JURNAL ILMIAH FLASH
Vol. 8 No. 2, Halaman: 109 - 117
Desember 2022

KOMPARASI ALGORITMA KLASIFIKASI UNTUK MEMPREDIKSI KELULUSAN MAHASISWA PROGRAM STUDI TEKNIK KOMPUTER JARINGAN

Petrisia Widyasari Sudarmadji[1]*, Nikson Fallo[2], dan Yohanes Suban Peli[3]
[1,2,3] Politeknik Negeri Kupang
[1,2,3] Jl. Adisucipto PO.BOX 139 Penfui – Kota Kupang - NTT
* E-mail: petrisia.pnk@gmail.com

Abstrak

Tingkat kelulusan mahasiswa dalam suatu institusi pendidikan sangatlah penting karena selain untuk tetap menjaga kredibilitas institusi tersebut, tingkat kelulusan juga berperan dalam menjaga rasio antara mahasiwa dengan dosen agar tetap dalam takaran yang tepat. Untuk itu, informasi yang cepat, tepat, dan akurat tentang klasifikasi tingkat kelulusan mahasiswa akan sangat dibutuhkan pihak institusi sehingga dapat dijadikan strategi ataupun solusi yang tepat dalam upaya meningkatkan trend positif terkait tingkat kelulusan mahasiswa. Jumlah kelulusan mahasiswa Program Studi Teknik Komputer dan Jaringan pada Politeknik Negeri Kupang cenderung mengalami penurunan jumlah wisudawan setiap tahun sehingga menjadi polemik internal. Sedangkan saat ini sebuah Perguruan Tinggi atau Universitas dituntut untuk selalu memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang dimilikinya. Selain sumber daya sarana, prasarana, dan manusia, sistem informasi adalah salah satu sumber daya yang dapat digunakan untuk meningkatkan keunggulan bersaing. Salah satu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar adalah Data Mining. Data mining adalah proses melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data. Ada 5 peranan utama data mining, yaitu: Estimasi, Prediksi, Klasifikasi, Klastering, dan Asosiasi. Algoritma yang di gunakan pada penelitian ini lebih dari satu model sehingga penulis Mengkomparasi Algoritma Klasifikasi Untuk Memprediksi Kelulusan Mahasiswa, serta memining *knowledge* dari dataset kelulusan mahasiswa untuk : membandingkan algoritma yang paling akurat dalam penentuan klasifikasi kelulusan mahasiswa. Algoritma- algoritma yang digunakan adalah *Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, SVC, Random Forest Classifier, Gradient Boosting Classifier* dan *GaussianNB*.

Kata kunci: data mining, algoritma, kelulusan, klasifikasi, mahasiswa

PENDAHULUAN

Mahasiswa adalah seseorang yang sedang dalam proses menimba ilmu ataupun belajar dan terdaftar sedang menjalani pendidikan pada salah satu bentuk perguruan tinggi yang terdiri dari akademik, politeknik, sekolah tinggi, institut dan universitas [9]. Menurut Siswoyo [13] mahasiswa dapat didefinisikan sebagai individu yang sedang menuntut ilmu di tingkat perguruan tinggi, baik negeri maupun swasta atau lembaga lain yang setingkat dengan perguruan tinggi. Mahasiswa di nilai memiliki tingkat intelektualitas yang tinggi, kecerdasan dalam berpikir dan kerencanaan dalam bertindak. Berpikir kritis dan bertindak dengan cepat dan tepat merupakan sifat yang cenderung melekat pada diri setiap mahasiswa, yang merupakan prinsip yang saling melengkapi. Ciri intelektualitas ini adalah kemampuan mahasiswa dalam menghadapi masalah dan mencari pemecahan masalahnya secara lebih sistematis [2].

Mengetahui tingkat kelulusan mahasiswa dalam suatu institusi pendidikan sangatlah penting. Selain untuk tetap menjaga kredibilitas institusi tersebut, tingkat kelulusan juga berperan dalam menjaga rasio antara mahasiwa dengan dosen agar tetap dalam takaran yang tepat. Untuk itu, informasi yang cepat, tepat, dan akurat tentang klasifikasi tingkat kelulusan mahasiswa akan sangat dibutuhkan supaya pihak institusi dapat membuat strategi ataupun solusi yang tepat agar dapat menjaga bahkan meningkatkan trend positif terkait tingkat kelulusan mahasiswa. Demikian juga dengan jumlah kelulusan mahasiswa Program Studi Teknik Komputer dan Jaringan pada Politeknik Negeri Kupang, yang cenderung mengalami penurunan jumlah wisudawan setiap tahun. Hal ini menjadi polemik internal, oleh karena rasio jumlah kelas yang banyak tidak sebanding dengan jumlah mahasiswa yang lulus setiap tahun. Sedangkan saat ini sebuah perguruan

tinggi atau Universitas dituntut untuk selalu memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang dimilikinya. Selain sumber daya sarana, prasarana, dan manusia, sistem informasi adalah salah satu sumber daya yang dapat digunakan untuk meningkatkan keunggulan bersaing. Teknologi yang berkembang sampai saat ini pun membuat sebuah sistem informasi berperan semakin penting dalam menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis. Salah satu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar adalah Data Mining. Data mining adalah proses melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data [15]. Data mining sering dianggap sebagai bagian dari *Knowledge Discovery in Database* (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. Selain itu data mining juga dikenal dengan nama *knowledge extraction, pattern analysis, information harvesting*, dan *Business intelligence*. Ada 5 peranan utama data mining, yaitu: Estimasi, Prediksi, Klasifikasi, Klastering, dan Asosiasi. Algoritma data mining yang sering digunakan dalam klasifikasi diantaranya adalah *Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression*, dan lain-lain.

Pada penelitian ini, akan di formulasikan menggunakan algoritma *Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, SVC, Random Forest Classifier, Gradient Boosting Classifier* dan *GaussianNB*. Hal ini didasarkan pada beberapa alasan, yaitu: Selain ketujuh algoritma tersebut sama-sama mudah di implementasikan dan sama-sama dapat memberikan hasil yang baik dalam kasus klasifikasi, ketujuh algoritma tersebut juga mempunyai beberapa keunggulan masing-masing, misalnya *Logistic Regression* merupakan algoritma klasifikasi pohon keputusan yang efisien dalam menangani atribut bertipe diskret dan numerik[8]. Algoritma *GaussianNB*,[8] menjelaskan bahwa algoritma ini hanya membutuhkan satu kali scan data training. Sedangkan algoritma *Random Forest Classifier*, didasarkan pada pernyataan[3] yang menyebutkan bahwa algoritma *Random Forest Classifier* dapat mengatasi data training dalam jumlah sangat besar secara efisien dan merupakan metode yang efektif dalam mengestimasi missing data. Merujuk atas uraian permasalahan di atas, maka pendekatan yang di gunakan pada penelitian ini untuk mendapatkan model algoritma yang paling baik, maka penulis membandingkan tujuh algoritma di atas dengan menggunakan dataset yang sama serta perlakuan yang sama pada dataset tersebut.

Penelitian ini bertujuan untuk : membandingkan algoritma yang paling akurat dalam penentuan klasifikasi kelulusan mahasiswa Program Studi Teknik Komputer dan Jaringan. Algoritma- algoritma yang digunakan adalah *Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, SVC, Random Forest Classifier, Gradient Boosting Classifier* dan *GaussianNB*. Urgensi penelitian :

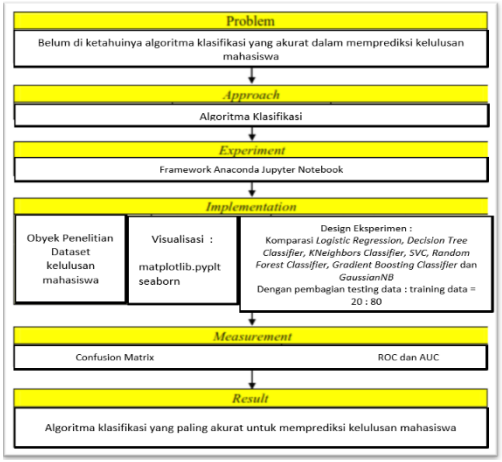
1. Dengan hasil yang akurat, memberikan kontribusi keilmuan pada penelitian bidang klasifikasi data mining bahwa model algoritma ini dapat digunakan untuk menentukan klasifikasi kelulusan mahasiswa.
2. Penelitian ini diharapkan dapat memberikan masukkan atas teori pemodelan *Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, SVC, Random Forest Classifier, Gradient Boosting Classifier* dan *GaussianNB,* khususnya untuk penelitian yang berhubungan dengan penelitian di bidang klasifikasi.
3. Penelitian ini diharapkan dapat digunakan untuk membantu para stakeholder institusi pendidikan (Politeknik Negeri Kupang) dalam mengambil keputusan atau strategi yang nantinya akan dipakai untuk meningkatkan tingkat kelulusan mahasiswanya.

METODE PENELITIAN

Beberapa peneliti sebelumnya telah melakukan penelitian tentang algoritma klasifikasi yaitu : Penelitian yang di lakukan oleh Gede Suwardika , I Ketut Putu Suniantara, pada Jurnal Barekeng – Jurnal Ilmu Matematika dan Terapan, Vol.13, No.3, Desember 2019 dengan judul “Analisis *Random Forest* Pada Klasifikasi Cart Ketidaktepatan Waktu Kelulusan Mahasiswa Universitas Terbuka” [6]; Penelitian yang di lakukan oleh Sarah Novia Hermawanti dkk, pada Jurnal Ilmiah Santika, Vol.9 No.1 Juni 2019 dengan judul “ Implementasi Algoritma C4.5 Untuk Prediksi Kelulusan Tepat Waktu (Studi kasus : Program Studi Teknik Informatika)” [11]; Penelitian yang di lakukan oleh Tias Mugi Rahayu dkk, pada Jurnal Media

Bina Ilmiah , Vol.15 No.10 Mei 2021 dengan judul "Klasifikasi Ketepatan Waktu Kelulusan Mahasiswa Dengan Metode *Naive Bayes*" [14]. Penelitian-penelitian sejenis tentang klasifikasi maupun prediksi kelulusan mahasiswa di atas, berbeda dengan penelitian yang di ajukan. Penelitian-penelitian tersebut rata-rata menggunakan satu metode algoritma, namun pada penelitian yang diajukan penulis adalah menggunakan formulasi tujuh jenis algoritma (komparasi) yang selanjutnya akan membandingkan algoritma yang paling akurat dalam penentuan klasifikasi kelulusan mahasiswa Program Studi Teknik Komputer dan Jaringan di Politeknik Negeri Kupang. Adapun tahapan penelitian di gambarkan dalam bentuk peta jalan penelitian :

Gambar 1. Peta Jalan Penelitian



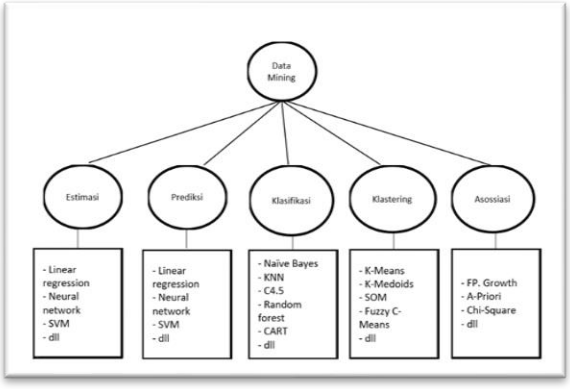
Data Mining :

Data mining adalah analisa terhadap data untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut. Data mining adalah metoda yang digunakan untuk mengekstraksi informasi prediktif tersembunyi pada database, ini adalah teknologi yang sangan potensial bagi perusahaan yang sangat potensial bagi perusahaan dalam memberdayakan data *warehouse*. Secara garis besar,data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

1. *Deskriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam satu basis data. Teknik data mining yang termasuk *descriptive mining* adalah *clustering, asosiation*, dan *sequential mining*.
2. *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan.Salah satu teknik yang terdapat dalam *predictive mining* adalah klasifikasi. Secara sederhana data mining biasa dikatakan sebagai proses penyaring atau "menambang" pengetahuan dari sejumlah data yang besar. Istilah lain untuk data mining adalah *Knowledge Discovery in Database* [4].

Terdapat lima peranan utama data mining, mengacu pada Larose [10], lima peranan tersebut yaitu:

1. Fungsi estimasi (*estimation*) :
   Fungi estimasi adalah fungsi untuk memperkirakan suatu hal yang sudah ada datanya. Fungsi estimasi terdiri dari dua cara yaitu Estimasi Titik dan Estimasi Selang Kepercayaan.
2. Fungsi prediksi (*prediction*) :
   Fungsi prediksi adalah memperkirakan hasil dari hal yang belum diketahui, untuk mendapatkan hal baru yang akan muncul selanjutnya. Cara memprediksi dalam fungsi ini adalah Regresi Linier.
3. Fungsi klasifikasi (*classification*) :
   Fungsi klasifikasi atau menggolongkan suatu data. Cara yang digunakan terdiri dari algoritma *Mean Vector*, algoritma K-nearset Neighbor, algoritma ID3 dan algoritma C4.5
4. Fungsi pengelompokan (*cluster*)
   Fungsi pengelompokan, data yang dikelompokan disebut objek catatan yang memiliki kemiripan atribut kemudian dikelompokan pada kelompok yang berbeda. Algoritma yang digunakan adalah algoritma *Hirarchical Clustering*, algoritma *Partitional Clustering*, algoritma *Single Linkage*, algoritma *Complete Linkage*, algoritma *Average Linkage*, algoritma *K-Means* dan lain-lain
5. Fungsi asosiasi (*association*) :
   Fungsi asosiasi adalah untuk menemukan aturan asosiasi (*association rule*) yang mampu



mengidentifikasi item-item yang menjadi objek. Algoritma yang digunakan adalah algoritma *Generalized Association Rule, Quantitative Association Rule, asynchronous Parallel Mining*.

Gambar 2. Taksonomi Peranan Data Mining
(sumber : Gorunescu, 2011)

Klasifikasi

Klasifikasi adalah proses menempatkan obyek atau konsep tertentu kedalam satu set kategori, berdasarkan sifat obyek atau konsep yang bersangkutan[7]. Dalam klasifikasi terdapat dua pekerjaan utama yang dilakukan: pertama, pembangunan model sebagai *prototype* untuk disimpan sebagai memori. Kedua, penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut berada. Proses klasifikasi didasarkan pada komponen [7]:

1. Kelas (*Class*)
   Variabel dependen dari model yang merupakan kategori variabel yang mewakili label-label yang diletakkan pada obyek setelah pengklasifikasian. Contoh: kelas bintang, kelas gempa bumi
2. Prediktor (*Predictor*)
   Variabel independen dari model yang diwakili oleh karakteristik atau atribut dari data yang diklasifikasikan berdasarkan klasifikasi yang dibuat. Contoh: tekanan darah, status perkawinan, musim
3. Dataset Pelatihan (*Training Dataset*)
   Merupakan dataset yang berisi dua komponen nilai yang digunakan untuk pelatihan mengenali model yang sesuai dengan kelasnya, berdasarkan predictor yang ada. Contoh: *database* penelitian gempa, database badai, database pelanggan supermarket
4. Database Pengujian (*Testing Database*) Merupakan dataset baru yang akan diklasifikasikan oleh model yang dibangun sehingga dapat dievaluasi hasil akurasi klasifikasi tersebut.

Algoritma Regresi C4.5

Salah satu metode klasifikasi yang melibatkan konstruksi pohon keputusan, koleksi node keputusan, terhubung oleh cabang-cabang, memperpanjang ke bawah dari simpul akar sampai berakhir di node daun. Dimulai dari node root, yang oleh konvensi ditempatkan dibagian atas dari diagram pohon keputusan, atribut diuji pada node keputusan, dengan setiap hasil yang mungkin menghasilkan cabang. Setiap cabang kemudian mengarah ke node lain baik keputusan atau ke node daun untuk mengakhiri. Algoritma C4.5 dan pohon keputusan (*decision tree*) merupakan dua mode yang tidak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5. Decision Tree merupakan algoritma pengklasifikasian yang sering digunakan dan mempunyai struktur yang sederhana dan mudah untuk diinterpretasikan). Pohon yang terbentuk menyerupai pohon terbalik, dimana akar (*root*) berada di bagian paling atas dan daun (*leaf*) berada di bagian paling bawah. Tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5[10] yaitu:

1. Mempersiapkan data training, data training biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung total entropy sebelum atau dicari masing-masing entropy class
   $H(T) = - \sum Pj \log2 (Pj)$
   Keterangan:
   H = Himpunan kasus
   T = Atribut
   Pj = Proporsi dari Hj terhadap H
3. Hitung nilai Gain dengan information gain dengan rata-rata:
   Gain Average = H(T) - Hsaving(T)
   Keterangan:
   H(T) = Total Entropy
   Hsaving(T) = Total Gain information untuk masing-masing atribut
4. Ulangi langkah ke 2 dan ke 3 hingga semua tupel terpartisi.
   Proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tupel dalam node N mendapatkan kelas yang sama
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi
   c. Tidak ada tupel di dalam cabang yang kosong

Naive Bayes

Naive Bayes merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris bernama Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Algoritma pengklasifikasi Naive Bayes adalah pengklasifikasi yang berdasarkan probabilitas bersyarat pada teorema Bayes [1]. Cara kerja metode pengklasifikasi Naive Bayes dapat diurutkan seperti langkah-langkah berikut [8]:

1. Diketahui D adalah dataset training

yang terdiri dari sekumpulan baris data dan label kelasnya. Setiap baris memiliki n dimensi vektor atribut, *X* = (x1,x2,...xn), menggambarkan n pengukuran dibuat atas baris dari n atribut, masing-masing sebagai berikut, A1, A2,...An.

2. Terdapat m kelas, C1, C2,...Cm memberikan sampel *X*, pengklasifikasi akan memprediksi bahwa *X* termasuk kelas yang memiliki probabilitas posteriori, dikondisikan pada *X*. Dimana *X* diperkirakan memiliki kelas Ci jika dan hanya jika:
   $P(C_i \mid X) > P(C_j \mid X)$untuk$1 \leq j \leq m, j \neq i$

   Dengan demikian ditemukan kelas yang maksimal $P(C_i \mid X)$. Kelas C_i untuk setiap $P(C_i \mid X)$ yang dimaksimalkan disebut hipotesisposteriori maksimum. Persamaan teorema bayes:

   $$P(C_i \mid X) = \frac{P(X \mid C_i)\, P(C_i)}{P(X)}$$

   Dengan:
   $P(C_i \mid X)$ = Probabilitas hipotesis kelas C_i berdasarkan kondisi *X*
   $P(X \mid C_i)$ = Probabilitas data *X* berdasarkan kondisi pada kelas C_i
   $P(C_i)$ = Probabilitas awal kelas C_i
   $P(X)$ = Probabilitas awal data *X*

3. P (*X*) adalah sama untuk semua kelas, hanya $P(X \mid C_i)\, P(C_i)$ yang perlu dimaksimalkan. Jika kelas apriori probabilitas, P(Ci) tidak diketahui, maka umumnya diasumsikan seperti ini $P(C_i) = P(C2) = ... = P(Cm)$ maka dari itu akan memaksimalkan $P(X \mid C_i)$. Tapi sebaliknya akan memaksimalkan $P(X \mid C_i)\, P(C_i)$. Dapat diperhatikan bahwa kelas probabilitas apriori dapat diperkirakan dengan $P(C_i) = |C_i,D| / |D|$, dimana $|C_i,D|$ merupakan jumlah pelatihan rangkap dari kelas C_i di dalam D.
4. Dataset dengan banyak atribut, akan menjadi perhitungan yang mahal untuk menghitung $P(X \mid C_i)$. Dalam rangka untuk mengurangi perhitungan dalam mengevaluasi $P(X \mid C_i)$. Pada asumsi naive bahwa kelas independen bersyarat dibuat. Ini menganggap bahwa nilai-nilai atribut yang independen bersyarat satu sama lain diberikan pada kelas sampel. Secara matematis berarti bahwa: Dataset dengan banyak atribut, akan menjadi perhitungan yang mahal untuk menghitung $P(X \mid C_i)$. Dalam rangka untuk mengurangi perhitungan dalam mengevaluasi $P(X \mid C_i)$. Pada asumsi naive bahwa kelas independen bersyarat dibuat. Ini menganggap bahwa nilai-nilai atribut yang independen bersyarat satu sama lain diberikan pada kelas sampel. Secara matematis berarti bahwa:

   $$P(X \mid C_i) = \Pi^{n}_{k=1} P(X_k \mid C_i)$$
   $$= P(x_1 \mid C_i) \times P(X_2 \mid C_i) \ldots n$$

5. Untuk memprediksi label kelas *X*, $P(X \mid C_i)\, P(C_i)$ merupakan evaluasi dari setiap kelas C_i. Pengklasifikasi memprediksi bahwa label kelas *X* adalah C_i jika dan hanya jika :

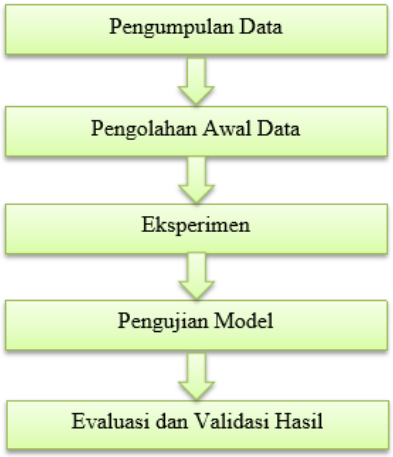
   $P(X \mid C_i)\, P(C_i) > P(X \mid C_j)$ untuk$1 \leq j \leq m, j \neq i$

Random Forest

Random Forest merupakan pengembangan dari *Decision Tree*, dimana setiap *Decision Tree* telah dilakukan training menggunakan sampel individu dan setiap atribut dipecah pada tree yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada vote dari suara terbanyak pada kumpulan populasi *tree*. *Random forest* adalah pengembangan dari metode *CART*, yaitu dengan menerapkan metode *bootstrap aggregating (bagging)* dan *random feature selection*[3]. Dalam penelitiannya, Breiman telah menunjukkan beberapa kelebihan *random forest* antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data training dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing* data. Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut.

Desain Penelitian

Penelitian adalah usaha mencari melalui proses yang metodis untuk menambahkan pengetahuan itu sendiri dan dengan yang lainnya, oleh penemuan fakta dan wawasan tidak biasa[5]. Untuk dapat menemukan fakta atau pengetahuan dari data, dibutuhkan suatu usaha ekstraksi yang disebut dengan data mining. Ekstraksi dilakukan untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui dari suatu data. Menurut Dawson dalam Setiyorini et al, terdapat beberapa metode penelitian yang dapat dipakai untuk mengatasi masalah penelitian yaitu *action research, experiment, case study* dan survey[12]. Dalam penelitian ini, metode penelitian yang digunakan adalah metode penelitian eksperimen dengan tahapan seperti diagram berikut :



Gambar 3. Tahapan Penelitian

Pengumpulan Data

Data yang di gunakan pada penelitian ini merupakan data sekunder yang di peroleh dari Bagian Akademik – Politeknik Negeri Kupang. Data yang peneliti ambil merupakan data kelulusan mahasiswa yang terdiri dari 9 atribut yaitu : ip1,ip2, ip3, ip4, ip5, ip6, ipk, lama semester dan outcome “tepat” dengan label 0 adalah lulus tidak tepat waktu dan 1 adalah lulus tepat waktu.

Pengolahan Data Awal

Tahapan selanjutnya adalah pengolahan data awal, setelah data terkumpul maka diperlukan preprocessing data terlebih dulu. Hal ini bertujuan untuk membersihkan dataset yang sudah ada dari data-data yang tidak perlu. Dataset yang digunakan dalam penelitian ini, masih ditemukan mempunyai missing value yang harus diperlakukan secara khusus. Adapun penanganan *missing value* menurut[8] adalah:

1. Mengabaikan *tuple* yang berisi *missing value*
2. Mengganti *missing value* secara manual
3. Mengganti *missing value* dengan konstanta global (misal “*unknown*” atau “∞”)
4. Mengganti *missing value* dengan nilai mean atau median dari atribut
5. Mengganti *missing value* dengan nilai mean atau median dari semua sampel
6. Mengganti *missing value* dengan nilai kemungkinan terbanyak dari dataset

Pada penelitian ini, perlakuan khusus yang diberikan untuk menangani *missing value* adalah dengan memberikan nilai rata-rata dari atribut. Teknik ini dapat diterapkan untuk atribut yang mempunyai nilai numerik.

Pengujian Model

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tujuh metode klasifikasi data mining. Algoritma yang akan digunakan adalah *Logistic Regression, Decision Tree Classifier, KNeighbors Classifier, SVC, Random Forest Classifier, Gradient Boosting Classifier* dan *GaussianNB*. Setelah diolah dan menghasilkan model, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menghitung akurasi dengan perbandingan antara data testing dan data training 20 : 80 dan mengulang pengujian tersebut beberapa kali.

Evaluasi dan Validasi Hasil

Tahap selanjutnya adalah melakukan evaluasi dan validasi hasil pengujian model tersebut dengan menggunakan *confussion matrix, ROC* dan *AUC*. *Confussion matrix* adalah alat (tools) visualisasi yang biasa digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda [8]. Sedangkan kurva *ROC* adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif. Selanjutnya setiap hasil akurasi dan *AUC* dari metode tujuh algoritma tersebut dibandingkan sehingga diperoleh model dari

metode klasifikasi kelulusan mahasiswa yang tertinggi.

Rencana Pengujian Sistem

Pengujian sistem akan di lakukan dengan analisis uji klasifikasi, yaitu membandingkan hasil klasifikasi yang di berikan oleh analisis metode komparasi algoritma data mining, dengan klasifikasi yang di berikan secara manual (hasil olah data Bagian Akademik – Politeknik Negeri Kupang).

HASIL DAN PEMBAHASAN

Persiapan dataset

Pengambilan dataset pada penelitian ini, berasal dari bagian Akademik – Politeknik Negeri Kupang. Data yang di ambil adalah indeks prestasi mahasiswa semester 1 sampai semester 6, indeks prestasi kumulatif (ipk), lama semester yang di lalui mahasiswa hingga lulus dan status kelulusan mahasiswa (tepat atau tidak tepat). Data mahasiswa yang di ambil adalah data mahasiswa Program Studi Teknik Komputer dan Jaringan – Angkatan tahun 2017, 2018 dan 2019. Selanjutnya dataset di import pada framework Anconda Jupyter Notebook serta mengimport libraries yang perlukan dalam proses klasifikasi data seperti di bawah ini :

```
# Dataframe/numerical libraries'
import pandas as pd
import numpy as np

# Visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns

# Machine Learning libraries
# Models
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.naive_bayes import GaussianNB

# Pre-processing
from sklearn.preprocessing import StandardScaler, binarize
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import train_test_split

# Evaluate
from sklearn.metrics import accuracy_score, confusion_matrix, recall_score, precision_score, roc_auc_score, roc_curve
```

Gambar 4. Proses Klasifikasi Data

Selanjutnya adalah proses membaca 10 data teratas dari dataset yang sudah di import. Selanjutnya melihat jumlah data dan kolom pada dataset yang sudah di import dalam entuk dimensi data. Dengan tag shape, maka terkonfirmasi dataset berjumlah 1687 baris dan 9 kolom. Setelah itu, di lanjutkan dengan melihat statistika dasar setiap kolom, karena dengan melihat statistika dasar maka kita bisa mendeteksi data yang tidak wajar/*missing value*. Dan terlihat jelas bahwa ada beberapa data missing *value* pada atribut ipk dan lama semester.

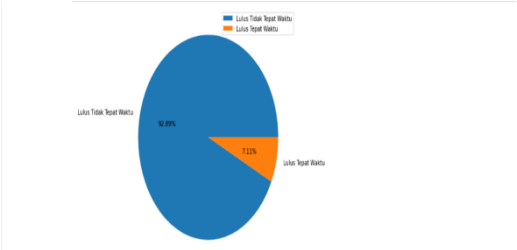
Gambar 5. *Load* 10 Data Teratas

```
Dimensi Data

In [4]: df.shape
Out[4]: (1687, 9)

In [5]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1687 entries, 0 to 1686
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ip1             1687 non-null   float64
 1   ip2             1687 non-null   float64
 2   ip3             1687 non-null   float64
 3   ip4             1687 non-null   float64
 4   ip5             1687 non-null   float64
 5   ip6             1687 non-null   float64
 6   ipk             1687 non-null   float64
 7   lama semester   1687 non-null   int64
 8   tepat           1687 non-null   int64
dtypes: float64(7), int64(2)
memory usage: 118.7 KB
```

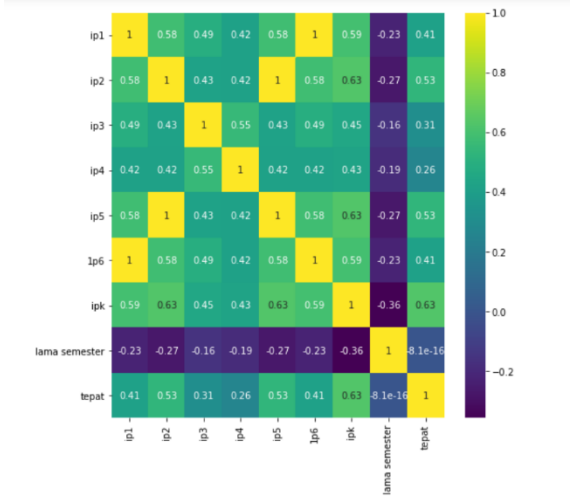
Untuk melihat informasi tersembunyi pada dataset yang ada, maka dataset bisa di explore terlebih dahulu sebelum di bersihkan. Dari proses explore di ketahui komposisi porsentase mahasiswa lulus tidak tepat waktu sebanyak 92,89% dan porsentase mahasiswa yang lulus



tepat waktu sebanyak 7,11%.

Gambar 6. Porsentase Kelulusan Mahasiswa

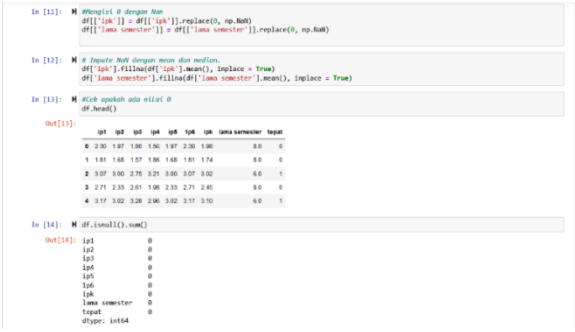
Selanjutnya, kita akan melihat korelasi antar atribut “lama semester” dan “tepat” melalui visualisasi seaborn seperti di bawah ini :



Gambar 7. Visualisasi Seaborn

Data Preparation

Pada deskripsi data menggunakan statistika dasar, terdapat 1 kolom yang mempunyai nilai 0, yaitu IPK dan lama semester. Kolom tersebut tidak dihapus karena dalam dunia pendidikan, faktor-faktor tersebut mempunyai korelasi dengan kelulusan mahasiswa. Oleh karena itu, data 0 pada tiap kolom akan diisi dengan nilai mean dan median. Median digunakan untuk data yang mempunyai std besar.

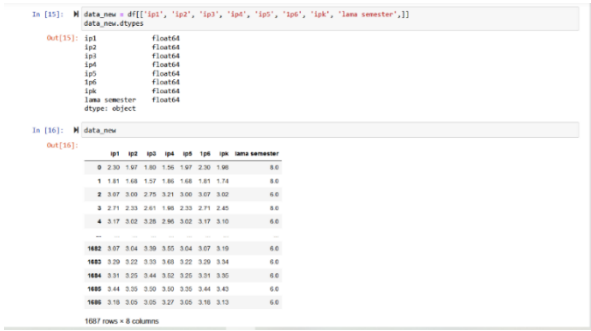


Gambar 8. Deskripsi Data

Setelah memformulasi imputasi pada kolom atribut ipk dan lama semester, dilanjutkan dengan mengecek apakah masih ada kolom yang missing value/berisi nilai NaN dengan tag isnull. Setelah semua kolom atribut terisi nilainya maka baru akan menjadi dataset yang siap di eksekusi menggunakan model yang di tentukan.

Modelling

Selanjutnya adalah membagi *features* dan target (*outcome*) untuk *sample train* dan *test* (80/20), untuk mendapatkan akurasi prediksi kelulusan mahasiswa dengan memilih *features* inputan dengan korelasi yang lebih tinggi.



Gambar 9. *Modelling*

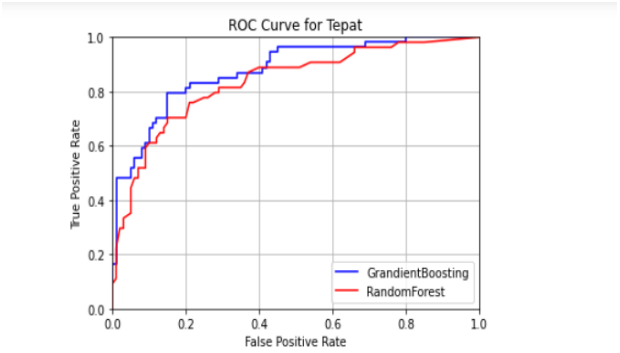
Terlihat, bahwa akurasi algoritma Logistic regression sebesar 75%, SVC 78%, KNeighbors Classifier 77%, GaussianNB 75%, Decision Tree 66%, Random Forest 75% dan Gradient Boosting 81%.

Out[22]:

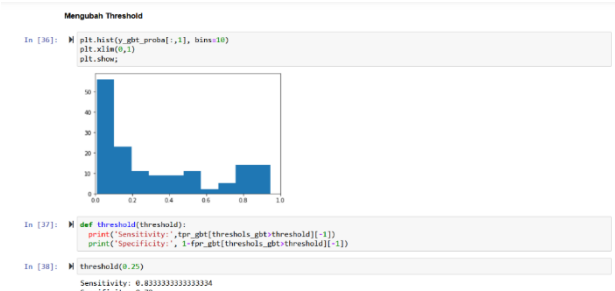
	Model	Accuracy
0	LogisticRegression(solver='liblinear')	0.759740
1	DecisionTreeClassifier()	0.668831
2	RandomForestClassifier()	0.785714
3	KNeighborsClassifier()	0.772727
4	SVC()	0.785714
5	GradientBoostingClassifier()	0.811688
6	GaussianNB()	0.753247

Gambar 10. Akurasi Semua Model

Gambar 11. ROC dan AUC



Gambar 12. ROC Curve For Outcome 1



Gambar 13. Mengubah *Tracehold*

Berdasarkan analisis yang telah dilakukan, dengan membandingkan akurasi dan *ROC*, *GradientBoostingClassifier* merupakan model terbaik untuk memprediksi kelulusan mahasiswa tepat waktu. Setelah mengubah parameter referensi yang berkorelasi kuat dengan status tepat waktu, akurasi prediksi model yang digunakan menjadi naik. Akurasi prediksi model *GradientBoostingClassifier*

mencapai 0.81%. Pada tahap akhir adalah mengukur tingkat akurasi. Dan hasil yang di dapatkan antara 0.81 persen. Berdasarkan hasil pengukuran tersebut, maka dapat di simpulkan bahwa hasil prediksi kelulusan mahasiswa termasuk dalam klasifikasi baik (*good classification*).

PENUTUP

Implementasi Komparasi Algoritma Klasifikasi Untuk Memprediksi Kelulusan Mahasiswa, menghasilkan akurasi sebesar 81 %, dan termasuk dalam klasifikasi baik (*good classification*) dan algoritma *GradientBoostingClassfier* merupakan model terbaik di antara 6 model klasifikasi lainnya untuk memprediksi kelulusan mahasiswa tepat waktu.

UCAPAN TERIMAKASIH

Ucapan terimakasih kepada Politeknik Negeri Kupang sebagai penyandang dana dan berkontribusi pada pengambilan data penelitian, serta Pusat Penelitian dan Pengabdian Masyarakat (P3M) Politeknik Negeri Kupang dengan no kontrak penelitian : Nomor 66f/PL23.PPK.2/ PL/2022, Tanggal 24 Mei 2022

DAFTAR PUSTAKA

[1] Anggarwal, Charu C, " Data Mining: The Textbook", New York: Springer, 2015

[2] Azwar, S, "Penyusunan Skala Psikologi", Yogyakarta: Pustaka Pelajar, 2004

[3] Breiman, L, "Bagging Predictors. Machine Learning", 1996

[4] Darmi, Y., & Setiawan, A, "Penerapan Metode Clustering K-Means" , (2016)

[5] Dawson, C. W, "Projects in Computing and Information Systems a student's guide", Harlow, UK: Addison-Wesley, 2009

[6] Gede Suwardika , I Ketut Putu Suniantara, "Analisis Random Forest Pada Klasifikasi Cart Ketidaktepatan Waktu Kelulusan Mahasiswa Universitas Terbuka", Jurnal Barekeng – Jurnal Ilmu Matematika dan Terapan, Vol.13, No.3, Desember 2019

[7] Gorunescu, Florin, " Data Mining: Concepts, Models, and Techniques" , Verlag Berlin Heidelberg: Springer, 2011

[8] Han, J., & Kamber., & Pei, J, " Data Mining Consepts and Techniques", San Fransisco: Morgan Kauffman, 2012

[9] Hartaji, Damar A, "Motivasi Berprestasi Pada Mahasiswa yang Berkuliah Dengan Jurusan Pilihan Orangtua", Fakultas Psikologi Universitas Gunadarma (tidak diterbitkan), 2012

[10] Larose, D. T, " Discovering Knowledge in Data", New Jersey: John Willey & Sons, Inc, 2005

[11] Sarah Novia Hermawanti dkk, " Implementasi Algoritma C4.5 Untuk Prediksi Kelulusan Tepat Waktu (Studi kasus : Program Studi Teknik Informatika)" , Jurnal Ilmiah Santika, Vol.9 No.1 Juni 2019

[12] Setiyorini, T., "Penerapan Metode Bagging untuk Mengurangi Data Noise pada Neural Network untuk Estimasi Kuat Tekan Beton", Journal of Intelligent Systems, Vol. 1, No. 1, February 2015

[13] Siswoyo, Dwi, "Ilmu Pendidikan", Yogyakarta: UNY Press, 2007

[14] Tias Mugi Rahayu dkk,"Klasifikasi Ketepatan Waktu Kelulusan Mahasiswa Dengan Metode Naïve Bayes", Jurnal Media Bina Ilmiah ,Vol.15 No.10 Mei 2021

[15] Witten, I. H., Frank, E., & Hall, M. A, "Data Mining: Practical Machine Learning and Tools", Burlington: Morgan Kaufmann Publisher, 2011