelajaran berharga kepada anakmu ini tentang makna perjalanan hidup yang sebenarnya sehingga kelak bisa menjadi orang yang bermanfaat bagi semua orang.

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, memberikan motivasi, memberikan arahan serta ilmu yang berharga.

Sahabat-sahabatku

Terimakasih kepada semua orang-orang baik yang telah memberikan pengalaman, semangat, motivasinya, serta doa-doanya dan senantiasa memberikan dukungan dalam hal apapun.

Almamater Tercinta Universitas Lampung

SANWACANA

Segala puji dan syukur penulis ucapkan kepada Allah SWT atas segala nikmat dan karunia-Nya yang tak terhingga sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Klasifikasi Nilai Mahasiswa Universitas Lampung Menggunakan Algoritma *Naïve Bayes* dan *Random Forest***". Penulisan skripsi ini tidak dapat diselesaikan tanpa adanya bimbingan, bantuan, dan dukungan dari berbagai pihak sehingga pada kesempatan ini penulis mengucapkan terima kasih kepada:

1. Ibu Dian Kurniasari, S.Si., M.Sc. selaku dosen pembimbing I yang senantiasa membimbing, memberi masukan dan saran, serta mendukung penulis dalam menyelesaikan skripsi ini.
2. Ibu Dr. Notiragayu, S.Si., M.Si. selaku dosen pembimbing II yang telah memberikan bimbingan, pengarahan, serta saran sehingga penulis dapat menyelesaikan skripsi ini.
3. Bapak Ir. Warsono, M.S., Ph.D. selaku dosen penguji yang telah memberikan kritik dan saran yang membangun sehingga skripsi ini dapat diselesaikan.
4. Bapak Drs. Nusyirwan, S.Si., M.Si. selaku dosen pembimbing akademik yang telah memberikan bimbingan dan arahan selama masa perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Kedua orang tuaku dan adik yang selalu memberikan motivasi serta dukunganya.
8. Seluruh dosen, staf, karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
9. Keluarga sekalian yang selalu memberikan semangat kepada penulis serta doa-

doanya.

10. Orang-orang baik yang namanya tidak bisa saya sebutkan satu persatu yang telah menjadi teman terbaik penulis yang selalu memberikan semangat dan menemani penulis dalam keadaan apapun serta telah memberikan pengalaman dan banyak cerita selama masa perkuliahan.
11. Teman-teman seperbimbingan yang selalu memberikan dukungan dan motivasi serta doa-doanya (Alifia, Dalfa, Farel, Febi, Ferzy, Joshua, Luthfia, Maydia, Oktin, Putri, Shavira, Sulis,Virda, Zaenal).
12. Semua teman-teman Jurusan Matematika 2018 dan teman kelas A yang telah membantu serta memberikan semangat kepada penulis.
13. Semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu penulis dalam menyelesaikan skripsi ini.

Penulis menyadari bahwa masih banyak kekurangan dalam penulisan skripsi ini. Oleh karena itu, penulis mengharapkan masukan serta saran untuk dijadikan pelajaran kedepannya.

Bandar Lampung, 18 Juli 2023
Penulis,

Rekti Nurul Hidayah

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Pendidikan merupakan usaha sadar untuk menggali potensi peserta didik dan mahasiswa untuk menjadi seorang yang profesional (Putra dan Yuniarti, 2020). Di perguruan tinggi, mahasiswa menerima hasil belajar untuk setiap mata kuliah yang mereka selesaikan pada setiap akhir semester. Tentu saja, hal ini terjadi berulang kali pada setiap akhir semester ketika mahasiswa menerima Indeks Prestasi Kumulatif (IPK). Semua perguruan tinggi, termasuk Universitas Lampung, selalu mengevaluasi hasil dari proses perkuliahan yang dilakukan.

Universitas Lampung (UNILA) adalah perguruan tinggi negeri yang terletak di Kota Bandar Lampung, Provinsi Lampung. Peraturan Universitas Lampung menyebutkan bahwa penilaian proses dan hasil pembelajaran dapat berupa kuis, tugas terstruktur, ujian praktik, Ujian Tengah Semester (UTS), Ujian Akhir Semester (UAS), dan observasi kelas. Penilaian hasil belajar untuk mahasiswa program diploma, sarjana, profesi, *magister*, dan *doctor* dinyatakan dengan huruf mutu dan angka mutu A (4), B+ (3,5), B (3), C+ (2,5), C (2), D (1), dan E (0) (Universitas Lampung, 2022).

Nilai akhir yang dicapai oleh seorang mahasiswa dalam suatu mata kuliah merupakan akumulasi dari nilai yang dicapai untuk setiap komponen pembelajaran dan Ujian Akhir Semester (UAS) dengan bobot yang ditentukan. Nilai akhir dinyatakan dalam huruf dan angka berdasarkan *range* nilai yang

diperoleh. Nilai akhir biasanya menjadi bencana bagi beberapa mahasiswa karena menentukan apakah mereka lulus atau tidak. Oleh karena itu perlu dilakukan analisis mengenai hasil belajar mahasiswa, sehingga dari hasil analisis tersebut dapat diperoleh metode terbaik untuk mengklasifikasi nilai mahasiswa dengan cara membandingkan tingkat akurasinya.

Klasifikasi merupakan metode *supervised learning*, metode yang mencoba menemukan hubungan antara atribut masukan dan atribut target. Tujuan klasifikasi adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data (Hendrian, 2018). Saat menganalisis kinerja beberapa metode klasifikasi, bandingkan metode *data mining*, baik metode *decision tree* maupun metode *machine learning*, untuk memilih metode terbaik dan paling akurat. Pada penelitian ini, algoritma *naïve bayes* dan *random forest* digunakan sebagai pembanding.

Naïve bayes adalah algoritma klasifikasi yang sangat efektif dan efisien. Tujuan dari algoritma ini adalah untuk mengklasifikasikan data ke dalam kategori tertentu (Yusuf, dkk., 2020). Pada penelitian sebelumnya yang dilakukan oleh Anggraini, dkk. (2019) tentang klasifikasi data blogger menyatakan bahwa hasil klasifikasi data blogger dibagi menjadi dua kelas klasifikasi yaitu kelas ya dan kelas tidak. Nilai sebaran dengan kelas sebesar 0,680 dan nilai sebaran tanpa kelas sebesar 0,320. Hasil pengolahan data menunjukkan akurasi klasifikasi data blogger mencapai 86,67%. Selain algoritma *naïve bayes*, ada juga algoritma *random forest* yang bertujuan untuk mengklasifikasikan kelas secara akurat (Yusuf, dkk., 2020). Penelitian yang dilakukan oleh Ramadhan, dkk. (2019) tentang mengidentifikasi faktor penting penilaian mutu pendidikan, mengatakan bahwa berdasarkan evaluasi model, nilai akurasi klasifikasi dalam pemodelan klasifikasi *random forest* dengan banyak kelas adalah 83,49%. Kemudian Momole, dkk. (2022) juga melakukan penelitian tentang perbandingan *naïve bayes* dan *random forest* dalam klasifikasi bahasa daerah, dan hasil yang diperoleh metode *naïve bayes* dalam pengenalan bahasa sangat baik, karena memberikan nilai akurasi lebih dari

0,90, dibandingkan dengan *random forest* yang hanya mendapatkan nilai akurasi kurang dari 0,70. Dalam menghitung *confusion matrix*, metode *naïve bayes* lebih baik dengan nilai akurasi sebesar 0,9922 dibandingkan dengan nilai akurasi random forest sebesar 0,6544.

Oleh karena itu, penulis akan mencoba menerapkan metode *naïve bayes* dan *random forest* untuk klasifikasi nilai mahasiswa Universitas Lampung agar mengetahui manakah dari kedua metode tersebut yang dapat dikatakan sebagai metode klasifikasi yang baik dengan cara membandingkan nilai akurasinya.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini yaitu:

1. Melakukan klasifikasi nilai mahasiswa Universitas Lampung menggunakan metode *naïve bayes* dan *random forest* untuk mendapatkan klasifikasi nilai mahasiswa dengan tingkat akurasi yang baik.
2. Membandingkan nilai akurasi metode *naïve bayes* dengan metode *random forest* agar dapat dikategorikan sebagai metode klasifikasi yang baik.

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini yaitu sebagai bahan rujukan penelitian dalam klasifikasi nilai mahasiswa di perguruan tinggi lainnya serta menjadi bahan pertimbangan dan informasi tambahan bagi peneliti selanjutnya.

1.4 Batasan Penelitian

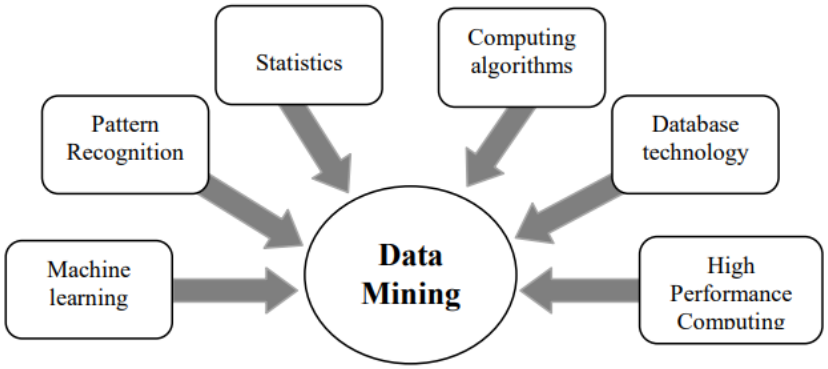
1. Data yang digunakan adalah data nilai mahasiswa Universitas Lampung semester ganjil tahun akademik 2021/2022.
2. Mata kuliah yang digunakan merupakan mata kuliah semester ganjil tahun akademik 2021/2022.
3. Klasifikasi hanya berdasarkan algoritma *naïve bayes* dan *random forest*, tidak berdasarkan peraturan akademik.

II. TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang berguna dan relevan dari basis data besar (Bramer, 2016). Banyak orang mendefinisikan *data mining* sebagai sinonim untuk istilah umum lainnya seperti seperti *Knowledge Discovery from Data* atau KDD (Han dan Kamber, 2006).

Data mining adalah serangkaian proses yang dirancang untuk menggali nilai tambah dari kumpulan data dalam bentuk informasi yang tidak dilaporkan secara manual (Syaputri, dkk., 2020). Selain itu, *data mining* mempunyai hubungan dengan berbagai bidang diantaranya statistik, *machine learning* (pembelajaran mesin), *pattern recognition*, *computing algorithms*, *database technology*, dan *high performance computing*. Diagram hubungan *data mining* disajikan pada Gambar 1.



Gambar 1. Diagram Hubungan *Data Mining*
(Sumber: Muslim, dkk., 2019)

Secara sistematik, tahapan utama dalam melakukan *data mining* terdiri dari tiga tahapan, yaitu sebagai berikut (Gonunescu, 2011):

1. Eksplorasi atau pemrosesan awal data

   Eksplorasi, atau pemrosesan data awal, terdiri dari pembersihan data, normalisasi data, transformasi data, pemrosesan nilai yang hilang, pengurangan dimensi, pemilihan subset fitur, dll.

2. Membangun model dan validasi

   Membangun dan validasi model, yaitu menganalisis model yang berbeda dan memilih model untuk kinerja terbaik. Metode seperti klasifikasi, regresi, analisis kluster, dan asosiasi digunakan dalam pengembangan model.

3. Penerapan

   Penerapan dilakukan dengan dengan menerapkan model terpilih pada data baru untuk mendapatkan hasil yang baik untuk masalah yang diteliti.

2.2 *Machine Learning*

Machine learning adalah disiplin kecerdasan buatan yang diprogram untuk memungkinkan komputer cerdas berperilaku seperti manusia dan secara

otomatis meningkatkan pemahaman mereka melalui pengalaman (Retnoningsih dan Pramudita, 2020). Bidang *machine learning* menjawab pertanyaan tentang bagaimana program komputer dibangun untuk meningkat secara otomatis berdasarkan pengalaman (Mitchell, 1997). Dalam *machine learning*, ada data latih dan data uji, data latih untuk melatih algoritma *machine learning*, dan data uji untuk mengetahui kinerja algoritma *machine learning* yang dilatih, yaitu menemukan data baru yang tidak pernah diberikan dalam pelatihan (Fikriya, dkk., 2017).

Studi terbaru menunjukkan bahwa *machine learning* terbagi menjadi tiga kategori, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning* (Roihan, dkk., 2020). Teknik yang digunakan oleh *supervised learning* adalah metode klasifikasi di mana dataset diberi label lengkap untuk klasifikasi ke dalam kelas yang tidak diketahui. Teknik *unsupervised learning* sering disebut *clustering* karena tidak memerlukan label pada data dan hasilnya tidak mengidentifikasi contoh dalam kelas yang telah ditentukan (Thupae, dkk., 2018). Sementara itu, biasanya *reinforcement learning* berada diantara *supervised learning* dan *unsupervised learning*, teknik ini bekerja di dalam lingkungan yang dinamis di mana konsepnya harus menyelesaikan tujuan tanpa adanya pemberitahuan dari komputer secara eksplisit jika tujuan tersebut telah tercapai (Roihan, dkk., 2020).

2.3 Klasifikasi

Klasifikasi merupakan sebuah proses yang bertujuan untuk menempatkan suatu objek ke dalam sebuah kelas atau kategori yang telah ditentukan sebelumnya (Puspitasari, dkk., 2018). Klasifikasi adalah adalah tahapan untuk menemukan pola atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan maksud untuk dapat memprediksi kelas dari suatu objek yang tidak memiliki label (Darwis, dkk., 2020).

Tujuan dari klasifikasi adalah untuk memprediksi kelas target dari setiap data. Misalnya, model klasifikasi dapat membantu mengidentifikasi aplikasi pinjaman bank apakah aman atau berisiko. Berbagai teknik klasifikasi yang digunakan dalam bidang *data mining* adalah *decision tree*, *rule-based method*, *memory-based learning*, *bayesian networks*, *neural networks* dan *support vector machines* (Gupta, dkk., 2017).

2.4 *Naïve Bayes*

Naïve bayes adalah metode klasifikasi berdasarkan teorema bayes (Asfi dan Fitrianingsih, 2020). *Naïve bayes* memprediksi kemungkinan masa depan berdasarkan pengalaman masa lalu. Rumusan umum prediksi bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{2.1}$$

Naïve bayes menggunakan asumsi yang sangat kuat tentang independensi setiap kondisi atau peristiwa yang tidak bergantung satu sama lain. Asumsi ini mengarah pada persamaan berikut (Pujianto, dkk., 2017):

$$P(H|X) = P(H)\prod_{i=1}^{n} P(X_i|H) \tag{2.2}$$

di mana:

X = *data testing* yang kelasnya belum diketahui

H = hipotesis data X yang merupakan suatu kelas yang lebih spesifik

$P(H|X)$ = probabilitas hipotesis H berdasar kondisi X (*posteriori probability*)

$P(X|H)$ = probabilitas hipotesis X berdasarkan kondisi H (*likelihood*)

$P(H)$ = probabilitas hipotesis H (*prior probability*)

$P(X)$ = probabilitas hipotesis X (*predictor prior probability*)

Algoritma *naïve bayes* sangat cocok untuk mengklasifikasikan data bertipe nominal. Untuk data nominal, persamaan digunakan dalam perhitungan (2.1). Apabila *dataset* bertipe numerik maka digunakan perhitungan distribusi *gaussian*. Perhitungan distribusi *gaussian* dapat dilihat dari persamaan (2.3), di mana dihitung terlebih dahulu nilai rata-rata μ sesuai pada persamaan (2.4), dan *standard deviasi* σ sesuai pada persamaan (2.5). Tipe data nominal adalah tipe data yang diperoleh dengan klasifikasi atau kategorisasi yang menunjukkan beberapa entitas yang berbeda seperti kode pos, jenis kelamin, nama kota, dll. Tipe data numerik adalah tipe data yang diperoleh dari pengukuran dimana jarak antara dua titik pada skala diketahui, seperti umur, berat badan, tinggi badan, dll (Asfi dan Fitrianingsih, 2020).

$$f(x) = \frac{1}{\sqrt{2\pi}.\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.3}$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.4}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \tag{2.5}$$

di mana:

$f(x)$ = nilai *gaussian*

x = nilai data

μ = nilai rata-rata (*mean*)

σ = standar deviasi

π = nilai *phi* (3,146 atau 22/7)

e = 2,7183

x_i = nilai data ke-i

n = jumlah data

2.5 *Random Forest*

Metode *random forest* merupakan salah satu metode yang dapat meningkatkan akurasi hasil karena pembangkitan simpul anak dilakukan secara acak untuk setiap node. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan cara mengambil atribut dan data secara acak sesuai dengan peraturan yang berlaku. *Root node* adalah simpul yang berada paling atas dari pohon keputusan, atau biasa disebut sebagai akar. *Internal node* adalah node bercabang, yang memiliki setidaknya dua *output* dan hanya satu *input*. *Leaf node* adalah node terakhir dengan hanya satu *input* dan tidak ada *output*. Pohon keputusan dimulai dengan cara menghitung nilai entropi sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain*. Untuk menghitung nilai entropi digunakan rumus seperti pada persamaan (2.6), sedangkan nilai *information gain* menggunakan persamaan (2.7) (Nugroho dan Emiliyawati, 2017).

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i\ .\ log_2\ p_i \qquad (2.6)$$

di mana:

S = himpunan kasus

n = jumlah partisi S

p_i = proporsi dari S_i terhadap S

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\ .\ Entropy(S_i) \qquad (2.7)$$

di mana:

S = himpunan kasus

A = atribut

n = jumlah partisi atribut A

$|S_i|$ = jumlah kasus pada partisi ke-i

$|S|$ = jumlah kasus dalam S

2.6 *Boosting*

Boosting merupakan algoritma yang berulang dan memberikan nilai pembobotan distribusi data latih yang berbeda setiap perulangan. Pada setiap pengulangan, algoritma *boosting* meningkatkan bobot pada contoh yang salah diklasifikasikan dan menurunkan bobot pada contoh yang diklasifikasikan dengan tepat. Hal ini efektif dalam mengubah distribusi pada data latih (Arrahimi, dkk., 2019). Secara teori, *boosting* bertujuan untuk meningkatkan kinerja algoritma klasifikasi hingga mencapai performa maksimal. Namun, ketika diterapkan pada dataset yang tidak seimbang, *boosting* tidak akan mengubah struktur dataset tersebut. Dengan kata lain, kondisi dataset tetap tidak seimbang (Pristyanto, 2019). Pada *data mining* terdapat beberapa jenis *boosting*, diantaranya sebagai berikut:

1. *Adaptive Boosting* (*AdaBoost*)
   Algoritma *AdaBoost* adalah algoritma *boosting* praktis pertama dan salah satu algoritma yang paling banyak digunakan dan dipelajari dibanyak bidang. *Boosting* dapat dikombinasikan dengan algoritma klasifikasi lainnya untuk meningkatkan kinerja klasifikasi (Pristyanto, 2019).
2. *Gradient Boosting* (GB)
   Gradient boosting termasuk *supervised learning* berbasis pohon keputusan yang dapat digunakan dalam klasifikasi. Algoritma *gradient boosting* bekerja secara berurutan, menambahkan prediksi sebelumnya yang tidak sesuai dengan prediksi dan memastikan kesalahan yang dibuat di masa lalu diperbaiki (Suryana, dkk., 2021).
3. *Extreme Gradient Boosting* (*XGBoost*)
   XGBoost adalah teknik pembelajaran mesin untuk menyelesaikan masalah regresi dan klasifikasi berdasarkan *Gradient Boosting Decision Tree* (GBDT). Secara esensial, *XGBoost* adalah teknik *ensemble* yang bergantung pada *gradient boosting tree* (Karo, 2020).

2.7 *K-Fold Cross Validation*

K-fold adalah salah satu teknik *cross validation* yang terkenal, yaitu melipat *k* data dan membuat percobaan secara berulang-ulang (iterasi) sebanyak *k* pula (Ratnawati, 2018). *Cross validation* merupakan salah satu cara terbaik untuk memvalidasi model. Salah satu teknik *cross validation* yang umum digunakan adalah *k-fold cross validation*, karena metode ini biasanya menghasilkan model yang tidak bias. Hal ini terjadi karena setiap pengamatan dalam data berpeluang menjadi data latih atau data uji (Widyaningsih, dkk., 2021).

Pertama, metode ini membagi (melipat) data menjadi *k* bagian yang sama. Nilai *k* diserahkan kepada peneliti, namun disarankan tidak terlalu besar atau terlalu kecil. Nilai *k* yang terlalu besar menghasilkan model yang tidak bias, tetapi dapat menyebabkan varians dan *overfitting* yang tinggi. Nilai *k* yang terlalu kecil akan membuat model mirip dengan metode *cross validation* biasa yang hanya membagi data dalam *training test* (yang dapat menimbulkan bias). Nilai *k* yang umum digunakan adalah $k = 5$ atau $k = 10$ (Kuhn dan Johnson, 2013).

2.8 *Confusion Matrix*

Confusion matrix adalah suatu metode untuk membuat perhitungan yang akurat tentang konsep *data mining*. Di mana evaluasi *confusion matrix* adalah matriks prediksi yang menguji skor item benar dan salah untuk menghasilkan nilai akurasi, presisi, dan *recall*. Presisi atau *confidence* adalah proporsi kasus yang diprediksi secara positif yang juga benar-benar positif dalam data aktual. *Recall* atau adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar (Sandag, 2020).

Tabel 1. Model *Confusion Matrix*

		Nilai Aktual	
		Possitive	**Negative**
Nilai Prediksi	**Possitive**	*True Positive* (TP)	*False Possitive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

Perhitungan *accuracy*, *precision*, *recall*, dan *Root Mean Squared Error* (RMSE) dengan tabel *confusion matrix* adalah sebagai berikut:

Rumus *accuracy*:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

Rumus *precision*:

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

Rumus *recall*:

$$Recall = \frac{TP}{TP + FN} \tag{2.10}$$

Rumus RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{2.11}$$

di mana:

Y_i = data awal (data sebenarnya)

$\hat{Y}_i$ = data akhir (data hasil estimasi)

n = jumlah data

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester ganjil tahun akademik 2022/2023, bertempat di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan pada penelitian ini adalah data nilai mahasiswa Universitas Lampung semester ganjil tahun akademik 2021/2022. Terdapat 7 variabel dari data tersebut, yaitu periode, NPM, Id Jadwal, kode MK, mata kuliah, nilai angka, dan nilai huruf yang terbagi menjadi 7 bagian penilaian yaitu nilai A (sangat baik) dengan jumlah 107711 data, nilai B+ dengan jumlah 47203 data, nilai B dengan jumlah 24259 data, nilai C+ dengan jumlah 7780 data, nilai C dengan jumlah 4301 data, nilai D dengan jumlah 2103 data, dan nilai E dengan jumlah 4511 data. Data pada penelitian ini berjumlah 197868 data.

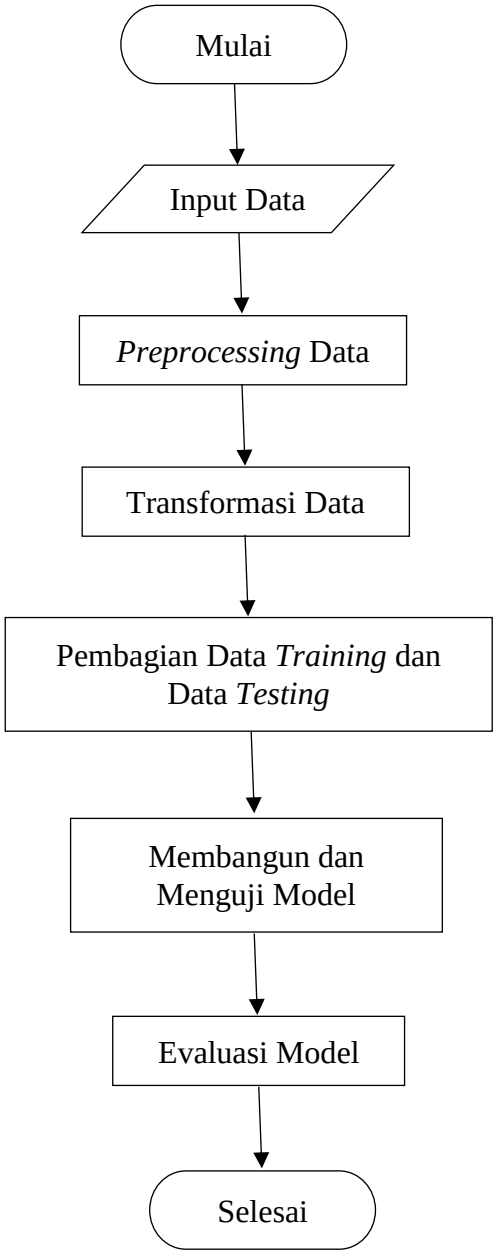
3.3 Metode Penelitian

Penelitian ini menggunakan metode *naïve bayes* dan *random forest* untuk klasifikasi nilai mahasiswa Universitas Lampung dengan menggunakan *software python*. Berikut adalah langkah-langkah yang dilakukan dalam penelitian ini:

a. Melakukan input data nilai mahasiswa Universitas Lampung semester ganjil tahun akademik 2021/2022.
b. Melakukan *pre-processing* data dengan melihat dan menangani *missing value*, melakukan *categorical encoding* untuk variabel mata kuliah dan nilai angka, serta melakukan transformasi data untuk variabel nilai angka.
c. Membagi data *training* dan data *testing* dengan skema:
    1) 60% data *training* dan 40% data *testing*
    2) 70% data *training* dan 30% data *testing*
    3) 80% data *training* dan 20% data *testing*
    4) 90% data *training* dan 10% data *testing*

    Selain dilakukan *splitting* data juga dilakukan metode *k-fold cross validation* dengan k = 5, 8, dan 10.
d. Membangun dan menguji model *naïve bayes* dan *random forest*.
e. Melakukan evaluasi model menggunakan *confusion matrix*.
f. Melakukan optimalisasi model dengan menggunakan algoritma *Gradient Boosting* (GB) untuk model *naïve bayes* dan *random forest*.
g. Membandingkan performa dari kedua metode.

Adapun *flowchart* pada algoritma *naïve bayes* dan *random forest* dapat digambarkan sebagai berikut:



Gambar 2. *Flowchart* Metode *Naïve Bayes* dan *Random Forest*.

V. KESIMPULAN

Telah dilakukan pengujian mengenai metode *naïve bayes* dan *random forest* untuk klasifikasi nilai mahasiswa Universitas Lampung. Berikut merupakan beberapa kesimpulan dari penelitian ini:

1. Berdasarkan hasil dari beberapa teknik pengujian, diperoleh bahwa metode *naïve bayes* dan *random forest* memiliki tingkat akurasi tertinggi jika menggunakan *splitting* data dibandingkan *k-fold cross validation*.
2. Berdasarkan hasil evaluasi model dengan *confusion matrix*, metode *random forest* merupakan metode terbaik untuk mengklasifikasi nilai mahasiswa Universitas Lampung jika dibandingkan dengan metode *naïve bayes* dengan tingkat akurasi sebesar 99,38%.
3. Selanjutnya untuk performa dari kedua metode lebih baik jika sesudah dilakukan *boosting* dengan *Gradient Boosting* (GB), yaitu untuk metode *naïve bayes* sebesar 99,45% dan metode *random forest* sebesar 99,89%.

DAFTAR PUSTAKA

Anggraini, R.A., Widagdo, G., Budi, A.S., dan Qomaruddin, M. 2019. Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes. *Jurnal Sisten dan Teknologi Informasi*. **7**(1): 47-51.

Arrahimi, A. R., Ihsan, M. K., Kartini, D., Faisal, M.R., dan Indriani, F. 2019. Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa. *Jurnal Sains dan Informatika*. **5**(1): 21-30.

Asfi, M. dan Fitrianingsih, N. 2020. Implementasi Algoritma Naïve Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi. *Jurnal Nasional Informatika dan Teknologi Jaringan*. **5**(1): 44-50.

Bramer, M. 2016. *Principles of Data Mining.* 3rd Edition. Springer, USA.

Darwis, D., Pratiwi, E. S., dan Pasaribu, A. F. O. 2020. Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Jurnal Ilmiah Edutic*. **7**(1): 1-11.

Fikriya, Z.A., Irawan, M.I., dan Soetrisno. 2017. Implementasi Extreme Learning Machine untuk Pengenalan Objek Citra Digital. *Jurnal Sains Dan Seni ITS.* **6**(1): 18-23.

Gorunescu, F. 2011. *Data Mining Concepts, Models and Techniques*. Springer, USA.

Gupta, B., Rawat, A., Jain, A., Arora, A., dan Dhami, M. 2017. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*. **163**(8): 15-19.

Han, J., dan Kamber, M. 2006. *Data Mining:Concepts and Techniques.* 2nd Edition. Morgan Kaufmann Publishers, San Francisco.

Hendrian, S. 2018. Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*. **11**(3): 266-274.

Karo, I. M. K. 2020. Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering, Information and Communication Technology*. **1**(1): 10-16.

Kuhn, M. dan Johnson, K. 2013. *Applied Predictive Modeling*. 2nd Edition. Springer, Berlin.

Mitchell, T. M. 1997. *Machine Learning*. McGraw Hill, New York.

Momole, G.M. dan Mailoa, E. 2022. Perbandingan Naïve Bayes dan Random Forest Dalam Klasifikasi Bahasa Daerah. *Jurnal Teknik Informatika dan Sistem Informasi*. **9**(2): 855-863.

Muslim, M.A., Prasetiyo, B., Mawarni, E.L.H., Herowati, A.J., Mirqotussa'adah, Rakmana, S.H., dan Nurzahputra, A. 2019. *Data Mining Algoritma C4.5*. Semarang.

Nugroho, Y.S. dan Emiliyawati, N. 2017. Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest. *Jurnal Teknik Elektro*. **9**(1): 24-29.

Pristyanto, Y. 2019. Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset. *Jurnal TEKNOINFO*. **13**(1): 11-16.

Pujianto, U., Widiyaningtyas, T., Prasetya, D.D., dan Romadhon, D. 2017. Penerapan Algoritma Naïve Bayes Classifier untuk Klasifikasi Judul Skripsi dan Tugas Akhir Berdasarkan Kelompok Bidang Keahlian. *Jurnal Teknologi Elektro dan Kejuruan*. **27**(1): 79-92.

Puspitasari, A.M., Ratnawati, D.E., dan Widodo, A.W. 2018. Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*. **2**(2): 802-810.

Putra, B.J.M. dan Yuniarti, D.A.F. 2020. Analisis Hasil Belajar Mahasiswa dengan Clustering Menggunakan Metode K-Means. *Jurnal POROS TEKNIK*. **12**(2): 49-58.

Ramadhan, A., Susetyo, B., dan Indahwati. 2019. Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan. *Jurnal Pendidikan dan Kebudayaan*. **4**(2): 169-182.

Ratnawati, F. 2018. Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *JURNAL INOVTEK POLBENG-SERI INFORMATIKA*. **3**(1): 50-59.

Retnoningsih, E. dan Pramudita, R. 2020. Mengenal Machine Learning dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python. *Bina Insani ICT Journal*. **7**(2): 156-165.

Roihan, A., Sunarya, P. A., dan Rafika, A. S. 2020. Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *Indonesian Journal on Computer and Information Technology*. **5**(1): 75-82.

Sandag, G.A. 2020. Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. *Cogito Smart Journal*. **6**(2): 167-178.

Suryana, S. E., Warsito, B., dan Suparti. 2021. Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank. *JURNAL GAUSSIAN*. **10**(4): 617-623.

Syaputri, A.W., Irwandi, E., dan Mustakim. 2020. Naïve Bayes Algorithm for Classification of Student Major's Spelization. *J Int Comp & He Inf*. **1**(1): 16-19.

Thupae, R., Isong, B., Gasela, N., dan Mahfouz, A. M. A. 2018. Machine Learning Techniques for Traffic Identification and Classification in SDWSN: A Survey. *IEEE Industrial Electronics Society*. 4645–4650.

Universitas Lampung. 2022. *Peraturan Akademik Universitas Lampung 2022*. Unila, Bandar Lampung.

Widyaningsih, Y., Arum, G. P., dan Prawira, K. 2021. Aplikasi K-Fold Cross Validation Dalam Penentuan Model Regresi Binomial Negatif Terbaik. *Jurnal Ilmu Matematika dan Terapan*. **15**(2): 315-322.

Yusuf, B., Qalbi, M., Basrul, Dwitawati, I., Malahayati, dan Ellyadi, M. 2020. Implementasi Algoritma Naive Bayes Dan Random Forest Dalam Memprediksi Prestasi Akademik Mahasiswa Universitas Islam Negeri Ar-Raniry Banda Aceh. *Jurnal Pendidikan Teknologi Informasi*. **4**(1): 50-58.