PERBANDINGAN MODEL *NAÏVE BAYES* DAN *RANDOM FOREST* DALAM PREDIKSI KLASIFIKASI MASA STUDI SARJANA MATEMATIKA UNIVERSITAS LAMPUNG

Skripsi

Oleh

SHELVIRA HESTINA PUTRI
NPM. 2117031035

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

COMPARISON OF NAÏVE BAYES AND RANDOM FOREST MODELS IN PREDICTING THE STUDY DURATION CLASSIFICATION OF MATHEMATICS UNDERGRADUATES AT THE UNIVERSITY OF LAMPUNG

By

Shelvira Hestina Putri

The advancement of information technology has encouraged the utilization of data mining across various fields, including education. Data mining enables the analysis of large-scale data to discover meaningful patterns and generate predictions based on those patterns. In the context of higher education, the timeliness of student graduation is a crucial indicator related to study program accreditation and institutional efficiency. This study aims to compare the performance of two machine learning-based classification algorithms, namely Naïve Bayes and Random Forest, in predicting the study duration of undergraduate students in the Mathematics Study Program at the University of Lampung. The data used consists of academic records that have been labeled based on whether the students graduated on time or not. Model evaluation was conducted using a confusion matrix and two validation techniques: data splitting and k-fold cross-validation. The test results indicate that the Random Forest algorithm achieved the highest accuracy of 94.44%, outperforming Naïve Bayes. Additionally, both models attained their highest accuracy when using the data splitting method. These findings suggest that Random Forest is a more effective method for classifying the study duration of undergraduate mathematics students at the University of Lampung and provides a strong foundation for predictive efforts to support improvements in on-time graduation rates.

Keywords: Data Mining, Classification, Naïve Bayes, Random Forest, Comparison, Study Duration, On-time Graduation.

ABSTRAK

PERBANDINGAN MODEL *NAÏVE BAYES* DAN *RANDOM FOREST* DALAM PREDIKSI KLASIFIKASI MASA STUDI SARJANA MATEMATIKA UNIVERSITAS LAMPUNG

Oleh

Shelvira Hestina Putri

Perkembangan teknologi informasi telah mendorong pemanfaatan data mining dalam berbagai bidang, termasuk dunia pendidikan. *Data mining* memungkinkan analisis terhadap data berukuran besar untuk menemukan pola yang bermakna dan menghasilkan prediksi berdasarkan pola tersebut. Dalam konteks pendidikan tinggi, ketepatan waktu kelulusan mahasiswa merupakan indikator penting yang berkaitan dengan akreditasi program studi dan efisiensi institusi. Penelitian ini bertujuan untuk membandingkan performa dua algoritma klasifikasi berbasis *machine learning*, yaitu *Naïve Bayes* dan *Random Forest*, dalam memprediksi masa studi mahasiswa Program Studi Sarjana Matematika Universitas Lampung. Data yang digunakan adalah data akademik mahasiswa yang telah diberi label berdasarkan status kelulusan tepat waktu atau tidak. Evaluasi model dilakukan dengan menggunakan *confusion matrix* dan dua teknik validasi, yaitu metode data *splitting* dan *k-fold cross validation*. Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memiliki akurasi tertinggi sebesar 94,44%, mengungguli *Naïve Bayes*. Selain itu, kedua model mencapai hasil akurasi tertinggi saat menggunakan metode data *splitting*. Hasil penelitian ini menunjukkan bahwa *Random Forest* merupakan metode yang lebih baik dalam klasifikasi masa studi sarjana Matematika Universitas Lampung.

Kata-kata kunci: Data Mining, klasifikasi, Naïve Bayes, Random Forest, Perbandingan, Masa Studi, Kelulusan Tepat Waktu.

PERBANDINGAN MODEL *NAÏVE BAYES* DAN *RANDOM FOREST* DALAM PREDIKSI KLASIFIKASI MASA STUDI SARJANA MATEMATIKA UNIVERSITAS LAMPUNG

SHELVIRA HESTINA PUTRI

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	***PERBANDINGAN MODEL NAÏVE BAYES DAN RANDOM FOREST DALAM PREDIKSI KLASIFIKASI MASA STUDI SARJANA MATEMATIKA UNIVERSITAS LAMPUNG***
Nama Mahasiswa	:	Shelvira Hestina Putri
Nomor Pokok Mahasiswa	:	**2117031035**
Program Studi	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Widiarti, S.Si.,M.Si.	**Dr. Aang Nuryaman, S.Si., M.Si.**
NIP 198005022005012003	NIP. 197403162005011001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si.,M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Widiarti, S.Si.,M.Si.** .....

Sekretaris : **Dr. Aang Nuryaman, S.Si., M.Si.** .....

Penguji
Bukan Pembimbing : **Prof. Drs. Mustofa Usman, M.A., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi : **03 Juni 2025**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	:	**Shelvira Hestina Putri**
Nomor Pokok Mahasiswa	:	**2117031035**
Jurusan	:	**Matematika**
Judul Skripsi	:	**Perbandingan Model *Naïve Bayes* dan *Random Forest* dalam Prediksi Klasifikasi Masa Studi Sarjana Matematika Universitas Lampung**

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

Bandar Lampung, 03 Juni 2025

Penulis.

Shelvira Hestina Putri

RIWAYAT HIDUP

Penulis memiliki nama lengkap Shelvira Hestina Putri yang lahir di Penumangan pada tanggal 27 Juli 2003. Penulis merupakan anak kedua dari tiga bersaudara dari pasangan Bapak Ali Basri dan Ibu Citra Trisnawati.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) RA. Nurul Muttaqim pada tahun 2009-2010. Kemudian menepuh pendidikan Sekolah Dasar (SD) di SDN 01 Penumangan Baru pada tahun 2010-2015. Kemudian melanjutkan ke Sekolah Menengah Pertama di SMP Bina Desa pada tahun 2015-2018, Sekolah Menengah Atas di SMAN 1 Tulang Bawang Tengah pada tahun 2018-2021. Pada tahun 2021 penulis terdaftar sebagai mahasiswa Program Studi S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN.

Pada tahun 2023 penulis melakukan Kuliah Kerja Praktik (KP) di Dinas Perhubungan Kota Bandar Lampung. Kemudian pada tahun 2024 penulis mengikuti Program Magang dan Studi Independen Bersertifikat (MSIB) Kampus Merdeka Cycle 6 di PT. Bank Rakyat Indonesia (Persero) Tbk. pada Regional Office Bandar Lampung serta mengikuti kegiatan Kuliah Kerja Nyata (KKN) di Desa Braja Harjosari, Lampung Timur.

KATA INSPIRASI

”Allah tidak mengatakan hidup ini mudah. Tapi Allah berjanji, bahwa sesungguhnya bersama kesulitan ada Kemudahan.”
(Qs.Al baqarah:5-6)

”Hatiku tenang mengetahui apa yang melewatkanku tidak akan pernah manjadi takdirku, dan apa yang ditakdirkan untukku tidak akan pernah melewatkanku”
(Umar bin Khattab)

”Jalani hidupmu seolah-olah semuanya dirancang untuk kebaikanmu”
(Jalaludin Rumi)

”Jika tidak hari ini mungkin minggu depan
jika tidak minggu ini mungkin bulan depan
jika tidak bulan ini mungkin tahun depan
segala harapan kan datang
yang kita impikan”
(Batas Senja - Kita Usahakan Lagi)

”Apapun yang terjadi, Pulanglah sebagai sarjana”

PERSEMBAHAN

Tidak ada lembar yang paling berarti dan paling indah dalam skripsi ini kecuali lembar persembahan. Dengan mengucap syukur Alhamdulillah sungguh sebuah perjuangan cukup panjang yang telah saya lalui untuk dapat menyelesaikan skripsi ini demi mendapatkan gelar yang sudah saya impikan dari lama. Rasa syukur dan bahagia yang saya rasakan ini akan saya persembahkan juga kepada orang-orang yang sangat berarti dalam proses perjalanan saya, karena berkat doa dan dukungan dari mereka saya bisa menyelesaikan skripsi ini dengan baik. Skripsi ini penulis persembahkan kepada:

Papa dan Mama Tercinta

Terimakasih kepada orang tuaku atas segala pengorbanan, motivasi, doa dan ridho 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 banyak orang.

Dosen Pembimbing dan Pembahas

Terimakasih 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

Alhamdulillah, puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul ”Perbandingan Model *Naïve Bayes* dan *Random Forest* dalam Prediksi Klasifikasi Masa Studi Sarjana Matematika Universitas Lampung” dengan baik dan lancar serta tepat pada waktu yang telah ditentukan. Shalawat serta salam semoga senantiasa tercurahkan kepada Nabi Muhammad SAW.

Dalam proses penyusunan skripsi ini, banyak pihak yang telah membantu memberikan bimbingan, dukungan, arahan, motivasi serta saran sehingga skripsi ini dapat terselesaikan. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Ibu Widiarti, S.Si., M.Si. selaku Pembimbing 1 yang senantiasa memberikan arahan, bimbingan, motivasi, saran serta dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
2. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Pembimbing II yang telah memberikan arahan, bimbingan dan dukungan kepada penulis sehingga dapat menyelesaikan skripsi ini.
3. Bapak Prof. Drs. Mustofa Usman, M.A., Ph.D. selaku Penguji yang telah bersedia memberikan kritik dan saran serta evaluasi kepada penulis sehingga dapat menjadi lebih baik lagi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Bapak Ir. Warsono, Ph.D. selaku dosen pembimbing akademik.
6. Seluruh dosen, staff dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Kedua orang tua tersayang, Papa Ali Basri dan Mama Citra Trisnawati. Terima kasih penulis ucapkan atas segala pengorbanan, kerja keras dan kasih sayang yang selalu diberikan. Doa, motivasi dan dukungan mereka menjadi kekuatan terbesar hingga saya berhasil menyelesaikan skripsi ini dan meraih gelar Sarjana Matematika. Besar harapan penulis semoga papa dan mama selalu sehat, panjang umur dan bisa menyaksikan keberhasilan lainnya yang akan penulis raih di masa yang akan datang.

8. Seluruh keluarga besar tercinta abang, ahun, yayik, yayik, binda, muda, pauda, bikcik, pakcik, walid, walida, sunda, manda, biksu dan paksu yang telah banyak memberikan dukungan, bantuan dan doa serta hiburan sehingga penulis dapat menyelesaikan skripsi ini dan lulus tepat waktu.

9. Teruntuk sahabat-sahabat tercinta Salma, Lola, Rhea dan Dewi terimakasih atas segala motivasi, dukungan, pengalaman, waktu dan ilmu yang dijalani bersama selama perkuliahan. Terima kasih selalu menjadi garda terdepan di masa-masa sulit penulis. Terima kasih selalu mendengarkan keluh kesah penulis. Ucapan syukur kepada Allah SWT karena telah memberikan sahabat terbaik seperti kalian. *see you on top, guys!*.

10. Teman-teman seperjuangan yaitu Bue, Irma, Dede, Windi, Dita, Siska dan Rosa yang telah menemani dan membersamai penulis dalam meyelesaikan skripsi ini, terimakasih atas petualangan yang luar biasa, kenangan canda tawa yang sangat menyenangkan dan berkesan bagi penulis.

11. Semua pihak yang tidak dapat disebutkan satu persatu yang telah banyak membantu memberikan pemikiran demi kelancaran dan keberhasilan penyusunan skripsi ini.

Semoga skripsi ini dapat bermanfaat bagi kita semua. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, sehingga penulis mengharapkan kritik dan saran yang membangun untuk menjadikan skripsi ini lebih baik lagi.

Bandar Lampung, 03 Juni 2025

Shelvira Hestina Putri

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi memberikan kemudahan bagi manusia dalam menyelesaikan berbagai permasalahan yang kompleks dan beragam seiring dengan kemajuan zaman yang semakin pesat. Salah satu contohnya adalah *data mining*, Witten *et al*. (2011) menjelaskan bahwa *data mining* adalah penerapan teknik *machine learning* pada data besar untuk menemukan pola atau struktur yang signifikan. *Data mining* berfokus pada analisis data yang besar, dengan tujuan menemukan informasi yang bermakna dan membuat prediksi berdasarkan pola yang ditemukan. Dalam *machine learning*, terdapat dua jenis teknik pembelajaran yaitu *supervised learning* dan *unsupervised learning*. *Supervised learning* adalah salah satu jenis *machine learning* di mana model dilatih menggunakan data yang sudah diberi label (Hastie *et al*, 2019). Klasifikasi merupakan salah satu kategori dalam metode *supervised learning*.

Klasifikasi dapat diartikan sebagai proses memetakan data ke kelas-kelas yang telah ditentukan berdasarkan aturan yang dipelajari dari data pelatihan (Bishop, 2006). Tujuan utama dari klasifikasi adalah membangun model atau algoritma yang mampu memprediksi kelas dari data baru yang sebelumnya belum pernah ditemui, dengan mengandalkan pola yang dipelajari dari data pelatihan. Dalam mengevaluasi kinerja berbagai metode klasifikasi, sangat penting untuk melakukan analisis perbandingan yang mendalam antar metode data mining guna mengetahui metode mana yang paling efisien dan akurat.

Naïve Bayes merupakan salah satu metode yang dapat digunakan untuk mengklasifikasikan data. Tujuan dari algoritma ini adalah untuk mengklasifikasi

kan data kedalam kategori tertentu (Yusuf, dkk., 2020). Pada penelitian sebelumnya yang dilakukan oleh Andhini, dkk. (2022) tentang klasifikasi penerima Program Keluarga Harapan (PKH) menyatakan bahwa hasil klasifikasi penerima Program Keluarga Harapan (PKH) dibagi menjadi dua kelas klasifikasi yaitu kelas layak dan tidak layak. Hasil pengolahan data menunjukkan akurasi klasifikasi penerima Program Keluarga Harapan (PKH) sebesar 88% yang masuk dalam kategori *Good Classification*. Selain algoritma *naïve bayes*, ada juga algoritma *random forest* yang bertujuan untuk mengklasifikasikan kelas secara akurat (Yusuf, dkk., 2020). Studi terdahulu yang dilakukan oleh Widya, dkk. (2021) tentang prediksi kemungkinan diabetes pada tahap awal, menyebutkan bahwa hasil evaluasi model menunjukkan akurasi klasifikasi pada pemodelan dengan metode *random forest* mencapai 97,88%.. Kemudian Ricky, dkk. (2020) juga melaksanakan studi terkait perbandingan *naïve bayes* dan *random forest* dalam prediksi keberhasilan klien telemarketing, dan hasil yang diperoleh metode *random forest* dalam prediksi keberhasilan klien telemarketing sangat baik, karena memberikan nilai akurasi sebesar 90%, jika dibandingkan dengan *naïve bayes* yang memperoleh tingkat akurasi sebesar 85%.

Kelulusan adalah hasil akhir dari proses kegiatan belajar mengajar selama mengikuti perkuliahan di perguruan tinggi (Dwi dan Abdul, 2016). Keberhasilan suatu perguruan tinggi dalam melaksanakan program pendidikan dapat diukur dari berbagai indikator, salah satunya adalah ketepatan waktu kelulusan mahasiswa. Ketepatan waktu kelulusan mahasiswa menjadi fokus utama karena berkaitan dengan akreditasi program studi dan efisiensi operasional. Selain itu, kelulusan tepat waktu juga memiliki dampak signifikan terhadap citra akademik universitas dan peluang mahasiswa di dunia kerja.

Universitas Lampung (UNILA) merupakan salah satu perguruan tinggi negeri yang berlokasi di Kota Bandar Lampung, Provinsi Lampung. Di Universitas Lampung, khususnya Program Studi Sarjana Matematika, berkomitmen untuk meningkatkan jumlah mahasiswa yang lulus tepat waktu. Namun, kenyataannya masih ada beberapa mahasiswa yang mengalami keterlambatan dalam menyelesaikan studi mereka. Situasi ini menimbulkan kekhawatiran karena rendahnya tingkat kelulusan tepat waktu dapat berdampak pada reputasi program studi, kualitas pendidikan, dan proses akreditasi institusi. Oleh sebab itu, diperlukan analisis mendalam mengenai faktor-faktor yang me-

mengaruhi kelulusan mahasiswa serta pengembangan model prediksi untuk mengidentifikasi potensi keterlambatan tersebut.

Kelulusan tepat waktu merupakan pencapaian yang dipengaruhi oleh banyak variabel. Beberapa faktor yang berpengaruh meliputi prestasi akademik, tingkat partisipasi dalam kegiatan perkuliahan, hingga faktor eksternal seperti kondisi ekonomi atau dukungan keluarga. Selain itu, faktor non-akademik seperti kesehatan mental dan kemampuan manajemen waktu juga turut berperan dalam menentukan ketepatan waktu kelulusan mahasiswa. Oleh karena itu, diperlukan upaya prediktif untuk mengidentifikasi mahasiswa yang berpotensi terlambat dalam menyelesaikan studi agar pihak Universitas dapat melakukan intervensi lebih awal.

Oleh sebab itu, penulis akan menggunakan metode *naïve bayes* dan *random forest* dalam melakukan klasifikasi terhadap masa studi sarjana Matematika di Universitas Lampung, dengan tujuan untuk mengetahui metode mana yang lebih unggul sebagai teknik klasifikasi melalui perbandingan nilai akurasinya.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

(a) Menerapkan metode *naïve bayes* dan *random forest* untuk melakukan klasifikasi guna memperoleh hasil dengan tingkat akurasi yang optimal.

(b) Melakukan perbandingan terhadap nilai akurasi dari metode *naïve bayes* dan *random forest* untuk menentukan metode mana yang dapat dianggap sebagai teknik klasifikasi yang lebih baik.

1.3 Manfaat Penelitian

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

1.4 Batasan Penelitian

(a) Data yang digunakan dalam penelitian ini berasal dari data yudisium alumni mahasiswa jurusan Matematika Universitas Lampung.

(b) Proses klasifikasi dilakukan semata-mata menggunakan algoritma *naïve bayes* dan *random forest*, tanpa mengacu pada peraturan akademik yang berlaku.

(c) Penelitian ini hanya berfokus pada aspek akademik dan tidak mencakup masalah internal mahasiswa.

BAB 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). Tujuan dari *data mining* adalah mengekstraksi informasi yang berguna dari himpunan data yang kompleks dan mengubahnya menjadi struktur yang dapat dipahami untuk penggunaan di masa depan (Gnanapriya, *et al*., 2010). *Data mining* kerap dimanfaatkan di berbagai sektor, seperti bisnis, kesehatan, pemasaran, dan penelitian ilmiah, guna mendukung proses pengambilan keputusan yang didasarkan pada data.

Menurut Han dan Kamber. (2006), *data mining* merupakan langkah penting dalam proses yang lebih luas yang disebut *Knowledge Discovery in Databases* (KDD), yang mencakup tahap-tahap berikut:

(a) Seleksi data: memilih data yang relevan dari basis data.

(b) Pra-pemrosesan data: membersihkan dan mengubah data untuk menghilangkan *noise* dan menangani data yang hilang.

(c) Transformasi data: mengonversi data ke dalam bentuk yang sesuai untuk ditambang, misalnya dengan normalisasi atau pengelompokan atribut.

(d) Penambangan data: menerapkan teknik guna mengidentifikasi pola atau model yang tersembunyi dalam data.

(e) Evaluasi pola: mengevaluasi pola yang ditemukan berdasarkan kriteria tertentu, seperti validitas statistik atau relevansi bisnis.

(f) Representasi pengetahuan: menyajikan pola atau model yang ditemukan dalam bentuk yang mudah dipahami oleh pengguna.

2.2 *Machine Learning*

Machine Learning adalah cabang dari kecerdasan buatan (AI) yang memungkinkan komputer untuk mempelajari dan membuat keputusan berdasarkan data. Alih-alih diprogram dengan instruksi spesifik, algoritma *machine learning* mengenali pola dari data untuk menyelesaikan tugas atau membuat prediksi (Alpaydin, 2021). Dalam *machine learning*, ada data latih dan data uji, data latih untuk melatih algotirma *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).

Machine Learning dibagi ke dalam tiga kategori utama berdasarkan cara sistem belajar nya yaitu, *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Pada *supervised learning*, model dilatih menggunakan dataset yang telah diberi label, yang berarti bahwa setiap input dalam data tersebut terkait dengan output yang diharapkan. Algoritma ini belajar dengan memetakan input ke output berdasarkan pola dalam data (Mitchell, 1997). Dalam *unsupervised learning*, data yang digunakan tidak memiliki label atau output yang diketahui. Algoritma mencoba menemukan pola atau struktur tersembunyi di dalam data tersebut (Bishop, 2006). Sementara itu, *reinforcement learning* berada di antara *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 dapat diartikan sebagai proses memetakan data ke kelas-kelas yang telah ditentukan berdasarkan aturan yang dipelajari dari data pelatihan (Bishop, 2006). Tujuan utama dari klasifikasi adalah membuat model atau algoritma yang dapat memprediksi kelas dari data baru yang belum pernah dilihat sebelumnya berdasarkan pola dari data pelatihan. Berbagai teknik klasifikasi yang digunakan dalam bidang *data mining* adalah *decision tree*, *rule-based method*, *memory-based learning*, *bayesian network*, *neural network*, dan *support vektor machines* (Gupta, dkk., 2017).

2.4 *Naïve Bayes*

Naïve bayes adalah metode klasifikasi probabilistik yang digunakan untuk memprediksi kemungkinan di masa depan berdasarkan pengalaman masa lalu. Metode ini didasarkan pada Teorema Bayes, yang menjelaskan tentang peluang sebuah kejadian berdasarkan pengetahuan awal (prior) dari kondisi yang berhubungan dengan kejadian tersebut atau dikenal sebagai teorema yang melakukan prediksi probabilitas di masa depan dengan menggunakan dasar dari pengalaman yang ada di masa sebelumnya (Sari, 2016).

Misalkan terdapat suatu kejadian dimana pada kejadian tersebut terdapat suatu percobaan yang memberikan hasil dua kemungkinan peristiwa yang akan terjadi, yaitu peristiwa H dan peristiwa X, dimana kedua peristiwa tersebut harus saling berhubungan *(dependent)*. Maka terjadinya peristiwa X akan memiliki pengaruh terhadap peluang terjadinya peristiwa H, yang dapat didefinisikan sebagai berikut (Larose, 2006):

$$P(H|X) = \frac{P(H \cap X)}{P(X)}; \quad P(X) > 0 \tag{2.4.1}$$

$P(H \cap X)$ adalah probabilitas interaksi H dan X lalu $P(X)$ adalah probabilitas X. Selanjutnya, untuk peluang bersyarat X, jika H diketahui maka dilambangkan dengan $P(X|H)$, didefinisikan sebagai berikut (Walpole, 1993):

$$P(X|H) = \frac{P(X \cap H)}{P(H)}; \quad P(H) > 0 \tag{2.4.2}$$

Persamaan (2.4.1), nilai $P(H \cap X) = P(H|X)P(X)$, dan pada persamaan (2.4.2) nilai $P(X \cap H) = P(X|H)P(H)$. Berdasarkan teori himpunan diketahui bahwa $P(H \cap X) = P(X \cap H)$ sehingga diperoleh:

$$\begin{aligned} P(H \cap X) &= P(X \cap H) \\ P(H|X)P(X) &= P(X|H)P(H) \\ P(H|X) &= \frac{P(X|H)P(H)}{P(X)} \end{aligned} \tag{2.4.3}$$

Persamaan (2.4.3) merupakan rumus Teorema Bayes, di mana $P(H|X)$ merupakan probabilitas hipotesis H berdasarkan kodisi X *(posterior probability)* (Han & Kamber 2006).

Diberikan teorema bayes sebagai dasar algoritma *naïve bayes classifier*:

$$P(H_k \mid X_i) = \frac{P(X_i \mid H_k) \cdot P(H_k)}{P(X_i)} \tag{2.4.4}$$

dimana:

X = *data testing yang kelasnya belum diketahui*

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

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

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

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

$P(X)$ = probabilitas total data X tanpa mempertimbangkan kelas (*evidence*)

Naïve bayes classifier mengasumsikan bahwa setiap variabel X_i bersifat independen secara kondisional terhadap variabel lainnya, jika diketahui kelas H_k. Ini berarti bahwa kemunculan satu variabel tidak mempengaruhi kemunculan variabel lainnya.
Tujuan dari algoritma *naïve bayes classifier* adalah mencari kelas H_k yang memaksimalkan probabilitas posterior $P(H_k \mid X_i)$. Oleh karena itu, proses klasifikasi dilakukan dengan memilih kelas yang memiliki nilai probabilitas posterior tertinggi (Raschka & Mirjalili, 2022).

$$\hat{H} = \arg\max_{H_k} P(H_k \mid X_i) \tag{2.4.5}$$

Artinya kita memilih kelas H_k yang memiliki nilai $P(H_k \mid X_i)$ paling tinggi.

Berdasarkan Teorema Bayes, persamaan tersebut dapat ditulis ulang sebagai:

$$\hat{H} = \arg\max_{H_k} \frac{P(X_i \mid H_k) \cdot P(H_k)}{P(X_i)} \tag{2.4.6}$$

Aturan probabilitas total menyatakan bahwa jika suatu himpunan peristiwa $H_1, H_2, ..., H_k$ membentuk partisi dari ruang sampel S artinya:

$$H_k \cap H_j = \emptyset \quad \text{jika } k \neq j \quad \text{(mutually exclusive)}$$

$$\bigcup_{k=1}^{n} H_k = S \quad \text{(collectively exhaustive)}$$

maka untuk sembarang peristiwa X_i, probabilitas dari X_i dapat dihitung sebagai:

$$P(X_i) = \sum_{k} P(X_i \mid H_k) \cdot P(H_k) \tag{2.4.7}$$

Karena $P(X_i)$ adalah probabilitas total dari data X_i, nilainya tidak tergantung pada kelas H_k maka nilainya sama dan bersifat konstan untuk semua kelas (Mitchell, 1997). Dengan demikian, pada proses pencarian nilai maksimum (*argmax*), komponen $P(X_i)$ tidak mempengaruhi hasil akhir. Jadi pada saat proses perbandingan kelas, kita dapat mengabaikannya.

sehingga untuk klasifikasi menggunakan algoritma *naïve bayes classifier* cukup digunakan:

$$P(H_k \mid X_i) \propto P(H_k) \cdot \prod_{i=1}^{n} P(X_i \mid H_k) \tag{2.4.8}$$

Dan klasifikasi *naïve bayes classifier* dilakukan dengan menghitung:

$$\hat{H} = \arg\max_{H_k} \left[P(H_k) \cdot \prod_{i=1}^{n} P(X_i \mid H_k) \right] \tag{2.4.9}$$

Penyederhanaan ini menjadi dasar pengambilan keputusan dalam algoritma *naïve bayes classiffier*, yaitu memilih kelas dengan nilai *probabilitas* tertinggi antara *prior* dan *likelihood*.

Langkah-langkah algortima *naïve bayes*:

(a) Memisahkan Data Berdasarkan Kelas Target
Data dibagi ke dalam kelompok berdasarkan nilai dari variabel target (kelas), yaitu kelas ”Tepat Waktu” dan ”Terlambat”.

(b) Menghitung Probabilitas Prior $P(H_k)$
Probabilitas awal dari masing-masing kelas dihitung berdasarkan proporsi jumlah data di setiap kelas terhadap total data. Probabilitas ini digunakan sebagai referensi awal sebelum mempertimbangkan fitur-fitur lainnya.

(c) Menghitung Probabilitas Likelihood $P(X_i|H_k)$
Untuk setiap fitur, dihitung probabilitas kemunculannya dalam masing-masing kelas.

(d) Menghitung Probabilitas Posterior $P(H_k|X_i)$
Dengan menggunakan Teorema Bayes, dihitung probabilitas bahwa suatu data termasuk ke dalam kelas tertentu berdasarkan nilai-nilai fitur yang dimilikinya.

(e) Menentukan Kelas Prediksi
Kelas dengan nilai probabilitas posterior tertinggi akan dipilih sebagai hasil prediksi akhir untuk data tersebut.

2.5 *Random Forest*

Random forest merupakan metode *supervised learning* yang menggabungkan banyak pohon keputusan (*decision trees*) untuk meningkatkan akurasi prediksi. *Random forest* adalah kombinasi pohon keputusan sedemikian rupa sehingga setiap pohon bergantung pada nilai vektor acak yang diambil sampelnya secara independen dan dengan distribusi yang sama untuk setiap pohon keputusan (Breiman, 2001). Algoritma ini menggunakan metode *bagging* (*bootstrap aggregating*), di mana sejumlah subset acak dari data pelatihan diambil untuk membangun masing-masing pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* (Bishop, 2006). *Root node* adalah node pertama atau tertinggi dalam sebuah pohon keputusan. *internal node* adalah node yang memiliki percabangan, yang setidaknya memiliki dua *output* dan hanya satu *input*. Sedangkan *leaf node* yang juga dikenal sebagai terminal node, adalah node terakhir dalam sebuah pohon dengan satu *input* dan tidak memiliki *output*.

Pohon keputusan dimulai dengan cara menghitung nilai entropi sebagai penentu tingkat ketidakmurniaan atribut dan nilai *information gain* (Nugroho dan Emiliyawati, 2017). Untuk menghitung nilai entropi menggunakan rumus pada persamaan (2.5.3), sedangkan untuk menghitung nilai *information gain* menggunakan persamaan (2.5.4).

$$Entropy(S) = \sum_{i=1}^{n} -p_i \cdot \log_2 p_i \tag{2.5.10}$$

dimana:

S = himpunan data

p_i = proporsi data pada kelas ke i

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot Entropy(S_i) \tag{2.5.11}$$

dimana:

$Entropy(S)$ = entropi awal seluruh data

A = atribut yang di uji

n = jumlah partisi atribut A

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

$|S|$ = jumlah data dalam S

Langkah-langkah algoritma *random forest*:

(a) Membuat Sejumlah *Bootstrap* Sample dari Data Latih
Dataset latih diambil secara acak dengan pengembalian (sampling with replacement) untuk membentuk beberapa subset data. Setiap subset akan digunakan untuk membangun satu pohon keputusan.

(b) Membentuk Pohon Keputusan dari Masing-Masing Sample
Untuk setiap subset data, dibentuk satu pohon keputusan. Pada setiap node, pemilihan fitur dilakukan secara acak dari sebagian kecil fitur yang tersedia, bukan dari keseluruhan fitur. Ini meningkatkan keanekaragaman antar pohon dan mengurangi korelasi.

(c) Mengulang Proses Hingga Terbentuk Sejumlah Pohon *(n estimators)*
Proses pembuatan pohon diulang hingga terbentuk sejumlah pohon sesuai parameter *n estimators*, misalnya 100 pohon.

(d) Melakukan Prediksi dengan Voting Mayoritas
Untuk data uji, prediksi dilakukan oleh setiap pohon, dan hasil akhir ditentukan berdasarkan voting mayoritas dari seluruh pohon.

(e) Menentukan Kelas Berdasarkan Suara Terbanyak
Kelas yang paling sering dipilih oleh semua pohon akan dijadikan prediksi akhir.

2.6 Metode Pembagian Data

Dalam proses pembangunan model klasifikasi, pembagian data merupakan tahapan penting untuk memastikan bahwa model tidak hanya bekerja baik pada data latih, tetapi juga mampu melakukan generalisasi pada data yang belum pernah dilihat sebelumnya. Dua metode umum yang digunakan adalah *Train-Test Split* dan *K-fold cross validation*.

2.6.1 *Train-Test Split*

Train-test split adalah metode pembagian data yang digunakan untuk mengevaluasi performa model pembelajaran mesin secara objektif. Dalam metode ini, dataset dibagi menjadi dua bagian utama, yaitu data latih *data training* dan data uji *data testing*. Data latih digunakan untuk membangun atau melatih model, sedangkan data uji digunakan untuk mengukur seberapa baik model dapat mengeneralisasi pola terhadap data yang belum pernah dilihat sebelumnya (Han & Kamber, 2006).

Pembagian ini biasanya dilakukan secara acak dengan rasio umum seperti 80:20 atau 70:30, yang berarti 80% atau 70% data digunakan untuk pelatihan, dan sisanya untuk pengujian. Pemilihan proporsi ini bergantung pada jumlah data yang tersedia dan kompleksitas permasalahan.

2.6.2 *K-fold cross validation*

K-fold cross validation adalah teknik validasi model yang digunakan untuk mengevaluasi performa algoritma pembelajaran mesin secara lebih andal. Menurut Geron, (2023) *k-fold cross validation* merupakan salah satu metode evaluasi yang paling sering diterapkan. karena memberikan keseimbangan antara bias dan variansi tanpa perlu meninggalkan data dalam jumlah besar untuk pengujian. Karena metode ini biasanya menghasilkan model yang tidak bias sehingga setiap pengamatan dalam data berpeluang menjadi data latih atau data uji (Widyaningsih, dkk., 2021).

Proses ini melibatkan pembagian dataset menjadi k subset atau ”folds” yang sama besar. Model dilatih k kali, setiap kali menggunakan $k-1$ subset untuk pelatihan dan satu subset yang berbeda untuk pengujian. Proses ini diulang

hingga semua subset telah digunakan sebagai set pengujian sekali. Nilai performa keseluruhan dihitung sebagai rata-rata dari performa model pada setiap iterasi pengujian. Biasanya nilai k yang umum digunakan adalah $k = 5$ atau $k = 10$ (kuhn dan Johnson, 2013).

2.7 *Confusion Matriks*

Confusion matriks merupakan alat yang digunakan untuk mengevaluasi performa model klasifikasi dengan cara membandingkan hasil prediksi model terhadap data sebenarnya. *Confusion matriks* memberikan gambaran jelas tentang bagaimana model klasifikasi melakukan kesalahan dalam prediksi dan membedakan antara prediksi yang benar dan salah (Geron, 2023). Matriks ini menampilkan prediksi yang benar dan salah dalam bentuk tabel yang dibagi menjadi empat bagian, yaitu:

(a) *True Positive* (TP): Kasus di mana model memprediksi kelas positif dan prediksinya benar.

(b) *True Negative* (TN): Kasus di mana model memprediksi kelas negatif dan prediksinya benar.

(c) *False Positive* (FP): Kasus di mana model memprediksi kelas positif, tetapi sebenarnya kelas tersebut negatif (dikenal juga sebagai *Type I Error*).

(d) *False Negative* (FN): Kasus di mana model memprediksi kelas negatif, tetapi sebenarnya kelas tersebut positif (*Type II Error*).

Tabel 2.1 Confusion Matriks

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Dengan menggunakan *confusion matriks*, kita dapat menghitung berbagai metrik untuk mengevaluasi performa model lebih mendalam, seperti:

(a) Akurasi: mengukur seberapa sering model membuat prediksi yang benar.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7.12}$$

(b) Presisi: mengukur seberapa akurat prediksi positif yang dilakukan oleh model.

$$\text{Presisi} = \frac{TP}{TP + FP} \tag{2.7.13}$$

(c) *Recall*: mengukur seberapa baik model dapat menemukan semua sampel positif dalam dataset.

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{2.7.14}$$

(d) *F1-Score*: rata-rata harmonis dari presisi dan recall. Ini sering digunakan sebagai metrik yang lebih seimbang antara presisi dan recall.

$$\textit{F1-Score} = 2 \times \frac{\text{Presisi} \times \textit{Recall}}{\text{Presisi} + \textit{Recall}} \tag{2.7.15}$$

Sementara itu, hasil evaluasi model menggunakan *confusion matriks* dapat diterapkan dalam konsep peluang bersyarat, berikut perhitungan peluang berdasarkan konsep peluang bersyarat:

(a) $P(P^+|A^+)$: peluang nilai prediksi positif saat nilai aktual positif *(True Positive)*

(b) $P(P^-|A^+)$: peluang nilai prediksi negatif saat nilai aktual positif *(False Negative)*

(c) $P(P^+|A^-)$: peluang nilai prediksi positif saat nilai aktual negatif *(False Positive)*

(d) $P(P^-|A^-)$: peluang nilai prediksi negatif saat nilai aktual negatif *(True Negative)*

BAB III

METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun ajaran 2024/2025 di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang beralamatkan di Jalan Prof. Dr. Ir. Soemantri Brojonegoro, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Lampung.

3.2 Data Penelitian

Penelitian ini menggunakan data sekunder, yaitu data yudisium alumni mahasiswa jurusan Matematika Universitas Lampung dari tahun 2020-2024 dengan jumlah data sebanyak 537 data yang diperoleh dari Badan Administrasi Akademik Jurusan Matematika Universitas Lampung.
Penelitian ini melibatkan dua tipe variabel yakni variabel terikat dan variabel bebas, variabel terikat berupa waktu kelulusan mahasiswa, sedangkan variabel bebas meliputi jalur masuk universitas, beasiswa, organisasi, jumlah SKS lulus dan nilai Indeks Prestrasi Kumulatif (IPK).

3.3 Metode Penelitian

Penelitian ini menerapkan metode *naïve bayes* dan *random forest* dalam klasifikasi masa studi sarjana Matematika Universitas Lampung dengan menggunakan bahasa pemrogramman *Python*. Langkah-langkah yang dilakukan pada penelitian ini sebagai berikut:

(a) melakukan input data yudisium mahasiswa/i alumni mahasiswa jurusan Matematika Universitas Lampung.

(b) melakukan *pre-processing* data dengan melihat dan menangani *missing value*, *data reduction*, *categorical encoding* serta melakukan transformasi data.

(c) Membagi data *training* dan data *testing* dengan skema:

   i. 60% data *training* dan 40% data *testing*

   ii. 70% data *training* dan 30% data *testing*

   iii. 80% data *training* dan 20% data *testing*

   iv. 90% data *training* dan 10% data *testing*

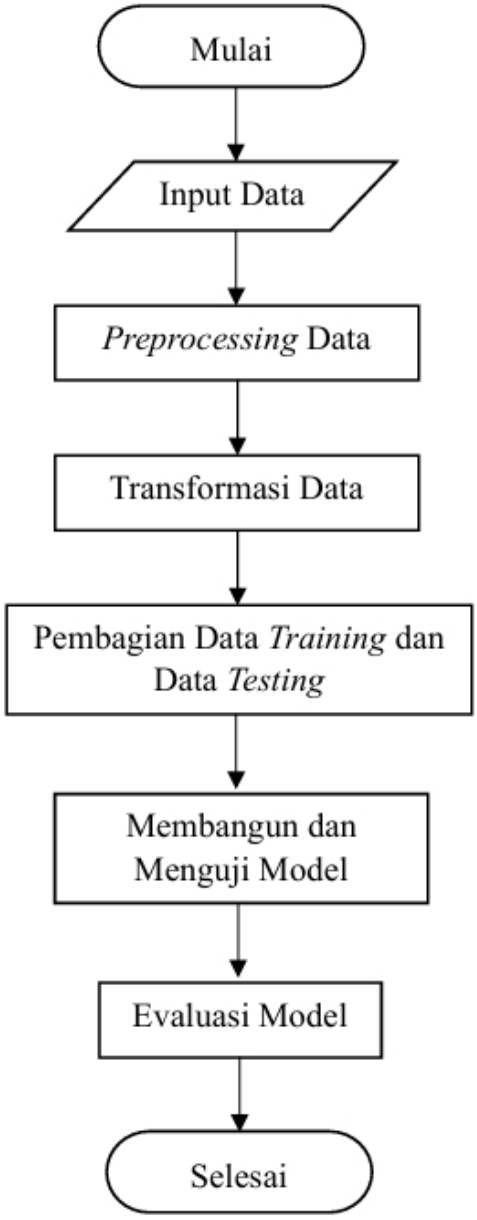
   setelah dilakukan *splitting* data selanjutnya dilakukan metode *k-fold cross validation* dengan k= 5, 8 dan 10.

(d) Membangun dan menguji model *naïve bayes* dan *random forest* menggunakan bahasa pemrogramman *Python*.

(e) Melakukan evaluasi model menggunakan *confusion matriks*

(f) Membandingkan nilai akurasi dan performa dari kedua model.

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



Gambar 3.1 Kerangka Penelitian Metode *Naïve Bayes* dan *Random Forest*.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil pengujian menggunakan model *naïve bayes* dan *random forest* untuk prediksi klasifikasi masa studi sarjana matematika Universitas Lampung, dapat disimpulkan bahwa model *random forest* menunjukkan tingkat akurasi terbaik berdasarkan evaluasi *confusion matrix*, dengan akurasi sebesar 94,44%, dibandingkan dengan model *naïve bayes*. Selain itu, hasil dari berbagai teknik pengujian menunjukkan bahwa kedua model tersebut mencapai tingkat akurasi tertinggi ketika menggunakan metode *splitting* data dibandingkan dengan metode *k-fold cross validation*.

5.2 Saran

Penelitian ini hanya menggunakan model *naïve bayes* dan *random forest* sebagai perbandingan. Oleh karena itu, penelitian selanjutnya disarankan untuk mencoba metode klasifikasi lainnya atau membandingkan lebih banyak model. Selain itu, jumlah data yang digunakan sebaiknya ditingkatkan agar dapat memperoleh nilai akurasi yang lebih optimal.

DAFTAR PUSTAKA

Alpaydin, E. (2021). *Introduction to Machine Learning*. 4^{th} *Edition*. MIT Press.

Andhini, A.A.A, Wiwin, H., dan Zulfan, E. (2022). Implementasi Metode *Naïve Bayes* Untuk Klasifikasi Penerima Program Keluarga Harapan. *Jurnal of Computer*. **2**(1): 21-26.

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.

Bramer, M. (2016). *Principles of Data Mining*. 3^{rd} *Edition*. Springer, USA.

Breiman, L. (2001). *Random Forests*. *Machine Learning Journal*. **45**(1): 5-32.

Dwi Ispriyanti., dan Abdul Hoyyi. (2016). Analisis Klasifikasi Masa Studi Mahasiswa Prodi Statistika UNDIP dengan Metode *Support Vector Machine* (SVM) dan ID3 *Iterative Dichotomiser 3*. *Jurnal Media Statistika*. **9**(1): 15-29.

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.

Geron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 3^{rd} *Edition*. O'Reilly Media.

Gnanapriya, S., Suganya, R., Devi, G., & Kumar, M. (2010). *Data Mining Concepts and Techniques*. *Data mining and knowledge engineering*. **2**: 256-263.

Gupta, B., Rawat, A., Jain, A., Aurora, A., & Dhami, M. (2017). *Analysis of Various Decision Tree Algorithm fot Classification in Data Mining*. *International Journal of Computer Application*. **163**(8): 15-19.

Han, J., dan Kamber, M. (2006). *Data Mining Concepts, Models and Techniques*. Springer, USA.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2^{nd} *Edition*. Springer, Berlin.

Kuhn, M. dan Johnson, K. (2013). *Applied Predictive Modeling*. 2^{nd} *Edition*. Springer, Berlin.

Larose, Daniel T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. *New Jersey: John Willey & Sons. Inc*.

Marsland, S. (2023). *Machine Learning: An Algorithmic Perspective*. 3^{rd} *Edition*. CRC Press.

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

Mulajati, M. (2017). Implemetasi Teknik Web *Scraping* dan Klasifikasi Sentimen Menggunakan Metode *Naïve Bayes Classifier* dan Asosiasi Teks. Skripsi Universitas Islam Indonesia.

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.

Pujianto,U., Widianingtyas, T., Prasetya, D.D., & 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.

Raschka, S., dan Mirjalili, V. (2022). *Python Machine Learning (3rd ed.)*. *Packt Publishing*.

Ricky, L., Janis, P., dan Chrisnatalis. (2020). Perbandingan Metode *Random Forest* dan *Naïve Bayes* Dalam Prediksi Keberhasilan Klien Telemarketing. *Jurnal Penelitian Teknik Informatika*. **3**(2): 455-459.

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

Sari, C. Riang. (2016). Teknik Data Mining Menggunakan *Classification* dalam Sistem Penunjang Keputusan Peminatan SMA Negeri 1 Polewali. *Indonesian Journal of Network & Security*. **5**(1): 48-54.

Walpole, Ronald E. (1993). Pengantar Statistika. Jakarta: PT Gramedia Pustaka Utama.

Widya, A., Ilham, K., Muhamad, B., & Tri, H. (2021). Prediksi Kemungkinan Diabetes Pada Tahap Awal Menggunakan Algotirma Klasifikasi *Random Forest*. *Jurnal Sistem Informasi*. **10**(1): 163-171.

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

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3^{rd} *Edition*. Morgan Kaufmann.

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