Data Mining untuk Meningkatkan Performa Akademik Mahasiswa: Studi Kasus dan Implikasinya

Muhammad Arifin, S.Kom, M.Kom, MCE
Prof. Dr. Widowati, S.Si., M.Si
Farikhin, M.Si, Ph.D

Data Mining untuk Meningkatkan Performa Akademik Mahasiswa: Studi Kasus dan Implikasinya

Oleh:
Muhammad Arifin, S.Kom, M.Kom, MCE
Prof. Dr. Widowati, S.Si., M.Si
Farikhin, M.Si, Ph.D

UNDIP PRESS
SEMARANG
2023

Data Mining untuk Meningkatkan Performa Akademik Mahasiswa: Studi Kasus dan Implikasinya

Oleh:
Muhammad Arifin, S.Kom, M.Kom, MCE
Prof. Dr. Widowati, S.Si., M.Si
Farikhin, M.Si, Ph.D

Uk. 15,5cm x 23cm (xii + 160hlm)
ISBN:

diterbitkan oleh :
UNDIP PRESS
SEMARANG

Anggota APPTI 003.151.1.3.2022

Edisi Revisi Februari 2022

Hak Cipta dilindungi undang-undang
Dilarang memperbanyak sebagian atau seluruh isi buku ini dalam bentuk apapun, baik secara elektronik maupun mekanik, termasuk memfotokopi, merekam atau dengan menggunakan system penyimpanan, tanpa izin tertulis dari Penulis.

KATA PENGANTAR

Puji dan Syukur penulis panjatkan kehadirat Allah SWT, atas limpahan Karunia, Taufik, Hidayah dan Ridho-Nya sehingga penulis bisa menyelesaikan penyusunan buku ini. Buku yang berjudul **Peran Educational Data Mining Dalam Meningkatkan Keberhasilan Mahasiswa**.

Penulis menyadari bahwa dalam proses penyusunan buku ini telah melibatkan berbagai pihak, baik secara langsung maupun tidak langsung, yang telah memberikan kontribusi dalam menyelesaikannya. Untuk itu dalam kesempatan ini penulis mengucapkan terima kasih dan penghargaan yang setinggi-tingginya kepada yang terhormat:

1. Dekan Sekolah Pascasarjana Universitas Diponegoro, Bapak Dr. R.B. Sularto, S.H., M.Hum
2. Ketua Program Studi Doktor Sistem Informasi Sekolah Pascasarjana Universitas Diponegoro, Bapak Prof. Dr. Drs. Rahmat Gernowo, M.Si
3. Bapak Dr. Maman Somantri, S.T., M.T.
4. Bapak Dr. Adian Fatchur Rochim, S.T., M.T.
5. Undip Press Semarang
6. Semua pihak yang tidak dapat penulis sebutkan satu persatu, yang telah membantu selama proses penyusunan sampai selesainya buku ini.

Penulis

DAFTAR ISI

Halaman

KATA PENGANTAR iii
DAFTAR ISI v
DAFTAR GAMBAR ix
DAFTAR TABEL xiii
BAB I PENDAHULUAN 1
1.1. Definisi dan tujuan Educational Data Mining 1
1.2. Pentingnya prediksi performa akademik mahasiswa 5
1.3. Kontribusi dan manfaat prediksi performa akademik dalam konteks pendidikan 6
1.4. Perkembangan Penelitian Prediksi Performa Akademik Mahasiswa 8
BAB II KONSEP DASAR PREDIKSI PERFORMA AKADEMIK MAHASISWA 13
2.1 Tinjauan konsep performa akademik 13
2.2 Faktor-faktor yang mempengaruhi performa akademik mahasiswa 16
2.3 Data yang digunakan untuk prediksi performa akademik 17
BAB III METODE DAN TEKNIK DALAM EDUCATIONAL DATA MINING 23
3.1. Pengumpulan dan Persiapan Data 23
a. Identifikasi jenis data yang diperlukan untuk prediksi performa akademik 23
b. Teknik pengumpulan data yang efektif 25
c. Pembersihan dan preprocessing data 26
3.2. Analisis Eksplorasi Data 28
a. Teknik eksplorasi data untuk memahami karakteristik dataset 28
b. Evaluasi kualitas data dan penanganan outliers 30
3.3. Pemilihan Fitur (Feature Selection) 32
3.4. Metode Prediksi Performa Akademik Mahasiswa 34

3.4.1. Gradient Boosted Trees (GBT)..... 36
3.4.2. Hyperparameter 45
3.5. Evaluasi model prediksi (MAE, MSE dan RMSE)..... 51
3.6. Validasi model..... 54
BAB IV STUDI KASUS DAN IMPLEMENTASI PREDIKSI PERFORMA AKADEMIK MAHASISWA 57
4.1. Obyek, Tempat, Waktu dan Data 57
4.2. Deskripsi data..... 61
4.3. Pengumpulan Data..... 67
4.3.1. Data LMS Moodle 67
4.3.2. Data Organisasi..... 73
4.3.3. Data Akademik 75
4.3.4. Data Ekonomi 76
4.3.5. Data Demografi..... 77
4.4. Data Preparation 78
4.4.1. Ekstraksi data LMS Moodle 79
4.4.2. Ekstraksi data Organisasi 92
4.4.3. Ekstraksi data Akademik..... 94
4.4.4. Ekstraksi data Ekonomi..... 94
4.4.5. Ekstraksi data Demografi 95
4.5. Data Visualisasi..... 100
4.6. Proses implementasi prediksi menggunakan model terpilih..... 103
4.7. Evaluasi dan analisis hasil prediksi 124
4.7.1. Evaluasi pengumpulan data..... 124
4.7.2. Evaluasi fitur..... 125
4.7.3. Evaluasi model prediksi 127
4.7.4. Evaluasi Kinerja Model..... 129
4.8. Implementasi dan Penerapan Hasil Prediksi..... 138
PENUTUP 141
1. Kesimpulan dan Tantangan di Masa Depan 141
a. Kesimpulan 141
b. Tantangan dan peluang yang dihadapi dalam prediksi performa akademik melalui Educational Data Mining 142

2. Pemikiran tentang arah penelitian dan pengembangan selanjutnya 144
DAFTAR PUSTAKA 147
BIOGRAFI PENULIS.....163

DAFTAR GAMBAR

Gambar 1.1 Desain EDM (Romero dkk., 2020) 4
Gambar 1.2 Grafik peningkatan jumlah artikel per tahun 8
Gambar 1.3 Perbandingan jumlah artikel dari berbagai negara... 9
Gambar 1.4 Sebaran disiplin ilmu terkait topik performa akademik siswa 10
Gambar 1.5 Perkembangan topik prediksi performa akademik siswa 10
Gambar 3.1 Cara kerja algoritma GBT 38
Gambar 3.2 Visualisasi iterasi model boosting 38
Gambar 3.3 Parameter mesin pembelajaran 47
Gambar 3.4 Ruang hiperparameter dua dimensi 48
Gambar 3.5 Representasi visual GS 50
Gambar 3.6 Parameter GBT 51
Gambar 3.7 Visualisasi pencarian parameter 51
Gambar 3.8 Visualisasi pembagian data untuk validasi silang.. 55
Gambar 4.1 Kerangka penelitian 58
Gambar 4.2 Model Kerangka Prediksi Performa Akademik Mahasiswa 60
Gambar 4.3 Halaman LMS Moodle 68
Gambar 4.4 Halaman bawah LMS Moodle 68
Gambar 4.5 Catatan LMS Moodle 69
Gambar 4.6 Kapasitas berkas catatan..... 69
Gambar 4.7 Grafik catatan periode mingguan..... 72
Gambar 4.8 Dokumen SK/Salinan organisasi 74
Gambar 4.9 Penyimpanan SK/Salinan organisasi 75
Gambar 4.10 Catatan LMS Moodle 79
Gambar 4.11 Potongan Kode Ekstraksi Catatan LMS (a)..... 80
Gambar 4.12 User full name..... 81
Gambar 4.13 Potongan kode ekstraksi catatan LMS (b)..... 82
Gambar 4.14 Potongan kode ekstraksi catatan LMS (c)..... 84
Gambar 4.15 Potongan kode ekstraksi catatan LMS (d)..... 85

Gambar 4.16 Potongan kode ekstraksi catatan LMS (d)..... 86
Gambar 4.17 Potongan kode ekstraksi catatan LMS (e)..... 87
Gambar 4.18 Potongan kode ekstraksi catatan LMS (e)..... 87
Gambar 4.19 Potongan kode ekstraksi catatan LMS (f) 88
Gambar 4.20 Gabungan variabel ekstraksi data LMS..... 89
Gambar 4.21 Gabungan data harian catatan LMS 90
Gambar 4.22 Gabungan data mingguan catatan LMS 91
Gambar 4.23 Salinan keputusan organisasi mahasiswa 92
Gambar 4.24 Pengkodean data ekonomi 95
Gambar 4.25 Kode gabungan data akademik dan LMS..... 98
Gambar 4.26 Kode gabungan data akademik, LMS dan organisasi 98
Gambar 4.27 Data EDM..... 99
Gambar 4.28 Visualisasi data IPS..... 100
Gambar 4.29 Visualisasi Data Jenis Kelamin..... 101
Gambar 4.30 Visualisasi Data Ekonomi 101
Gambar 4.31 Visualisasi Data Kota Tinggal Mahasiswa..... 102
Gambar 4.32 Visualisasi Data Mahasiswa Berorganisasi 102
Gambar 4.33 Visualisasi Data LMS..... 103
Gambar 4.34 Perbandingan algoritma prediksi 104
Gambar 4.35 Hasil perbandingan algoritma mesin pembelajaran 105
Gambar 4.36 Perbandingan algoritma regresi 107
Gambar 4.37 Nilai bawaan parameter GBT 112
Gambar 4.38 Pengambilan data prediksi performa akademik. 113
Gambar 4.39 Pencarian parameter dengan GridSearchCV 114
Gambar 4.40 Proses pengujian validasi 119
Gambar 4.41 Hasil pencarian parameter terbaik..... 120
Gambar 4.42 Input data pengujian..... 120
Gambar 4.43 Kode pengujian GBT tanpa optimasi..... 121
Gambar 4.44 Kode pengujian GBT dengan optimasi GS 123
Gambar 4.45 Bobot fitur data EDM..... 125

Gambar 4.46 Pengaruh fitur data EDM terhadap fitur target (IPS) 126
Gambar 4.47 Perbandingan model klasifikasi dan regresi 127
Gambar 4.48 Grafik perbandingan pengujian model 129
Gambar 4.49 Pengujian menggunakan data baru 130
Gambar 4.50 Proses muat data EDM 133
Gambar 4.51 Penentuan tipe data dan kolom id serta kolom target 134
Gambar 4.52 Pemilihan model 135
Gambar 4.53 Evaluasi Hasil Prediksi 135
Gambar 4.54 Visualisasi hasil prediksi bentuk pohon keputusan 136
Gambar 4.55 Visualisasi hasil prediksi bentuk deskripsi 136
Gambar 4.56 Hasil model GBT dalam bentuk deskripsi 137
Gambar 4.57 Hasil lain yang dihasilkan oleh model GBT 137
Gambar 4.58 Model simulator 138

DAFTAR TABEL

Tabel 2.1 Variabel-variabel prediksi performa akademik mahasiswa 16
Tabel 2.2 Perbandingan karakteristik pembelajaran 18
Tabel 2.3 Variabel Moodle untuk prediksi performa akademik siswa 20
Tabel 3.1 Contoh soal algoritma GBT 41
Tabel 3.2 Nilai residu pertama 42
Tabel 3.3 Nilai residu kedua 44
Tabel 3.4 Nilai residu ketiga 44
Tabel 3.5 Predisksi data baru nilai tinggi <1.5 44
Tabel 3.6 Predisksi data baru nilai tinggi >1.5 45
Tabel 3.7 Contoh soal untuk menghitung MAE, MSE dan RMSE 52
Tabel 3.8 Contoh perhitungan MSE 52
Tabel 3.9 Contoh Perhitungan RMSE 53
Tabel 4.1 Tahapan Penelitian 59
Tabel 4.2 Jumlah catatan berkas LMS Moodle minggu 1-8..... 70
Tabel 4.3 Jumlah catatan berkas LMS Moodle minggu 5-14.... 71
Tabel 4.4 Jumlah catatan berkas LMS Moodle minggu 17-19.. 72
Tabel 4.5 Data akademik 75
Tabel 4.6 Data organisasi 93
Tabel 4.7 Data ekonomi 94
Tabel 4.8 Data demografi kota tinggal 96
Tabel 4.9 Perbandingan waktu pembangunan model 106
Tabel 4.10 Hasil perbandingan akurasi model regresi 108
Tabel 4.11 Parameters for GridSearchCV 113
Tabel 4.12 Nilai validasi model 131

BAB I
PENDAHULUAN

1.1. Definisi dan tujuan Educational Data Mining

Pendidikan memberikan dampak yang sangat penting terhadap pertumbuhan ekonomi suatu bangsa (Deda, Pacukaj, dan Vardari 2021). Faktor tenaga kerja, modal dan kemajuan teknologi merupakan faktor penentu pertumbuhan ekonomi. Pendidikan berperan besar dalam menentukan kualitas tenaga kerja, dengan kata lain semakin tinggi pendidikan yang dimiliki oleh tenaga kerja maka semakin berkualitas kinerjanya. Dengan demikian dapat disimpulkan bahwa semakin tinggi kualitas tenaga kerja suatu bangsa semakin tinggi pula pertumbuhan ekonominya.

Pendidikan terdiri dari Pendidikan dasar, menengah dan Pendidikan tinggi. Pendidikan tinggi memainkan peranan yang sangat penting terhadap daya saing ekonomi lokal, negara bagian, dan nasional (Labini 2015). Pendidikan tinggi juga memainkan peranan yang penting dalam perkembangan masyarakat (Pinheiro, Wangenge-ouma, dan Balbachevsky 2015). Diera perkembangan teknologi yang sangat pesat ini Pendidikan dituntut untuk selalu mengikuti perkembangan teknologi.

Peran teknologi dalam dunia Pendidikan saat ini dipergunakan dalam semua proses dimulai dari proses pendaftaran, seleksi, proses belajar mengajar, penilaian, pembayaran sampai dengan proses kelulusan. Di dalam pendidikan peran teknologi menghasilkan sejumlah data besar yang berkaitan dengan mahasiswa dalam bentuk elektronik. Sangat penting bagi pemangku kepentingan untuk secara efektif mengubah kumpulan data yang sangat besar ini menjadi pengetahuan yang akan membantu pengajar, administrator, dan pembuat kebijakan untuk menganalisisnya guna meningkatkan kualitas pengambilan keputusan. Selain itu,

teknologi dapat dipergunakan untuk meningkatkan kualitas proses pendidikan dengan memberikan informasi yang tepat waktu kepada pemangku kepentingan (Asif dkk. 2017). Perguruan tinggi kebanjiran data namun minim untuk mengubahnya menjadi informasi yang berguna (Morris 2016). Proses merubah data yang besar tersebut menjadikan informasi atau pengetahuan tidak mungkin dilakukan menggunakan cara-cara manual, oleh karena itu dibutuhkan suatu teknik atau metode yang mampu menggali informasi atau pengetahuan dari data tersebut secara cepat dan tepat. Metode yang dipergunakan untuk mencari informasi atau pengetahuan dari kumpulan data dapat dilakukan dengan *Data Mining* (DM). DM digunakan untuk menganalisis data historis dari organisasi manapun dengan menggunakan algoritma, yang selanjutnya menemukan informasi tersembunyi dari data tersebut yang tidak mungkin ditemukan secara manual (Kumar dan Salal 2019).

Penerapan DM pada data pendidikan disebut sebagai *Educational Data Mining* (EDM) (Baker dan Yacef 2009). EDM adalah disiplin ilmu yang berkembang, berkaitan dengan perluasan metode DM klasik dan mengembangkan metode baru untuk menemukan data yang berasal dari sistem pendidikan (Alsuwaiket 2018). EDM dalam dua dekade terakhir ini mengalami perkembangan sangat pesat (Hernández-Blanco dkk. 2019; Maphosa dan Maphosa 2020; Romero dan Ventura 2020). Pesatnya penerapan ini dipengaruhi dengan data yang melimpah dan kecepatan perkembangan teknologi komputer yang memudahkan pengguna dalam melakukan pengolahan data. Bahkan EDM menjadi area penelitian tersendiri didalam ilmu DM, dimana peneliti dari belahan dunia membentuk asosiasi yang terhubung didalam portal website www.educationaldatamining.org, setiap tahunnya diadakan

pertemuan melalui konferensi dari diberbagai negara yang tergabung didalamnya.

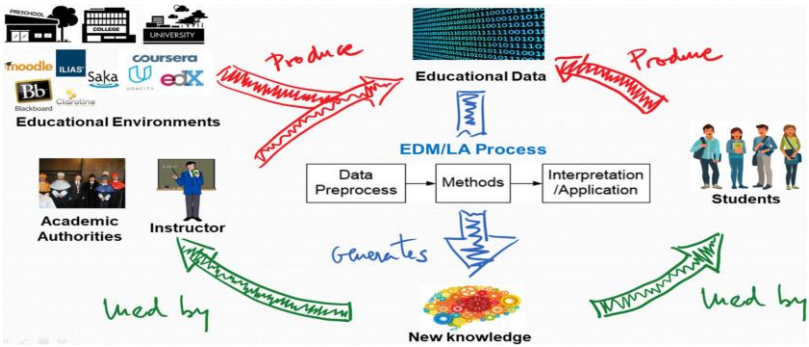
EDM menggunakan proses penemuan pengetahuan yang iteratif Gambar 1.1. Proses dimulai dan berlangsung di lingkungan pendidikan, seperti ruang kelas tradisional, sistem e-learning, dan sistem pendidikan berbasis Web yang adaptif dan cerdas. Lingkungan pendidikan ini menghasilkan data mentah, termasuk data penggunaan dan interaksi mahasiswa, informasi kursus, dan data akademik (Mais Haj Qasem, Raneem Qaddoura 2017).

Proses EDM terdiri dari formulasi hipotesis, pengujian, dan penyempurnaan. Hipotesis dikembangkan dari berbagai lingkungan pendidikan yang menciptakan data dalam volume besar. Proses utama EDM dimulai dengan validasi data (yaitu, menemukan hubungan antara variabel, parameter, dan item data) yang juga dikenal sebagai pra-pemrosesan data. Setelah pra-pemrosesan selanjutnya diolah menggunakan data mining, dan hasil/interpretasi diberikan kepada berbagai pengguna pendidikan. Rekomendasi lebih lanjut disarankan untuk mengatasi masalah ataupun penyempurnaan tugas pendidikan (Romero dan Ventura 2013).

Tujuan EDM di Perguruan Tinggi adalah; 1) Memprediksi perilaku belajar siswa di masa depan dengan penggunaan model siswa, tujuan ini dapat dicapai dengan menciptakan model siswa yang menggabungkan karakteristik siswa, termasuk informasi terperinci seperti pengetahuan, perilaku, dan motivasi mereka untuk belajar. 2) Menemukan atau meningkatkan domain model melalui berbagai metode dan aplikasi EDM, dimungkinkan untuk menemukan model baru dan meningkatkan model yang ada. 3) Mempelajari efek dukungan pendidikan terhadap capaian sistem pembelajaran. 4) Mendapatkan pengetahuan ilmiah tentang pembelajaran dengan membangun dan menggabungkan model

siswa, teknologi dan perangkat lunak yang digunakan (Suthar, Patel, dan Patel 2019).

Penerapan EDM merupakan siklus berulang dari pembentukan hipotesis, pengujian, dan penyempurnaan dalam proses pendidikan. Komponen utama EDM adalah sistem pendidikan, teknik data mining dan pengguna. Sistem pendidikan terdiri dari ruang kelas tradisional, sistem *e-learning*, LMS, sistem adaptif berbasis web, *intelligent tutoring systems*, kuesioner dan kuis. Teknik data mining yang digunakan dalam EDM yaitu; *statistics, visualization, clustering, classification, association rule mining* dan *sequence mining, text mining.* Pengguna EDM adalah mahasiswa atau pelajar, instruktur, guru atau dosen, administrasi, peneliti dan pejabat sekolah (Romero, Ventura, Sebastian Pechenizkiy, dan Baker 2010).



Gambar 1.1 Desain EDM (Romero dkk., 2020)

EDM dikelompokkan kedalam 11 area yaitu; *1) Analysis and Visualization of Data, 2) Providing Feedback for Supporting Instructors, 3) Recommendations for Students, 4) Predicting Student's Performance, 5) Student Modeling, 6) Detecting Undesirable Student Behaviors, 7) Grouping Students, 8) Social Network Analysis, 9) Developing Concept Maps, 10) Constructing*

Courseware, 11) Planning and Scheduling (Romero dan Ventura 2010).

1.2. Pentingnya prediksi performa akademik mahasiswa

Prediksi performa akademik mahasiswa merupakan bagian dari prediksi performa mahasiswa (Romero dkk. 2010). Performa akademik juga disebut hasil akademik, prestasi akademik (Seow dan Pan 2014). Sebagian besar institusi dan universitas menggunakan nilai akhir siswa sebagai kriteria performa akademik mahasiswa. Performa mahasiswa tergantung pada berapa banyak nilai yang diperoleh mahasiswa dalam ujian akhir (Kumar, Singh, dan Handa 2017).

Performa akademik mahasiswa perlu dipantau dengan cermat untuk membantu lembaga mengidentifikasi mahasiswa yang berisiko gagal, mencegah mereka DO atau lulus terlambat (Nachouki dan Abou Naaj 2022). Performa akademik yang buruk menjadi indikator mahasiswa kesulitan dalam menyesuaikan diri dengan perguruan tinggi dan berpeluang besar untuk putus dekolah/*drop out* (DO) (Lau 2003). Beberapa dampak negatif dari buruknya performa akademik dan tingginya DO diantaranya yaitu rendahnya tingkat kelulusan tepat waktu (Delen 2010), membuang biaya Pendidikan yang sia-sia (Costa, Bispo, dan Pereira 2018), pengurangan kesempatan hidup secara profesional dan sosial yang berdampak negatif kepada keluarga dan masyarakat (Casanova dkk. 2021), sulitnya perguruan tinggi dalam mendapatkan akreditasi yang baik (Delen, Topuz, dan Eryarsoy 2020), hilangnya kepercayaan dan motivasi untuk melanjutkan Pendidikan (Coussement dkk. 2020), dan lain sebagainya. Aluko dkk. (2018); Valli Jayanthi dkk., (2014) dan para peneliti pada umumnya menggunakan IPK untuk mengukur performa akademik mahasiswa, mahasiswa dengan IPK baik maka

performa mereka baik dan mahasiswa dengan IPK buruk maka performa mereka adalah buruk.

1.3. Kontribusi dan manfaat prediksi performa akademik dalam konteks pendidikan

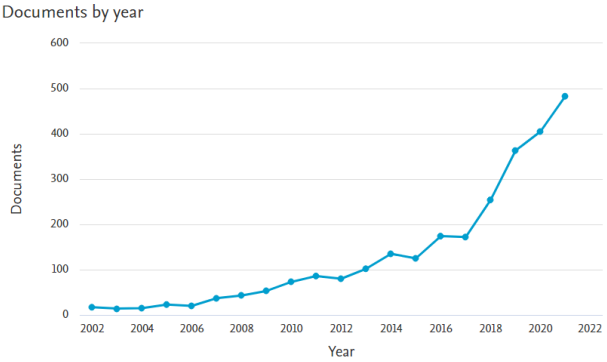
Kontribusi EDM semakin hari semakin berkembang, saat ini EDM merupakan metode yang telah digunakan dalam meningkatkan kualitas Pendidikan (Abu Saa, Al-Emran, dan Shaalan 2019; Al-Emran dkk. 2018; Asif dkk. 2017). EDM telah digunakan dalam mengevaluasi kualitas proses pembelajaran (Bakhshinategh dkk. 2018; Ibitoye, Borokini, dan Alabi 2019). EDM sendiri memiliki banyak area, setidaknya terdapat 11 area di dalam EDM. Salah satu area EDM yang paling populer dan tertua adalah prediksi performa mahasiswa dimana tujuannya adalah untuk memperkirakan nilai yang tidak diketahui dari variabel yang menggambarkan mahasiswa. Dalam pendidikan, nilai-nilai yang biasanya diprediksi adalah performa, pengetahuan, skor, atau nilai (Romero dan Ventura 2010). Prediksi performa mahasiswa sangat penting dalam sistem pendidikan saat ini. Prediksi performa mahasiswa dapat membantu mahasiswa dan pengajar dalam melacak kemajuan mahasiswa (Ragab dkk. 2021). Memprediksi performa mahasiswa dapat menciptakan peluang untuk meningkatkan hasil pendidikan. Dengan pendekatan prediksi performa yang efektif, instruktur dapat mengalokasikan sumber daya dan instruksi yang lebih akurat. Prediksi awal performa mahasiswa dapat membantu pengambil keputusan untuk memberikan tindakan pada saat yang tepat dan untuk merencanakan pembelajaran yang tepat dalam meningkatkan tingkat keberhasilan mahasiswa (Alyahyan dan Düştegör 2020). Penelitian dalam memprediksi performa mahasiswa berusaha untuk menentukan fitur yang saling terkait dan untuk mengidentifikasi alasan yang mendasari mengapa fitur tertentu

lebih berpengaruh daripada fitur yang lain (Hellas dkk. 2018). Prediksi performa akademik mahasiswa merupakan bagian dari prediksi performa mahasiswa (Romero dkk. 2010). Performa akademik mahasiswa merupakan komponen terpenting pada institusi pendidikan tinggi, hal ini dikarenakan lembaga pendidikan tinggi dituntut untuk menghasilkan lulusan yang terampil dengan nilai akademik yang tinggi (Jenitha, Santhi, dan Monisha Privthy Jeba 2021).

Pendekatan-pendekatan dalam prediksi performa akademik mahasiswa telah dilakukan oleh para peneliti sebelumnya diantaranya yaitu (Amrieh, Hamtini, dan Aljarah 2016) mengusulkan model baru untuk prediksi performa mahasiswa berdasarkan teknik data mining dengan atribut/fitur data baru, yang disebut fitur perilaku siswa. Jenis fitur ini terkait interaksi mahasiswa dengan *Learning Management System* (LMS). Aluko dkk (2018) mengusulkan prediksi performa akademik mahasiswa dengan memanfaatkan informasi yang terkandung dalam prestasi akademik sebelumnya. IPK digunakan sebagai metrik untuk mengukur performa akademik mahasiswa. IPK dibagi menjadi 2 (lulus/gagal), IPK lebih besar dari 2,4 dinyatakan lulus dan IPK kurang dari 2,4 dinyatakan gagal. Helal dkk (2018) mengusulkan model klasifikasi yang berbeda untuk memprediksi performa mahasiswa, menggunakan data yang dikumpulkan dari data pendaftaran mahasiswa dan data aktivitas yang dihasilkan dari sistem manajemen pembelajaran universitas (LMS). Data pendaftaran berisi informasi mahasiswa seperti fitur sosio-demografis, dasar masuk universitas (misalnya melalui ujian masuk atau tanpa ujian) dan jenis kehadiran (misalnya penuh waktu vs paruh waktu). Ramaswami dkk (2019) mengumpulkan data interaksi siswa dengan LMS (Xorro-Q) selama satu semester dan satu mata kuliah, data dibagi dua yaitu partisipasi di dalam kelas dan diluar kelas.

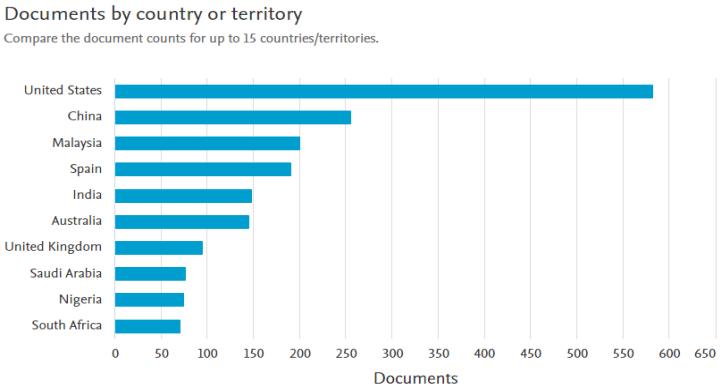
1.4. Perkembangan Penelitian Prediksi Performa Akademik Mahasiswa

Perkembangan penelitian tentang performa akademik mahasiswa terus mengalami peningkatan. Sampai dengan saat ini terdapat 13.653 dokumen terindeks scopus yang membahas terkait performa akademik siswa. Penelitian terkait performa akademik siswa pertama kali pada tahun 1954 oleh Reed M. Merrill dalam rangka mengevaluasi kinerja akademik mahasiswa pada masa percobaan. Namun perkembangan penelitian berkaitan tentang topik ini dari tahun-ketahun tidaklah signifikan. Sampai dengan tahun 1981 penelitian terkait topik ini tidak mencapai angka 10 artikel per tahunnya. Selanjutnya pada tahun 1983-1992 artikel terkait selalu mengalami peningkatan namun belum mencapai angka ratusan di setiap tahunnya. Dimulai tahun 1993-2013 artikel selalu selalu bertambah tetapi jumlah artikel di bawah 500 artikel di setiap tahunnya. Tahun 2014 sampai dengan sekarang jumlah artikel terus meningkat, peningkatan signifikan dimulai dari tahun 2018 sampai dengan sekarang, bahkan di tahun 2021 artikel terkait tentang performa akademik siswa berjumlah 2.485 artikel. Perkembangan jumlah penelitian dari tahun 2002 sampai dengan 2021 ditunjukkan pada gambar 1.2.



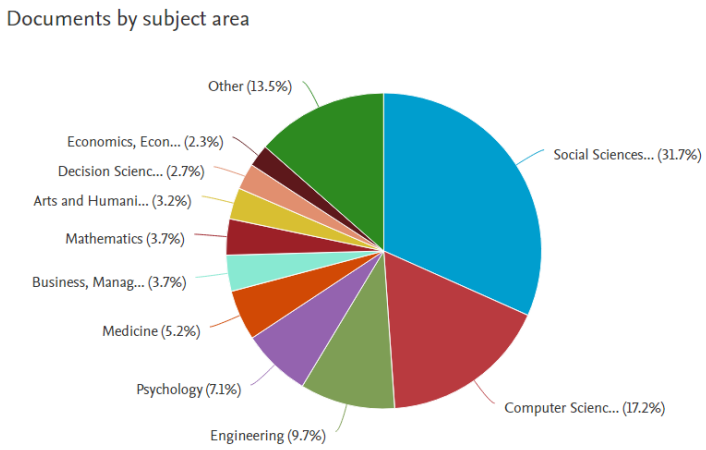
Gambar 1.2 Grafik peningkatan jumlah artikel per tahun

Artikel terkait performa akademik siswa dari tahun 2002 sampai dengan 2021 berjumlah 13.016 artikel. Negara dengan jumlah artikel terbanyak terkait topik performa akademik siswa adalah Amerika Serikat dengan 4.098 artikel, selanjutnya India dengan 815 artikel, China, Spanyol, Australia dengan jumlah masing-masing 700san artikel. Sementara itu Indonesia berada di peringkat ke 8 dengan jumlah artikel 470. Gambar 1.3 memperlihatkan perbandingan jumlah artikel dari setiap negara.



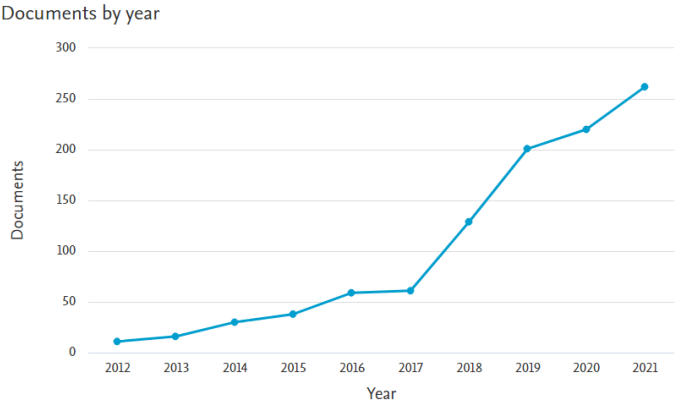
Gambar 1.3 Perbandingan jumlah artikel dari berbagai negara

Artikel-artikel yang membahas tentang performa akademik siswa terdiri dari berbagai disiplin ilmu. Ilmu sosial merupakan disiplin ilmu dengan jumlah artikel terbanyak yaitu 1465 artikel (31,7%), sementara itu disiplin ilmu komputer dengan jumlah terbanyak kedua yaitu 796 artikel (17,2%). Ilmu teknik 450 artikel (7,9%), Psikologi 327 artikel (7,1%), Kesehatan 239 artikel (5,2%), Bisnis manajemen dan akuntansi 171 artikel (3,7%), matematika 171 artikel (3,7%) sisanya terdiri dari disiplin ilmu lainnya. Gambar 1.4 menunjukkan sebaran disiplin ilmu tentang topik penelitian.



Gambar 1.4 Sebaran disiplin ilmu terkait topik performa akademik siswa

Perkembangan penelitian dengan fokus pada prediksi performa akademik siswa dalam sepuluh tahun terakhir juga mengalami peningkatan luar biasa. Tercatat 1027 artikel yang terindex di Scopus dari tahun 2012-2021. Gambar 1.5 memperlihatkan perkembangan topik prediksi performa akademik siswa.



Gambar 1.5 Perkembangan topik prediksi performa akademik siswa

Bidang ilmu yang paling banyak menghasilkan artikel pada topik ini adalah Ilmu Komputer dengan total 682 artikel (33.7%), selanjutnya Ilmu Teknik 376 artikel (18,6%), Ilmu Sosial berada di urutan ketiga yang sebelumnya pada topik umum di urutan pertama yaitu 314 artikel (15,5%), Ilmu Matematika 149 artikel (7,4%) sisanya terdiri dari disiplin ilmu lainnya.

BAB II
KONSEP DASAR PREDIKSI PERFORMA AKADEMIK MAHASISWA

2.1 Tinjauan konsep performa akademik

Amrieh dkk (2016) mengusulkan model baru untuk prediksi performa siswa berdasarkan teknik data mining dengan atribut/fitur data baru, yang disebut fitur perilaku siswa. Jenis fitur ini terkait interaksi pelajar dengan sistem manajemen *e-learning* yang diperoleh dari data mahasiswa semester satu dan dua. Metode yang digunakan adalah ensemble yaitu menggunakan sekumpulan model, kemudian menggabungkannya untuk mengambil yang terbaik. Algoritma klasifikasi yang digunakan adalah (ANN, NB and DT) dan menggunakan metode ensemble (Bagging, Boosting dan RF). Pada penelitian ini disampaikan bahwa fitur perilaku sumber yang diakses merupakan fitur yang paling efektif dalam memprediksi performa siswa. Metode ensembel DT dengan Boosting merupakan metode terbaik dengan tingkat akurasi sebesar 82,2%.

Sementara itu (Aluko dkk. 2018) mengusulkan prediksi performa akademik mahasiswa dengan memanfaatkan informasi yang terkandung dalam prestasi akademik sebelumnya. IPK digunakan sebagai metrik untuk mengukur performa akademik mahasiswa. IPK dibagi menjadi 2 (lulus/gagal) IPK 2,4 lulus dan IPK di bawahnya gagal. Algoritma yang digunakan adalah Logistic Regression (LR) dan Support Vector Machine (SVM). Model SVM mengungguli model LR dalam hal akurasi. Hasil penelitian juga menunjukkan bahwa prestasi akademik sebelumnya merupakan prediktor yang baik untuk performa akademik mahasiswa.

Hellas dkk (2018) mengusulkan model klasifikasi yang berbeda untuk memprediksi performa siswa, menggunakan data yang dikumpulkan dari data pendaftaran mahasiswa dan data aktivitas yang dihasilkan dari sistem manajemen pembelajaran universitas (LMS) periode 2011 sampai dengan 2013. Data pendaftaran berisi informasi siswa seperti fitur sosio-demografis, dasar masuk universitas (misalnya melalui ujian masuk atau tanpa ujian) dan jenis kehadiran (misalnya penuh waktu vs paruh waktu). Data LMS merekam keterlibatan siswa dengan aktivitas pembelajaran online mereka. Kontribusi penting dari penelitian ini adalah pertimbangan heterogenitas siswa dalam membangun model prediksi. Hal ini didasarkan pada pengamatan bahwa siswa dengan fitur sosio-demografis atau model belajar yang berbeda dapat menunjukkan motivasi belajar yang berbeda-beda. Lebih lanjut, eksperimen mengungkapkan bahwa mempertimbangkan fitur pendaftaran dan aktivitas perkuliahan membantu dalam mengidentifikasi siswa yang rentan secara lebih tepat. Empat algoritma digunakan untuk memprediksi performa akademik mahasiswa yaitu Naive-Bayes, J48, SMO, JRip. Eksperimen mengungkapkan bahwa tidak ada metode tunggal yang menunjukkan kinerja unggul dalam semua aspek untuk memprediksi performa siswa. Namun, sangat penting untuk menggunakan pengetahuan yang ditemukan untuk memprediksi siswa yang rentan. Kombinasi J48 dan JRip, berkontribusi secara signifikan dengan menghasilkan output yang dapat dipahami masing-masing dalam bentuk pohon dan aturan. Selanjutnya, dengan mempertimbangkan fitur gabungan, diperoleh hasil prediksi yang lebih unggul dalam mengidentifikasi siswa yang tidak berhasil dibandingkan dengan mempertimbangkan fitur secara terpisah.

Ramaswami dkk (2019) mengumpulkan data interaksi siswa dengan LMS (Xorro-Q) selama satu semester dan satu mata

kuliah, data dibagi dua yaitu partisipasi di dalam kelas dan diluar kelas. Selanjutnya mereka membandingkan algoritma Naïve Bayes, Logistic Regression, k-Nearest Neighbor dan Random Forest untuk menemukan algoritma yang paling baik dalam hal akurasi untuk memprediksi performa akademik mahasiswa. Hasil penelitian memperlihatkan bahwa fitur partisipasi diluar kelas memberikan dampak yang baik namun tidak signifikan, selain itu dari hasil penelitian ini algoritma Random Forest adalah algoritma terbaik dari algoritma lainnya.

Algoritma yang digunakan untuk memprediksi performa akademik mahasiswa berbeda-beda antar peneliti oleh karena itu (Arifin dkk. 2021) membandingkan beberapa algoritma yang sering digunakan dalam memprediksi performa akademik mahasiswa. Mereka membandingkan 5 algoritma *Generalized Linear Model* (GLM), *Deep learning* (DL), *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosted Trees* (GBT), *Support Vector Machine* (SVM). Data yang dipergunakan merupakan kumpulan data yang terdiri dari data keuangan, data sosial, dan data akademik dari 12.411 siswa. Hasil percobaan menunjukkan bahwa GBT adalah algoritma yang paling efektif dalam memprediksi kinerja siswa, dengan RMSE paling rendah. Algoritma terbaik diperlukan dalam memprediksi performa akademik. Sebuah perbandingan diperlukan sebelum menentukan algoritma yang akan digunakan. Beberapa peneliti menyatakan bahwa algoritma terbaik dari penelitian mereka bervariasi. Tidak ada satu algoritma yang terbaik dalam memprediksi performa akademik siswa. Hal ini dipengaruhi oleh data yang diperoleh dan proses preprocessing. Oleh karena itu, peneliti perlu membandingkan algoritma sebelum dipergunakan dalam memprediksi performa akademik.

2.2 Faktor-faktor yang mempengaruhi performa akademik mahasiswa

Pencarian faktor-faktor yang mendukung performa akademik telah dilakukan oleh banyak peneliti. Wolaver (2002) menganalisa hubungan performa akademik dengan mahasiswa yang mengkonsumsi alkohol. Mekonen dkk. (2017) meneliti hubungan prestasi akademik dengan penggunaan penggunaan zat (alkohol, tembakau dan khat). Romer (1993); Cohn dkk (1998) meneliti hubungan antara tingkat kehadiran mahasiswa terhadap nilai akademik. Abu (2016) menemukan ada banyak fitur lain yang memiliki pengaruh yang sama dengan lebih signifikan termasuk fitur demografis, seperti jenis kelamin, dan pekerjaan ibu, serta fitur pra-pendaftaran, seperti nilai sekolah menengah, dan diskon biaya Universitas. Sementara itu Alsinaidi dkk. (2021) dalam penelitiannya menyampaikan bahwa ekstrakurikuler mempengaruhi performa akademik mahasiswa. Arola Anderson dkk. (2021) menyampaikan bahwa aktifitas organisasi kampus pada semester pertama memiliki pengaruh positif terhadap optimis mahasiswa, membantu mengatasi perubahan (Bohnert, Fredricks, dan Randall 2010), berpengaruh positif terhadap perkembangan dan perilaku anak (Morris 2015), memiliki pengaruh terhadap masa tunggu lulusan dalam mendapatkan pekerjaan (Arifin, Widowati, dan Farikhin 2022), tabel 2.1 merupakan variabel-variabel yang telah digunakan oleh para peneliti untuk memprediksi performa akademik.

Tabel 2.1 Variabel-variabel prediksi performa akademik mahasiswa

NO	PENULIS	VARIABEL
1	Amrieh dkk., 2016	(1) Fitur demografis seperti jenis kelamin dan kebangsaan. (2) Fitur latar belakang akademik seperti tahap pendidikan, tingkat kelas dan

		bagian. (3) Fitur perilaku, seperti mengangkat tangan di kelas, sumber yang akses dll.
2	Aluko dkk., 2018	Prestasi akademik sebelumnya, Tes masuk, IPK
3	Helal dkk., 2018	(1) LMS, (2) fitur sosio-demografis, (3) Dasar penerimaan (ujian masuk atau nilai SLTA) dan (4) Jenis kehadiran (misalnya penuh waktu vs paruh waktu).
4	Yaacob dkk., 2019	Transkrip nilai dan data mahasiswa
5	Sudani dan Ramaswamy, 2019	Kelembagaan, akademik, demografi, psikologis dan ekonomi
6	Leo & Yulia (2020)	Demografi, Akademik, LMS (lentera).
7	Arifin dkk, 2021	Informasi akademik, sosial, ekonomi dari 12.411 siswa di Bogota Colombia
8	Halit Karalar dkk, 2021	IPK, LMS (Moodle) dan Conference Management Software (Adobe Connect) selama satu semester

2.3 Data yang digunakan untuk prediksi performa akademik

Prediksi performa akademik mahasiswa sebagian besar menggunakan dua jenis kumpulan data akademik yaitu database perguruan tinggi/universitas dan data platform pembelajaran online (Albreiki, Zaki, dan Alashwal 2021). Para peneliti umumnya menggunakan data *E-learning* dari sebuah mesin pembelajaran (LMS) dalam melakukan prediksi performa akademik mahasiswa, sementara itu beberapa peneliti menyampaikan bahwa perlu ditambahkan fitur-fitur lain untuk memprediksi performa akademik mahasiswa. Fitur-fitur non akademik juga perlu diidentifikasi, termasuk penilaian demografis dan eksternal mahasiswa, kegiatan ekstrakurikuler, latar belakang sekolah menengah, dan jaringan interaksi sosial (Abu Saa dkk. 2019).

E-learning adalah sebuah proses pembelajaran yang difasilitasi dan didukung dengan penggunaan teknologi informasi dan komunikasi (Jenkins dan Hanson 2003). Sun dkk (2009) mendefinisikan *e-learning* merupakan sebuah paradigma baru

dalam dunia pendidikan modern, dimana proses pembelajaran dan pelatihan disampaikan dengan menggunakan media teknologi telekomunikasi yang memungkinkan adanya kebebasan berinteraksi, baik antara peserta didik dan pendidik maupun sesama peserta didik tanpa ada batas ruang dan waktu dengan menggunakan model belajar *sinkron* dan *asinkron*. Materi dan instruksi pembelajaran pada *e-learning* diberikan melalui pemanfaatan media elektronik, termasuk Internet, intranet, ekstranet, broadcast satelit, audio/video tape, CD-ROM, dan televisi interaktif. Terdapat beberapa karakteristik yang dapat membedakan *E-learning* dengan pembelajaran konvensional. Tabel 2.2 menyajikan perbandingan karakteristik antara pembelajaran konvensional dengan *e-learning* (Du, Liu, dan Wei 2010).

Tabel 2.2 Perbandingan karakteristik pembelajaran

Aspek Karakteristik	**Pembelajaran konvensional**	**E-learning**
Waktu belajar	Tetap	Dinamis
Tempat belajar	Ruang kelas	Dimana saja
Jumlah peserta didik	Terbatas	Tidak terbatas
Konten dan sumber pembelajaran	Sedikit dan tidak mudah untuk melakukan pembaruan	Banyak dan tersedia untuk semua peserta
Efek dan efektivitas	Tidak tinggi dan tidak efektif untuk setiap peserta	Tinggi dan lebih efektif untuk setiap peserta
Komunikasi	Tatap muka	Mudah berkomunikasi, tidak tatap muka
Evaluasi dan pemantauan	Mudah mengevaluasi dan memantau	Sulit mengevaluasi dan memantau

(Sumber: (Du, Liu & Wei, 2010)

Platform e-learning menawarkan fitur dan layanan yang hampir sama antar *platform*. Saat ini, yang paling banyak dipergunakan untuk pembelajaran adalah Moodle *(modular*

object-oriented developmental learning environment), sistem manajemen pembelajaran gratis yang memungkinkan pembuatan *e-learning* yang powerful, fleksibel, dan menarik.

Sistem Moodle menyimpan banyak informasi rinci tentang konten pembelajaran, pengguna, dan penggunaan dalam database relasional. Moodle terdiri dari kuis, tugas, dan forum. Kuis adalah alat yang berguna bagi siswa untuk menguji tingkat pengetahuan mereka dan meninjau setiap mata pelajaran yang dipelajari. Mereka adalah alat yang hebat untuk memberi siswa umpan balik yang cepat tentang kinerja mereka dan untuk mengukur pemahaman mereka tentang materi. Tugas adalah alat untuk mengumpulkan pekerjaan siswa. Modul tugas memberikan cara mudah untuk memungkinkan siswa mengunggah konten digital untuk penilaian. Mereka dapat diminta untuk mengirimkan esai, spreadsheet, presentasi, halaman web, foto, atau klip audio atau video kecil. Forum adalah alat komunikasi yang kuat dalam Moodle. Mereka memungkinkan pendidik dan siswa untuk berkomunikasi satu sama lain kapan saja, dari mana saja dengan koneksi Internet. Siswa tidak harus login pada saat yang sama untuk berkomunikasi dengan guru atau teman sekelasnya. Dengan demikian, siswa dapat meluangkan waktu untuk menyusun balasan, yang dapat mengarah pada diskusi yang lebih bijaksana. Banyak penelitian menunjukkan bahwa lebih banyak siswa yang bersedia berpartisipasi dalam forum asinkron daripada yang bersedia berbicara di kelas. Forum menciptakan banyak kesempatan untuk meniru percakapan yang Anda lakukan di kelas, merumuskan diskusi tugas antara kelompok siswa, atau membawa ide dan pertanyaan terbaik dari forum ke dalam kelas (Romero dkk. 2010).

Bravo-Agapito dkk (2021) mengekstrak variabel-variabel yang digunakan untuk memprediksi performa akademik mahasiswa dari LMS Moodle, dijabarkan dalam tabel 2.3.

Tabel 2.3 Variabel Moodle untuk prediksi performa akademik siswa

No	Nama	Deskripsi
1	*Total_logins*	Jumlah total login ke platform Moodle
2	*N_access_forum*	Frekuensi akses mahasiswa ke semua forum
3	*N_added_messages_forum*	Jumlah total pesan yang ditambahkan oleh mahasiswa di forum
4	*N_access_didactic_units*	Frekuensi akses siswa terhadap bahan ajar
5	*N_access_glossaries*	Frekuensi akses siswa ke semua glosarium
6	*Total_assignments*	Jumlah tugas kursus
7	*N_assignments_consulted*	Frekuensi siswa berkonsultasi tugas
8	*N_assignments_submitted*	Jumlah tugas yang diserahkan
9	*N_access_questionnaires*	Frekuensi siswa mengakses semua kuesioner
10	*N_attempts_questionnaires*	Frekuensi siswa mencoba untuk memecahkan kuesioner
11	*N_answered_questions*	Jumlah pertanyaan yang dijawab di semua kuesioner
12	*N_questionnaire_views*	Frekuensi siswa mengamati angket
13	*N_questionnaires_submitted*	Frekuensi siswa menyerahkan angket
14	*N_reviews_questionnaire*	Frekuensi siswa upaya untuk merevisi kuesioner
15	*Days_first_access*	Jumlah hari hingga akses pertama ke kelas LMS
16	*N_entries course*	Jumlah total entri ke kursus

Conijn dkk., (2017) menganalisis 17 mata kuliah campuran dengan 4.989 siswa menggunakan log LMS Moodle. Tujuan mereka adalah untuk memprediksi nilai akhir mahasiswa dari variabel prediktor LMS dan nilai tugas, menggunakan model logistik (lulus/gagal) dan regresi standar (nilai akhir). Mereka memprediksi performa mahasiswa selama 10 minggu pertama. Akurasi sedikit meningkat saat minggu pertama, dengan peningkatan yang mencolok setelah minggu ke-5, saat nilai tugas tersedia. Pada minggu ke-5, model regresi menunjukkan 0,43

penyesuaian R2, dan pengklasifikasi biner lulus/gagal memperoleh akurasi 67% pada minggu ke-3.

Gerritsen menggunakan file log Moodle dari 17 mata pelajaran untuk meramalkan apakah seorang mahasiswa akan lulus atau gagal dalam suatu mata pelajaran (klasifikasi biner) (Gerritsen 2017). Dari tujuh model, *multilayer perceptron* mengungguli pengklasifikasi lainnya, berhasil mendeteksi siswa berisiko dengan akurasi 66,1%.

BAB III
METODE DAN TEKNIK DALAM EDUCATIONAL DATA MINING

3.1. Pengumpulan dan Persiapan Data

a. Identifikasi jenis data yang diperlukan untuk prediksi performa akademik

Untuk melakukan prediksi performa akademik dalam educational data mining (EDM), ada beberapa jenis data yang dapat diidentifikasi dan dikumpulkan. Data ini mencakup informasi tentang siswa, lingkungan belajar, dan konteks pendidikan. Berikut adalah beberapa jenis data yang penting untuk prediksi performa akademik:

1. Data Demografis: Data demografis meliputi informasi tentang identitas siswa, seperti usia, jenis kelamin, latar belakang etnis, dan latar belakang sosioekonomi. Data ini dapat memberikan konteks tentang siswa dan faktor-faktor demografis yang dapat mempengaruhi performa akademik.
2. Data Akademik Sebelumnya: Data akademik sebelumnya mencakup informasi tentang prestasi akademik siswa di masa lalu, seperti nilai tes atau ujian, rapor, dan catatan absensi. Data ini membantu dalam memahami sejarah akademik siswa dan dapat memberikan indikasi awal tentang performa masa depan.
3. Data Keterlibatan Belajar: Data keterlibatan belajar mencakup informasi tentang kehadiran siswa di kelas, partisipasi dalam diskusi, pengumpulan tugas, dan interaksi dengan materi pembelajaran. Data ini dapat mencerminkan tingkat keterlibatan siswa dalam proses pembelajaran dan dapat mempengaruhi performa akademik.

4. Data Interaksi Online: Jika ada penggunaan teknologi dalam proses pembelajaran, data interaksi online dapat dikumpulkan dari platform pembelajaran elektronik atau sistem manajemen pembelajaran (LMS). Data ini mencakup aktivitas siswa seperti mengakses materi pembelajaran, berinteraksi dengan rekan sekelas atau guru melalui forum diskusi, atau mengikuti ujian online. Data ini memberikan wawasan tentang pola belajar dan preferensi siswa.
5. Data Perilaku Belajar: Data perilaku belajar mencakup informasi tentang strategi belajar siswa, tingkat motivasi, kepercayaan diri, dan gaya belajar. Data ini dapat dikumpulkan melalui survei, kuesioner, atau instrumen penilaian psikologis. Informasi ini membantu memahami faktor psikologis yang memengaruhi performa akademik.
6. Data Lingkungan Belajar: Data lingkungan belajar mencakup informasi tentang konteks fisik dan sosial di mana pembelajaran terjadi. Ini termasuk informasi tentang jenis sekolah, ukuran kelas, fasilitas, dukungan sosial, dan iklim kelas. Data ini membantu memahami bagaimana faktor-faktor lingkungan dapat mempengaruhi performa akademik siswa.
7. Data Evaluasi Formatif: Data evaluasi formatif mencakup hasil pengukuran formatif selama proses pembelajaran, seperti tes formatif, tugas, atau proyek. Data ini memberikan pemahaman tentang pemahaman siswa tentang materi pembelajaran saat ini dan dapat digunakan sebagai indikator performa akademik di masa depan.

Penting untuk merencanakan dan mengumpulkan data dengan hati-hati, memastikan bahwa data yang dikumpulkan relevan, akurat, dan sesuai dengan tujuan prediksi performa akademik yang

b. Teknik pengumpulan data yang efektif

Teknik pengumpulan data yang efektif dalam educational data mining (EDM) adalah suatu pendekatan yang memungkinkan pengumpulan data yang relevan, representatif, dan berkualitas tinggi untuk tujuan analisis pendidikan. Berikut ini beberapa teknik yang dapat digunakan untuk mencapai tujuan tersebut:

1. Pengumpulan data langsung: Teknik ini melibatkan pengumpulan data secara langsung dari sumber utama, seperti siswa, guru, atau sistem pembelajaran. Contohnya adalah penggunaan survei, wawancara, atau observasi langsung di dalam kelas. Data langsung ini memberikan pemahaman yang mendalam tentang perilaku dan pengalaman individu dalam konteks pendidikan.
2. Pengumpulan data otomatis: Dalam EDM, penggunaan teknologi dapat memungkinkan pengumpulan data otomatis. Misalnya, penggunaan sistem manajemen pembelajaran (LMS) yang merekam aktivitas online siswa secara otomatis, seperti interaksi dengan materi pembelajaran, waktu yang dihabiskan di platform, atau hasil tes yang diperoleh. Dengan memanfaatkan teknologi ini, data dapat dikumpulkan secara terus-menerus dan dalam skala yang lebih besar.
3. Penggunaan instrumen pengukuran: Penggunaan instrumen pengukuran, seperti tes atau kuesioner, dapat menjadi teknik efektif untuk mengumpulkan data dalam EDM. Instrumen ini dapat memberikan data kuantitatif yang dapat diolah dan dianalisis dengan mudah. Penting untuk merancang instrumen yang valid dan reliabel agar data yang dikumpulkan dapat menghasilkan informasi yang akurat dan dapat diandalkan.
4. Pemantauan sensor: Dalam beberapa konteks pendidikan, penggunaan sensor atau perangkat canggih lainnya dapat

memberikan data yang berharga. Misalnya, penggunaan sensor gerakan untuk melacak aktivitas fisik siswa atau perangkat penjejak mata untuk mempelajari pola perhatian saat belajar. Pengumpulan data melalui sensor ini dapat memberikan wawasan tambahan tentang karakteristik dan tingkah laku siswa yang tidak dapat diperoleh melalui metode tradisional.

5. Penggunaan data yang sudah ada: Terkadang, data yang relevan sudah tersedia dalam bentuk yang terstruktur, seperti basis data sekolah atau rekaman hasil tes. Menggunakan data yang sudah ada dapat menghemat waktu dan sumber daya. Namun, perlu dipastikan bahwa data tersebut valid, akurat, dan relevan dengan tujuan EDM yang ingin dicapai.

Dalam pengumpulan data untuk EDM, penting untuk mempertimbangkan isu privasi dan etika. Pastikan kegiatan pengumpulan data dilakukan dengan izin yang diperlukan dan mengikuti pedoman dan peraturan yang berlaku terkait perlindungan data pribadi.

c. Pembersihan dan preprocessing data

Pembersihan dan preprocessing data dalam educational data mining (EDM) merujuk pada serangkaian langkah yang dilakukan untuk membersihkan, mempersiapkan, dan mengorganisir data sebelum dilakukan analisis. Langkah-langkah ini penting untuk memastikan bahwa data yang digunakan dalam analisis EDM berkualitas tinggi, konsisten, dan siap untuk dieksplorasi. Berikut adalah beberapa tahapan yang umum dilakukan dalam pembersihan dan preprocessing data EDM:

1. Identifikasi dan penanganan missing values: Missing values adalah nilai yang hilang atau tidak ada dalam data. Pada tahap ini, langkah-langkah diambil untuk mengidentifikasi dan

menangani missing values, seperti menghapus baris atau kolom yang memiliki banyak missing values atau menggantikan nilai yang hilang dengan nilai yang sesuai berdasarkan konteks data.

2. Pendeteksian dan penanganan outliers: Outliers adalah data yang jauh berbeda dari pola umum dalam dataset. Outliers dapat mempengaruhi analisis EDM dan perlu ditangani dengan hati-hati. Pada tahap ini, outliers diidentifikasi menggunakan metode statistik atau teknik lainnya, dan keputusan diambil untuk membuang outliers atau mengolahnya secara khusus tergantung pada konteks analisis.
3. Normalisasi dan transformasi data: Normalisasi dan transformasi data dilakukan untuk memastikan data memiliki skala yang seragam atau untuk mengubah distribusi data menjadi lebih normal. Contohnya adalah mengubah rentang nilai data ke dalam rentang yang sama atau melakukan transformasi logaritmik pada data yang memiliki distribusi yang condong.
4. Encoding variabel kategorikal: Jika dataset mengandung variabel kategorikal, langkah-langkah encoding diperlukan untuk mengubah variabel tersebut menjadi bentuk numerik yang dapat diproses oleh algoritma analisis EDM. Encoding dapat dilakukan dengan metode seperti one-hot encoding, label encoding, atau penggunaan variabel dummy.
5. Integrasi dan penggabungan data: Dalam beberapa kasus, data EDM dapat berasal dari sumber yang berbeda atau terpisah. Pada tahap ini, langkah-langkah diambil untuk mengintegrasikan dan menggabungkan data dari sumber yang berbeda agar dapat digunakan secara bersama-sama dalam analisis.
6. Seleksi fitur: Seleksi fitur adalah proses memilih subset variabel yang paling relevan dan signifikan untuk tujuan

analisis EDM. Dengan memilih fitur yang tepat, dapat mengurangi kompleksitas analisis dan meningkatkan interpretabilitas hasil.

7. Reduksi dimensi: Reduksi dimensi melibatkan mengurangi jumlah dimensi atau variabel dalam dataset sambil mempertahankan sebagian besar informasi yang relevan. Metode seperti analisis komponen utama (PCA) atau analisis faktor dapat digunakan untuk mengurangi dimensi data.

Setelah pembersihan dan preprocessing data selesai, data siap untuk digunakan dalam analisis EDM. Penting untuk dicatat bahwa tahapan-tahapan ini bersifat iteratif, dan terkadang diperlukan percobaan dan evaluasi ulang untuk memastikan bahwa data yang diolah memenuhi persyaratan dan tujuan analisis yang ditetapkan.

3.2. Analisis Eksplorasi Data

a. Teknik eksplorasi data untuk memahami karakteristik dataset

Teknik eksplorasi data merupakan serangkaian metode yang digunakan untuk memahami karakteristik dataset dalam educational data mining (EDM). Tujuan utama dari eksplorasi data adalah untuk mendapatkan wawasan awal tentang data yang ada, mengidentifikasi pola, dan menemukan hubungan atau tren yang mungkin tersembunyi di dalamnya. Berikut adalah beberapa teknik eksplorasi data yang umum digunakan dalam EDM:

1. Statistik Deskriptif: Statistik deskriptif digunakan untuk merangkum dan menggambarkan karakteristik dasar dari dataset. Ini termasuk menghitung statistik seperti rata-rata, median, modus, simpangan baku, kuartil, dan rentang data. Statistik ini memberikan informasi tentang distribusi, pemusatan, dan variabilitas data.

2. Visualisasi Data: Visualisasi data melibatkan penggunaan grafik dan diagram untuk mewakili data secara visual. Diagram seperti histogram, diagram batang, scatter plot, atau box plot dapat digunakan untuk memvisualisasikan distribusi data, pola hubungan antar variabel, atau perbedaan antara kelompok data. Visualisasi ini membantu dalam memahami pola dan tren yang mungkin terdapat dalam dataset.
3. Korelasi dan Analisis Regresi: Analisis korelasi digunakan untuk mengukur hubungan antara dua variabel dalam dataset. Hal ini dapat membantu mengidentifikasi keterkaitan antara variabel-variabel tertentu dengan performa akademik. Analisis regresi dapat digunakan untuk memodelkan hubungan antara variabel independen dan variabel dependen, memungkinkan prediksi performa akademik berdasarkan variabel yang ada.
4. Analisis Kelompok dan Pembandingan: Dalam EDM, sering kali ada kebutuhan untuk membandingkan kelompok atau subkelompok dalam dataset. Analisis ini dapat melibatkan pembandingan performa akademik antara kelompok yang berbeda berdasarkan variabel seperti jenis kelamin, tingkat pendidikan, atau latar belakang etnis. Ini dapat memberikan wawasan tentang faktor-faktor yang memengaruhi performa akademik.
5. Analisis Klasifikasi dan Klastering: Analisis klasifikasi digunakan untuk mengklasifikasikan data menjadi kelompok atau kategori berdasarkan pola yang ada dalam dataset. Ini dapat membantu dalam mengidentifikasi profil siswa atau pola perilaku yang berkaitan dengan performa akademik. Analisis klastering digunakan untuk mengelompokkan data menjadi klaster atau kelompok yang memiliki kesamaan tertentu. Ini dapat membantu dalam mengungkap pola dan karakteristik siswa yang serupa.

6. Analisis Urutan Waktu: Jika dataset berisi data yang dikumpulkan seiring waktu, analisis urutan waktu dapat digunakan untuk memahami tren dan pola perubahan seiring waktu. Ini dapat melibatkan analisis deret waktu, model prediksi, atau penggunaan grafik garis untuk memvisualisasikan perubahan dalam performa akademik seiring waktu.
7. Analisis Frekuensi dan Asosiasi: Analisis frekuensi digunakan untuk mengidentifikasi kejadian yang paling umum atau jarang dalam dataset. Ini dapat memberikan wawasan tentang pola atau tren yang muncul dalam data. Analisis asosiasi digunakan untuk mengidentifikasi hubungan atau keterkaitan antara variabel-variabel dalam dataset, misalnya, apakah ada asosiasi antara jenis tugas dan hasil ujian siswa.

Teknik eksplorasi data ini membantu peneliti dan praktisi EDM untuk memahami karakteristik dataset dengan lebih baik dan membimbing tahapan analisis yang lebih lanjut untuk tujuan prediksi performa akademik.

b. Evaluasi kualitas data dan penanganan outliers

Evaluasi kualitas data dan penanganan outliers merupakan dua langkah penting dalam educational data mining (EDM) untuk memastikan bahwa data yang digunakan dalam analisis memiliki kualitas yang baik dan tidak terdapat nilai-data yang ekstrim (outliers) yang dapat mempengaruhi hasil analisis. Berikut adalah penjelasan lebih lanjut mengenai kedua aspek tersebut:

1. Evaluasi Kualitas Data:

   Evaluasi kualitas data melibatkan penilaian terhadap keandalan, kevalidan, kelengkapan, dan ketepatan data yang digunakan dalam EDM. Beberapa langkah yang dapat

dilakukan untuk mengevaluasi kualitas data adalah sebagai berikut:

- Pemeriksaan kesalahan atau kesalahan pengetikan: Memeriksa apakah ada kesalahan pengetikan atau kesalahan dalam data seperti angka yang salah atau huruf yang terbalik. Pengecekan ini penting untuk memastikan bahwa data yang digunakan akurat.
- Identifikasi dan penanganan missing values: Mengidentifikasi apakah ada data yang hilang atau missing values. Jika ditemukan missing values, langkah-langkah penanganan yang tepat dapat dilakukan, seperti menghapus baris atau kolom yang memiliki banyak missing values, atau menggantikan nilai yang hilang dengan estimasi yang tepat.
- Verifikasi konsistensi data: Memeriksa konsistensi data antara variabel yang saling terkait. Misalnya, memeriksa apakah nilai-nilai dalam variabel yang seharusnya saling terkait memiliki hubungan yang konsisten.
- Validasi terhadap sumber data lain: Membandingkan data dengan sumber data lain yang tersedia untuk memastikan konsistensi dan keandalan data.
- Pemeriksaan outlier: Melakukan identifikasi dan penanganan outliers, yang merupakan nilai-data ekstrim yang jauh berbeda dari nilai-data lainnya. Outliers dapat mempengaruhi hasil analisis dan perlu ditangani secara khusus (seperti dijelaskan di bawah).

2. Penanganan Outliers:

Outliers adalah nilai-data yang ekstrim dan tidak umum dalam dataset. Penanganan outliers adalah proses mengidentifikasi, memverifikasi, dan menentukan apakah outlier tersebut harus dihapus atau diperlakukan secara

berbeda. Beberapa teknik yang digunakan untuk penanganan outliers antara lain:

- Identifikasi outliers: Menggunakan metode statistik seperti analisis box plot atau deviasi standar untuk mengidentifikasi outlier dalam dataset.
- Penanganan outliers: Setelah outliers diidentifikasi, keputusan dapat diambil apakah outlier tersebut harus dihapus atau diproses secara khusus tergantung pada konteks analisis. Dalam beberapa kasus, outliers mungkin merupakan data yang valid dan signifikan, dan oleh karena itu tidak boleh dihapus tanpa alasan yang kuat.
- Penggunaan teknik penggantian nilai: Jika outliers tidak dihapus, teknik penggantian nilai dapat digunakan untuk menggantikan outlier dengan nilai yang lebih representatif, seperti menggunakan nilai rata-rata atau median.
- Melaporkan hasil dengan dan tanpa outliers: Kadang-kadang, penting untuk melaporkan hasil analisis dengan dan tanpa outliers untuk memahami dampak outliers terhadap temuan dan kesimpulan.

Dalam kedua aspek ini, penting untuk memiliki pemahaman yang baik tentang konteks analisis, serta menggunakan metode yang tepat untuk mengevaluasi kualitas data dan menangani outliers dengan bijak agar hasil analisis menjadi lebih andal dan akurat dalam EDM.

3.3. Pemilihan Fitur (Feature Selection)

Metode seleksi fitur merupakan proses pemilihan subset fitur yang paling relevan dan informatif dari suatu dataset untuk digunakan dalam prediksi performa akademik dalam educational data mining (EDM). Tujuan dari seleksi fitur adalah untuk

mengurangi dimensi data, meningkatkan keakuratan prediksi, mengurangi overfitting, dan mempercepat proses analisis. Berikut adalah beberapa metode seleksi fitur yang efektif yang dapat digunakan dalam prediksi performa akademik:

1. Analisis Korelasi: Metode ini mengukur hubungan antara setiap fitur dengan variabel target (misalnya, nilai akademik) menggunakan metrik korelasi, seperti korelasi Pearson. Fitur-fitur yang memiliki korelasi tinggi dengan variabel target dianggap lebih relevan dan dipilih untuk digunakan dalam model prediksi.
2. Analisis Informasi: Metode ini menggunakan metrik seperti gain informasi, rasio informasi, atau keuntungan Gini untuk mengukur tingkat informasi yang disediakan oleh setiap fitur dalam mengklasifikasikan variabel target. Fitur-fitur yang memberikan informasi yang lebih tinggi dianggap lebih penting dan dipilih untuk digunakan dalam model prediksi.
3. Analisis Statistik: Metode ini melibatkan penggunaan metrik statistik, seperti uji-t atau uji-F, untuk menguji signifikansi statistik dari hubungan antara setiap fitur dengan variabel target. Fitur-fitur yang memiliki signifikansi statistik yang tinggi dianggap lebih relevan dan dipilih untuk digunakan dalam model prediksi.
4. Wrapper Methods: Metode ini melibatkan penggunaan algoritma pembungkus (wrapper algorithm) yang menggabungkan proses seleksi fitur dengan proses pembangunan model prediksi. Ini melibatkan membangun model prediksi dengan subset fitur tertentu, menguji kinerjanya, dan menggunakan kriteria seperti akurasi atau validasi silang untuk menentukan subset fitur yang optimal.
5. Embedded Methods: Metode ini melibatkan penggunaan algoritma pembelajaran yang memiliki mekanisme bawaan untuk memilih fitur yang relevan selama proses

pembelajaran. Contohnya, algoritma seperti Regresi Lasso atau Pohon Keputusan dapat memberikan bobot nol pada fitur yang tidak relevan atau tidak informatif.

6. Pengurangan Dimensi: Metode ini melibatkan penggunaan teknik pengurangan dimensi, seperti Analisis Komponen Utama (PCA) atau Analisis Diskriminan Linier (LDA), untuk mengubah ruang fitur menjadi ruang yang lebih rendah dimensi yang mempertahankan sebagian besar variasi atau informasi yang relevan dari dataset.
7. Pemilihan Fitur Berdasarkan Pengetahuan Domain: Metode ini melibatkan pengetahuan domain ahli yang digunakan untuk memilih fitur-fitur yang dianggap penting dan informatif berdasarkan pemahaman tentang domain dan faktor-faktor yang mempengaruhi performa akademik.

Pilihan metode seleksi fitur yang efektif akan tergantung pada karakteristik dataset, jumlah fitur, kompleksitas model prediksi, dan konteks analisis. Kombinasi dari beberapa metode juga

3.4. Metode Prediksi Performa Akademik Mahasiswa

Para peneliti dalam memprediksi performa akademik mahasiswa umumnya menyertakan perbandingan algoritma, hal ini dilakukan untuk mencari algoritma terbaik yang akan digunakan dalam memprediksi. Kumar dkk. (2012) menyampaikan bahwa tidak ada satu algoritma terbaik dari algoritma lainnya di semua dataset, sehingga perlu membandingkan beberapa algoritma sebelum menggunakannya. Perbandingan algoritma diperlukan sebelum menentukan algoritma terbaik yang akan digunakan dalam memprediksi performa akademik (Arifin dkk. 2021). Beberapa peneliti telah membandingkan algoritma untuk mencari algoritma yang tepat diantaranya adalah Aluko dkk. (2018) mengatakan bahwa *Support*

Vector Machine (SVM) lebih baik dari *Logistic Regression* (LR) dalam hal akurasi untuk memprediksi performa akademik mahasiswa. Kumar dan Salal (2019) menyebutkan bahwa *Decision Tree* (DT) merupakan algoritma terbaik dengan nilai akurasi 98,86% dibandingkan dengan *Neural Network* (NN), *Naïve Bayes* (NB), *K-Nearest Neighbor* (K-NN), dan SVM. Sementara itu, Mengash (2020) mengatakan bahwa *Artificial Neural Network* (ANN) adalah algoritma yang terbaik dibandingkan dengan algoritma DT, SVM, dan NB. Yaacob dkk. (2019) menemukan algoritma NB merupakan algoritma terbaik dibandingkan dengan KNN, DT dan LR. Arifin dkk. (2021) membandingkan algoritma *Generalized Linear Model* (GLM), *Deep Learning* (DL), DT, *Random Forest* (RF), dan SVM dan *Gradient Boosted Trees* (GBT) untuk memprediksi performa akademik mahasiswa, adapun hasil perbandingan menunjukkan bahwa algoritma GBT memiliki tingkat kesalahan terkecil dibanding dengan algoritma lainnya. GBT merupakan teknik *machine learning* untuk masalah *regression* dan *classification*, yang menghasilkan model prediksi dalam bentuk *ensemble (Boosting)* dari model prediksi yang lemah seperti *decision tree*, GBT telah menunjukkan keberhasilan besar dalam berbagai penerapan (Fernandes 2019; Friedman 2001; Hew dkk. 2020; Ketui, Wisomka, dan Homjun 2019; Sawant 2019; Zhang dkk. 2019).

GBT bekerja dengan sangat baik pada kumpulan data berukuran sedang. GBT adalah pendekatan ansambel dengan melatih banyak pelajar individu, biasanya pohon keputusan. Pohon dilatih secara berurutan, dengan setiap pohon belajar dari kesalahan sebelumnya, tidak seperti di RF di mana pohon dilatih secara bersamaan. *Gradient Descent* digunakan untuk mengumpulkan kontribusi dari setiap individu yang dipelajari selama pelatihan untuk membuat satu pembelajar ansambel yang

kuat dari ratusan atau ribuan pembelajar yang lemah (bobot dari masing-masing pohon karenanya akan menjadi parameter model).

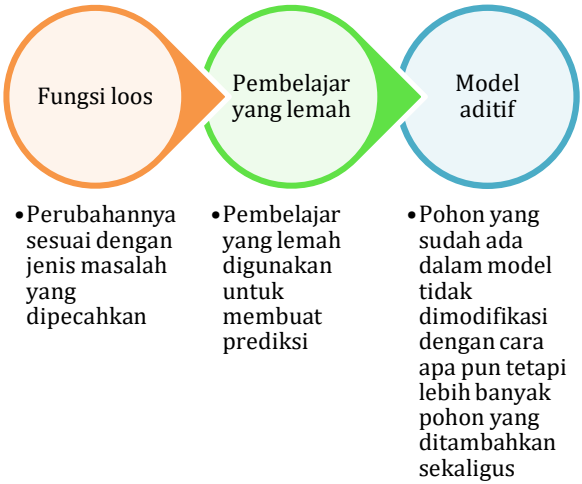
3.4.1. Gradient Boosted Trees (GBT)

Model *Gradient Boosted Trees* (GBT) dipilih sebagai metode atau algoritma prediksi yang handal menangani data numerik, nominal, maupun data yang hilang (missing data) dalam membangun sebuah model prediksi dibandingkan dengan metode prediksi lain (Friedman 2002). GBT merupakan metode ensembel yang membangun sejumlah decision tree. Setiap tree yang dibangun mengacu pada kelemahan tree sebelumnya. Dasar pemikiran GBT adalah model terbaik berikutnya akan meminimalkan kesalahan prediksi keseluruhan jika dikombinasikan dengan model sebelumnya. Tiga elemen utama dari metode boosting ini adalah *loss function*, *weak learner*, dan *additive model*.

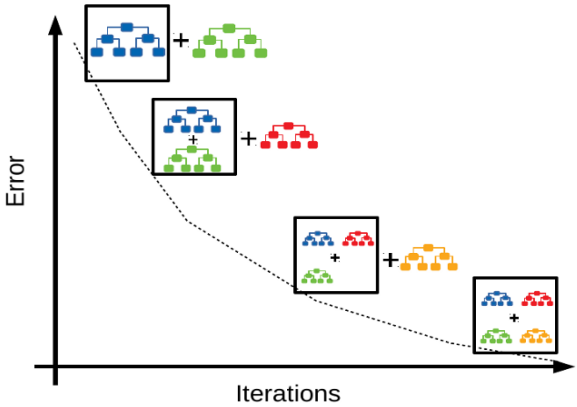
Tujuan dasarnya adalah untuk mengoptimalkan *loss function*. *Loss function* berubah dengan berbagai jenis masalah. Seseorang dapat dengan mudah menentukan fungsi kerugian standar mereka sendiri, tetapi harus dapat dibedakan. Sebagai contoh, kita dapat mengatakan bahwa regresi dapat menggunakan kesalahan kuadrat & klasifikasi dapat menggunakan kerugian algoritmik. Salah satu hal terbaik tentang peningkatan gradien adalah bahwa dengan setiap kerangka kerja, algoritma peningkatan baru tidak diperlukan untuk setiap fungsi kerugian yang dipermasalahkan. Dengan demikian, kerangka kerja yang lebih umum sudah cukup. Sedangkan *weak learner* adalah untuk tujuan membuat prediksi. Pohon keputusan pada dasarnya adalah *weak learner*. Pohon regresi khusus digunakan untuk nilai output riil yang digunakan untuk pemisahan. Kita dapat mengoreksi pengingat dalam model prediksi. Skor kemurnian seperti Gini memilih titik pisah terbaik, yang selanjutnya membangun pohon. Ketika datang ke teknik

peningkatan lain yang disebut Adaboost, seseorang menggunakan pohon keputusan tumpahan tunggal. Dalam hal jumlah level yang lebih tinggi (katakanlah 5 sampai 10), kita dapat menggunakan pohon yang lebih besar. Lebih baik membatasi atau membatasi peserta didik yang lemah dalam menggunakan jumlah simpul daun atau jumlah lapisan atau jumlah belahan, atau bahkan jumlah lapisan. Adapun *additive model* mendefinisikan bahwa pada model ini tidak ada modifikasi pohon yang sudah ada sebelumnya dalam model, tetapi ada penambahan jumlah pohon yang lebih banyak sekaligus. Pada saat penambahan pohon, prosedur penurunan gradien meminimalkan kerugian. Ini meminimalkan jumlah parameter yang ditetapkan. Untuk mengurangi kesalahan, ada pembaruan bobot hanya setelah menghitung kesalahan. Sub-model pembelajar yang lemah menggantikan parameter. Setelah menghitung kerugian, kita harus menambahkan pohon ke model dengan cara mengurangi kerugian sehingga kita dapat melakukan prosedur penurunan gradien. Pada akhirnya, kita dapat menambahkan output ke urutan pohon.

Cara kerja algoritma GBT yaitu bekerja dengan menggabungkan beberapa model yang lemah menjadi sebuah model yang lebih kuat. Algoritma ini menggunakan pendekatan iteratif, di mana setiap iterasi bertujuan untuk meningkatkan model sebelumnya dengan menambahkan model baru. Proses ini dilakukan secara berulang-ulang hingga model yang dihasilkan memenuhi kriteria tertentu, seperti nilai loss function yang cukup kecil.



Gambar 3.1 Cara kerja algoritma GBT



Gambar 3.2 Visualisasi iterasi model boosting

Proses iteratif dalam algoritma Gradient Boosting terdiri dari beberapa tahap, yaitu:

1. Inisialisasi model: Tahap pertama dalam algoritma Gradient Boosting adalah inisialisasi model. Pada tahap ini, model awal dibuat sebagai model konstan yang merupakan rata-rata atau median dari target variable.

2. Membuat weak model: Pada tahap ini, weak model dibuat sebagai model yang mampu memprediksi error dari model sebelumnya. Model lemah biasanya berupa decision tree yang dangkal dengan satu atau dua percabangan.
3. Menghitung residual error: Setelah model lemah dibuat, residual error dihitung sebagai selisih antara nilai prediksi dari model sebelumnya dan nilai asli dari target variable.
4. Menyusun kembali data training: Pada tahap ini, data training diubah dengan menggunakan residual error sebagai target variable.
5. Membuat model baru: Pada tahap ini, model baru dibuat dengan memprediksi residual error yang dihasilkan dari model sebelumnya.
6. Menggabungkan model: Model baru yang dibuat pada tahap sebelumnya digabungkan dengan model sebelumnya untuk membentuk model yang lebih baik.
7. Iterasi berulang: Tahap-tahap di atas diulang berulang-ulang hingga mencapai kondisi berhenti yang ditentukan, seperti jumlah iterasi yang telah ditentukan atau ketika model tidak mengalami peningkatan yang signifikan lagi.

Setelah iterasi selesai dilakukan, model yang dihasilkan akan digunakan untuk memprediksi nilai target pada data testing yang baru.

Beberapa kelebihan GBT yaitu 1). Akurasi yang tinggi; GBT sering menghasilkan model yang akurat dan kuat, terutama ketika digunakan pada data yang kompleks dan tidak terstruktur. 2). Tidak memerlukan persyaratan data yang ketat; Algoritma ini dapat digunakan pada berbagai jenis data tanpa memerlukan asumsi yang ketat, seperti asumsi tentang distribusi data atau homoskedastisitas. 3). Kecepatan komputasi yang cukup cepat. Sedangkan kelemahan dari GBT adalah 1). Memerlukan tuning yang cermat; Algoritma ini memerlukan tuning parameter yang

cermat untuk mendapatkan model yang optimal. Hal ini dapat memakan waktu dan mengharuskan penggunaan cross-validation dan teknik tuning parameter lainnya. 2). Mudah overfitting; GBT dapat cenderung overfit pada data training jika tidak dilakukan pengaturan parameter yang baik. Overfitting terjadi ketika model terlalu kompleks dan terlalu menyesuaikan dengan data training, sehingga tidak dapat melakukan generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. 3). Memerlukan data yang besar; GBT memerlukan jumlah data yang besar untuk memperoleh model yang akurat dan stabil. Jika jumlah data terlalu sedikit, algoritma ini dapat menjadi tidak stabil dan menghasilkan model yang tidak akurat.

Tahapan pengujian algoritma GBT yang dilakukan dengan menggunakan prosedur berikut:

1. ***Input***. Data $\{(X_i, Y_i)\}_{i=1}^{n}$, dan
   perbedaan ***Loss Function*** $L(Y_i, F(x))$

2. **Langkah pertama**: Inisialisasi model dengan nilai konstan
   $F_0(X) = \frac{argmin}{\gamma} \sum_{i=1}^{n} L(Y_i, \gamma)$

3. **Langkah kedua**: Untuk m = 1 ke M:
   (A) Hitung $r_{im} = -\left[\frac{\partial L(y_i, F(X_i))}{\partial F(X_i)}\right]_{F(x)=F_{m-1}(x)}$ untuk i= 1,…,n
   (B) Paskan pohon regresi dengan nilai r_{im} dan buat daerah terminal
   R_{jm} untuk j =1… J_m
   (C) Untuk j = 1...J_m
   Hitung $\gamma_{jm} = \frac{argmin}{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$
   (D) Perbarui $F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$
4. **Langkah ketiga**: Keluaran $F_M(x)$

Contoh perhitungan GBT:

Tabel 3.1 Contoh soal algoritma GBT

Tinggi (m)	Berat (kg)
1.6	88
1.6	76
1.5	56

Input. Data $\{(X_i, Y_i)\}_{i=1}^{n}$, dan perbedaan ***Loss Function*** $L(Y_i, F(x))$

Loss Function = ½ (nilai yang diamati-nilai prediksi)^2

$\frac{d}{d\ nilai\ prediksi}$ ½ (nilai yang diamati-nilai prediksi)^2

= 2/2 (nilai yang diamati-nilai prediksi) x -1

= - (nilai yang diamati-nilai prediksi)

Langkah pertama:

Inisialisasi model dengan nilai konstan:

$$F_0(X) = \frac{argmin}{\gamma} \sum_{i=1}^{n} L(Y_i, \gamma)$$

Loss function = ½ (88 – prediksi)^2 + ½ (76 – prediksi)^2 + ½ (56 – prediksi)^2

= - (88 – prediksi) + - (76 – prediksi) + - (56 – prediksi)

Tentukan jumlah turunan = 0

Maka 0 = - (88 – prediksi) + - (76 – prediksi) + - (56 – prediksi)

= - 88 + prediksi) + - 76 + prediksi + - 56 + prediksi

88+76+56 = prediksi + prediksi +prediksi

88+76+56 = 3 x prediksi

Prediski = $\frac{88+76+56}{3}$

$$F_0(x) = \frac{88 + 76 + 56}{3}$$

= 73.3

Langkah kedua:

Untuk m = 1 ke M:

(a) Hitung $r_{im} = -\left[\frac{\partial L(y_i, F(X_i))}{\partial F(X_i)}\right]_{F(x)=F_{m-1}(x)}$ untuk i= 1,…,n

$r_{1,1}$ = (88 – 73.3) = 14.7

$r_{1,1}$ = (76 – 73.3) = 2.7

$r_{1,1}$ = (58 – 73.3) = -17.3

Tabel 3.2 Nilai residu pertama

Tinggi (m)	Berat (kg)	$r_{i,1}$
1.6	88	14.7
1.6	76	2.7
1.5	56	-17.3

(b) Paskan pohon regresi dengan nilai r_{im} dan buat daerah terminal

R_{jm} untuk j =1… J_m

Membuat Pohon Keputusan

Tinggi < 1.55:

Cabang kiri = -17.3 (R1,1)

Cabang kanan = 14.7 dan 2.7 (R2,1)

(c) untuk j = 1...J_m

Hitung $\gamma_{jm} = \frac{argmin}{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

$$\gamma_{1,1} = \frac{argmin}{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$$

$L(y_i, F_{m-1}(x_i) + \gamma)$ adalah Loss Function

$$\gamma_{1,1} = \frac{argmin}{\gamma} \sum_{x_i \in R_{ij}} \frac{1}{2}(y_i - (F_{m-1}(x_i) + \gamma))^2$$

$\sum_{x_i \in R_{ij}}$ = Penjumlahan ini diubah kedalam individu

$$\gamma_{1,1} = \frac{argmin}{\gamma} \frac{1}{2}(y_3 - (F_{m-1}(x_3) + \gamma))^2$$

$$\gamma_{1,1} = \frac{argmin}{\gamma} \frac{1}{2}(56 - (73.3 + \gamma))^2$$

$$\gamma_{1,1} = \frac{argmin}{\gamma} \frac{1}{2}(-17.3 - \gamma)^2$$

$\frac{d}{d\gamma} \frac{1}{2}(-17.3 - \gamma)^2$ menjadi 17.3 + γ = 0

$$\gamma = -17.3$$

$\gamma_{1,1}$= -17.3

$$\gamma_{2,1} = \frac{argmin}{\gamma} \left[\frac{1}{2}(14.7 - \gamma))^2 + \frac{1}{2}(2.7 - \gamma))^2\right]$$

$$\frac{d}{d\gamma} \left[\frac{1}{2}(14.7 - \gamma))^2 + \frac{1}{2}(2.7 - \gamma))^2\right]$$

Diturunkan: -14.7 + γ + -2.7 + γ = 0

-14.7 + -2.7 + 2γ = 0

2γ = 14.7 + 2.7

γ = $\frac{14.7 + 2.7}{2}$ = 8.7

Pohon Pertama ($F_1(x)$)

Tinggi < 1.55:

Cabang kiri = -17.3 (R1,1), $\gamma_{1,1}$= -17.3

Cabang kanan = 14.7 dan 2.7 (R2,1), $\gamma_{2,1}$= 8.7

(d) Perbarui $F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$

$$F_1(x) = 73.3 + 0.1 * 8.7 = 74.2$$

$$F_2(x) = 73.3 + 0.1 * 8.7 = 74.2$$

$$F_3(x) = 73.3 + 0.1 * -17.3 = 71.6$$

Tabel 3.3 Nilai residu kedua

Tinggi (m)	Berat (kg)	$r_{i,2}$
1.6	88	13.8
1.6	76	1.8
1.5	56	-15.6

Pohon Kedua ($\boldsymbol{F_2(x)}$)

Tinggi < 1.55:

Cabang kiri = 15.6 (R1,2), $\gamma_{1,2}$= -15.6

Cabang kanan = 13.8 dan 1.8 (R2,2), $\gamma_{2,2}$= 7.8

(e) Perbarui $F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$

$$F_1(x) = 73.3 + 0.1 * 7.8 = 74.1$$

$$F_2(x) = 73.3 + 0.1 * 7.8 = 74.1$$

$$F_3(x) = 73.3 + 0.1 * -15.6 = 71.7$$

Tabel 3.4 Nilai residu ketiga

Tinggi (m)	Berat (kg)	$r_{i,3}$
1.6	88	13.9
1.6	76	1.9
1.5	56	-15.7

Langkah ke 3: Keluaran $F_M(x)$

Jika ***M* = 2** maka $F_2(x)$ = $F_0(x)$ + 0.1 x $F_1(x)$ + 0.1 x $F_2(x)$

Contoh 1:

Jika kita ingin memprediksi sebuah data baru sebagai berikut:

Tabel 3.5 Predisksi data baru nilai tinggi <1.5

Tinggi (m)	Berat (kg)
1.4	…?

Kalau M= 2 maka $F_2(x)$ = $F_0(x)$ + 0.1 x $F_1(x)$ + 0.1 x $F_2(x)$
Karena nilai Tinggi yang diprediksi 1.4 atau <1.5 maka cabang yang digunakan adalah sebelah kiri sehingga:
$F_2(x)$ =73.3 + (0.1 x -17.3) + (0.1 x -15.6) = 70
Jadi prediksi dari berat tersebut adalah 70.

Contoh 2:
Jika kita ingin memprediksi sebuah data baru sebagai berikut:

Tabel 3.6 Predisksi data baru nilai tinggi >1.5

Tinggi (m)	Berat (kg)
1.7	…?

Kalau M= 2 maka $F_2(x)$ = $F_0(x)$ + 0.1 x $F_1(x)$ + 0.1 x $F_2(x)$
Karena nilai Tinggi yang diprediksi 1.7 atau >1.5 maka cabang yang digunakan adalah sebelah kanan sehingga:
$F_2(x)$ =73.3 + (0.1 x 8.7) + (0.1 x 7.8) = 75
Jadi prediksi dari berat tersebut adalah 75.

3.4.2. Hyperparameter

Banyak *hyperparameter* tersedia di GBT yang dapat disesuaikan untuk mempengaruhi ansambel umum (seperti *learning rate*) dan pohon keputusan individu (seperti *number of leaves* atau *maximum depth*). Karena ada interaksi yang rumit antara *hyperparameter*, maka sulit untuk menentukan kombinasi hyperparameter mana yang akan berfungsi paling baik hanya berdasarkan teori. Dengan demikian, penyetelan *hyperparameter* diperlukan karena tidak ada metode lain untuk menentukan nilai hyperparameter terbaik selain menguji sejumlah besar kemungkinan kombinasi. Penyetelan *hyperparameters* sangat

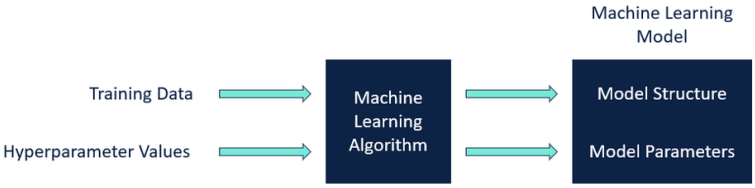
penting sebelum menerapkan algoritma prediksi (Probst, Boulesteix, dan Bischl 2019; Victoria dan Maragatham 2021; Weerts, Mueller, dan Vanschoren 2020; Zhang, Wang, dan Shen 2022). Kinerja model tergantung pada pemilihan *hiperparameter* (Shekhar, Bansode, dan Salim 2021). Pilihan konfigurasi *hyperparameter* diketahui berdampak signifikan terhadap kinerja model mesin pembelajaran (Zhang dkk. 2022) Untuk menentukan kinerja terbaik dari algoritma mesin pembelajaran, para peneliti telah melakukan penyetelan *hiperparameter* (Duarte dan Wainer 2017; Elgeldawi dkk. 2021; Lorraine, Vicol, dan Duvenaud 2019; MacKay dkk. 2019; Probst, Wright, dan Boulesteix 2019; Rijsdijk dkk. 2021; Schratz dkk. 2019; Victoria dan Maragatham 2021; Wong dkk. 2019; Zhang dkk. 2022).

Untuk membuat model pembelajaran mesin terlatih, ada tiga kategori besar; data pelatihan, parameter, dan hyperparameter. Model memperoleh nilai parameter selama pelatihan dengan belajar dari data yang diberikan, nilai ini tidak dapat ditetapkan secara manual. Data pelatihan (input) adalah kumpulan data yang dipergunakan untuk melatih model. Data pelatihan adalah apa yang dimanfaatkan algoritma (instruksi untuk membangun model) untuk mengidentifikasi pola dalam data dan mengeksploitasi pola untuk membuat prediksi. Data pelatihan (biasanya) tidak menjadi bagian dari model secara langsung; mereka hanya digunakan untuk mengajarkan model seperti apa data itu, dan bagaimana menangani data untuk membuat perkiraan yang berarti dari variabel target.

Parameter dipelajari dari data pelatihan, mereka adalah bobot, ambang batas, atau koefisien yang memungkinkan model membuat prediksi. Ketika algoritma "mempelajari" data pelatihan yang disediakan untuk membuat model, itu benar-benar menyesuaikan parameter model akhir untuk membuat prediksi terbaik dari variabel target berdasarkan data pelatihan. Parameter

pada akhirnya disimpan sebagai bagian dari model yang dipelajari. Parameter adalah bagian dari model yang secara otomatis disesuaikan agar sesuai dengan data spesifik dan kasus penggunaan Anda.

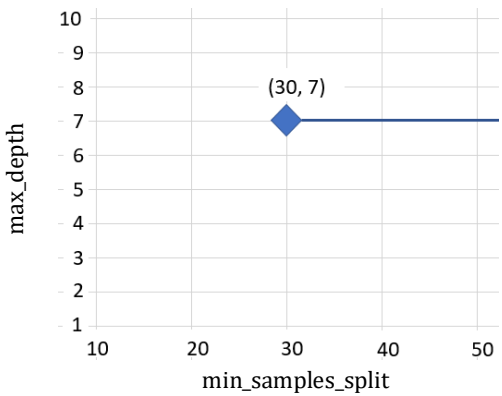
Komponen ketiga dari model terlatih adalah hyperparameter, dimana variabel yang mengatur proses pelatihan model. Hyperparameter biasanya menentukan kapan model selesai dilatih, atau berapa banyak rekaman yang dipertimbangkan algoritma pada satu waktu selama pelatihan. Tidak seperti parameter, hyperparameter bersifat konstan selama proses pelatihan, ditetapkan sebelum pelatihan model dan tidak disesuaikan selama proses pelatihan. Hyperparameter berdampak langsung pada nilai akhir parameter model, dan beberapa hyperparameter (misalnya, jumlah lapisan tersembunyi di jaringan saraf) juga berdampak langsung pada struktur model pembelajaran mesin yang dilatih. Karena hyperparameter menentukan struktur sebenarnya dari algoritma pembelajaran mesin dan proses pelatihan model, tidak ada cara untuk "mempelajari" nilai-nilai ini menggunakan fungsi kerugian dan data pelatihan. Kita dapat menganggap hyperparameter sebagai kenop dan tuas yang diputar dan tarik untuk membuat algoritma mengembalikan sinyal sejelas mungkin dalam model terlatih.



Gambar 3.3 Parameter mesin pembelajaran

Parameter model meliputi koefisien model dalam model regresi linier (Ma dkk. 2018). Di sisi lain, model *Hyperparameter*

tidak memperoleh nilainya dari data. Oleh karena itu, peneliti harus mengaturnya secara manual. Pada pembangunan model tertentu, peneliti selalu menetapkan nilai pada *hyperparameter* (misalnya, sebelum proses pelatihan) (Ma dkk. 2018). Parameter model dapat dikendalikan oleh *hiperparameter* model. Dengan kata lain, performa model dapat dipengaruhi oleh *hiperparameter* model. Oleh karena itu, kita harus memodifikasi nilai *hyperparameter* model untuk menghasilkan output model yang optimal atau terbaik. Bahkan model sederhana sering memiliki dua atau lebih *hiperparameter*. Kita harus melihat semua nilai ini untuk mendapatkan nilai ideal untuk setiap hiperparameter. Optimasi hiperparameter adalah proses menentukan set model hyperparameter (tuning) yang sempurna. Karena sejumlah besar *hyperparameter* dan berbagai nilai untuk masing-masing, kami tidak dapat melakukan ini secara manual. Ruang pencarian hyperparameter sangat penting untuk proses penyetelan *hyperparameter*. Semua kemungkinan kombinasi nilai *hyperparameter* yang ditentukan pengguna dapat ditemukan di ruang pencarian. Ruang pencarian 2 dimensi untuk dua *hyperparameter* terpisah *max_depth* dan *min_samples_split* ditunjukkan pada gambar 3.4.



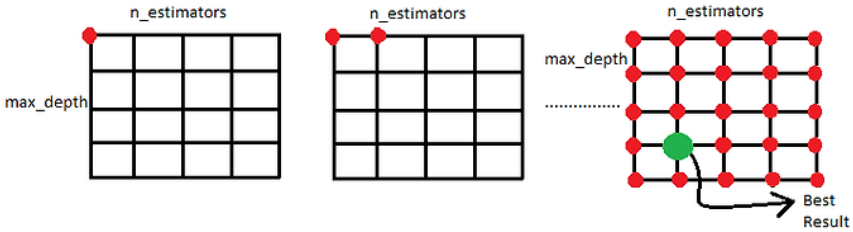
Gambar 3.4 Ruang hiperparameter dua dimensi

Ruang pencarian adalah tiga dimensi jika ada tiga hyperparameter yang berbeda. Demikian juga, ketika jumlah hiperparameter meningkat, ruang pencarian dapat menjadi dimensi tinggi. Secara umum, kita harus menyadari bahwa jumlah hyperparameter tergantung pada jumlah dimensi dalam ruang pencarian (misalnya, dua dimensi dan dua hyperparameter). Titik tertentu di ruang pencarian mendefinisikan setiap kombinasi nilai hyperparameter. Poin (30, 7) tentukan nilai max_depth 7 dan nilai min_samples_split 30.

Tuning hyperparameters sangat penting sebelum menerapkan algoritma prediksi (Probst, Boulesteix, dkk. 2019; Victoria dan Maragatham 2021; Weerts dkk. 2020; Zhang dkk. 2022). Kinerja model tergantung pada pemilihan model hiperparameter (Shekhar dkk. 2021). Pilihan konfigurasi hyperparameter diketahui berdampak signifikan terhadap kinerja model pembelajaran mesin (Zhang dkk. 2022) Untuk menentukan kinerja terbaik dari algoritma pembelajaran mesin, para peneliti telah melakukan penyetelan hiperparameter (Duarte dan Wainer 2017; Elgeldawi dkk. 2021; Lorraine dkk. 2019; MacKay dkk. 2019; Probst, Wright, dkk. 2019; Rijsdijk dkk. 2021; Schratz dkk. 2019; Victoria dan Maragatham 2021; Wong dkk. 2019; Zhang dkk. 2022).

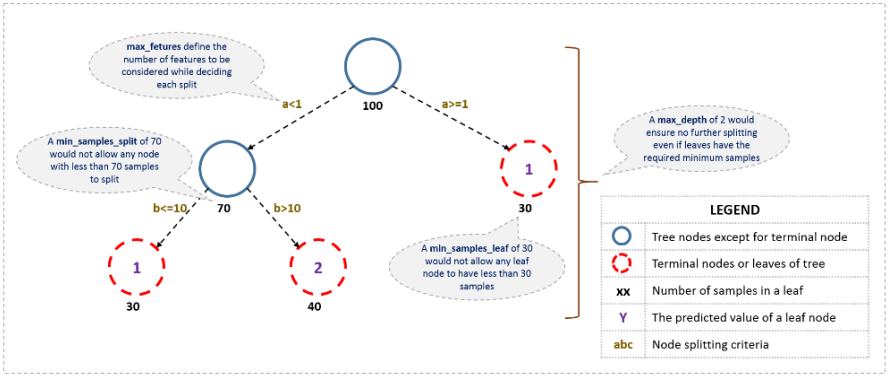
Para peneliti telah mencari hyperparameter terbaik secara manual dan otomatis. Pencarian manual adalah salah satu metode untuk optimasi hyperparameter. Namun, ini membutuhkan waktu yang cukup lama (Putatunda dan Rama 2019). Banyak metode untuk optimasi hyperparameter tetapi penulis memilih GridSearchCV (GS). GS adalah metode sederhana dan mudah digunakan (Bergstra dan Bengio 2012). Metode ini paling sering digunakan oleh peneliti untuk melakukan optimasi dan terbukti meningkatkan akurasi (Andonie dan Florea 2020; Duarte dan Wainer 2017; Florea dan Andonie 2019; Joy dan Selvan 2022;

Liashchynskyi dan Liashchynskyi 2019; MacKay dkk. 2019; Putatunda dan Rama 2018; Shekar dan Dagnew 2019; Villalobos-Arias dkk. 2020; Villalobos-Arias dan Quesada-López 2021; Wong dkk. 2019). GS adalah teknik untuk mencari ruang pencarian untuk semua kemungkinan kombinasi hyperparameter yang diberikan oleh pengguna. Ketika ruang pencarian berdimensi tinggi dan memiliki banyak kombinasi nilai, umumnya akan membutuhkan banyak sumber daya pemrosesan dan waktu eksekusi yang lama. Pendekatan ini bekerja paling baik ketika ada beberapa hyperparameter dan jumlah nilai terbatas (tetap) untuk masing-masing. GS adalah metode yang secara tradisional telah digunakan dengan melihat melalui semua kombinasi parameter potensial. Seluruh ruang parameter dipertimbangkan dan dipartisi menjadi kotak saat melakukan pencarian kisi. Titik-titik grid kemudian masing-masing dinilai sebagai hyperparameter. GS mencoba setiap kombinasi dari daftar nilai-nilai tersebut dari hyperparameter dan mengevaluasi model untuk setiap kombinasi. Pola yang diikuti mirip dengan kisi-kisi, di mana semua nilai ditempatkan dalam bentuk matriks. Setiap set parameter dipertimbangkan dan akurasi dicatat. Setelah semua kombinasi dievaluasi, model dengan kumpulan parameter yang memberikan akurasi tertinggi dianggap yang terbaik, representasi visual dari GS ditunjukkan pada Gambar 3.5.

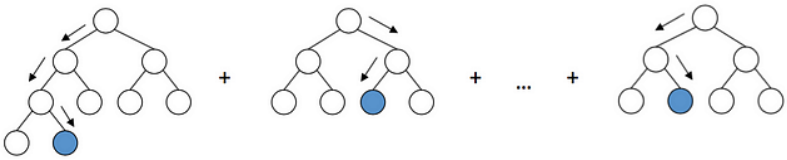


Gambar 3.5 Representasi visual GS

(Sumber: https://www.numpyninja.com/post/hyper-parameter-tuning-using-grid-search-and-random-search)



Gambar 3.6 Parameter GBT
(Sumber: https://www.kaggle.com/hyperparameter/gbt)



Gambar 3.7 Visualisasi pencarian parameter

3.5. Evaluasi model prediksi (MAE, MSE dan RMSE)

Langkah penting dalam setiap model pembelajaran mesin adalah mengevaluasi keakuratan model. *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) dan *Root Mean Squared Error* (RMSE) digunakan untuk mengevaluasi kinerja model dalam analisis regresi.

MAE mewakili rata-rata perbedaan absolut antara nilai aktual dan prediksi dalam kumpulan data. Ini mengukur rata-rata residu dalam kumpulan data.

$$MAE = \sum_{i=1}^{n} \frac{|y-\hat{y}|}{n}$$

Dimana:

$\sum$ = Jumlah keseluruhan nilai

y = Nilai aktual

$\hat{y}$ = Nilai hasil prediksi

i = Urutan data pada database

n = Banyaknya data

Tabel 3.7 Contoh soal untuk menghitung MAE, MSE dan RMSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	Kesalahan	$\lvert y - \hat{y} \rvert$
1.6	88	74.2	13.8	13.8
1.6	76	74.2	1.8	1.8
1.5	56	71.6	-15.6	15.6
			Jumlah	31.2

$$MAE = \frac{31.2}{3} = 10.4$$

Sementara itu MSE mewakili rata-rata perbedaan kuadrat antara nilai asli dan prediksi dalam kumpulan data. Ini mengukur varian dari residual.

$$MSE = \sum_{i=1}^{N} \frac{(y-\hat{y})^2}{n}$$

Tabel 3.8 Contoh perhitungan MSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	$(y - \hat{y})$	$(y - \hat{y})^2$
1.6	88	74.2	13.8	190.44
1.6	76	74.2	1.8	3.24
1.5	56	71.6	-15.6	243.36
			Jumlah	437.04

$$MSE = \frac{437.04}{3} = 145.68$$

Adapun RMSE adalah uji untuk mengukur kinerja model dengan cara mengukur besarnya kesalahan pendugaan antara nilai observasi dan nilai prediksi. Semakin rendah nilai RMSE maka model semakin akurat (Chai dan Draxler 2014). RMSE juga disebut sebagai akurasi dari sebuah sistem rekomendasi. RMSE memberikan hasil yang memiliki penalti yang lebih besar untuk perbedaan yang lebih besar antara hasil aktual dengan prediksi (Gunawan, Tania, dan Suhartono 2016) yang dapat disesuaikan dengan persamaan sebagai berikut:

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{N} \frac{(y-\hat{y})^2}{n}}$$

Tabel 3.9 Contoh Perhitungan RMSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	$(y - \hat{y})$	$(y - \hat{y})^2$
1.6	88	74.2	13.8	190.44
1.6	76	74.2	1.8	3.24
1.5	56	71.6	-15.6	243.36
			Jumlah	437.04

$$MSE = \frac{437.04}{3} = 145.68$$

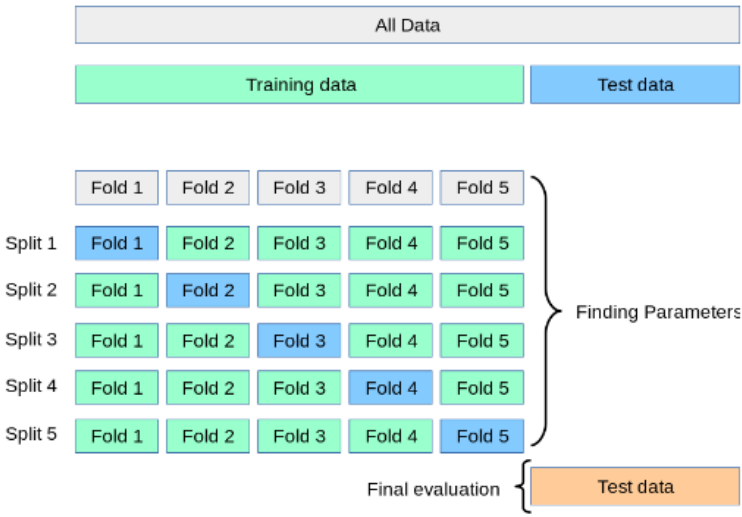
$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{145.68} = 12.07$$

MSE dan RMSE menghukum kesalahan prediksi yang besar dibandingkan dengan MAE. Namun, RMSE lebih banyak digunakan daripada MSE untuk mengevaluasi kinerja model regresi dengan model acak lainnya karena memiliki unit yang sama dengan variabel dependen (sumbu Y). MSE adalah fungsi yang dapat dibedakan yang memudahkan untuk melakukan operasi matematika dibandingkan dengan fungsi yang tidak dapat dibedakan seperti MAE. Oleh karena itu, dalam banyak model, RMSE digunakan sebagai metrik default untuk menghitung Loss Function meskipun lebih sulit untuk diinterpretasikan daripada MAE. Nilai MAE, MSE, dan RMSE yang lebih rendah menyiratkan akurasi model regresi yang lebih tinggi. RMSE menunjukkan seberapa baik suatu model regresi dapat memprediksi nilai suatu variabel respon secara absolut. Berikut ini contoh evaluasi keakuratan prediksi.

3.6. Validasi model

C*ross validation* adalah metode pengamatan dalam kumpulan data yang akan digunakan untuk pengujian dan pelatihan, tetapi tidak pada waktu yang sama. Metode untuk pengujian hasil prediksi menggunakan *x-validation (k-fold validation)* dalam 5 langkah (5 *folds cross-validation*). Dengan *x-validation*, pengukuran hasil dapat lebih akurat karena data yang ada dibagi kedalam 5 data dengan jumlah yang sama, kemudian satu persatu diambil untuk test, dan 4 bagian lainnya digunakan untuk training (Witten, Frank, dan Hall 2011). Dengan *x-validation* akurasi dari hasil pengukuran data akan lebih terjamin karena mengurangi kemungkinan data yang tidak konsisten dalam tahap prediksi.



Gambar 3.8 Visualisasi pembagian data untuk validasi silang (Sumber: https://scikit-learn.org/stable/modules/cross_validation.html)

BAB IV
STUDI KASUS DAN IMPLEMENTASI PREDIKSI PERFORMA AKADEMIK MAHASISWA

4.1. Obyek, Tempat, Waktu dan Data

Langkah awal dalam penerapan prediksi performa akademik mahasiswa maka kita harus menentukan obyek, tempat, waktu dan data yang akan digunakan. Obyek penelitian adalah mahasiswa sementara itu tempat penelitian dilakukan di laboratorium dan di sebuah Universitas di Jawa Tengah, sedangkan data penelitian yang digunakan data akademik dan data non-akademik. Penelitian ini dilaksanakan pada bulan Februari 2020 hingga bulan Januari 2023. Adapun alat penelitian yang digunakan pada penelitian ini terbagi kedalam dua kategori yaitu perangkat keras dan perangkat lunak. Perangkat keras (*Hardware*) adalah seluruh komponen peralatan yang membentuk suatu sistem dan peralatan lainnya yang memungkinkan komputer dapat melaksanakan tugasnya secara fisik dan dapat terlihat secara jelas dan nyata, bagian-bagian pokok perangkat keras meliputi masukan (*Input*), pemprosesan yang disebut dengan CPU (*Central Processing Unit*), tempat penyimpanan (*Secondary Memory*), dan keluaran (*Output*). Sementara itu Perangkat lunak (*software*) adalah komponen data processing yang berupa program-program dan teknik-teknik lainnya untuk mengontrol sistem komputer. Pada penelitian ini kebutuhan peralatan perangkat keras yang diperlukan adalah:

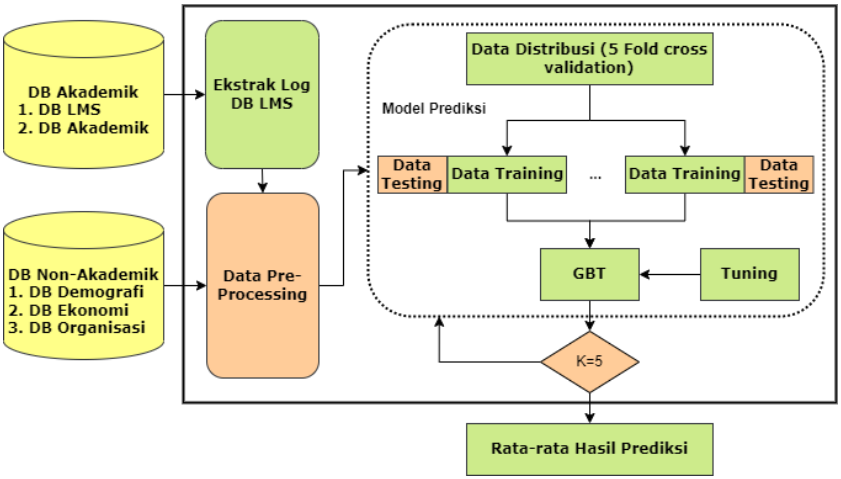
a. Processor: Intel Core i5 (1.5Ghz. 1MB L2 cache)
b. Memory Size (RAM): 8 GB
c. Harddisk: 320 GB
d. Keyboard / Mouse: 102 Keys / Standart Mouse
e. Monitor: Acer Aspire 17 inc

Sedangkan peralatan perangkat lunak yang diperlukan adalah sebagai berikut:

b. Sistem operasi: Microsoft windows 10
c. Pengolahan data: Microsoft excel, Jupyter Notebook dan RapidMiner

Langkah selanjutnya adalah menentukan infrastruktur sistem mengacu pada kerangka penelitian yang akan digunakan pada studi kasus ini, seperti ditunjukkan pada gambar 4.1.

Gambar 4.1. secara singkat menjelaskan kegiatan yang dilakukan selama penelitian untuk memprediksi performa akademik mahasiswa. Adapun detail tahapan dan hasil per tahapan penelitian dijelaskan pada tabel 4.1.

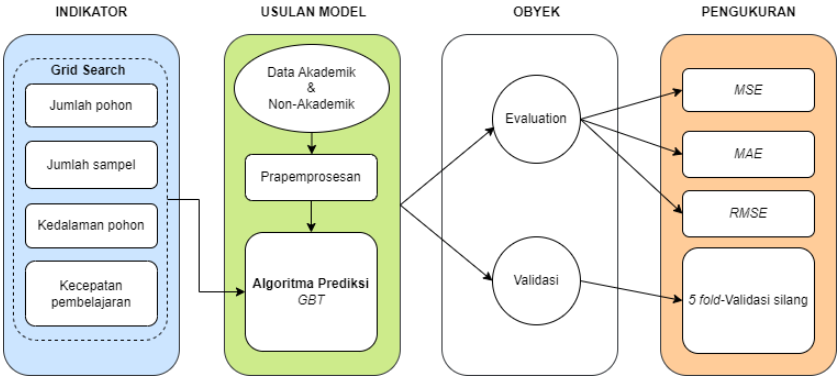


Gambar 4.1 Kerangka penelitian

Tabel 4.1 Tahapan Penelitian

Tahapan	**Aktivitas**	**Alat/Metode**	**Hasil**
Pengumpulan data	Mengumpulkan data log Moodle, demografi, akademik, ekonomi dan organisasi	Menggunakan API / query dalam proses pengambilan data	Data perilaku interaksi dengan LMS, demografi, akademik, ekonomi dan data organisasi
Ekstrak log	Melakukan ekstraksi data log interaksi siswa dengan Moodle	Menggunakan metode NLP atau sejenisnya	Data interaksi siswa dengan *Moodle*
Data preprocessing	Proses pembersihan data	Missing value	Data hasil preprocessing
Model Prediksi	Data terdistribusi	Data didistribusi ke dalam 10 bagian, selanjutnya dilakukan pengujian selama sepuluh kali	Rata-rata pengujian selama 10 kali
Tuning	Melakukan *tuning* pada GBT dengan learning rate, jumlah sampel, kedalaman pohon dan jumlah pohon	*Tuning Hyperparameter*	Nilai RMSE terbaik

Usulan model kerangka pada yang digunakan adalah penerapan metode GBT dengan modifikasi *learning rate*, kedalaman pohon, jumlah sampel dan jumlah pohon untuk memprediksi performa akademik mahasiswa berdasarkan data demografi, data LMS, data akademik, data ekonomi dan data organisasi. Ditunjukkan pada gambar 4.2.



Gambar 4.2 Model Kerangka Prediksi Performa Akademik Mahasiswa

Penjelasan usulan model kerangka prediksi performa akademik mahasiswa sebagai berikut:

1. Penggunaan data (variabel penelitian)
   Pada penelitian ini, data set yang digunakan adalah data set dari data demografi, data LMS, ekonomi dan data organisasi.
2. Usulan model
   Model yang diusulkan adalah GBT yang diobservasi nilai learning rates, kedalaman pohon, jumlah sampel dan jumlah pohonnya.
3. Tujuan model
   Tujuan (obyek) pada model ini adalah adanya pengurangan kesalahan prediksi algoritma GBT dalam membuat model prediksi performa akademik mahasiswa.
4. Validasi
   Proses validasi menggunakan *x-validation*, dimana data dibagi kedalam 5 data dengan jumlah yang sama, kemudian satu persatu diambil untuk test, dan 4 bagian lainnya digunakan untuk training, hal ini dilakukan agar akurasi dari hasil pengukuran data akan lebih terjamin karena kemungkinan data yang tidak konsisten dalam tahap prediksi.

5. Evaluasi

   Evaluasi model ini adalah dengan menggunakan MAE, MSE dan RMSE, dimana RMSE merupakan sebuah metode pengukuran akurasi peramalan, yang digunakan untuk mengukur perbedaan antara nilai-nilai peramalan oleh model atau sebuah estimator. Selain RMSE pada tahap ini dilakukan evaluasi kesalahan model yang nantinya akan dibandingkan antara penerapan model GBT tanpa optimasi dengan model GBT teroptimasi.

4.2. Deskripsi data

Proses pemahaman data diawali dari proses pengumpulan data, pembersihan, penggabungan dan dilanjutkan dengan visualisasi data. Sebelum proses pengumpulan data maka perlu mengetahui data-data yang akan dipergunakan diantaranya adalah:

1. Data Akademik

   Data ini merupakan data yang bersumber dari kegiatan-kegiatan akademik yaitu data LMS dan data hasil studi di akhir semester (IPS).

2. Data Non-akademik

   Data yang merupakan data yang dihasilkan dari kegiatan non akademik dan data profil mahasiswa. Data ini meliputi; pendaftaran mahasiswa baru dan data keikutsertaan mahasiswa dalam berorganisasi kampus.

Adapun variabel yang akan dipergunakan untuk prediksi performa akademik mahasiswa adalah; LMS, Organisasi, Akademik, Ekonomi, Jenis kelamin, Kota tinggal.

a. Variabel LMS

Pertumbuhan penggunaan LMS di beberapa tahun terakhir mengalami peningkatan yang cukup signifikan (Li, dkk., 2018). LMS mewakili alat yang berharga untuk mengubah kursus tatap

muka tradisional menjadi program campuran dan *online* (Conijn dkk. 2017). LMS pertama kali dimunculkan oleh Universitas Illinois dengan nama *Programmed Logic for Automated Teaching Operation* (PLATO). Saat itu PLATO merupakan LMS pertama yang dikategorikan sebagai *Computer Assisted Instruction* (CAI) yang mampu menyediakan fitur-fitur konsep utama dari LMS seperti fitur komunikasi online (*online* forum, papan pesan, email, *chat rooms, picture languages*, pesan singkat), ujian *online*, dan permainan *online* dengan banyak pemain untuk menunjang pembelajaran (Kumar, Gankotiya, dan Dutta 2011). (Chu dkk. 2010) mendefinisikan bahwa LMS merupakan ruangan kelas yang terbuka 24 jam dan tujuh hari dalam seminggu. Definisi secara luas mengenai LMS yaitu sebuah aplikasi yang melakukan otomatisasi terhadap administrasi, penelusuran, dan pelaporan dari berbagai peristiwa pelatihan dalam pembelajaran. Bentuk LMS yang paling sederhana adalah berupa media penyimpanan untuk materi pembelajaran *(learning object)* yang dapat diakses oleh peserta didik ketika masuk kedalam sebuah portal pembelajaran yaitu *e-learning.* Sementara itu LMS yang canggih menyediakan berbagai fitur seperti pilihan tes dan kuis, materi audio atau video, forum interAktif, kemampuan tinggi dalam pencarian file, dan pelaporan nilai. LMS dapat memberikan akses bagi pendidik untuk menentukan akses-akses lain dalam pembelajaran bagi peserta didiknya atau bagi pendidik lain. LMS memberikan fitur yang dapat memfasilitasi peserta didik belajar dalam group secara berkompetisi maupun berkolaborasi. Pemberian umpan balik dalam pembelajaran juga salah satu fitur LMS yang dapat menunjang pembelajaran *e-learning* berlangsung tahan lama.

Kaewkiriya, Utakrit, dan Tsuji (2011) mendefinisikan bahwa LMS merupakan sebuah sistem perangkat lunak yang bertanggung jawab dalam mengelola pembelajaran dan tersedia

dalam berbagai bentuk seperti portal website yang mengoleksi semua *link* dan direktori video, sebagian bentuk yang lain ditambahkan dengan fitur tes untuk memandu peserta didik tahapan pembelajaran selanjutnya, dan sebagian LMS yang lain mampu menyimpan rekaman pembelajaran untuk tujuan tertentu. Albarrak, Aboalsamh, dan Abouzahra (2010) menjelaskan beberapa fitur yang harus dipertimbangkan pada LMS diantaranya adalah fleksibilitas, dimana LMS mampu beradaptasi dengan kapabilitas, kebutuhan, dan tujuan dari tiap peserta didik; kemudahan penggunaan, dimana LMS mampu memfokuskan peserta didik pada materi yang dipelajari dan bukan pada bagaimana cara menggunakan LMS; interaktivitas, dimana LMS mampu menyediakan berbagai fasilitas komunikasi yang baik sehingga meningkatkan interaksi baik sesama peserta didik maupun antara pendidik dengan peserta didiknya.

LMS menghasilkan file log yang berisi data tentang interaksi pengguna (misalnya, tampilan kursus dan sumber daya, pengiriman dan evaluasi tugas, serta interaksi kuis dan forum). Informasi tersebut telah digunakan untuk membuat model prediktif untuk tujuan yang berbeda seperti memprediksi performa siswa (Conijn dkk. 2017).

b. Data Organisasi

Organisasi kemahasiswaan merupakan organisasi mahasiswa yang berada di lingkungan perguruan tinggi. Sesuai dengan peraturan pemerintah republik Indonesia nomor 60 tahun 1999 pada pasal 111 telah dijelaskan bahwa “untuk melaksanakan peningkatan kepemimpinan, penalaran, minat, kegemaran dan kesejahteraan mahasiswa dalam kehidupan kemahasiswaan pada perguruan tinggi dibentuk organisasi kemahasiswaan”. Secara umum pedoman organisasi kemahasiswaan di perguruan tinggi telah diatur oleh pemerintah melalui Keputusan Menteri

Pendidikan dan Kebudayaan Republik Indonesia Nomor 155/U/1998, menurut pasal 1 ayat 1 disampaikan bahwa "Organisasi kemahasiswaan intra perguruan tinggi adalah wahana dan sarana pengembangan diri mahasiswa ke arah perluasan wawasan dan peningkatan kecendekiawanan serta integritas kepribadian untuk mencapai tujuan pendidikan tinggi", sedangkan pada pasal 5 disampaikan bahwa fungsi dari organisasi kemahasiswaan diantaranya adalah 1. perwakilan mahasiswa tingkat perguruan tinggi untuk menampung dan menyalurkan aspirasi mahasiswa, menetapkan garis-garis besar program dan kegiatan kemahasiswaan; 2. pelaksanaan kegiatan kemahasiswaan; 3. komunikasi antar mahasiswa; 4. pengembangan potensi jatidiri mahasiswa sebagai insan akademis, calon ilmuwan dan intelektual yang berguna di masa depan; 5. pengembangan pelatihan keterampilan organisasi, manajemen dan kepemimpinan mahasiswa; 6. pembinaan dan pengembangan kader-kader bangsa yang berpotensi dalam melanjutkan kesinambungan pembangunan nasional; 7. untuk memelihara dan mengembangkan ilmu dan teknologi yang dilandasi oleh norma-norma agama, akademis, etika, moral, dan wawasan kebangsaan. Organisasi mahasiswa yang dibentuk berdasarkan kesamaan minat, baik di bidang olahraga, seni atau lainnya sering disebut dengan Unit Kegiatan Mahasiswa (UKM) sedangkan Badan Eksekutif Mahasiswa (BEM) merupakan perwakilan mahasiswa dari semua jurusan maupun fakultas. Kedudukan UKM, karena beranggotakan berbagai disiplin ilmu, maka kedudukannya ada di Universitas/Rektorat begitu pula BEM.

Menurut berbagai sumber, organisasi kemahasiswaan dianggap memiliki manfaat bagi mahasiswa diantaranya yaitu; (1) Membangun relasi, umumnya, di dalam organisasi kampus merupakan kumpulan dari berbagai mahasiswa atau mahasiswa

dari berbagai jurusan, mereka akan bertemu dengan orang-orang yang berbeda jurusan bahkan fakultas dan pasti juga bertemu dengan orang-orang yang berbeda daerah maupun provinsi. Interaksi dengan teman-teman baru tersebut akan mendapatkan pengalaman yang lebih luas dan menambah relasi. (2) Menambah pengalaman baru, dengan banyaknya kegiatan yang dilakukan di organisasi, sesama anggota akan mendapatkan pengalaman yang baru yang berasal dari kegiatan sosial atau acara di organisasi. (3) Belajar berkepemimpinan, di dalam berorganisasi masing-masing anggota dapat membangun jiwa-jiwa berkepemimpinan karena mereka akan selalu dituntut agar setiap kegiatan yang dilaksanakan dapat terlaksana dengan sukses. (4) Memanfaatkan waktu dengan baik, setiap anggota dituntut untuk dapat memanfaatkan waktunya dengan baik hal ini dikarenakan kegiatan organisasi bisa jadi berbarengan dengan kegiatan perkuliahan dan lain sebagainya. (5) Membentuk pola pikir yang baik, dalam sebuah organisasi pasti kita akan mendapatkan banyak pelajaran yang secara tidak langsung meningkatkan kemampuan berorganisasi kita, keterampilan sosial, menghargai sesuatu, dan yang terpenting dapat membiasakan diri menggunakan pola pikir yang baik karena apa yang dilakukan berhubungan dengan orang lain.

Fitri Oviyanti (2016) menyampaikan bahwa organisasi-organisasi kemahasiswaan merupakan sarana bagi pengembangan diri mahasiswa yang memiliki banyak manfaat, diantaranya memperluas wawasan mahasiswa, meningkatkan kecendekiawanan, serta meningkatkan integritas pribadi mahasiswa dalam menyikapi permasalahan kampus, masyarakat dan bangsa. Manfaat lain yang dapat diperoleh mahasiswa dengan aktif pada organisasi kemahasiswaan adalah berkembangnya kemampuan bekerja sama, keterampilan berorganisasi serta kemampuan berkomunikasi dan keterampilan

khusus yang sesuai dengan bidang atau jenis organisasi kemahasiswaan yang diikuti.

Beberapa peneliti telah melakukan penelitian terkait pengaruh organisasi kampus terhadap mahasiswa diantaranya; Arola Anderson dkk. (2021) menyampaikan bahwa aktivitas organisasi kampus pada semester pertama memiliki pengaruh positif terhadap optimis mahasiswa, membantu mengatasi perubahan (Bohnert dkk. 2010), berpengaruh positif terhadap perkembangan dan perilaku anak (Morris 2015), memiliki pengaruh terhadap masa tunggu lulusan dalam mendapatkan pekerjaan (Arifin dkk. 2022).

c. Data Akademik

Data akademik merupakan data yang bersumber dari informasi akademik mahasiswa diantaranya adalah data indek prestasi, nomor induk mahasiswa, nilai dari berbagai mata kuliah, dan lain sebagainya. Data ini dapat dipergunakan untuk menganalisis tren, pola, dan hubungan antara faktor-faktor akademik dan performa mahasiswa. Analisis data akademik dapat membantu dalam pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi performa akademik mahasiswa dan mendukung pengambilan keputusan pendidikan yang informasional dan terarah.

d. Data Demografi

Data ini mencakup informasi demografis siswa, seperti usia, jenis kelamin, etnis, latar belakang sosioekonomi, dan informasi pribadi lainnya. Data ini dapat memberikan konteks tentang mahasiswa dan faktor-faktor demografis yang dapat mempengaruhi performa akademik.

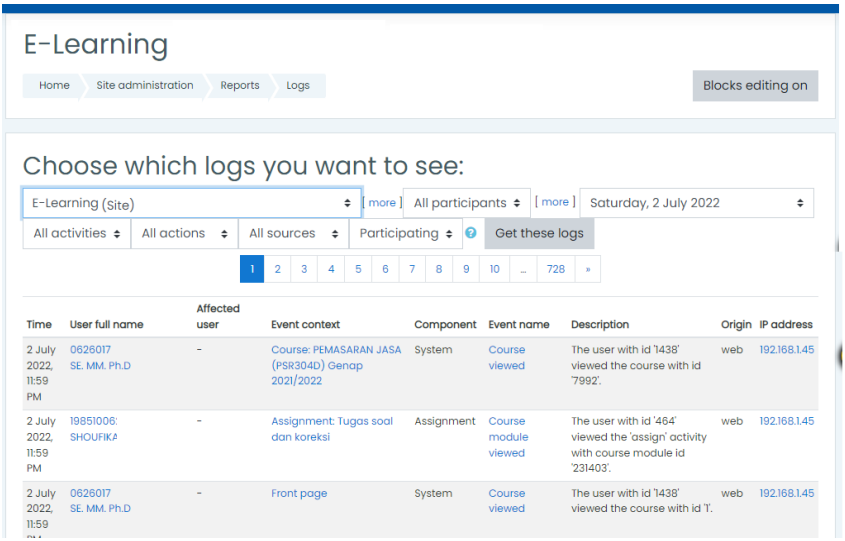
4.3. Pengumpulan Data

Skema pengumpulan data terdiri dari beberapa skema diantaranya adalah statis dan terjadwal. Untuk data LMS dilakukan secara terjadwal yaitu dikumpulkan dari LMS setelah proses perkuliahan selesai sesuai dengan kalender akademik. Adapun data ekonomi dan demografi menggunakan metode statis yaitu dikumpulkan dari sistem informasi pendaftaran mahasiswa baru (SISFO-PMB). Sementara itu data organisasi juga menggunakan metode statis dimana data dikumpulkan dari surat keputusan (SK) organisasi mahasiswa yang ada di universitas tempat penelitian. Sementara itu data akademik dikumpulkan menggunakan metode terjadwal yaitu dikumpulkan dari sistem informasi akademik (SIAKAD) pada akhir semester setelah masa entry nilai berakhir.

4.3.1. Data LMS Moodle

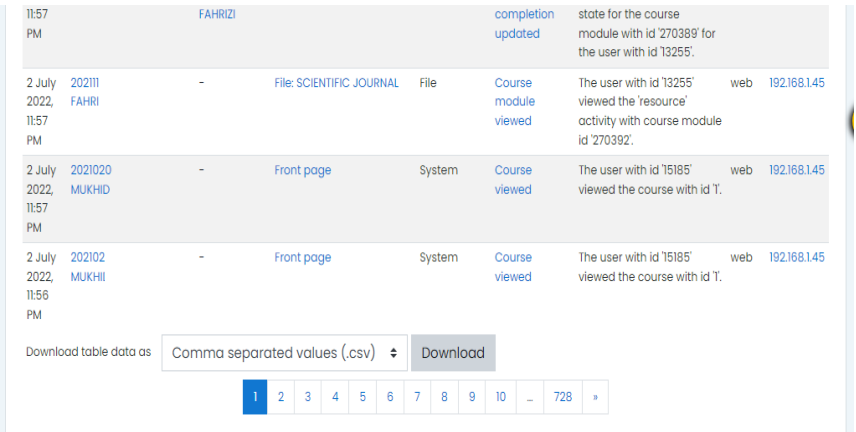
Data LMS diperoleh dari aplikasi LMS Moodle. Data ini terdiri dari 20.231.298 catatan dari 8.500 mahasiswa aktif yang mengikuti perkuliahan selama 19 minggu sesuai dengan kalender pendidikan yang berlaku di lokasi penelitian, dimana masa perkuliahan dilakukan selama enam bulan yang dimulai dari bulan Februari tanggal 21 tahun 2022 s/d bulan Juli tanggal 2 tahun 2022.

Proses pengumpulan data LMS dilakukan dengan metode mengunduh catatan kegiatan di LMS Moodle ditunjukkan pada Gambar 4.3. Proses ini dilakukan di akhir semester dan dikumpulkan dari catatan kegiatan siswa per harinya. Proses pengumpulan data ini dilakukan dengan cara login kedalam aplikasi LMS Moodle, selanjutnya pada menu Logs perlu ditentukan tanggal, peserta, aktivitas, aksi dan sumber. Setelah semuanya ditentukan maka perlu di klik menu yang bertuliskan “Get these logs”.

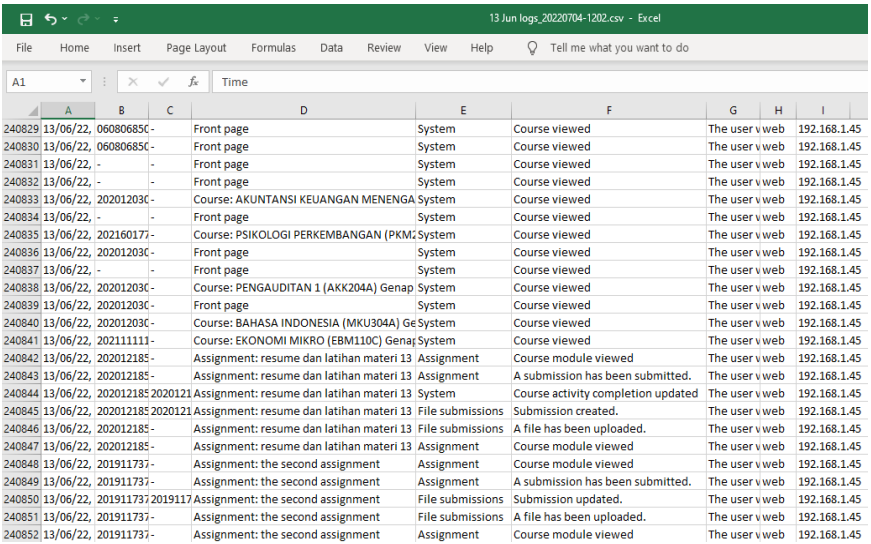


Gambar 4.3 Halaman LMS Moodle

Setelah menu tersebut diklik maka akan ditunjukkan semua catatan aktivitas pada tanggal tersebut, termasuk jumlah halaman yang memuat data catatan. Untuk mengunduh catatan maka tombol unduh berapa pada bagian paling bawah di halaman tersebut ditunjukkan pada Gambar 4.4.



Gambar 4.4 Halaman bawah LMS Moodle



13 Jun logs_20220704-1202.csv - Excel

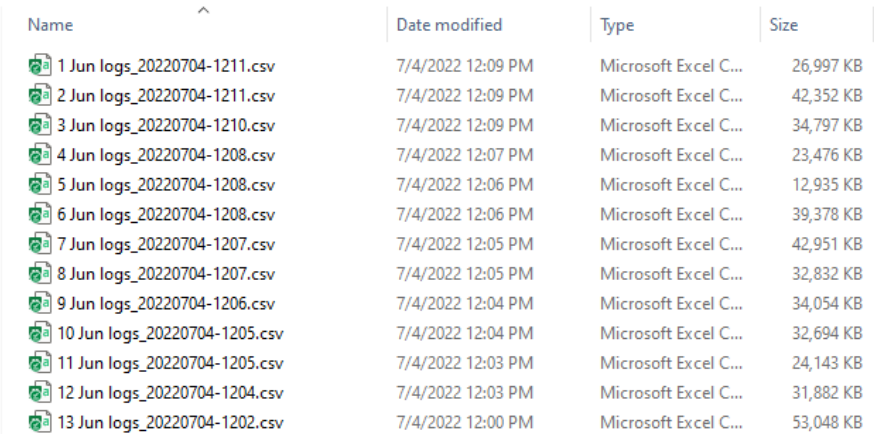
File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

A1 Time

	A	B	C	D	E	F	G	H	I
240829	13/06/22,	06080685(	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240830	13/06/22,	06080685(	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240831	13/06/22,	-	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240832	13/06/22,	-	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240833	13/06/22,	20201203(	-	Course: AKUNTANSI KEUANGAN MENENGA	System	Course viewed	The user v	web	192.168.1.45
240834	13/06/22,	-	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240835	13/06/22,	202160177	-	Course: PSIKOLOGI PERKEMBANGAN (PKM2	System	Course viewed	The user v	web	192.168.1.45
240836	13/06/22,	20201203(	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240837	13/06/22,	-	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240838	13/06/22,	20201203(	-	Course: PENGAUDITAN 1 (AKK204A) Genap	System	Course viewed	The user v	web	192.168.1.45
240839	13/06/22,	20201203(	-	Front page	System	Course viewed	The user v	web	192.168.1.45
240840	13/06/22,	20201203(	-	Course: BAHASA INDONESIA (MKU304A) Ge	System	Course viewed	The user v	web	192.168.1.45
240841	13/06/22,	202111111	-	Course: EKONOMI MIKRO (EBM110C) Genap	System	Course viewed	The user v	web	192.168.1.45
240842	13/06/22,	202012185	-	Assignment: resume dan latihan materi 13	Assignment	Course module viewed	The user v	web	192.168.1.45
240843	13/06/22,	202012185	-	Assignment: resume dan latihan materi 13	Assignment	A submission has been submitted.	The user v	web	192.168.1.45
240844	13/06/22,	202012185	2020121	Assignment: resume dan latihan materi 13	System	Course activity completion updated	The user v	web	192.168.1.45
240845	13/06/22,	202012185	2020121	Assignment: resume dan latihan materi 13	File submissions	Submission created.	The user v	web	192.168.1.45
240846	13/06/22,	202012185	-	Assignment: resume dan latihan materi 13	File submissions	A file has been uploaded.	The user v	web	192.168.1.45
240847	13/06/22,	202012185	-	Assignment: resume dan latihan materi 13	Assignment	Course module viewed	The user v	web	192.168.1.45
240848	13/06/22,	201911737	-	Assignment: the second assignment	Assignment	Course module viewed	The user v	web	192.168.1.45
240849	13/06/22,	201911737	-	Assignment: the second assignment	Assignment	A submission has been submitted.	The user v	web	192.168.1.45
240850	13/06/22,	201911737	2019117	Assignment: the second assignment	File submissions	Submission updated.	The user v	web	192.168.1.45
240851	13/06/22,	201911737	-	Assignment: the second assignment	File submissions	A file has been uploaded.	The user v	web	192.168.1.45
240852	13/06/22,	201911737	-	Assignment: the second assignment	Assignment	Course module viewed	The user v	web	192.168.1.45

Gambar 4.5 Catatan LMS Moodle

Berkas hasil unduhan catatan aktivitas mahasiswa pada LMS tersebut per berkasnya memiliki ratusan ribu catatan salah satu contoh jumlah catatan ditunjukkan pada Gambar 4.5, adapun rata-rata per berkas memiliki kapasitas 40 Mb ditunjukkan pada Gambar 4.6.



Name	Date modified	Type	Size
1 Jun logs_20220704-1211.csv	7/4/2022 12:09 PM	Microsoft Excel C...	26,997 KB
2 Jun logs_20220704-1211.csv	7/4/2022 12:09 PM	Microsoft Excel C...	42,352 KB
3 Jun logs_20220704-1210.csv	7/4/2022 12:09 PM	Microsoft Excel C...	34,797 KB
4 Jun logs_20220704-1208.csv	7/4/2022 12:07 PM	Microsoft Excel C...	23,476 KB
5 Jun logs_20220704-1208.csv	7/4/2022 12:06 PM	Microsoft Excel C...	12,935 KB
6 Jun logs_20220704-1208.csv	7/4/2022 12:06 PM	Microsoft Excel C...	39,378 KB
7 Jun logs_20220704-1207.csv	7/4/2022 12:05 PM	Microsoft Excel C...	42,951 KB
8 Jun logs_20220704-1207.csv	7/4/2022 12:05 PM	Microsoft Excel C...	32,832 KB
9 Jun logs_20220704-1206.csv	7/4/2022 12:04 PM	Microsoft Excel C...	34,054 KB
10 Jun logs_20220704-1205.csv	7/4/2022 12:04 PM	Microsoft Excel C...	32,694 KB
11 Jun logs_20220704-1205.csv	7/4/2022 12:03 PM	Microsoft Excel C...	24,143 KB
12 Jun logs_20220704-1204.csv	7/4/2022 12:03 PM	Microsoft Excel C...	31,882 KB
13 Jun logs_20220704-1202.csv	7/4/2022 12:00 PM	Microsoft Excel C...	53,048 KB

Gambar 4.6 Kapasitas berkas catatan

Tabel 4.2 memperlihatkan jumlah catatan per hari dari setiap berkas. Data catatan di setiap minggunya rata-rata memiliki jutaan catatan kecuali pada minggu pertama dan kedua hal ini dikarenakan perkuliahan masih baru berjalan sehingga aktivitas pembelajaran masih sedikit.

Tabel 4.2 Jumlah catatan berkas LMS Moodle minggu 1-8

Minggu Ke	Tanggal	Bulan	Jumlah		Minggu Ke	Tanggal	Bulan	Jumlah
1	21	Februari	170,168		5	21	Maret	183,865
	22	Februari	143,781			22	Maret	197,081
	23	Februari	157,042			23	Maret	197,021
	24	Februari	143,281			24	Maret	209,192
	25	Februari	107,110			25	Maret	179,719
	26	Februari	75,064			26	Maret	107,816
	27	Februari	23,312			27	Maret	58,969
		Jumlah	**819,758**				**Jumlah**	**1,133,663**
2	28	Februari	111,424		6	28	Maret	174,690
	1	Maret	109,847			29	Maret	184,718
	2	Maret	196,652			30	Maret	191,273
	3	Maret	151,732			31	Maret	195,811
	4	Maret	174,048			1	April	175,189
	5	Maret	114,321			2	April	114,710
	6	Maret	48,222			3	April	83,455
		Jumlah	**906,246**				**Jumlah**	**1,119,846**
3	7	Maret	202,778		7	4	April	241,499
	8	Maret	200,431			5	April	235,914
	9	Maret	215,105			6	April	232,611
	10	Maret	222,699			7	April	223,853
	11	Maret	180,784			8	April	232,845
	12	Maret	127,587			9	April	203,075
	13	Maret	61,647			10	April	100,897
		Jumlah	**1,211,031**				**Jumlah**	**1,470,694**
4	14	Maret	205,640		8	11	April	272,137
	15	Maret	196,059			12	April	283,569
	16	Maret	193,557			13	April	273,114
	17	Maret	203,716			14	April	262,234
	18	Maret	190,271			15	April	134,025
	19	Maret	122,190			16	April	182,092
	20	Maret	58,790			17	April	91,105
		Jumlah	**1,170,223**				**Jumlah**	**1,498,276**

Tabel 4.3 Jumlah catatan berkas LMS Moodle minggu 5-14

Minggu Ke	Tanggal	Bulan	Jumlah	Minggu Ke	Tanggal	Bulan	Jumlah
9	18	April	219,631	13	16	Mei	106,320
	19	April	197,865		17	Mei	207,692
	20	April	210,776		18	Mei	200,861
	21	April	197,134		19	Mei	192,090
	22	April	167,020		20	Mei	155,154
	23	April	144,057		21	Mei	117,848
	24	April	70,377		22	Mei	47,732
		Jumlah	**1,206,860**			**Jumlah**	**1,027,697**
10	25	April	209,758	14	23	Mei	183,165
	26	April	213,549		24	Mei	184,669
	27	April	207,313		25	Mei	199,018
	28	April	213,309		26	Mei	144,889
	29	April	166,105		27	Mei	159,517
	30	April	116,908		28	Mei	94,199
	1	Mei	32,670		29	Mei	47,658
		Jumlah	**1,159,612**			**Jumlah**	**1,013,115**
11	2	Mei	27,205	15	30	Mei	213,377
	3	Mei	32,897		31	Mei	172,534
	4	Mei	39,316		1	Juni	128,837
	5	Mei	43,966		2	Juni	201,647
	6	Mei	43,484		3	Juni	166,781
	7	Mei	32,620		4	Juni	114,893
	8	Mei	16,599		5	Juni	62,138
		Jumlah	**236,087**			**Jumlah**	**1,060,207**
12	9	Mei	92,827	16	6	Juni	190,980
	10	Mei	106,718		7	Juni	203,463
	11	Mei	157,630		8	Juni	161,379
	12	Mei	168,858		9	Juni	167,147
	13	Mei	136,247		10	Juni	155,497
	14	Mei	101,540		11	Juni	117,112
	15	Mei	32,085		12	Juni	136,816
		Jumlah	**795,905**			**Jumlah**	**1,132,394**

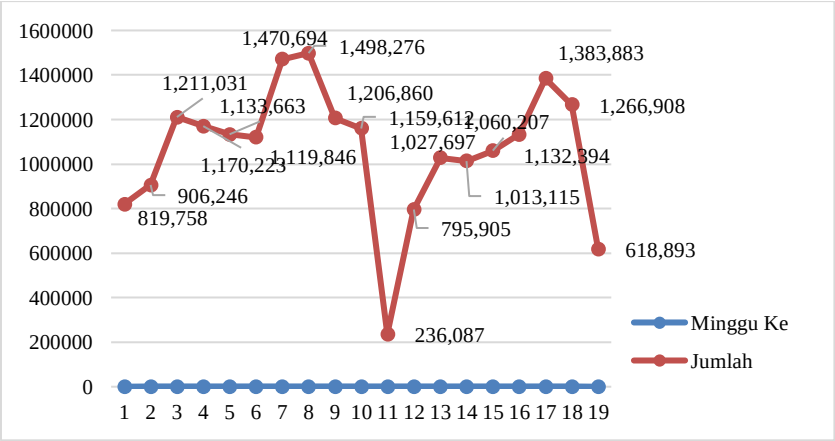
Tabel 4.3 menunjukkan catatan aktivitas mahasiswa di LMS Moodle dari minggu ke-9 sampai dengan minggu ke-16. Aktivitas menurun terletak pada minggu ke-11 dan ke-12. Hal ini terjadi karena pada minggu tersebut merupakan minggu setelah ujian tengah semester dan minggu pertama setelah tengah semester.

Tabel 4.4 merupakan catatan aktivitas mahasiswa di tiga minggu terakhir perkuliahan, pada tabel tersebut diperlihatkan

bahwa pada minggu terakhir mengalami penurunan catatan hal ini dikarenakan masa penilaian.

Tabel 4.4 Jumlah catatan berkas LMS Moodle minggu 17-19

Minggu Ke	Tanggal	Bulan	Jumlah		Minggu Ke	Tanggal	Bulan	Jumlah
17	13	Juni	240,851		19	27	Juni	115,008
	14	Juni	229,338			28	Juni	104,195
	15	Juni	200,984			29	Juni	96,230
	16	Juni	260,553			30	Juni	128,225
	17	Juni	257,605			1	Juli	102,485
	18	Juni	85,040			2	Juli	72,750
	19	Juni	109,512				**Jumlah**	**618,893**
		Jumlah	**1,383,883**				**Total**	**20,231,298**
18	20	Juni	276,132					
	21	Juni	203,468					
	22	Juni	182,813					
	23	Juni	195,262					
	24	Juni	211,391					
	25	Juni	142,649					
	26	Juni	55,193					
		Jumlah	**1,266,908**					



Gambar 4.7 Grafik catatan periode mingguan

Pada gambar 4.7 menunjukkan bahwa terjadi fluktuasi catatan dalam satu semester perkuliahan. Fluktuasi jumlah catatan

ini terjadi karena variasi model di dalam proses perkuliahan antar pengampu berbeda-beda.

4.3.2. Data Organisasi

Data organisasi mahasiswa dikumpulkan dari bagian administrasi akademik dan kemahasiswaan (BAAK). Data dikumpulkan berdasarkan surat keterangan (SK) asli ataupun salinan dari masing-masing organisasi. Data ini mewakili keikutsertaan mahasiswa dalam berorganisasi, data organisasi tersebut terdiri dari data organisasi universitas, fakultas, program studi (Prodi) maupun unit kegiatan mahasiswa (UKM). Organisasi di tingkat universitas diantaranya adalah badan eksekutif mahasiswa (BEM) Universitas dan dewan perwakilan mahasiswa (DPM). Sementara itu organisasi fakultas terdiri dari BEM fakultas dan teater fakultas. Sedangkan himpunan mahasiswa prodi (Himapro/Hima) merupakan organisasi program studi (Prodi). Organisasi di tingkat prodi beranggotakan mahasiswa pada prodi tersebut, begitu pula organisasi di tingkat fakultas anggotanya merupakan mahasiswa-mahasiswa pada fakultas tersebut pula. Organisasi yang anggotanya berasal dari semua prodi yang ada di universitas adalah BPM dan BEM Universitas, selain itu ada beberapa organisasi yang berfungsi untuk menyalurkan minat maupun bakat mahasiswa dari semua prodi yaitu UKM.

Dari hasil pengumpulan data ditemukan bahwa organisasi di tingkat universitas terdiri dari dua organisasi yaitu BEM Universitas dan DPM, sedangkan organisasi di tingkat fakultas terdiri dari sepuluh organisasi yaitu BEM Fakultas Ekonomi dan Bisnis (FEB), BEM Fakultas Hukum (FH), BEM Fakultas Teknik (FT), BEM Fakultas Keguruan dan Ilmu Pendidikan (FKIP), BEM Fakultas Pertanian (FP), BEM Fakultas Psikologi (FPsi), Obeng (Teater FT), Teater Coin FEB, Teater Tiga Koma FKIP dan SEC

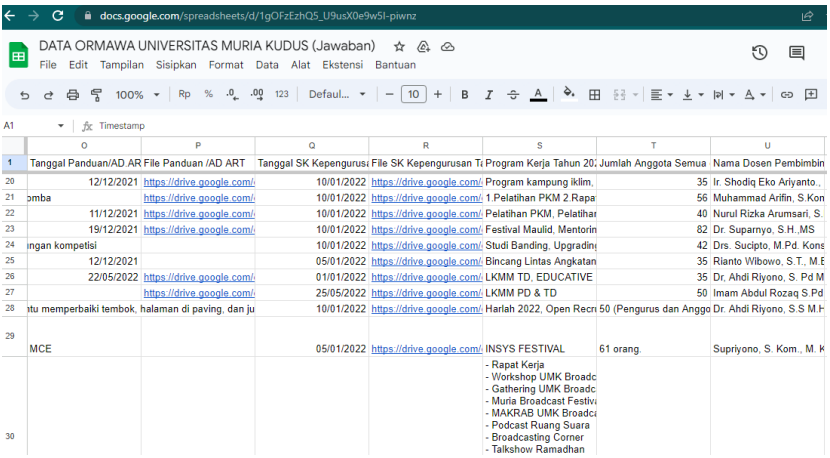
FEB. Sementara itu organisasi di tingkat prodi terdiri dari sebelas organisasi yaitu Himapro Manajemen, Himapro Teknik Elektro, Himapro Akuntansi, Hima Pendidikan Bahasa Inggris, Hima Pendidikan Bimbingan Konseling, Hima Sistem Informasi, Hima Pendidikan Guru Sekolah Dasar, Hima Teknik Informatika, Hima Matematika, Hima Teknik Industri dan Hima Teknik Mesin. UKM di kampus ini terdiri dari 17 UKM yaitu UKM Tabula Rasa, Mapala, Formi, Resimen Mahasiswa (Menwa), Dewan Racana, Forum Mahasiswa Ilmiah (FIMA), Paduan Suara Pitaloka, Broadcasting, Korps Sukarela (KSR), Persekutuan Mahasiswa Katolik (PMK), Debat, Jurnalistik, Koperasi Mahasiswa, Ningnong, Olah Raga (Olga), Seni Kampus (Sekam) dan UKM pramuka yaitu Racana. Gambar 4.8 menunjukkan SK maupun salinan SK yang telah dikumpulkan dari BAAK.

Name	Date modified	Type	Size
SALINAN SK BEM FEB UMK 2022 - ARDEGHA HIERO AKBAR.pdf	6/20/2022 9:55 AM	Microsoft Edge P...	533 KB
SALINAN SK PENGURUS HIMAPRO AKT - HIMAPRO AKUNTANSI.pdf	6/20/2022 9:55 AM	Microsoft Edge P...	912 KB
Salinan SK Pitaloka2022 - SEPTIAN JULIANTO.pdf	6/20/2022 9:48 AM	Microsoft Edge P...	1,723 KB
SK BEM Fakultas Psikologi 2022 - AULA ARIQURROHMAN.pdf	6/20/2022 10:06 AM	Microsoft Edge P...	244 KB
SK BEM FH 2022 - Hendrik Cahyono.pdf	6/20/2022 9:51 AM	Microsoft Edge P...	1,102 KB
SK BEM FKIP 2022 - Haytsam Abdillah.pdf	6/20/2022 9:59 AM	Microsoft Edge P...	676 KB
SK ESA 2022 - Nazarudddin Ghozali.pdf	6/20/2022 10:08 AM	Microsoft Edge P...	480 KB
SK HIMA BK 22 - Maulinda Tri Muliasetyani.pdf	6/20/2022 9:45 AM	Microsoft Edge P...	1,205 KB
SK HIMA SI 2022 - AGUS CAHYATI.pdf	6/20/2022 10:05 AM	Microsoft Edge P...	995 KB
SK HIMAPRO PGSD 2022 - Raka Ady.pdf	6/20/2022 10:13 AM	Microsoft Edge P...	1,181 KB
SK Himapro TI - Vera Nurviana.pdf	6/20/2022 9:47 AM	Microsoft Edge P...	341 KB
SK HIMATIKA 2022 - Melly kn.pdf	6/20/2022 10:12 AM	Microsoft Edge P...	966 KB
SK KEPENGURUSAN DARI KAMPUS - Arga Dahana UMK.pdf	6/20/2022 9:56 AM	Microsoft Edge P...	784 KB
SK Kepengurusan HIMAPROTIND - MUHAMMAD MIFTAKHUL HUDA.docx	6/20/2022 9:54 AM	Microsoft Word D...	773 KB
SK KEPENGURUSAN UKMF TABULA RASA - SITI HAMIDAH.pdf	6/20/2022 9:46 AM	Microsoft Edge P...	1,005 KB
SK Kepengurusan UNIV KSR UMK .pdf	6/20/2022 9:46 AM	Microsoft Edge P...	238 KB
SK OBENG.pdf	7/12/2022 7:20 PM	Microsoft Edge P...	606 KB
SK Teater Coin - Umahatun qurotu Ayuni.pdf	6/20/2022 9:48 AM	Microsoft Edge P...	2,242 KB
SK TIGA KOMA - RIZKY ANITASARI.pdf	6/20/2022 10:00 AM	Microsoft Edge P...	305 KB
SK TIM SEC FEB (2) - idha rachmawati.pdf	6/20/2022 10:11 AM	Microsoft Edge P...	358 KB
SK TM 2021 - Himapro Teknik Mesin UMK.pdf	6/20/2022 9:58 AM	Microsoft Edge P...	500 KB
SK UKM BROADCASTING PERIODE 2022 - Broadcasting Universitas Muria Kudu...	6/20/2022 10:06 AM	Microsoft Edge P...	1,434 KB
SK UKM FORMI PERIODE 2022 - AFRA GHAIDA YUSFIKA FAUZIYA.pdf	6/20/2022 9:57 AM	Microsoft Edge P...	1,545 KB

Gambar 4.8 Dokumen SK/Salinan organisasi

Metode penyimpanan SK maupun salinannya selain disimpan dalam bentuk fisik, SK ini juga disimpan di media penyimpanan Google dan web BAAK. Metode penyimpanan ini bertujuan untuk memudahkan dalam mengakses SK maupun

salinan tersebut. Gambar 4.9 menunjukkan lokasi penyimpanan SK maupun salinan di Google.



	O	P	Q	R	S	T	U
1	Tanggal Panduan/AD AR	File Panduan /AD ART	Tanggal SK Kepengurusa	File SK Kepengurusan Tı	Program Kerja Tahun 20	Jumlah Anggota Semua	Nama Dosen Pembimbin
20	12/12/2021	https://drive.google.com/	10/01/2022	https://drive.google.com/	Program kampung iklim,	35	Ir. Shodiq Eko Ariyanto.,
21	omba	https://drive.google.com/	10/01/2022	https://drive.google.com/	1 Pelatihan PKM 2.Rapa	56	Muhammad Arifin, S.Kon
22	11/12/2021	https://drive.google.com/	10/01/2022	https://drive.google.com/	Pelatihan PKM, Pelatihar	40	Nurul Rizka Arumsari, S.
23	19/12/2021	https://drive.google.com/	10/01/2022	https://drive.google.com/	Festival Maulid, Mentorin	82	Dr. Suparnyo, S.H.,MS
24	ingan kompetisi		10/01/2022	https://drive.google.com/	Studi Banding, Upgradin	42	Drs. Sucipto, M.Pd. Kons
25	12/12/2021		05/01/2022	https://drive.google.com/	Bincang Lintas Angkatan	35	Rianto Wibowo, S.T., M.E
26	22/05/2022	https://drive.google.com/	01/01/2022	https://drive.google.com/	LKMM TD, EDUCATIVE	35	Dr, Ahdi Riyono, S. Pd M
27		https://drive.google.com/	25/05/2022	https://drive.google.com/	LKMM PD & TD	50	Imam Abdul Rozaq S.Pd
28	ıtu memperbaiki tembok, halaman di paving, dan ju		10/01/2022	https://drive.google.com/	Harlah 2022, Open Recr	50 (Pengurus dan Anggo	Dr. Ahdi Riyono, S.S M.H
29	MCE		05/01/2022	https://drive.google.com/	INSYS FESTIVAL	61 orang.	Supriyono, S. Kom., M. K
30					- Rapat Kerja - Workshop UMK Broadc - Gathering UMK Broadc - Muria Broadcast Festiv - MAKRAB UMK Broadca - Podcast Ruang Suara - Broadcasting Corner - Talkshow Ramadhan		

Gambar 4.9 Penyimpanan SK/Salinan organisasi

4.3.3. Data Akademik

Data akademik merupakan data yang diperoleh dari sistem informasi akademik (SIAKAD) pada akhir semester di semester genap tahun ajaran 2020-2021. Data yang diambil adalah data identitas mahasiswa yaitu nomor induk mahasiswa (NIM) dan indek prestasi sementara mahasiswa (IPS). Data ini terdiri dari 10.044, yang berasal dari 15 program studi dan 6 fakultas. Tabel 4.5 menunjukkan sebaran data akademik dari masing-masing fakultas maupun prodi.

Tabel 4.5 Data akademik

NO	FAKULTAS	PRODI	JUMLAH
1	Psikologi	Psikologi	649
2	Teknik	Teknik Industri	185
3		Teknik Mesin	458
4		Sistem Informasi	701
5		Teknik Elektronika	156
6		Teknik Informatika	985
7	Agroteknologi	Agroteknologi	304
8		Pendidikan Matematika	107

9	Keguruan dan Ilmu Pendidikan	Pendidikan Bahasa Indonesia	128
10		Pendidikan Guru Sekolah Dasar	1264
11		Pendidikan Bahasa Inggris	425
12		Pendidikan Konseling	165
13	Hukum	Hukum	783
14	Ekonomi dan Bisnis	Akuntansi	960
15		Manajemen	2774
			10044

Pada tabel 4.5 tersebut ditunjukkan bahwa prodi yang memiliki jumlah mahasiswa terbanyak adalah mahasiswa dari prodi Manajemen dengan 2774 mahasiswa dan selanjutnya dari prodi PGSD dengan 1264 mahasiswa, sedangkan prodi dengan jumlah mahasiswa paling sedikit adalah prodi baru yaitu Pendidikan Matematika yang memiliki 107 mahasiswa.

4.3.4. Data Ekonomi

Data ini diperoleh dari data informasi pendapatan orang tua atau wali mahasiswa saat mahasiswa baru melengkapi data pendaftaran ulang. Data pendaftaran ulang tersebut memiliki 88 kolom yang terdiri dari; NO, NIM, Nama Lengkap, Jenis Kelamin, Tempat/Tanggal Lahir, Program Studi, Periode Masuk Angkatan, Status Masuk, Jalur Masuk, Tanggal Daftar, Gelombang Pendaftaran, Status Akademik, Status Mahasiswa, Agama, Status Nikah, Kewarganegaraan, Status Domisili, Kota Tinggal, Alamat, Kode Pos. Negara, No. Telp/Hp, Email, Hubungan Biaya, Sumber Dana Beasiswa, Jumlah Saudara, Jumlah Saudara Laki, Jumlah Saudara Perempuan, Status Bekerja, Pekerjaan, Institusi/Kantor, Alamat Institusi/Kantor, No Asuransi, Hoby, NIM Lama, PT Asal, Tahun Masuk PT Asal, Nama Ayah, Tanggal Lahir Ayah, Status Ayah, Tanggal Meninggal Ayah, Pendidikan Ayah, Pendidikan Terakhir Ayah, Pekerjaan Ayah, Nama Ibu, Tanggal Lahir Ibu, Status Ibu, Tanggal Meninggal Ibu, Pendidikan Ibu, Pendidikan

Terakhir Ibu, Pekerjaan Ibu, Agama Orang Tua, Warga Negara Orang Tua, Alamat Orang Tua, Kota Orang Tua, Kode Pos Orang Tua, No. Telp Orang Tua, Email Orang Tua, Orang Tua Mampu, Penghasilan Orang Tua, Jumlah Tanggungan, Nama Wali, Tanggal Lahir Wali, Status Wali, Tanggal Meninggal Wali, Alamat Wali, Kota Wali, Kode Pos Wali, No Telp Wali, Email Wali, Pendidikan Wali, Pendidikan Terakhir Wali, Pekerjaan Wali, Tahun Daftar SMTA, Tahun Lulus SMTA, Jurusan SMTA, Jenis SMTA, Nama SMTA, Alamat SMTA, No Ijazah SMTA, Ijazah SMTA, Tanggal Ijazah SMTA, Status SMTA, Akreditasi SMTA, Nilai Ujian AKhir SMTA, NIK (KTP).

Penulis menyadari bahwa sebenarnya masih banyak variabel yang perlu diteliti lebih lanjut namun setelah dicermati ditemukan bahwa variabel-variabel tersebut tidak semuanya dilengkapi oleh mahasiswa baru. Adapun penelitian ini memfokuskan variabel ekonomi dan demografi untuk dijadikan salah satu variabel dalam memprediksi performa akademik mahasiswa. Adapun data ekonomi dikumpulkan dari kolom penghasilan orang tua, kolom ini memiliki 7 kategori yaitu 100.000-500.000, 500.000-1.000.000, 1.000.000-2.500.000, 2.500.000-5.000.000, 5.000.000-7.500.000, 7.500.000-10.000.000 dan >10.000.000.

4.3.5. Data Demografi

Data demografi merupakan data tempat tinggal mahasiswa, seperti halnya data ekonomi data ini juga diperoleh dari data pendaftaran ulang mahasiswa baru dengan mengambil kolom Kota Tinggal. Berdasarkan data yang diperoleh, data ini memiliki 53 variasi yang artinya mahasiswa di kampus ini berasal dari berbagai daerah. Kota-kota asal mahasiswa tersebut yaitu; Kudus, Grobogan, Jepara, Pati, Rembang, Blora, Demak, Semarang, Batang, Bekasi, Bogor, Brebes, Depok, Jakarta Barat, Jakarta

Timur, Karanganyar, Karawang, Kedalon, Batangan, Kendal, Klaten, Magelang, Majalengka, Ngawi, Pemalang, Purbalingga, Purwokerto, Sukoharjo, Surakarta, Tangerang, Tegal, Tuban, Wonosobo, Balikpapan, Batam, Bengkulu, Denpasar, Dompu, Jambi, Kepulauan Selayar, Ketapang, Kotawaringin Barat, Kotawaringin Timur, Lampung Barat, Lampung Selatan, Makassar, Merangin, Mesuji, Ogan Komering Ilir, Palu, Pringsewu, Rokan Hulu, Sukamara, Tanjung Pinang.

Tujuan penggunaan data ini adalah untuk mengetahui apakah ada korelasi antara prestasi akademik dengan asal mahasiswa tersebut. Selain data domisili, data demografi juga memanfaatkan data jenis kelamin yang juga dikumpulkan dari data pendaftaran ulang mahasiswa baru dengan mengambil kolom Jenis Kelamin. Data ini terdiri dari dua kategori yaitu laki-laki dan perempuan.

4.4. Data Preparation

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Tahap ini merupakan bagian penting dari proses data mining, yaitu proses yang memiliki peran 60-80%. Setelah proses pengumpulan data, langkah selanjutnya adalah ekstraksi data, proses ini merupakan proses untuk mengekstrak data mentah kedalam data yang yang mudah dipahami oleh manusia maupun mesin. Hal ini dikarenakan beberapa data mentah khususnya data catatan aktivitas mahasiswa dalam LMS Moodle, data SK organisasi berbentuk tidak terstruktur sehingga diperlukan ekstraksi data untuk dapat membacanya. Selain itu data-data lainnya juga perlu diekstraksi agar mudah digunakan dan siap untuk digunakan dalam prediksi performa akademik mahasiswa.

4.4.1. Ekstraksi data LMS Moodle

Sebelum melakukan ekstraksi dari sebuah data maka diperlukan informasi dari data yang akan diekstrak. Pada Gambar 4.10. ditunjukkan data catatan LMS dari Moodle, yang terdiri dari sembilan kolom. Kolom pertama, ketiga, kedelapan, dan kesembilan merupakan informasi tentang waktu akses, pengguna, media akses, dan alamat IP pengguna. Kolom kedua adalah id dan nama siswa dimana kolom ini adalah kunci utama data. Kolom kedua memiliki data yang sama dengan kolom id hal ini dikarenakan kolom ini mencatat semua aktivitas pengguna. Adapun kolom keempat hingga ketujuh merupakan detail aktivitas yang dilakukan oleh pengguna di LMS Moodle tersebut.

Proses ekstraksi data menggunakan bahasa pemrograman Python dengan menggunakan alat bantu Jupyter. Jupyter merupakan aplikasi web gratis yang digunakan untuk membuat dan membagikan dokumen yang memiliki kode, hasil hitungan, visualisasi, dan teks. Jupyter adalah singkatan dari tiga bahasa pemrograman Julia (Ju), Python (Py), dan R. Jupyter berfungsi untuk membuat narasi komputasi yang menjelaskan makna dari data di dalamnya dan memberikan insight mengenai data tersebut.

Time	User full name	Affected user	Event context	Component	Event name	Description	Origin	IP address
6/06/22, 23:59	202111597 QONITA ELV	-	Assignment: Pengum	Assignment	Course module v	The user with i	web	192.168.1.45
6/06/22, 23:59	202051172 YAASINTA	-	Course: PENGEMBAN	System	Course viewed	The user with i	web	192.168.1.45
6/06/22, 23:59	201911341 DITA AYU W	-	Front page	System	Course viewed	The user with i	web	192.168.1.45
6/06/22, 23:59	202051172 YAASINTA	-	Front page	System	Course viewed	The user with i	web	192.168.1.45
6/06/22, 23:59	202111597 QONITA ELV	-	File: Latihan Soal Uji	File	Course module v	The user with i	web	192.168.1.45
6/06/22, 23:59	202051162 DANENDRA	-	Assignment: Decisio	Assignment	Course module v	The user with i	web	192.168.1.45
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
6/06/22, 00:00	202151157 VIAN NIBRA	-	Assignment: Collecti	Assignment	Course module v	The user with i	web	192.168.1.45
6/06/22, 00:00	201911542 RIZQINAZIL	-	Course: ETIKA BISNIS	System	Course viewed	The user with i	web	192.168.1.45
6/06/22, 00:00	202052004 MOCHAMM	-	Course: PROGRAMM	System	Course viewed	The user with i	web	192.168.1.45

Gambar 4.10 Catatan LMS Moodle

```
import pandas as pd
import numpy as np

df = pd.read_csv('D:Data Disertasi/Dataku/Juli/2 Jul logs_20220704-1146.csv

#Hapus dan milih
df.drop(["IP address","Affected user","Time",
         "Event context","Description","Origin"], axis=1, inplace=True)

#memilih user fulname yg ada carakter tertentu
df1 = df[df['User full name'].str.contains(r'2019|2020')]

#Memisahkan Nim dan Nama
first_names_list=[]
last_names_list=[]

#Fungsi perulangan pemisahan
for name in df1['User full name']:
  first_name, last_name= name.split(' ',1)
  first_names_list.append(first_name)
  last_names_list.append(last_name)

#Menambahkan data ke df
df1.insert(0,"NIM", first_names_list)
df1.insert(1,"NAMA", last_names_list)
del df1['User full name']
```

Gambar 4.11 Potongan Kode Ekstraksi Catatan LMS (a)

Proses ekstraksi diawali dengan membuka aplikasi Jupiter, selanjutnya pada menu New pilih Notebook dan pilih Python 3 (ipykernel) maka akan muncul file baru. Langkah selanjutnya adalah mulai melakukan penulisan kode-kode ekstraksi seperti yang ditunjukkan pada gambar 4.11. Baris pertama dan kedua merupakan kode untuk memanggil Pustaka Pandas dan Numpy yang dilanjutkan dengan memberikan nama alias agar memudahkan dalam pemanggilan didalam kode-kode selanjutnya. Pandas merupakan pustaka open source pada Python yang sering digunakan untuk memproses data yang meliputi pembersihan data, manipulasi data, hingga melakukan analisis data. Pustaka Pandas ini dapat mengolah data berbentuk csv, txt, excel, html, dan sebagainya. Sementara itu NumPy (Numerical Python) adalah pustaka Python yang fokus pada scientific computing. Numpy menyediakan fungsi yang siap pakai untuk memudahkan dalam melakukan perhitungan saintifik seperti matriks, aljabar, statistik,

dan sebagainya. Pada baris ke 4 diawali dengan membuat variabel dengan nama df yang digunakan untuk menampung kode setelahnya, kemudian memanggil pandas untuk membaca file catatan LMS Moodle yang berada pada folder tempat penyimpanan catatan. Kode pada baris keenam dan semua kode yang diawali dengan # merupakan keterangan dan tidak diproses oleh aplikasi, selanjutnya pada kode baris ketujuh dan kedelapan merupakan kode untuk menghapus kolom-kolom pada catatan yang tidak diperlukan, yaitu kolom; IP address, Affected user, Time, Event context, Description dan kolom Origin.

B
User full name
202151061 AHMAD PUJI
201912227 AYU FAJERIN
202151061 AHMAD PUJI
202151091 MUHAMMAD FARID
201912227 AYU FAJERIN
202111342 HELMY YAHYA
202111150 AMANDA

Gambar 4.12 User full name

Selanjutnya pada baris kesebelas merupakan proses pemilihan data pada kolom “User full name” yang mengandung kata 2019 dan 2020. Proses ini merupakan proses ekstraksi untuk mendapatkan data catatan LMS Moodle dari pengguna dengan sasaran mahasiswa angkatan 2019 dan 2020 sesuai dengan tujuan awal penelitian. Baris ke-13 dan ke-14 merupakan kode untuk menampung hasil ekstraksi dari kolom ‘User full name’, kolom ini memiliki dua nilai yaitu nomor induk mahasiswa (NIM) dan nama mahasiswa ditunjukkan pada gambar 4.12, sehingga untuk mendapatkan kedua nilai tersebut maka diperlukan pemisahan. Baris kode ke-18 sampai dengan ke-21 merupakan kode perulangan untuk memisahkan NIM dan nama dan selanjutnya

menyimpan ke dalam variabel first_names_list dan last_names_list. Proses ekstraksi NIM dan nama berakhir pada baris ke-24 sampai dengan baris ke-26 yang merupakan proses menambahkan kolom NIM dan NAMA hasil pemisahan kolom ‘User full name’ kedalam data catatan aktivitas LMS Moodle sekaligus menghapus kolom “User full name”.

```
28 #memilih hanya colum ini
29 df1[['NIM','Component','Event name']]
30
31 #rename columns
32 df2 = df1.rename({'Event name':'EventName'}, axis=1)
33
34 #memilih Nim 9
35 df3 = df2[df2['NIM'].str.len()==9]
36
37 #Mengambil 1 nim
38 colum1 = df3.drop_duplicates(subset=['NIM'], keep='first')
39
40 #Menambahkan colum 'Total Login' dengan nilai = 1
41 colum1.insert(1, 'Total_login', 1)
42
43 #Seleksi NIM dan Total Login
44 Tl = colum1[['NIM','Total_login']]
45
46 ##########################################################################
47 #N_access_forum
48 #seleksi
49 colum2 = df3[df3['Component'].str.contains('Forum')]
50 #menghitung
51 gt = colum2.groupby(['NIM']).count().reset_index()
52 #merubah nama colom
53 gl = gt.rename({'NAMA':'N_access_forum'}, axis=1)
54 #seleksi sesuai kebutuhan
55 Naf = gl[['NIM','N_access_forum']]
```

Gambar 4.13 Potongan kode ekstraksi catatan LMS (b)

Langkah selanjutnya dalam proses ekstraksi data LMS ditunjukkan pada gambar 4.13, pada baris ke-29 merupakan kode untuk mengambil hanya kolom 'NIM', 'Component' dan 'Event name'. dilanjutkan dengan kode baris ke-32 yaitu membuat variabel dengan nama df2 yang digunakan untuk menampung kolom NIM dan sekaligus menghapus spasi untuk memastikan nilai NIM benar-benar valid yang terdiri dari 9 karakter. Baris ke-

38 mendefinisikan variabel dengan nama colum1 yang digunakan untuk menghapus NIM yang sama sehingga nantinya didapatkan satu NIM dalam data tersebut. Pada kode baris ke-41 merupakan kode yang berfungsi untuk menambahkan kolom ‘Total Login” dengan nilai satu kedalam variabel kolom 1. Baris ke-44 mendefinisikan sebuah variabel dengan nama T1 untuk menampung data dari variabel colom1 yang berisi data 'NIM' dan 'Total_login'. Masih pada potongan kode gambar 4.13 yaitu pada baris ke-49 sampai dengan baris ke-55 merupakan ekstraksi untuk mendapatkan nilai dari variabel jumlah pengguna dalam melakukan akses di dalam forum (N_access_forum). Baris ke-49 pada kode program dimulai dengan membuat variabel dengan nama kolom 2 yang digunakan untuk menampung variabel df3 dan mengambil kolom 'Component' yang mengandung kata 'Forum'. Baris ke-51 dimulai dengan membuat variabel dengan nama gt untuk menampung jumlah NIM dari variabel kolom 2, dilanjutkan dengan kode baris ke-53 yaitu membuat variabel gl untuk mengubah kolom NAMA menjadi kolom ‘N_access_forum dan pada baris ke-55 merupakan variabel dengan nama naf yang digunakan untuk menampung data NIM dan data ‘N_access_forum’.

Gambar 4.14 merupakan proses ekstraksi dengan mengekstrak data dari kolom EventName untuk mendapatkan data jumlah pengguna dalam mengirimkan pesan ke forum (N_added_messages_forum) hal ini ditunjukkan pada gambar baris ke-60 sampai dengan baris ke-66. Proses ini hampir sama dengan proses untung mengekstrak data N_access_forum, pada baris ke-60 dimulai dengan membuat variabel dengan nama colum3 untuk menampung variabel df3 dari kolom EventName yang mengandung kata ‘message’, baris ke-62 merupakan kode untuk menghitung jumlah message dari setiap NIM, kode baris ke-64 untuk merubah nama kolom dan kode pada baris ke-66

merupakan variabel dengan nama Namf yang digunakan untuk menampung hasil ekstraksi N_added_messages_forum.

```
57 ##########################################################################
58 #N_added_messages_forum
59 #seleksi
60 colum3 = df3[df3['EventName'].str.contains('message')]
61 #menghitung
62 gt1 = colum3.groupby(['NIM']).count().reset_index()
63 #merubah nama colom
64 gl1 = gt1.rename({'NAMA':'N_added_messages_forum'}, axis=1)
65 #seleksi sesuai kebutuhan
66 Namf = gl1[['NIM','N_added_messages_forum']]
67
68 ##########################################################################
69 #N_access_didactic_units
70 colum4 = df3[df3['Component'].str.contains('File') &
71              df3['EventName'].str.contains('Course module viewed')]
72 #menghitung
73 gt2 = colum4.groupby(['NIM']).count().reset_index()
74 #merubah nama colom
75 gl2 = gt2.rename({'NAMA':'N_access_didactic_units'}, axis=1)
76 #seleksi sesuai kebutuhan
77 Nadu = gl2[['NIM','N_access_didactic_units']]
78
```

Gambar 4.14 Potongan kode ekstraksi catatan LMS (c)

Kode ekstraksi pada baris ke-70 digunakan untuk mengekstrak data pengguna dalam mengakses materi pada LMS (N_access_didactic_units), kode dan metode yang digunakan sama dengan kode dan metode ekstraksi sebelumnya, bedanya adalah terletak pada pengambilan nilai dari variabel df3 yaitu mengambil kolom Component dan Eventname. Kedua kolom tersebut diambil nilainya jika kolom Component mengandung kata ‘file’ dan kolom Eventname mengandung kata ‘Course module viewed’. Adapun proses selanjutnya yaitu merubah nama kolom dan menampung hasilnya pada variabel Nadu.

Pada gambar 4.15 kode baris ke-81 merupakan variabel yang digunakan untuk menampung data dari variabel df3 kolom Component yang mengandung kata Glossary. Kode baris ke-81 sampai dengan baris ke-87 merupakan kode untuk mengekstrak data catatan LMS untuk mendapatkan jumlah pengguna dalam mengakses Glossary. Glossary adalah suatu daftar alfabetis istilah

dalam suatu ranah pengetahuan tertentu yang dilengkapi dengan definisi untuk istilah-istilah tersebut. Kode selanjutnya adalah kode merubah nama kolom dan menyimpan hasil ekstraksi Glossary kedalam variabel Nag. Kode pada baris ke-91 sampai dengan baris ke-97 merupakan kode ekstraksi untuk mendapatkan nilai dari jumlah tugas yang diberikan kepada pengguna LMS. Data ini diambil dari kolom Component yang mengandung kata Assignment, data hasil ekstraksi disimpan kedalam variabel Tass.

```
79  ##########################################################################
80  #N_access_glossaries
81  colum5 = df3[df3['Component'].str.contains('Glossary')]
82  #menghitung
83  gt3 = colum5.groupby(['NIM']).count().reset_index()
84  #merubah nama colom
85  gl3 = gt3.rename({'NAMA':'N_access_glossaries'}, axis=1)
86  #seleksi sesuai kebutuhan
87  Nag = gl3[['NIM','N_access_glossaries']]
88
89  ##########################################################################
90  #Total_assignments
91  colum6 = df3[df3['Component'].str.contains('Assignment')]
92  #menghitung
93  gt4 = colum6.groupby(['NIM']).count().reset_index()
94  #merubah nama colom
95  gl4 = gt4.rename({'NAMA':'Total_assignments'}, axis=1)
96  #seleksi sesuai kebutuhan
97  Tass = gl4[['NIM','Total_assignments']]
98
```

Gambar 4.15 Potongan kode ekstraksi catatan LMS (d)

Gambar 4.16 Merupakan kode ekstraksi untuk mendapatkan catatan pengguna LMS Moodle ke dalam data jumlah mahasiswa dalam berkonsultasi tugas (N_assignments_consulted), jumlah tugas yang diserahkan (N_assignments_submitted) dan jumlah pengguna dalam mengakses kuesioner (N_access_questionnaires). N_assignments_consulted diekstrak dari variabel df3 kolom Component yang mengandung kata ‘Assignment’ dan kolom EventName yang mengandung kata ‘consul’, sementara itu untuk mendapatkan data N_assignments_submitted diperoleh dari variabel df3 kolom

Component yang mengandung kata ‘Assignment’ dan kolom EventName yang mengandung kata ‘A submission has been submitted’, sedangkan data N_access_questionnaires diperoleh dari kolom Component yang mengandung kata ‘Quiz’, masing-masing data tersebut disimpan kedalam variabel Nac, Nasu dan Naqu.

```
99  ##########################################################################
100 #N_assignments_consulted
101 colum7 = df3[df3['Component'].str.contains('Assignment') &
102              df3['EventName'].str.contains('consul')]
103 #menghitung
104 gt5 = colum7.groupby(['NIM']).count().reset_index()
105 #merubah nama colom
106 gl5 = gt5.rename({'NAMA':'N_assignments_consulted'}, axis=1) #bagian ir
107 #seleksi sesuai kebutuhan
108 Nac = gl5[['NIM','N_assignments_consulted']]
109
110 ##########################################################################
111 #N_assignments_submitted
112 colum8 = df3[df3['Component'].str.contains('Assignment') &
113              df3['EventName'].str.contains('A submission has been subm
114 #menghitung
115 gt6 = colum8.groupby(['NIM']).count().reset_index()
116 #merubah nama colom
117 gl6 = gt6.rename({'NAMA':'N_assignments_submitted'}, axis=1)
118 #seleksi sesuai kebutuhan
119 Nasu = gl6[['NIM','N_assignments_submitted']]
120
121 ##########################################################################
122 #N_access_questionnaires
123 colum9 = df3[df3['Component'].str.contains('Quiz')]
124 #menghitung
125 gt7 = colum9.groupby(['NIM']).count().reset_index()
126 #merubah nama colom
127 gl7 = gt7.rename({'NAMA':'N_access_questionnaires'}, axis=1)
128 #seleksi sesuai kebutuhan
129 Naqu = gl7[['NIM','N_access_questionnaires']]
```

Gambar 4.16 Potongan kode ekstraksi catatan LMS (d)

Gambar 4.17 merupakan potongan kode untuk mengekstrak jumlah mahasiswa dalam berkonsultasi yang diperoleh dari variabel df3 kolom EventName yang mengandung kata ‘Quiz attempt’ lalu disimpan didalam variabel dengan nama Natqu.

```
130
131 ##########################################################################
132 #N_attempts_questionnaires
133 colum10 = df3[df3['EventName'].str.contains('Quiz attempt')]
134 #menghitung
135 gt8 = colum10.groupby(['NIM']).count().reset_index()
136 #merubah nama colom
137 gl8 = gt8.rename({'NAMA':'N_attempts_questionnaires'}, axis=1)
138 #seleksi sesuai kebutuhan
139 Natqu = gl8[['NIM','N_attempts_questionnaires']]
140
141 ##########################################################################
142 #N_answered_questions
143 colum11 = df3[df3['EventName'].str.contains('answer')]
144 #menghitung
145 gt9 = colum11.groupby(['NIM']).count().reset_index()
146 #merubah nama colom
147 gl9 = gt9.rename({'NAMA':'N_answered_questions'}, axis=1)
148 #seleksi sesuai kebutuhan
149 Nanqu = gl9[['NIM','N_answered_questions']]
150
151 ##########################################################################
```

Gambar 4.17 Potongan kode ekstraksi catatan LMS (e)

Sementara data jumlah tugas yang diserahkan dikumpulkan dari variabel df3 kolom EventName yang mengandung kata ‘answer’ lalu disimpan kedalam variabel dengan nama Nanqu. Sedangkan data jumlah pengguna dalam mengakses kuesioner diperoleh dari variabel df3 kolom Component yang mengandung kata ‘Quiz” dan kolom EventName yang mengandung kata ‘viewed’ lalu disimpan kedalam variabel dengan nama Nqv ditunjukan pada gambar 4.18.

```
151 ##########################################################################
152 #N_questionnaire_views
153 colum12 = df3[df3['Component'].str.contains('Quiz') &
154               df3['EventName'].str.contains('viewed')]
155 #menghitung
156 gt10 = colum12.groupby(['NIM']).count().reset_index()
157 #merubah nama colom
158 gl10 = gt10.rename({'NAMA':'N_questionnaire_views'}, axis=1)
159 #seleksi sesuai kebutuhan
160 Nqv = gl10[['NIM','N_questionnaire_views']]
```

Gambar 4.18 Potongan kode ekstraksi catatan LMS (e)

Proses ekstraksi data LMS Moodle telah selesai dengan mendapatkan sebelas data, namun masih ada empat data yang perlu diekstrak, proses ekstraksi ini ditunjukkan pada gambar 4.19. Data tersebut adalah data jumlah mahasiswa dalam menyelesaikan quiz (N_questionnaires_submitted), frekuensi mahasiswa dalam merevisi kuesioner (N_reviews_questionnaire), hari pertama mengakses LMS (Days_first_access), dan jumlah total masuk kedalam LMS (N_entries course).

```
162 ##########################################################################
163 #N_questionnaires_submitted
164 colum13 = df3[df3['EventName'].str.contains('Quiz attempt submitted')]
165 #menghitung
166 gt11 = colum13.groupby(['NIM']).count().reset_index()
167 #merubah nama colom
168 gl11 = gt11.rename({'NAMA':'N_questionnaires_submitted'}, axis=1)
169 #seleksi sesuai kebutuhan
170 Nqs = gl11[['NIM','N_questionnaires_submitted']]
171
172 ##########################################################################
173 #N_reviews_questionnaire
174 colum14 = df3[df3['EventName'].str.contains('updated')]
175 #menghitung
176 gt12 = colum14.groupby(['NIM']).count().reset_index()
177 #merubah nama colom
178 gl12 = gt12.rename({'NAMA':'N_reviews_questionnaire'}, axis=1)
179 #seleksi sesuai kebutuhan
180 Nrq = gl12[['NIM','N_reviews_questionnaire']]
181
182 ##########################################################################
183 #Days_first_access
184 dayfi = colum1.rename({'Total_login':'Days_first_access'}, axis=1)
185 Dfa = dayfi[['NIM','Days_first_access']]
186
187 ##########################################################################
188 #N_entries_course
189 Neco = colum1.rename({'Total_login':'N_entries_course'}, axis=1)
190 Nec = Neco[['NIM','N_entries_course']]
191
```

Gambar 4.19 Potongan kode ekstraksi catatan LMS (f)

Proses ekstraksi untuk mendapatkan data N_questionnaires_submitted diperoleh dari variabel df3 kolom EventName yang mengandung kata ‘Quiz attempt submitted’ lalu disimpan kedalam variabel dengan nama Nqs, proses yang sama dilakukan untuk mendapatkan data frekuensi mahasiswa dalam

merevisi kuesioner yaitu didapatkan dari variabel df3 kolom EventName yang mengandung kata ‘update’ lalu disimpan kedalam variabel dengan nama Nrq. Sementara itu untuk data hari pertama mengakses LMS dan jumlah total masuk kedalam LMS diperoleh dari data login yang pada tahap selanjutnya akan diekstrak lagi untuk mengetahui kapan mahasiswa tersebut pertama kali mengakses LMS.

Ekstraksi kedalam data-data sesuai kebutuhan prediksi dari satu dokumen catatan telah dilakukan, proses ini merupakan proses ekstraksi dokumen catatan pada tanggal 2 juli, proses yang sama dilakukan untuk seluruh dokumen. Setelah melakukan proses ekstraksi awal dari dokumen ini langkah selanjutnya adalah menggabungkan variabel-variabel yang sudah didefinisikan tadi ke dalam satu dokumen.

Gambar 4.20 memperlihatkan potongan kode program dimana pada baris kode ke-194 sampai dengan baris ke-201 merupakan kode untuk menggabungkan seluruh variabel dengan memanfaatkan kata kunci NIM dan menggunakan fungsi outer sehingga data yang dihasilkan adalah data dengan satu NIM dan data dari semua variabel yang telah diekstrak.

```
192 ##############################################################################
193 #Merger dataframe Semua
194 gabnaqu = Tl.merge(Naf, on='NIM', how="outer").merge(Namf, on='NIM', how="outer") \
195           .merge(Nadu, on='NIM', how="outer").merge(Nag, on='NIM', how="outer") \
196           .merge(Tass, on='NIM', how="outer").merge(Nac, on='NIM', how="outer") \
197           .merge(Nasu, on='NIM', how="outer").merge(Naqu, on='NIM', how="outer") \
198           .merge(Natqu, on='NIM', how="outer").merge(Nanqu, on='NIM', how="outer") \
199           .merge(Nqv, on='NIM', how="outer").merge(Nqs, on='NIM', how="outer") \
200           .merge(Nrq, on='NIM', how="outer").merge(Dfa, on='NIM', how="outer") \
201           .merge(Nec, on='NIM', how="outer")
202
203 gabnaqu = gabnaqu.replace(np.nan, 0, regex=True)
204
205 gabnaqu.to_csv('2_juli.csv', index=False, encoding='utf8')
```

Gambar 4.20 Gabungan variabel ekstraksi data LMS

```
1  import pandas, sys
2  import pandas as pd
3
4
5  a = pd.read_csv('M1/21_februari.csv')
6  b = pd.read_csv('M1/22_februari.csv')
7  c = pd.read_csv('M1/23_februari.csv')
8  d = pd.read_csv('M1/24_februari.csv')
9  e = pd.read_csv('M1/25_februari.csv')
10 f = pd.read_csv('M1/26_februari.csv')
11 g = pd.read_csv('M1/27_februari.csv')
12
13
14 merged = pd.concat([a,b,c,d,e,f,g])
15
16 gab = merged.groupby('NIM').sum()
17
18 gab.to_csv('gabungan_m1.csv')
```

Gambar 4.21 Gabungan data harian catatan LMS

Proses yang dijelaskan sebelumnya telah menghasilkan data catatan kegiatan mahasiswa dalam sehari pada LMS Moodle, proses selanjutnya adalah menggabungkan data-data hasil ekstraksi tersebut kedalam data mingguan, adapun potongan kode penggabungan dari berbagai data harian ditunjukkan pada gambar 4.21, pada gambar tersebut baris pertama dan kedua merupakan proses mengambil pustaka pandas dan mengaliaskan agar mudah dalam melakukan pemanggilan di kode-kode selanjutnya, baris ke-5 sampai dengan baris ke-11 merupakan kode yang mengambil data dari data harian dan masing-masing data ditampung pada variabel a-g, baris ke-14 merupakan proses penggabungan dari variabel-variabel sebelumnya dan ditampung pada variabel merged, selanjutnya pada baris ke-16 merupakan variabel dengan nama gab yang berfungsi untuk menampung gabungan dari berbagai variabel sebelumnya dengan menjumlahkan isi dari masing-masing kolom berdasarkan NIM, sedangkan baris terakhir

merupakan perintah untuk menyimpan penggabungan dari berbagai data harian tersebut kedalam data mingguan.

```
import pandas, sys
import pandas as pd

m1= pd.read_csv('Gabungan Per-M/gabungan_m1.csv')
m2= pd.read_csv('Gabungan Per-M/gabungan_m2.csv')
m3= pd.read_csv('Gabungan Per-M/gabungan_m3.csv')
m4= pd.read_csv('Gabungan Per-M/gabungan_m4.csv')
m5= pd.read_csv('Gabungan Per-M/gabungan_m5.csv')
m6= pd.read_csv('Gabungan Per-M/gabungan_m6.csv')
m7= pd.read_csv('Gabungan Per-M/gabungan_m7.csv')
m8= pd.read_csv('Gabungan Per-M/gabungan_m8.csv')
m9= pd.read_csv('Gabungan Per-M/gabungan_m9.csv')
m10= pd.read_csv('Gabungan Per-M/gabungan_m10.csv')
m11= pd.read_csv('Gabungan Per-M/gabungan_m11.csv')
m12= pd.read_csv('Gabungan Per-M/gabungan_m12.csv')
m13= pd.read_csv('Gabungan Per-M/gabungan_m13.csv')
m14= pd.read_csv('Gabungan Per-M/gabungan_m14.csv')
m15= pd.read_csv('Gabungan Per-M/gabungan_m15.csv')
m16= pd.read_csv('Gabungan Per-M/gabungan_m16.csv')
m17= pd.read_csv('Gabungan Per-M/gabungan_m17.csv')
m18= pd.read_csv('Gabungan Per-M/gabungan_m18.csv')
m19= pd.read_csv('Gabungan Per-M/gabungan_m19.csv')

merged = pd.concat([m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,
                    m11,m12,m13,m14,m15,m16,m17,m18,m19])

gab = merged.groupby('NIM').sum()
gab.to_csv('gabunganall.csv')
```

Gambar 4.22 Gabungan data mingguan catatan LMS

Data catatan LMS Moodle sampai saat ini telah terkumpul kedalam data mingguan, selanjutnya data perlu dijadikan satu data yang mana data ini merupakan data hasil penggabungan dari data mingguan, kode program untuk menggabungkan catatan data mingguan ini ditunjukkan pada gambar 4.22, kode program tersebut tidak jauh beda dengan kode untuk menggabungkan data catatan harian kedalam catatan mingguan, gabungan data catatan mingguan ini disimpan dalam format csv (*comma separated values*) dengan nama gabunganall.csv

4.4.2. Ekstraksi data Organisasi

Sumber data organisasi berasal dari kumpulan dokumen salinan keputusan, dokumen ini berisi tentang catatan struktur organisasi yang terdiri dari nama-nama dan NIM dari pengurus serta anggota pengurusnya. Metode untuk mengekstraksi dokumen ini adalah dengan cara mencatat kembali NIM dan nama pengurus organisasi tersebut. Gambar 4.23 merupakan contoh salah satu salinan keputusan organisasi mahasiswa.

LAMPIRAN : SALINAN KEPUTUSAN REKTOR UNIVERSITAS MURIA KUDUS
NOMOR : 03/AK.UMK/Kep/D.17.02/I/2022
TENTANG : PENGANGKATAN PENGURUS UNIT KEGIATAN MAHASISWA FORUM ILMIAH MAHASISWA UNVERSITAS MURIA KUDUS PERIODE 2022

SUSUNAN KEPENGURUSAN
UNIT KEGIATAN MAHASISWA FORUM ILMIAH MAHASISWA
UNIVERSITAS MURIA KUDUS PERIODE 2022

Pelindung	: Rektor Universitas Muria Kudus	
Penasehat	: Wakil Rektor III Universitas Muria Kudus	
Pembina	: Muhammad Arifin, S.Kom. M. Kom	NIDN. 0621048301
Pembimbing	: Slamet Budi Cahyono, M.Sc	
Ketua Umum	: Dimas Apriadi	NIM. 201954027
Wakil Ketua	: Avivah Septyaningrum	NIM. 201911081
Sekretaris	: Nickyta Yudhisaputri	NIM. 202032091
	Rista Amanda Putri	NIM. 202011396
Bendahara	: Dyah Diana Pertiwi	NIM. 201911504
	Putri Ayuningsih	NIM. 202020054

DEPARTEMEN PENGEMBANGAN SUMBER DAYA MANUSIA

Koordinator	: Idha Rachmawati	NIM. 201912244
Anggota	:	
	1. Wahyu Utomo	NIM. 201941063
	2. Ira Nur Azizah	NIM. 201935025
	3. Silla Anggi Amalia	NIM. 202012144
	4. Joko Saputro	NIM. 201941004
	5. Musfirotul Khusna	NIM. 202051194

Gambar 4.23 Salinan keputusan organisasi mahasiswa

Hasil ekstraksi dokumen salinan keputusan organisasi mahasiswa ditunjukkan pada tabel 4.6, data tersebut berjumlah 1395 yang terdiri dari 50 orang berorganisasi di tingkat universitas, 305 orang berorganisasi tingkat fakultas, 369 orang

berorganisasi di tingkat prodi, sedangkan jumlah orang yang mengikuti organisasi UKM merupakan jumlah terbanyak yaitu 671 orang yang berasal dari 17 organisasi UKM.

Tabel 4.6 Data organisasi

No	Kriteria Organisasi	Nama Organisasi	Jumlah Anggota
1	Universitas	BEM Universitas	30
2		DPM	20
3	Fakultas	BEM Fakultas Ekonomi dan Bisnis (FEB)	47
4		BEM Fakultas Hukum (FH)	40
5		BEM Fakultas Teknik (FT)	31
6		BEM Fakultas Keguruan dan Ilmu Pendidikan (FKIP)	35
7		BEM Fakultas Pertanian (FP)	25
8		BEM Fakultas Psikologi (FPsi)	18
9		Obeng (Teater FT)	42
10		Teater Coin FEB	24
11		Teater Tiga Koma FKIP	28
12		SEC FEB	15
13	Prodi	Himapro Manajemen	40
14		Himapro Teknik Elektro	37
15		Himapro Akuntansi	50
16		Hima Pendidikan Bahasa Inggris	29
17		Hima Bimbingan Konseling	27
18		Hima Sistem Informasi	43
19		Hima Pendidikan Guru Sekolah Dasar	36
20		Hima Teknik Industri	32
21		Hima Matematika	23
22		Hima Teknik Industri	28
23		Hima Teknik Mesin	24
25	UKM	Mahasiswa Pecinta Alam	15
26		Forum Mahasiswa Islam	82
27		Resimen Mahasiswa (Menwa)	23
28		Dewan Racana	16
29		Forum Mahasiswa Ilmiah (FIMA)	27
30		Paduan Suara Pitaloka	48
31		Broadcasting	76
32		Korp Suka Rela	52
33		Persekutuan Mahasiswa Katolik	19
34		Debat	15

35		Jurnalistik	29
36		Koperasi Mahasiswa	22
37		Ningnong	91
38		Olah Raga	106
39		Seni Kampus	26
40		Unit Kegiatan Racana	10
Jumlah			1395

4.4.3. Ekstraksi data Akademik

Data yang berasal dari sistem informasi akademik ini terdiri dari enam kolom yaitu; NIM, nama, fakultas, prodi, IPS dan IPK. Ekstraksi data akademik ini dilakukan dengan cara sederhana yaitu dengan mengambil kolom NIM, nama dan IPS, selain mengambil kolom-kolom tersebut data diseleksi dengan memilih NIM yang memiliki empat awalan kata ‘2019’ dan ‘2020’. Proses pemilihan NIM yang memiliki awalan kata tersebut merupakan proses seleksi mahasiswa angkatan 2019 dan 2020, dimana empat karakter awal dari NIM menunjukkan tahun masuk mahasiswa tersebut. Seleksi data lain adalah hanya mengambil mahasiswa yang memiliki IPS 1 sampai dengan 4, dari hasil seleksi tersebut didapatkan data akademik sebanyak 4435.

4.4.4. Ekstraksi data Ekonomi

Proses ekstraksi data ekonomi dimulai dari pengambilan data akademik yang dihasilkan dari pendaftaran mahasiswa baru di kolom ‘Penghasilan Orang Tua’. Data ekonomi ini diseleksi berdasarkan NIM yang telah dihasilkan oleh ekstraksi data akademik. Selanjutnya data dikumpulkan berdasarkan kategori dan diberikan kode seperti yang ditunjukkan pada tabel 4.7.

Tabel 4.7 Data ekonomi

No	Kategori	Kode	Jumlah
1	100.000 - 600.000	1	394
2	500.000 - 1.000.000	2	868
3	1.000.000 - 2.500.000	3	1832

4	2.500.000 - 5.000.000	4	1087
5	5.000.000 - 7.500.000	5	179
6	7.500.000 - 10.000.000	6	50
7	> 10.000.000	7	25

Untuk mempersiapkan data agar nantinya dapat digabungkan dengan data-data lain maka perlu dilakukan pengkodean, kode program untuk melakukan pengkodean data ekonomi ditunjukkan pada potongan kode di gambar 4.24. Kode program tersebut diawali dengan memanggil pustaka pandas, pada kode program baris selanjutnya adalah membaca data, kemudian melakukan pengkodean dan di baris terakhir merupakan kode untuk merubah data ekonomi kedalam kode yang sesuai dengan kode pada tabel 4.7.

```
1 import pandas as pd
2 #Imputdata
3 df = pd.read_csv('Data_ekonomi.csv')
4
5 #encoding Economy
6 obj_dict = {
7           '100.000 - 600.000': 1,
8           '500.000 - 1.000.000': 2,
9           '1.000.000 - 2.500.000': 3,
10          '2.500.000 - 5.000.000': 4,
11          '5.000.000 - 7.500.000': 5,
12          '7.500.000 - 10.000.000': 6,
13          '> 10.000.000': 7,
14          }
15 df['Economy'] = df['Economy'].replace(obj_dict)
```

Gambar 4.24 Pengkodean data ekonomi

4.4.5. Ekstraksi data Demografi

Data demografi yang dipergunakan dalam penelitian ini adalah data jenis kelamin dan data kota tinggal, metode ekstraksi data ini sama dengan ekstraksi data ekonomi yaitu hanya mengumpulkan data yang NIM mahasiswanya berada pada data hasil ekstraksi data akademik. Hasil ekstraksi data demografi

untuk data jenis kelamin berjumlah 4.435 yang terdiri dari 2.621 berjenis kelamin perempuan dan 1814 berjenis kelamin laki-laki. Sedangkan data kota tinggal sejumlah 4.435 data yang terdiri dari kota 2138 dikota yang sama dengan lokasi kampus sedangkan lainya berada diluar kota kampus itu berada. Tahap selanjutnya adalah mengelompokkan kota-kota tersebut kedalam beberapa kelompok, kelompok 1 adalah kota yang sama dengan kampus berada, kelompok 2 merupakan kota yang bersebelahan dengan kota kampus, kelompok 3 adalah kota yang terletak berjarak satu kota dari kota kampus, kelompok 4 yaitu kota yang berada di wilayah pulau Jawa yang jarak dari kota kampus lebih dari dua atau lebih, sedangkan kelompok ke-5 adalah kota yang berada di luar pulau Jawa. Pada tabel 4.8 memperlihatkan bahwa mahasiswa paling banyak berasal dari kota yang sama dengan kampus, adapun kota yang berada di kelompok ke-2 berjumlah 1.932, kelompok ke-3 berjumlah 303, kelompok 4 berjumlah 37 dan kelompok ke-5 berjumlah 23 yang berasal dari 20 kota berbeda.

Tabel 4.8 Data demografi kota tinggal

No	Kota Tinggal	Kode	Jumlah
1	Kudus	1	2138
2	Grobogan	2	70
3	Jepara	2	788
4	Pati	2	943
5	Rembang	2	131
6	Blora	3	43
7	Demak	3	253
8	Semarang	3	7
9	Batang	4	2
10	Bekasi	4	1
11	Bogor	4	1
12	Brebes	4	2
13	Depok	4	1
14	Jakarta Barat	4	3
15	Jakarta Timur	4	1
16	Karanganyar	4	2
17	Karawang	4	1

18	Kedalon, Batangan	4	2
19	Kendal	4	3
20	Klaten	4	1
21	Magelang	4	1
22	Majalengka	4	1
23	Ngawi	4	1
24	Pemalang	4	1
25	Purbalingga	4	1
26	Purwokerto	4	1
27	Sukoharjo	4	1
28	Surakarta	4	1
29	Tangerang	4	5
30	Tegal	4	2
31	Tuban	4	2
32	Wonosobo	4	2
33	Balikpapan	5	1
34	Batam	5	1
35	Bengkulu	5	1
36	Denpasar	5	1
37	Dompu	5	1
38	Jambi	5	2
39	Kepulauan Selayar	5	1
40	Ketapang	5	1
41	Kotawaringin Barat	5	1
42	Kotawaringin Timur	5	1
43	Lampung Barat	5	1
44	Lampung Selatan	5	1
45	Makassar	5	1
46	Merangin	5	1
47	Mesuji	5	1
48	Ogan Komering Ilir	5	1
49	Palu	5	1
50	Pringsewu	5	1
51	Rokan Hulu	5	2
52	Sukamara	5	1
53	Tanjung Pinang	5	1
		Jumlah	4.435

Penggabungan data merupakan proses menggabungkan dari berbagai data hasil ekstraksi untuk kebutuhan prediksi. Data LMS, akademik, demografi dan data organisasi disatukan kedalam sebuah data. Proses penggabungan menggunakan kode program

komputer yang dipresentasikan pada gambar 4.25, metode penggabungan sama dengan metode untuk menggabungkan data LMS mingguan kedalam satu data LMS, kode dimulai dengan memanggil pustaka pandas dan dialiaskan, kemudian kode selanjutnya adalah kode untuk mengambil data LMS dengan nama gabunganall.csv dan data akademik dengan nama OlahDataPerformaKu.csv, adapun hasil penggabungan data LMS dan data akademik ini selanjutnya disimpan dan diberi nama Data_gabungan_nim.csv.

```
import pandas, sys
import pandas as pd


a = pd.read_csv('OlahDataPerformaKu.csv')
b = pd.read_csv('gabunganall.csv')


gabnaqu = a.merge(b, on='NIM', how="outer")


gabnaqu.to_csv('Data_gabungan_nim.csv', index=False,
               encoding='utf8')
```

Gambar 4.25 Kode gabungan data akademik dan LMS

```
import pandas, sys
import pandas as pd


a = pd.read_csv('Data_gabungan_nim.csv')
b = pd.read_csv('Nm1.csv')


gabnaqu = a.merge(b, on='NIM', how="outer")


gabnaqu.to_csv('OlahDT.csv', index=False, encoding='utf8')
```

Gambar 4.26 Kode gabungan data akademik, LMS dan organisasi

Langkah berikutnya adalah menggabungkan data gabungan LMS dan data akademik digabungkan dengan data organisasi, metode dan proses penggabungannya menggunakan cara yang sama dengan penggabungan data sebelumnya, hasil dari gabungan ketiga data ini selanjutnya disimpan dan diberi nama OlahDT.csv, kode penggabungan ditunjukkan pada potongan kode pada gambar 4.26.

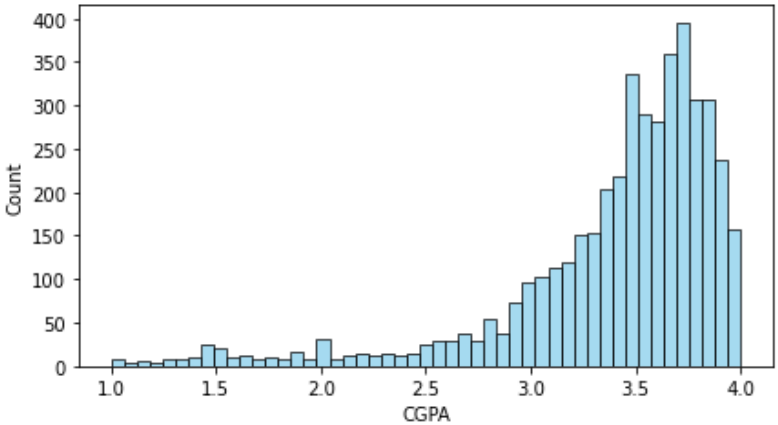
Data LMS, akademik dan data organisasi telah digabungkan, selanjutnya menggabungkan data tersebut dengan data ekonomi dan data demografi, metode dan kode program untuk menggabungkan ketiga data tersebut sama dengan cara sebelumnya yaitu dimulai dengan memasukkan pustaka pandas dilanjutkan dengan mengambil data yang akan digabungkan dan selanjutnya digabungkan lalu disimpan, hasil dari gabungan semua data tersebut disimpan dan diberi nama Data_prediksi_performa.csv, adapun data hasil gabungan dari semua data-data itu ditunjukkan pada gambar 4.27 dan selanjutnya data ini disebut data EDM.

No	NIM	Gender	Domicile	Economy	Campus_organization	Total_login	N_access_forum	N_access_didactic_units	Total_assignments	N_assignments_submitted	N_access_questionnaires	N_attempts_questionnaires	N_answered_questions	N_questionnaire_views	N_questionnaires_submitted	N_reviews_questionnaire	Days_first_access_x	N_entries_course_x	CGPA
1	2011233	0	2	1	0	97	715	216	228	26	142	122	14	136	3	284	1	444	3.43
2	1911631	1	1	3	0	63	2	24	122	7	1	0	3	1	0	278	1	146	2.62
3	2011216	0	2	3	0	121	1490	296	569	42	225	192	15	221	2	439	1	865	3.55
4	2011036	0	1	3	0	111	208	261	384	41	225	214	15	221	2	116	1	645	3.61
5	2011199	0	2	3	2	105	824	182	376	32	172	158	15	168	2	434	1	558	3.64
6	2011509	0	2	3	0	113	468	173	344	35	301	287	15	297	2	350	1	517	3.7
7	1911171	1	1	4	0	103	140	111	200	17	158	135	3	150	4	29	1	311	3.08
8	2060018	1	1	3	4	35	0	93	153	9	121	94	0	108	6	173	2	246	2.81
9	2011628	1	2	2	0	107	538	171	428	30	271	254	13	267	2	301	1	599	3.55
10	2011220	0	1	2	0	106	486	148	347	31	179	170	15	175	2	225	1	495	3.41
11	2011400	0	1	6	0	108	549	155	550	43	324	280	14	314	5	46	1	705	3.68
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4429	1933229	0	1	7	1	61	0	20	56	6	26	11	0	21	2	9	2	76	2.5
4430	2054079	1	2	6	0	27	1	23	18	2	56	29	0	48	4	117	1	41	1
4431	1957017	1	3	3	3	85	0	48	385	52	71	42	7	62	3	324	1	433	2.8
4432	1951008	0	2	6	0	103	2	7	521	53	112	107	9	108	2	83	1	528	4
4433	1951079	1	2	7	2	102	2	45	488	51	66	57	0	62	2	61	1	533	3.55
4434	2053164	1	2	6	0	105	0	84	580	43	784	633	0	746	19	266	1	664	3.87
4435	2033353	0	2	2	0	43	13	16	52	7	0	0	0	0	0	45	1	68	1.47

Gambar 4.27 Data EDM

4.5. Data Visualisasi

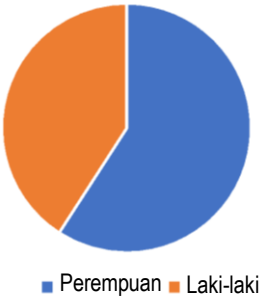
Data visualisasi merupakan tampilan berupa grafis atau visual dari informasi data, dengan kata lain visualisasi data adalah proses mewakili informasi atau data dalam bentuk grafis atau visual, sehingga memungkinkan pengguna untuk melihat pola, tren, dan hubungan antara data yang berbeda. Visualisasi data dapat membantu pengguna untuk memahami informasi yang rumit dan besar dengan cara yang lebih mudah dipahami, dan memungkinkan mereka untuk membuat keputusan yang lebih baik dan lebih cepat. Visualisasi data biasanya menggunakan berbagai jenis grafik, seperti diagram batang, grafik garis, diagram lingkaran, dan sebagainya. Setiap jenis grafik memiliki kegunaan dan keuntungan yang berbeda-beda tergantung pada jenis data dan tujuan visualisasi. Manfaat lain data visual ini dipercaya lebih cepat dan efektif karena orang lebih mudah mencerna informasi dalam bentuk visual.



Gambar 4.28 Visualisasi data IPS

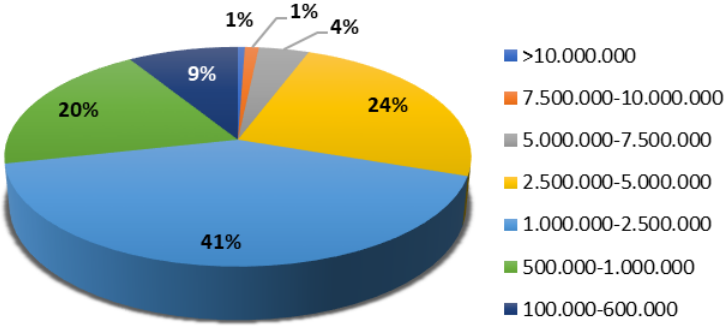
Data IPS yang berasal dari data akademik bernilai antara 1 hingga 4, dengan data paling banyak berada pada nilai 3 sampai dengan 4, data IPS ini divisualisasikan didalam gambar 4.28,

sementara itu pada Gambar 4.29 menunjukkan proporsi mahasiswa laki-laki dan perempuan, di mana jumlah mahasiswa perempuan lebih banyak daripada mahasiswa laki-laki.



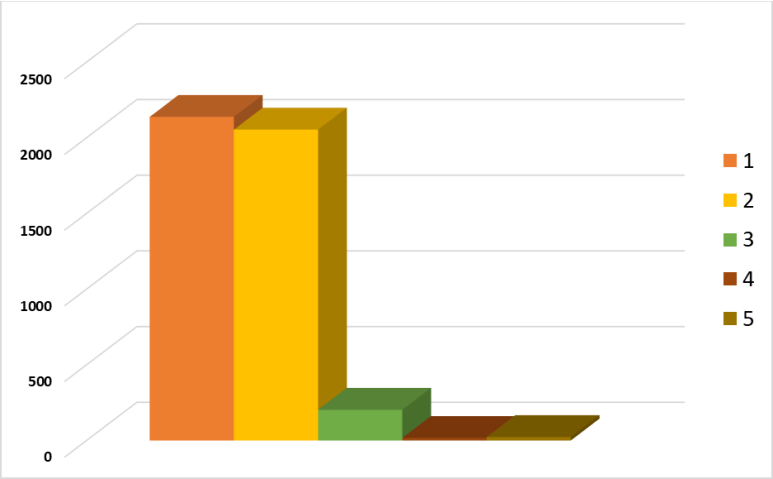
Gambar 4.29 Visualisasi Data Jenis Kelamin

Gambar 4.30 menunjukkan bahwa orang tua mahasiswa yang memiliki pendapatan antara 1.000.000–2.500.000 merupakan jumlah mayoritas yaitu 1.832 orang tua dari total 4.435, orang tua wali mahasiswa dengan pendapatan 2.500.000-5.000.000 berada di urutan kedua dengan prosentase 24%, sementara jumlah terbanyak ketiga dari kategori pendapatan orang tua wali mahasiswa ini adalah 500.000-1.000.000 atau 20% dari total keseluruhan dari kategori pendapatan wali mahasiswa. Data dari pendapatan ini tidak seimbang di seluruh kategori kelompok ekonomi.

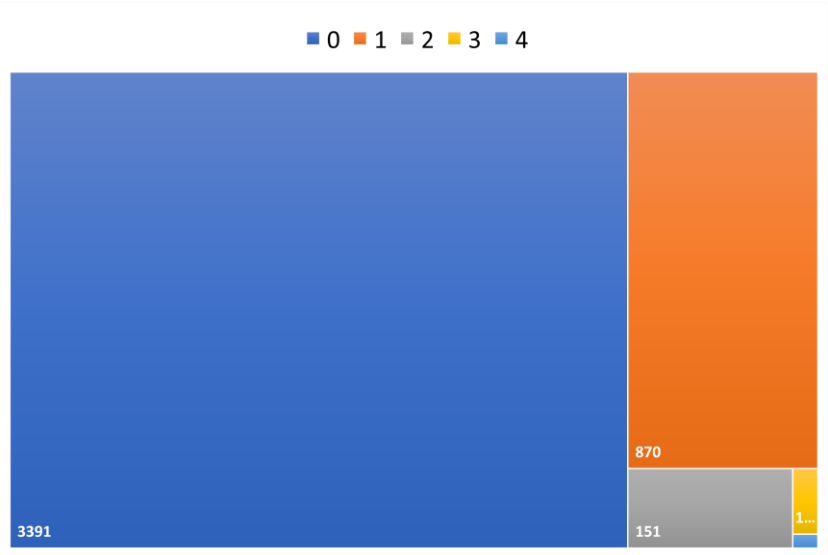


Gambar 4.30 Visualisasi Data Ekonomi

Fitur domisili pada Gambar 4.31 menunjukkan bahwa siswa yang berasal dari kota tinggal di ring 1 dan 2 merupakan yang tertinggi dibandingkan ring lainnya.

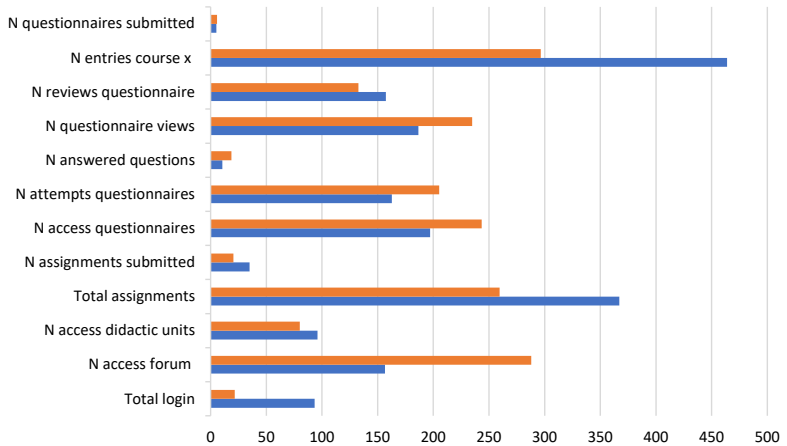


Gambar 4.31 Visualisasi Data Kota Tinggal Mahasiswa



Gambar 4.32 Visualisasi Data Mahasiswa Berorganisasi

Gambar 4.32 menunjukkan fitur organisasi mahasiswa dimana mahasiswa yang bukan anggota organisasi kampus lebih dari 77% dari jumlah keseluruhan, 20% hanya mengikuti satu organisasi, dan 3% mengikuti dua organisasi.



	Total login	N access forum	N access didactic units	Total assignments	N assignments submitted	N access questionnaires	N attempts questionnaires	N answered questions	N questionnaire views	N reviews questionnaire	N entries course x	N questionnaires submitted
Stdv	21.6	287.9	80.1	259.6	20.5	243.5	205.2	18.7	234.9	132.8	296.5	5.7
Avg	93.4	156.6	95.95	367	34.9	197.2	162.7	10.4	186.6	157.4	463.9	5.1

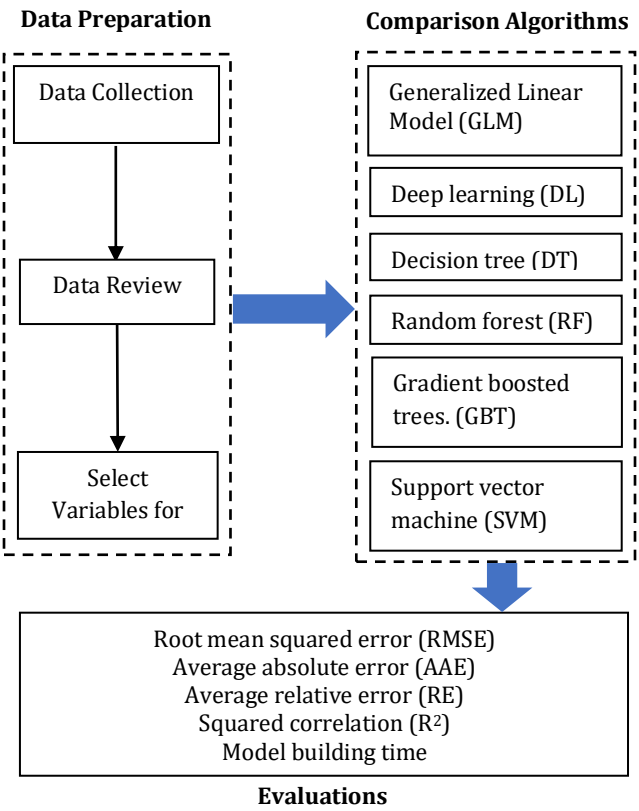
Gambar 4.33 Visualisasi Data LMS

Visualisasi data log LMS pada Gambar 4.33 menunjukkan bahwa variabel N_entries_course memiliki angka paling signifikan, diikuti oleh variabel Total_assignments. Secara bersamaan, variabel n_questionnaires_submitted memiliki jumlah terkecil.

4.6. Proses implementasi prediksi menggunakan model terpilih

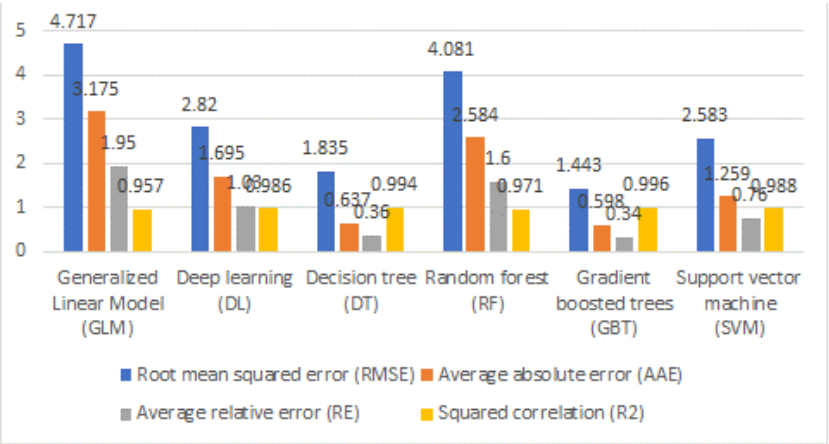
Sebelum melakukan prediksi maka langkah awal adalah menentukan teknik data mining yang digunakan, menentukan

tools data mining, algoritma data mining, menentukan parameter dengan nilai yang optimal. Bagian penentuan tools untuk kebutuhan data mining telah disampaikan pada awal bab IV khususnya pada bagian perangkat lunak yang digunakan. Sedangkan untuk menentukan algoritma yang dibutuhkan dalam penelitian ini, langkah awal yang dilakukan adalah melakukan studi pustaka. Hasil dari studi pustaka tersebut menunjukkan bahwa ada berbagai metode yang tersedia, namun tidak ada satu metode yang dianggap terbaik secara mutlak.



Gambar 4.34 Perbandingan algoritma prediksi

Pemilihan metode tergantung pada faktor-faktor seperti jumlah data, jenis data, dan proses pra-pemrosesan yang digunakan. Berdasarkan hal tersebut untuk menentukan metode yang akan digunakan penulis telah melakukan percobaan dengan membandingkan beberapa metode untuk memprediksi performa akademik mahasiswa. Data yang digunakan merupakan data yang terdiri dari data keuangan, data sosial, dan data akademik dari 12.411 siswa yang diperoleh dari *Mendeley Data Repository* (https://data.mendeley.com/datasets/83tcx8psxv/1) yang bersifat publik yaitu data yang dapat diakses oleh semua orang. Penulis telah membandingkan enam algoritma yang sering digunakan dalam memprediksi performa akademik mahasiswa yaitu algoritma *Generalized Linear Model* (GLM), *Deep learning* (DL), *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosted Trees* (GBT), *Support Vector Machine* (SVM) ditunjukkan pada gambar 4.34.



Gambar 4.35 Hasil perbandingan algoritma mesin pembelajaran

Hasil percobaan menunjukkan bahwa GBT adalah algoritma yang paling efektif dalam memprediksi kinerja siswa, dengan nilai RMSE paling rendah. Hasil studi ini menemukan bahwa GBT

merupakan algoritma yang paling efektif dalam memprediksi kinerja mahasiswa, dengan nilai RMSE sebesar 1,443, yang merupakan algoritma paling minor tingkat kesalahannya dibandingkan algoritma lainnya hal ini ditunjukkan pada gambar 4.35. Semakin kecil nilai RMSE, semakin baik algoritma dalam prediksi.

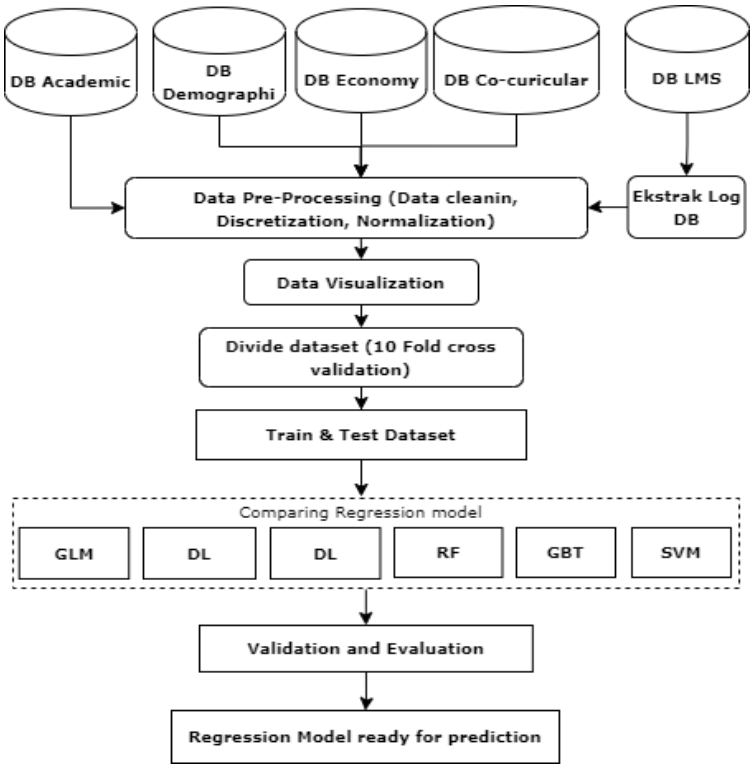
Selain membandingkan tingkat kesalahan dalam membuat prediksi, penulis juga membandingkan waktu pembangunan model dari algoritma yang diusulkan. Hasil penelitian menunjukkan bahwa GLM adalah algoritma tercepat dengan waktu 1,609 s dalam membuat prediksi tetapi memiliki tingkat kesalahan tertinggi. Sementara itu algoritma GBT, yang merupakan algoritma terbaik dengan kesalahan terkecil, memiliki waktu eksekusi 100,57 s atau nomor dua setelah SVM. Detail perbandingan ditunjukkan pada tabel 4.9.

Tabel 4.9 Perbandingan waktu pembangunan model

Algorithm	**Model building time**
Generalized Linear Model (GLM)	1.609 s
Deep learning (DL)	19.199 s
Decision tree (DT)	2.062 s
Random forest (RF)	47.222 s
Gradient boosted trees (GBT)	100.57 s
Support vector machine (SVM)	812.768 s

Pada percobaan pertama data yang digunakan adalah data publik yang bersumber dari Mendelay, untuk memastikan bahwa algoritma GBT ini memang merupakan algoritma yang paling tepat untuk prediksi performa akademik mahasiswa maka penulis melakukan percobaan ulang dengan menggunakan data yang dikumpulkan untuk penelitian ini yaitu fitur akademik dan non-akademik, adapun desain percobaannya ditunjukkan pada gambar 4.36. Proses pengumpulan dan pengolahan data serta pencarian

algoritma terbaik telah dipublikasikan di jurnal internasional bereputasi.



Gambar 4.36 Perbandingan algoritma regresi

Selain membandingkan beberapa algoritma untuk menemukan model terbaik dalam memprediksi performa akademik mahasiswa, pada artikel ini juga mengusulkan prediksi performa akademik dengan menggunakan model regresi. Umumnya para peneliti menggunakan model klasifikasi untuk memprediksi fitur target yaitu IPK/IPS, mereka mengkategorikannya kedalam kelas-kelas yang berbeda seperti baik dan buruk, lulus dan gagal, sangat baik, baik, rata-rata, dan buruk menggunakan pendekatan yang disebut diskritisasi atau

binning. Namun, pendekatan tersebut dapat menyebabkan hilangnya informasi penting dari data. Untuk menghindari hilangnya informasi tersebut penulis mengusulkan menggunakan model regresi untuk memprediksi performa akademik mahasiswa. berdasarkan hasil percobaan ditemukan bahwa GBT tetap paling akurat jika dibandingkan dengan algoritma lainnya, hal ini ditunjukkan pada tabel 4.10.

Tabel 4.10 Hasil perbandingan akurasi model regresi

Model	**RMSE**	**AE**	**RE**	**SE**	**Corr**
Generalized Linear Model	0.446	0.32	0.09	0.20	0.50
Deep Learning	0.413	0.29	0.09	0.17	0.61
Decision Tree	0.462	0.32	0.09	0.21	0.51
Random Forest	0.408	0.30	0.09	0.17	0.63
Gradient Boosted Trees	**0.379**	**0.26**	**0.08**	**0.14**	**0.68**
Support Vector Machine	0.434	0.29	0.08	0.19	0.56

Hasil percobaan menunjukkan bahwa model regresi GBT memiliki nilai RMSE terkecil sebesar 0,374%, AE sebesar 0,258%, dan SE sebesar 0,140%. Berdasarkan hasil ini, model GBT dapat dianggap sebagai model regresi yang paling tepat untuk memprediksi performa akademik menggunakan fitur akademik dan non-akademik.

Proses pencarian model terbaik telah ditemukan yaitu GBT, selanjutnya adalah proses menemukan parameter yang tepat untuk meningkatkan akurasi prediksi. Adapun dalam upaya meningkatkan akurasi prediksi pada penelitian ini mengusulkan metode penyetelan hyperparameter dari algoritma GBT. Penyetelan hyperparameter algoritma GBT ini menggunakan metode Grid Search, adapun upaya ini telah dipublikasikan pada jurnal internasional bereputasi. Dalam percobaan meningkatkan akurasi prediksi algoritma GBT, penulis menggunakan data akademik LMS, hal ini dilakukan agar antara artikel ilmiah ini

dengan artikel ilmiah sebelumnya tidak menggunakan data yang sama.

Proses penemuan hyperparameter terbaik untuk model GBT diawali dengan memahami kode GBT dan parameter apa saja yang ada didalamnya, GBT merupakan model ansambel adapun parameter model ansambel dibagi menjadi tiga kategori: pertama adalah parameter khusus pohon yang menentukan *min samples split, min samples leaf, min weight fraction leaf, max depth, max leaf nodes,* and *max features.* Kedua adalah parameter boosting untuk mengatur boosting yaitu *learning_rate, n_estimator, dan subsamples*. Sedangkan yang ketiga adalah parameter lainnya. Berbagai parameter mempengaruhi fungsionalitas secara keseluruhan: *loss, init, random state, verbose, warm start, dan presort*. Pada penelitian ini akan dilakukan tuning pada empat parameter yaitu *n_estimators, learning_rate, max_depth,* dan *subsample* yang merupakan parameter paling sering digunakan oleh para peneliti untuk tuning GBT. Berikut adalah langkah-langkah umum dan cara kerja algoritma GBT:

1. Inisialisasi model: Mengatur parameter-parameter yang diperlukan untuk GBT, seperti jumlah pohon (*n_estimators*), tingkat pembelajaran (*learning_rate*), kedalaman maksimum (*max_depth*), dll.
2. Menghitung prediksi awal: Inisialisasi prediksi awal dengan nilai rata-rata target atau dengan nilai lainnya.
3. Iterasi pembelajaran:
   - ✓ Hitung residual: Hitung selisih antara nilai target aktual dan prediksi saat ini. Residual ini akan menjadi target untuk model berikutnya.
   - ✓ Bangun pohon keputusan: Buat pohon keputusan berdasarkan residual yang dihitung sebelumnya. Pada setiap iterasi, pohon keputusan baru akan dilatih untuk memprediksi residual.

- ✓ Skala kontribusi pohon: Skala kontribusi pohon keputusan dengan mengalikan nilai pembelajaran (learning_rate). Hal ini memungkinkan kontrol terhadap kontribusi masing-masing pohon dalam model akhir.
- ✓ Update prediksi: Tambahkan kontribusi pohon keputusan terbaru ke prediksi sebelumnya.
- ✓ Iterasi selanjutnya: Ulangi langkah-langkah di atas dengan residual yang diperbarui sebagai target hingga mencapai jumlah iterasi (n_estimators) yang ditentukan.

4. Model akhir: Prediksi akhir dihasilkan dengan menggabungkan kontribusi semua pohon keputusan.
5. Evaluasi model: Evaluasi performa model menggunakan matrik evaluasi yang sesuai untuk masalah regresi, seperti MAE, MSE, atau RMSE
6. Penyetelan parameter: Jika diperlukan, lakukan penyetelan parameter untuk meningkatkan performa model dengan menggunakan teknik seperti validasi silang.
7. Penggunaan model: Gunakan model yang telah dilatih untuk melakukan prediksi pada data baru.

GBT secara iteratif memperbaiki prediksi dengan mengarahkan pohon keputusan berikutnya untuk fokus pada residu yang masih ada. Proses ini menghasilkan *ansambel* pohon yang saling memperbaiki, menghasilkan prediksi akhir yang lebih baik dari prediksi awal.

GBT memiliki parameter *n_estimators* yang digunakan untuk mengatur jumlah pohon keputusan yang akan dibangun dalam *ensemble*. *n_estimators* menentukan jumlah iterasi atau jumlah pohon yang akan digunakan dalam proses pelatihan. Semakin besar nilai *n_estimators*, semakin kompleks model dan semakin banyak iterasi yang dilakukan untuk meminimalkan kesalahan residu. Pengguna dapat menyesuaikan nilai

n_estimators sesuai kebutuhan. Peningkatan jumlah pohon keputusan (*n_estimators*) biasanya akan meningkatkan kinerja model, tetapi dengan biaya peningkatan waktu komputasi. Penting untuk menemukan keseimbangan antara kompleksitas model dan efisiensi komputasi. Kita juga dapat menggunakan teknik penyetelan parameter seperti validasi silang untuk menentukan nilai *n_estimators* yang optimal untuk dataset Anda.

Parameter *learning_rate* digunakan untuk mengatur tingkat pembelajaran atau kontribusi setiap pohon terhadap model akhir. *learning_rate* adalah faktor yang mengontrol seberapa besar penyesuaian yang dilakukan pada setiap iterasi untuk memperbarui prediksi model. Nilai *learning_rate* yang lebih kecil menghasilkan kontribusi yang lebih kecil dari setiap pohon, membuat model lebih konservatif dan mengurangi risiko overfitting. Di sisi lain, nilai *learning_rate* yang lebih tinggi dapat menghasilkan kontribusi yang lebih besar dari setiap pohon, membuat model lebih kompleks dan mungkin memiliki risiko overfitting yang lebih tinggi. Kita dapat menyesuaikan nilai *learning_rate* sesuai dengan preferensi dan karakteristik dataset Anda. Biasanya, mengurangi *learning_rate* dan meningkatkan jumlah *n_estimators* (jumlah pohon) secara bersamaan dapat memberikan hasil yang lebih baik. Namun, perlu diingat bahwa memilih *learning_rate* yang terlalu kecil dapat mengakibatkan pelatihan yang lambat atau konvergensi yang buruk, sementara *learning_rate* yang terlalu tinggi dapat menyebabkan model tidak stabil atau *overfitting*. Oleh karena itu, perlu dilakukan penyetelan parameter yang hati-hati untuk menemukan nilai *learning_rate* yang optimal.

Parameter lain yaitu *max_depth* yang digunakan untuk mengatur kedalaman maksimum dari setiap pohon keputusan dalam *ensemble*. *max_depth* mengontrol sejauh mana setiap pohon dapat tumbuh dalam struktur hirarkis. Kedalaman

maksimum yang lebih tinggi memungkinkan pohon untuk mempelajari hubungan yang lebih kompleks dalam data, tetapi juga dapat menyebabkan *overfitting* jika tidak diatur dengan benar. Kita dapat menyesuaikan nilai *max_depth* sesuai dengan karakteristik dan kompleksitas dataset. Peningkatan nilai *max_depth* bisa memungkinkan model untuk mempelajari pola yang lebih rumit, tetapi perlu diingat bahwa nilai *max_depth* yang terlalu tinggi dapat menyebabkan *overfitting*, terutama jika ukuran dataset relatif kecil. Penting untuk menemukan keseimbangan antara kompleksitas model dan generalisasi yang baik terhadap data baru.

```
1 from sklearn.ensemble import GradientBoostingRegressor
2 gbr = GradientBoostingRegressor()
3 print(gbr)
4 GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse',
               init=None, learning_rate=0.1, loss='ls',
               max_depth=3, max_features=None,
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=100, presort='auto', random_state=None,
            subsample=1.0, verbose=0, warm_start=False)
```

Gambar 4.37 Nilai bawaan parameter GBT

Parameter *subsample* dapat digunakan untuk mengatur ukuran *subsample* yang digunakan dalam pelatihan setiap pohon keputusan. *Subsample* mengacu pada fraksi dari total sampel yang digunakan untuk melatih setiap pohon. Secara default, nilai subsample diatur sebagai 1.0, yang berarti menggunakan seluruh dataset pelatihan untuk melatih setiap pohon. Namun, kita dapat menyesuaikan nilai subsample sesuai kebutuhan. Misalnya, jika kita mengatur subsample sebagai 0.8, maka setiap pohon akan dilatih dengan menggunakan 80% data acak dari dataset pelatihan. Ini dapat membantu mengurangi varians dan memperkenalkan sedikit regularisasi pada model. Secara default nilai-nilai parameter pada algoritma GBT ditunjukkan pada potongan kode

pada gambar 4.37, dimana nilai *learning_rate*=0.1, *max_depth*=3, *n_estimators*=100 dan *subsample*=1. Sedangkan untuk usulan parameter yang diusulkan ditunjukkan pada tabel 4.11.

Tabel 4.11 Parameters for GridSearchCV

Parameters	**Values/Ranges**	
estimator	*GradientBoostingRegressor*	
scoring	*neg_mean_squared_error*	
n_jobs	-1	
cv	5	
param_grid	*n_estimators*	[50, 100, 200, 300]
	Subsample	[0.9, 0.5, 0.2, 0.1]
	learning_rate	[0.1, 1, 1.5, 2]
	max_depth	[2, 8, 12, 16]

Imput Data

```
1  import pandas as pd
2  #Imputdata
3  df = pd.read_csv('Data_prediksi_performa.csv')
4
5  X = df.iloc[:, 6:19]
6  y = df.iloc[:, 19]
7
8  from sklearn.preprocessing import StandardScaler
9  sc = StandardScaler()
10 X_sc = sc.fit_transform(X)
11
```

Gambar 4.38 Pengambilan data prediksi performa akademik

Metode pencarian parameter terbaik dari GBT ini menggunakan bahasa pemrograman Python. Adapun potongan kode program ditunjukkan pada gambar 4.38. Kode baris pertama dari gambar tersebut merupakan kode untuk memanggil pustaka dari pandas, selanjutnya untuk baris ketiga adalah kode untuk mengambil data yang dibutuhkan. Selanjutnya pada baris kelima dan keenam merupakan kode untuk membuat variabel X dan y dari dataset dengan mengambil kolom 6 sampai dengan kolom 18 untuk variabel X dan kolom 19 untuk variabel y yang merupakan variabel variabel target. Adapun kode untuk melakukan

penskalaan data, dimana StandardScaler bertujuan untuk mengubah distribusi data sehingga memiliki mean (rata-rata) 0 dan standar deviasi (simpangan baku) 1. Fungsi StandardScaler umumnya digunakan dalam pra-pemrosesan data sebelum melakukan pemodelan atau analisis lebih lanjut.

```
1  from sklearn.ensemble import GradientBoostingRegressor
2  from sklearn.model_selection import train_test_split
3  from sklearn.model_selection import GridSearchCV
4  from sklearn.metrics import mean_absolute_error
5
6  X_train, X_test, y_train, y_test=train_test_split(X_sc, y, random_state=12, test_size=0.2)
7
8  model = GradientBoostingRegressor()
9
10 params = {'learning_rate' : [0.01,0.02,0.03,0.04],
11              'subsample'    : [0.9, 0.5, 0.2, 0.1],
12              'n_estimators':[50,100,200,300],
13              'max_depth'  : [2,8,12,16]
14             }
15
16 grid = GridSearchCV(estimator=model, param_grid = params, scoring='neg_mean_absolute_error',
17                     cv=5, n_jobs=-1)
18
19 grid.fit(X_train, y_train)
20
21 ypred = grid.predict(X_test)
22
23 print(" Results from Grid Search " )
24 print("\n The Best MSE across all searched param:\n", grid.best_score_)
25 print("\n The Best parameters across all searched param:\n", grid.best_params_)
```

Gambar 4.39 Pencarian parameter dengan GridSearchCV

Gambar 4.39 merupakan kode program untuk menemukan parameter terbaik dengan menggunakan metode GridSearchCV. Diawali pada baris pertama sampai dengan kode baris keempat yang merupakan kode untuk memanggil pustaka yang dibutuhkan yaitu GradientBoostingRegressor, train_test_split, GridSearchCV, dan mean_absolute_error. Selanjutnya pada kode baris keenam merupakan kode untuk membagi data kedalam data pelatihan dan data pengujian dengan rasio 80:20. Proses pendefinisian X_train sebagai data X_sc yang ingin kita latih (train), X_test sebagai data X_sc yang ingin kita tes, y_train adalah variabel dependen yang dilatih, dan y_test merupakan variabel dependen yang diuji. Kita mendefinisikan keempat variabel ini secara bersamaan. Pada perintah train_test_split diisi

dengan parameter arrays X_sc, y yang menunjukkan arrays yang dipakai adalah X_sc sebagai variabel independen dan y sebagai variabel dependen. Parameter random_state=12 digunakan untuk memastikan bahwa pemilihan acak yang digunakan dalam pembagian dataset akan tetap konsisten setiap kali program dijalankan dengan random_state yang sama. Parameter random_state pada fungsi train_test_split dalam pustaka scikit-learn (sklearn) di Python digunakan untuk mengontrol pemilihan data secara acak saat membagi dataset menjadi subset pelatihan dan subset pengujian. Secara umum, saat membagi dataset menjadi subset pelatihan dan pengujian, yang diinginkan adalah hasil pembagian tersebut dapat direproduksi. Hal ini penting agar pemodelan dan evaluasi dapat dilakukan dengan konsistensi. Dalam kasus ini, kita dapat menggunakan parameter random_state untuk menentukan angka sembarang yang digunakan sebagai seed (benih) dalam pemilihan acak. Dengan menggunakan seed yang sama, kita dapat memastikan bahwa pemilihan acak yang dilakukan akan konsisten di setiap eksekusi program. Perlu diketahui bahwa nilai sebenarnya yang digunakan untuk random_state tidak penting, asalkan nilai tersebut sama saat menjalankan program yang sama.

Kode kedelapan merupakan variabel yang digunakan untuk memudahkan dalam pemanggilan model GradientBoostingRegressor. Sedangkan baris kesepuluh sampai dengan baris keempat belas merupakan variabel yang digunakan untuk menampung nilai-nilai dari parameter yang diusulkan untuk dilakukan percobaan. Kode baris keenam belas dan tujuh belas merupakan kode pencarian parameter dengan menggunakan pustaka GridSearchCV, pada kode tersebut diawali dengan membuat variabel yang diberi nama grid, selanjutnya GridSearchCV digunakan untuk mencari kombinasi terbaik dari parameter param untuk model GradientBoostingRegressor.

Parameter cv=5 menunjukkan bahwa validasi yang digunakan adalah 5 x validasi silang, parameter scoring= neg_mean_absolute_error merupakan parameter untuk memilih evaluasi yang digunakan pada saat proses pencarian parameter terbaik dari model GBT dalam hal ini evaluasi yang dipilih adalah MAE. Sedangkan parameter n_job= -1 digunakan pada GridSearchCV untuk memanfaatkan semua core CPU yang tersedia untuk pemrosesan paralel. Hal ini dapat mempercepat proses pelatihan model atau pencarian grid pada hiperparameter. Jika n_jobs diatur sebagai nilai lainnya, seperti n_jobs=1, maka pemrosesan paralel tidak akan dilakukan. Nilai yang dapat diberikan pada parameter n_jobs adalah sebagai berikut; Jika n_jobs bernilai -1 (default), maka semua core CPU yang tersedia akan digunakan untuk mengeksekusi pekerjaan secara paralel. Jumlah core CPU yang digunakan akan tergantung pada sistem yang digunakan. Jika n_jobs bernilai 1, maka hanya satu core CPU yang digunakan untuk mengeksekusi pekerjaan, sehingga tidak ada pemrosesan paralel yang dilakukan. Jika n_jobs bernilai k, di mana k adalah bilangan bulat positif, maka sebanyak k core CPU akan digunakan untuk mengeksekusi pekerjaan secara paralel. Namun, jumlah core CPU yang dapat digunakan tergantung pada sistem yang digunakan. Kode pada baris kesembilan belas merupakan kode untuk melatih (mengajarkan) model pada dataset pelatihan. Dalam konteks pembelajaran mesin (machine learning), metode fit() digunakan untuk menyesuaikan model dengan data yang diberikan agar dapat belajar pola atau hubungan yang ada dalam data. Metode fit(X_train, y_train) digunakan untuk melatih model pada dataset pelatihan yang terdiri dari fitur-fitur (X_train) dan label-label (y_train). Metode ini akan menghasilkan model yang telah belajar dari data pelatihan dan siap digunakan untuk melakukan prediksi pada data baru.

Pada baris kedua puluh satu digunakan untuk melakukan prediksi dengan menggunakan model regresi (GBT) yang telah dilatih pada data pelatihan. Dalam konteks regresi, metode predict() digunakan untuk menghasilkan nilai prediksi berdasarkan fitur-fitur (X_test) dari dataset pengujian. Dalam pustaka scikit-learn (sklearn) di Python, metode predict(X_test) umumnya digunakan setelah model regresi dilatih dengan menggunakan metode fit() pada data pelatihan. Metode predict() akan menghasilkan prediksi berdasarkan fitur-fitur pada dataset pengujian (X_test) dengan menggunakan model yang telah dilatih sebelumnya. Metode predict() dipanggil dengan memasukkan X_test sebagai argumen untuk menghasilkan prediksi nilai target yang sesuai. Hasil dari predict(X_test) adalah sebuah array atau struktur data yang berisi nilai prediksi yang dihasilkan oleh model regresi untuk setiap sampel pada dataset pengujian. Nilai prediksi ini mencerminkan estimasi atau perkiraan nilai target berdasarkan fitur-fitur yang diberikan.

Ketiga baris terakhir dari gambar 4.39 merupakan kode untuk menampilkan skor terbaik dari prediksi menggunakan model GBT dan parameter terbaik dari kombinasi parameter yang telah diusulkan. Setelah kode dalam blok itu dijalankan maka model GBT akan diuji sebanyak 1280 yang berasal dari 256 kandidat dikalikan 5 kali validasi. 25 kandidat tersebut berasal dari 4 nilai n_estimator, 4 nilai subsample, 4 nilai learning_rate dan 4 nilai max_depth sehingga 4 x 4 x 4 x 4= 256. Secara umum cara pengujian ditunjukkan pada blok kode pada gambar 4.40.

Pada blok kode diatas ditunjukkan bahwa untuk pencarian parameter terbaik menggunakan metode GridSearchCV, kemungkinan dari semua kombinasi parameter yang kita usulkan akan dicoba satu-persatu dan divalidasi 5 kali validasi. Hasil dari blok kode program tersebut setelah dijalankan adalah sebagai berikut:

```
#> [CV 1/5] END learning_rate=0.01, subsample=0.9,
n_estimators=50, max_depth= 2;, score=0.993 total
time=   0.0s
#> [CV 2/5] END learning_rate=0.01, subsample=0.9,
n_estimators=50, max_depth= 2;, score=0.999 total
time=   0.0s
#> [CV 3/5] END learning_rate=0.01, subsample=0.9,
n_estimators=50, max_depth= 2;, score=1.000 total
time=   0.0s
#> [CV 4/5] END learning_rate=0.01, subsample=0.9,
n_estimators=50, max_depth= 2;, score=0.993 total
time=   0.0s
#> [CV 5/5] END learning_rate=0.01, subsample=0.9,
n_estimators=50, max_depth= 2;, score=0.994 total
time=   0.0s
#> [CV 1/5] END learning_rate=0.01, subsample=0.5,
n_estimators=100, max_depth= 8;, score=0.973 total
time=   0.0s
#> [CV 2/5] END learning_rate=0.01, subsample=0.5,
n_estimators=100, max_depth= 8;, score=1.893 total
time=   0.0s
#> [CV 3/5] END learning_rate=0.01, subsample=0.5,
n_estimators=100, max_depth= 8;, score=0.903 total
time=   0.0s
#> [CV 4/5] END learning_rate=0.01, subsample=0.5,
n_estimators=100, max_depth= 8;, score=0.972 total
time=   0.0s
#> [CV 5/5] END learning_rate=0.01, subsample=0.5,
n_estimators=100, max_depth= 8;, score=0.963 total
time=   0.0s
#> [CV 1/5] END learning_rate=0.01, subsample=0.2,
n_estimators=200, max_depth= 12;, score=0.973 total
time=   0.0s
#> [CV 2/5] END learning_rate=0.01, subsample=0.2,
n_estimators=200, max_depth= 12;, score=1.893 total
time=   0.0s
#> [CV 3/5] END learning_rate=0.01, subsample=0.2,
n_estimators=200, max_depth= 12;, score=0.903 total
time=   0.0s
#> [CV 4/5] END learning_rate=0.01, subsample=0.2,
n_estimators=200, max_depth= 12;, score=0.972 total
time=   0.0s
```

```
#> [CV 5/5] END learning_rate=0.01, subsample=0.2,
n_estimators=200, max_depth= 12;, score=0.963 total
time=   0.0s
#> [CV 1/5] END learning_rate=0.01, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.973 total
time=   0.0s
#> [CV 2/5] END learning_rate=0.01, subsample=0.1,
n_estimators=300, max_depth= 16;, score=1.893 total
time=   0.0s
#> [CV 3/5] END learning_rate=0.01, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.903 total
time=   0.0s
#> [CV 4/5] END learning_rate=0.01, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.972 total
time=   0.0s
#> [CV 5/5] END learning_rate=0.01, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.822 total
time=   0.0s
.
.
.
#> [CV 1/5] END learning_rate=0.04, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.973 total
time=   0.0s
#> [CV 2/5] END learning_rate=0.04, subsample=0.1,
n_estimators=300, max_depth= 16;, score=1.893 total
time=   0.0s
#> [CV 3/5] END learning_rate=0.04, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.903 total
time=   0.0s
#> [CV 4/5] END learning_rate=0.04, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.972 total
time=   0.0s
#> [CV 5/5] END learning_rate=0.04, subsample=0.1,
n_estimators=300, max_depth= 16;, score=0.822 total
time=   0.0s
```

Gambar 4.40 Proses pengujian validasi

Hasil percobaan gambar 4.41 menunjukkan bahwa nilai MAE terbaik dari perpaduan kombinasi parameter tersebut adalah -0, 2491447879620859 sedangkan untuk parameter terbaiknya adalah 'learning_rate': 0.02, 'max_depth': 16, 'n_estimators': 300, dan 'subsample': 0.5.

```
Results from Grid Search

The Best MAE across all searched param:
 -0.2491447879620859

 The Best parameters across all searched param:
 {'learning_rate':     0.02,     'max_depth':     16,
'n_estimators': 300, 'subsample': 0.5}
```

Gambar 4.41 Hasil pencarian parameter terbaik

Parameter terbaik untuk model GBT sudah ditemukan, langkah selanjutnya adalah melakukan pengujian nilai tersebut terhadap data EDM, hasil pengujian akan dibandingkan dengan parameter bawaan dari model GBT. Adapun kode program untuk pengujian model GBT dengan optimasi parameter dan tanpa optimasi ditunjukkan pada gambar 4.43 dan 4.44.

```
1  #Imputdata
2  import pandas as pd
3
4  df = pd.read_csv('Data_prediksi_performa.csv')
5
6  X = df.iloc[:, 2:19]
7  y = df.iloc[:, 19]
```

Gambar 4.42 Input data pengujian

Gambar 4.42 merupakan potongan kode program untuk mengambil data, Kode baris pertama dari gambar tersebut merupakan komentar atau keterangan, selanjutnya kode baris kedua berfungsi untuk memanggil pustaka dari pandas, sedangkan untuk baris keempat adalah kode untuk mengambil data yang dibutuhkan yaitu data EDM dengan format csv dari dokumen dengan nama Data_prediksi_performa.csv. Selanjutnya pada baris keenam dan ketujuh merupakan kode untuk membuat variabel X

dan y dengan mengambil kolom 2 sampai dengan kolom 18 untuk variabel X dan kolom 19 untuk variabel y yang merupakan variabel variabel target.

```
1  from sklearn.metrics import make_scorer, mean_squared_error, mean_absolute_error
2  from sklearn.model_selection import cross_validate
3  from numpy import sqrt, mean, absolute
4  from sklearn.model_selection import cross_val_score, KFold
5  from sklearn.ensemble import GradientBoostingRegressor
6  import numpy as np
7
8  crossvalidation= KFold(n_splits= 5, shuffle= True, random_state= 12)
9
10 gbr = GradientBoostingRegressor()
11
12 scoring_metrics = {'mse' : make_scorer(mean_squared_error),
13                    'mae': make_scorer(mean_absolute_error)}
14
15 score= cross_validate(gbr, X, y, scoring=scoring_metrics, cv=crossvalidation, n_jobs=-1)
16
17 print('Nilai MSE: %.2f' % np.mean(score['test_mse']))
18 print('Nilai MAE: %.2f' % np.mean(score['test_mae']))
19 print('Nilai RMSE: %.2f' % np.sqrt(mean(score['test_mse'])))
20
21 df = pd.DataFrame()
22 print('')
23 print('Nilai 5 x Coss-Validasi')
24 df['MSE'] = score ['test_mse']
25 df['MAE'] = score ['test_mae']
26 df['RMSE'] = sqrt(score ['test_mae'])
27 print(df)
```

```
Nilai MSE: 0.15
Nilai MAE: 0.27
Nilai RMSE: 0.39

Nilai 5 x Coss-Validasi
        MSE       MAE      RMSE
0  0.145899  0.268346  0.518022
1  0.142196  0.263924  0.513736
2  0.150438  0.264521  0.514316
3  0.168252  0.279177  0.528372
4  0.148391  0.261592  0.511460
```

Gambar 4.43 Kode pengujian GBT tanpa optimasi

Proses pemanggilan data telah dilakukan, selanjutnya adalah proses pengujian yang ditunjukkan pada gambar 4.43, dimana pada awal potongan kode program baris pertama sampai dengan baris keenam merupakan pemanggilan pustaka dan sekaligus penamaan alias. Pada kode baris kedelapan merupakan kode untuk membagi data kedalam lima kelompok, empat dari kelompok akan dijadikan data pelatihan dan data satu kelompok akan dijadikan data pengujian, proses ini dilakukan lima kali

sampai setiap bagian pernah menjadi data uji. Baris kesepuluh merupakan variabel untuk memanggil model GBT, sedangkan baris kedua belas dan tiga belas digunakan untuk memanggil metode pengujian yaitu MSE dan MAE, kode baris ke lima belas merupakan baris untuk melakukan pengujian data EDM menggunakan model GBT dengan metode pengukuran MAE dan MSE serta validasi silang lima kali, proses tersebut dilakukan dengan memanfaatkan semua pemrosesan yang ada di komputer. Baris ketujuh belas sampai dengan baris kesembilan belas merupakan kode untuk menampilkan hasil skor pengujian, sedangkan baris kedua puluh satu sampai dengan baris terakhir merupakan kode untuk menampilkan nilai skor dari masing-masing validasi silang. Hasil dari pengujian menggunakan model GBT tanpa optimasi diperoleh nilai MSE sebesar 0,15, nilai MAE sebesar 0,27 dan nilai 0,39. Sedangkan nilai dari validasi silangnya rata-rata memiliki nilai yang sama.

Proses pengujian model GBT tanpa optimasi telah dilakukan selanjutnya adalah proses pengujian model dengan menggunakan optimasi GS. Kode program untuk pengujian ini ditunjukkan pada potongan kode program gambar 4.44, dimana pada awal kode tidak lagi memanggil pustaka dan penamaan alias dari pustaka yang dibutuhkan hal ini karena pada kode pengujian sebelumnya pustaka ini telah dipanggil dan diberikan nama alias. Baris pertama merupakan kode komentar atau keterangan, kode baris kedua sampai dengan baris keenam merupakan kode untuk memberikan nilai parameter dimana nilai ini diperoleh dari pencarian GS yang dilakukan sebelumnya. Kode baris kedelapan merupakan pembuatan variabel model GBT dengan nilai parameter yang telah ditentukan oleh kode sebelumnya. Kode baris kesepuluh dan sebelas merupakan kode untuk mengukur keakuratan model, sedangkan kode baris ketiga belas sampai dengan kode baris terakhir penjelasannya sama dengan kode pada

gambar 4.43. Hasil dari pengujian menggunakan model GBT dengan menggunakan nilai parameter yang ditemukan oleh GS diperoleh nilai MSE sebesar 0,13, nilai MAE sebesar 0,25 dan nilai 0,37, adapun nilai dari validasi silangnya rata-rata memiliki nilai yang sama. Hasil pengujian dari kedua model memiliki nilai MSE dan MAE terpaut 2 persen sedangkan nilai RMSE terpaut 3%, dengan demikian penggunaan optimasi hyperparameter GS pada model GBT terbukti dapat meningkatkan akurasi model.

```
#GBT-GridSearchCV
gbr_params = {'n_estimators': 300,
              'max_depth': 16,
              'subsample': 0.5,
              'learning_rate': 0.02
              }

gbro = GradientBoostingRegressor(**gbr_params)

scoring_metrics = {'mse' : make_scorer(mean_squared_error),
                   'mae': make_scorer(mean_absolute_error)}

score= cross_validate(gbro, X, y, scoring=scoring_metrics, cv=crossvalidation, n_jobs=-1)

print('Nilai MSE: %.2f' % np.mean(score['test_mse']))
print('Nilai MAE: %.2f' % np.mean(score['test_mae']))
print('Nilai RMSE: %.2f' % np.sqrt(mean(score['test_mse'])))

df = pd.DataFrame()
print('')
print('Nilai 5 x Coss-Validasi')
df['MSE'] = score ['test_mse']
df['MAE'] = score ['test_mae']
df['RMSE'] = sqrt(score ['test_mae'])
print(df)
```

```
Nilai MSE: 0.13
Nilai MAE: 0.25
Nilai RMSE: 0.36

Nilai 5 x Coss-Validasi
        MSE       MAE      RMSE
0  0.133465  0.255838  0.505804
1  0.126115  0.245215  0.495192
2  0.127191  0.241923  0.491857
3  0.141722  0.249895  0.499895
4  0.134315  0.245863  0.495846
```

Gambar 4.44 Kode pengujian GBT dengan optimasi GS

4.7. Evaluasi dan analisis hasil prediksi

4.7.1. Evaluasi pengumpulan data

a. Pengumpulan data

Data bersumber dari empat sumber berbeda yaitu SIAKAD, Sisfo PMB, LMS dan SK Organisasi. SIAKAD menghasilkan data akademik berupa IPS, sementara Sisfo PMB menghasilkan data geografi yang terdiri dari jenis kelamin dan kota tinggal serta data ekonomi, adapun LMS menghasilkan data aktivitas mahasiswa pada LMS, sedangkan SK Organisasi menghasilkan data mahasiswa yang berperan aktif dalam organisasi kampus.

b. Ekstraksi data

Proses ekstraksi data dilakukan untuk mengambil variabel-variabel yang sesuai dengan kebutuhan dari data yang telah dikumpulkan pada proses sebelumnya. Proses ekstraksi data LMS dilakukan untuk memperoleh data dari data catatan aktivitas mahasiswa di LMS. Proses ini diawali dengan menentukan sasaran prediksi yaitu mahasiswa angkatan 2019 dan 2020. Proses ekstraksi dilakukan secara bertahap per variabel yang selanjutnya dari masing-masing variabel yang telah diekstraksi dikumpulkan kembali menjadi satu dokumen, dari masing-masing dokumen perhari digabungkan menjadi dokumen perminggu dan dokumen perminggu digabungkan untuk menjadi data LMS. Adapun ekstraksi data SIAKAD, Sisfo PMB, dan SK Organisasi dilakukan dengan cara mengambil variabel-variabel yang dibutuhkan untuk prediksi sesuai dengan sasaran prediksi.

c. Penggabungan data

Penggabungan data merupakan proses untuk menjadikan data dari empat sumber yang berbeda tersebut menjadi data yang siap untuk digunakan (data EDM). Proses penggabungan dilakukan dengan memanfaatkan kode pemrograman Python,

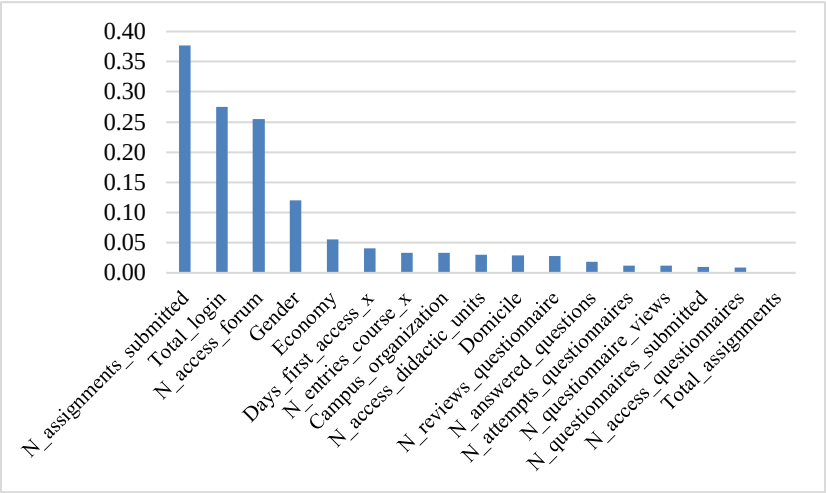
adapun kunci utama dari penggabungan ini adalah NIM dari mahasiswa sasaran.

d. Pemisahan data

Memisahkan data menjadi dua bagian: data pelatihan (training data) dan data pengujian (testing data). Pembagian ini dengan porsi 80% untuk data pelatihan dan 20% untuk data pengujian.

4.7.2. Evaluasi fitur

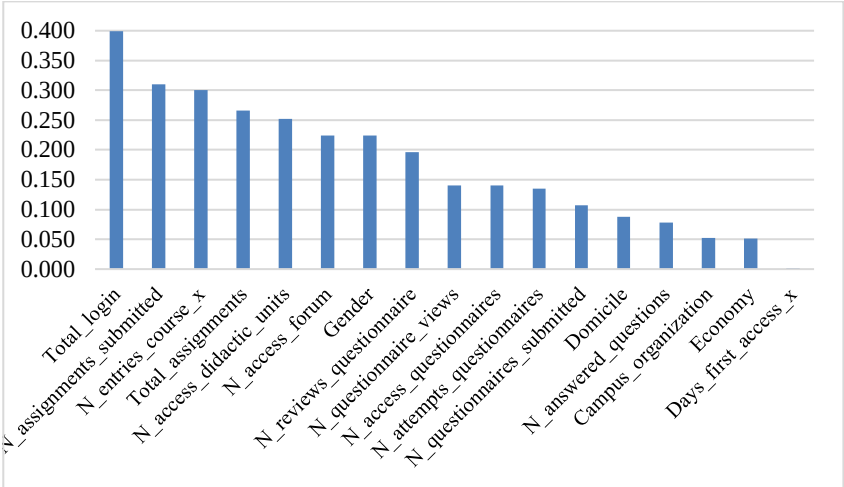
Data EDM ini terdiri dari 20 fitur, dua diantaranya tidak digunakan untuk melakukan prediksi yaitu NIM yang merupakan kunci dan Nama_Lengkap. 18 fitur satu diantaranya adalah target yaitu IPS, sehingga total fitur untuk variabel x terdiri dari 17 fitur dan variabel y terdiri dari 1 fitur. Bobot dari masing-masing fitur ditunjukkan pada gambar 4.45.



Gambar 4.45 Bobot fitur data EDM

Gambar 4.45 menunjukkan bahwa tiga fitur dari data LMS mendominasi di urutan pertama sampai ketiga, sementara data demografi fitur gender berada pada posisi keempat dilanjutkan

dengan data ekonomi, dua fitur berikutnya adalah fitur dari data LMS lagi. Fitur dari data organisasi mahasiswa berada pada posisi kedelapan dilanjutkan kembali fitur dari data LMS kemudian data demografi fitur domisili berada pada peringkat kesepuluh. Fitur peringkat kesebelas hingga ketujuh belas kembali ditempati fitur-fitur dari data LMS. Berdasarkan hal tersebut diatas dapat disimpulkan bahwa data non-akademik yang diusulkan pada penelitian ini memiliki bobot yang baik jika digunakan untuk memprediksi performa akademik mahasiswa.



Gambar 4.46 Pengaruh fitur data EDM terhadap fitur target (IPS)

Gambar 4.46 adalah visualisasi dari pengaruh fitur data EDM terhadap fitur target (IPS). Terdapat beberapa perbedaan antara bobot fitur dengan pengaruh fitur terhadap fitur target dimana enam fitur yang paling berpengaruh merupakan fitur yang bersumber dari data LMS, sementara fitur gender berada pada posisi ketujuh, fitur domisili berada pada posisi ketiga belas, fitur organisasi kampus dan ekonomi masing-masing berada pada posisi kelima belas dan enam belas. Berdasarkan pada visualisasi data yang disampaikan pada pembahasan sebelumnya

disampaikan bahwa fitur-fitur non akademik memiliki data yang tidak seimbang, hal ini sangat mempengaruhi signifikansi pengaruh fitur terhadap fitur target

4.7.3. Evaluasi model prediksi

a. Pemilihan model

Para peneliti prediksi performa akademik mahasiswa umumnya menggunakan model klasifikasi untuk memprediksi fitur target (IPK/IPS), dengan cara mengkategorikan fitur target kedalam klasifikasi yang berbeda seperti baik dan buruk (Aluko dkk. 2018; Jenitha dkk. 2021), baik, rata-rata, dan buruk (Mengash 2020), (Karalar, Kapucu, dan Gürüler 2021), (Oreski, Visnjic, dan Kadoic 2022), (Oreski dkk. 2022). Namun, pendekatan ini dapat menyebabkan hilangnya informasi penting dari data (García dkk. 2013).

Total_login	N_access_forum	N_access_didactic_units	Total_assignments	N_assignments_submitted	N_access_questionnaires	N_attempts_questionnaires	N_answered_questions	N_questionnaire_views	N_questionnaires_submitted	N_reviews_questionnaire	Days_first_access_x	N_entries_course_x	CGPA	Regression	Classification
60	1	44	58	7	0	0	0	0	0	16	1	102	1.5	1.554503	Fail
121	1490	296	569	42	225	192	15	221	2	439	1	865	3.55	3.559111	Pass
78	1	8	231	23	338	290	5	324	7	43	1	239	2.36	2.368069	Fail
51	1	10	61	4	6	0	5	6	0	50	1	71	1	1.032266	Fail
57	6	18	90	8	0	0	0	0	0	8	2	108	2.45	2.437308	Pass
117	1847	140	315	29	327	310	16	321	3	142	1	455	3.95	3.935609	Pass
58	2	61	186	21	0	0	0	0	0	60	1	247	2.32	2.371394	Fail
73	102	38	318	17	16	12	14	14	1	23	1	356	2.93	2.917675	Pass

Gambar 4.47 Perbandingan model klasifikasi dan regresi

Kelemahan dari model klasifikasi adalah hilangnya informasi penting dari data, untuk itu penulis mengusulkan model regresi untuk memprediksi performa akademik mahasiswa. Adapun hasil perbandingan prediksi menggunakan model regresi dan klasifikasi ditunjukkan pada gambar 4.47. Model klasifikasi yang diusulkan oleh (Aluko dkk. 2018) mengklasifikasikan IPS menjadi lulus atau gagal berdasarkan nilai ambang batas > 2,4. Baris 5 dan 6 dalam model klasifikasi memiliki nilai kelulusan

yang sama, meskipun memiliki skor IPS yang berbeda secara signifikan. Studi ini menunjukkan bahwa perbedaan kecil 0,05 pada baris 5 dapat mengubah hasil dari lulus menjadi gagal, menunjukkan informasi yang hilang yang perlu dipertimbangkan untuk mencapai kinerja yang baik. Sebaliknya, baris 3 dan 4 menunjukkan bahwa baris 3 hanya membutuhkan peningkatan kecil 0,04 atau peningkatan salah satu nilai variabel evaluasi untuk mencapai nilai kelulusan, sedangkan baris 4 membutuhkan lebih banyak upaya untuk mencapai nilai kelulusan. Temuan menunjukkan bahwa model regresi lebih efektif daripada model klasifikasi karena tidak mengakibatkan hilangnya informasi penting, dan nilai prediksi mewakili nilai aktual.

Penentuan model prediksi performa akademik mahasiswa telah ditentukan yaitu model regresi, setelah itu proses selanjutnya adalah penentuan model terbaik dari model regresi yang telah banyak digunakan oleh para peneliti sebelumnya. Model-model regresi yang sering digunakan diantaranya adalah GLM, DL, DT, RF, GBT, dan SVM. Model-model tersebut selanjutnya dibandingkan dan hasil perbandingan dari model-model tersebut ditemukan GBT merupakan model terbaik untuk memprediksi performa akademik mahasiswa.

b. Penyetelan model (model tuning)

GBT merupakan model ensembel dimana model ini terbentuk dari beberapa model didalamnya, model ini memiliki banyak parameter dan hyperparameter, berdasarkan tinjauan pustaka dari para peneliti sebelumnya bahwa penyetelan hyperparameter dapat meningkatkan akurasi prediksi dari model yang digunakan. Pada penelitian ini penulis menggunakan metode penyetelan hyperparameter menggunakan GS. Dari hasil percobaan yang telah dilakukan metode penyetelan GS menemukan nilai parameter terbaik yaitu nilai *learning_rate*

sebesar 0.02, *max_depth:* 16, *n_estimators:* 300, dan *subsample* sebesar 0.5.

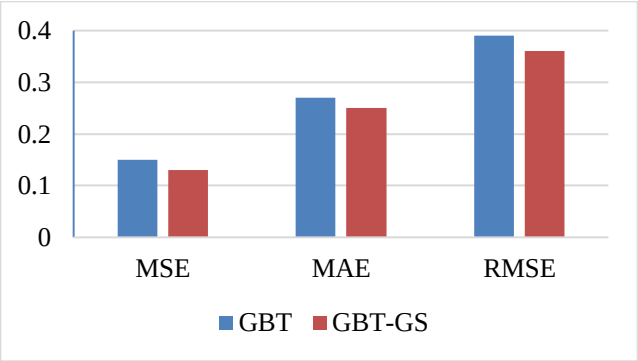
c. Pelatihan model

Pada fase evaluasi pelatihan model ditemukan bahwa model dilatih menggunakan data pelatihan yang telah dipisahkan sebelumnya. Model akan "belajar" dari data pelatihan untuk menemukan pola atau hubungan yang dapat digunakan untuk melakukan prediksi performa akademik.

4.7.4. Evaluasi Kinerja Model

a. Metrik evaluasi performa model prediksi

Evaluasi model dilakukan dengan menggunakan metrik MAE, MSE dan RMSE. Metrik tersebut digunakan untuk mengetahui akurasi dari model. Akurasi model selanjutnya dibandingkan antara model tanpa optimasi dengan model yang teroptimasi. Nilai MSE sebesar 0,15, nilai MAE sebesar 0,27 dan nilai 0,39 untuk model GBT tanpa optimasi dan nilai MSE sebesar 0,13, nilai MAE sebesar 0,25 dan nilai 0,37 untuk model GBT dengan optimasi, nilai MSE dan MAE antar model terpaut 2 persen sedangkan nilai RMSE terpaut 3%, perbandingan kedua model tersebut ditunjukkan pada gambar 4.48.



Gambar 4.48 Grafik perbandingan pengujian model

Evaluasi model selanjutnya adalah dengan membandingkan hasil prediksi model dengan nilai sebenarnya pada data pengujian ditunjukkan pada gambar 4.49.

```
#Membuat inputan
input_data = np.array([[0,1,3,1,96,481,148,284,32,210,201,13,206,2,218,1,432]])

prediction = gbr.predict(input_data)
print("Nilai IPK adalah: %.2f" % prediction)
```

```
Nilai IPK adalah: 3.71
```

```
#Membuat inputan
input_data = np.array([[0,1,3,1,96,481,148,284,32,210,201,13,206,2,218,1,432]])

prediction = gbro.predict(input_data)
print("Nilai IPK adalah: %.2f" % prediction)
```

```
Nilai IPK adalah: 3.70
```

Gambar 4.49 Pengujian menggunakan data baru

Gambar 4.49 menunjukkan dua potongan kode program, dimana pada kode pertama merupakan kode pengujian GBT tanpa optimasi dan potongan kedua adalah GBT teroptimasi GS dengan data baru. Hasil pengujian ditemukan bahwa prediksi nilai IPK sebesar 3.71 sedangkan pada potongan kode kedua memprediksi nilai IPS sebesar 3.70. Data baru ini merupakan data yang diambil dari data EDM, dimana nilai IPS dari data tersebut sebesar 3.68.

b. Validasi silang dan evaluasi statistik

Untuk memvalidasi model metode yang digunakan adalah validasi silang dengan 5 kali validasi, yaitu data dibagi menjadi 5 bagian dimana satu bagian sebai data pengujian dan empat bagian sebagai data pelatihan. Hasil dari validasi 5 kali ditunjukkan pada tabel 4.12.

Berdasarkan tabel 4.12 diatas ditunjukkan bahwa masing-masing matrik penilaian memiliki nilai yang hampir sama antar validasi, nilai MSE dan RMSE paling kecil berada pada validasi pertama sementara masing-masing dengan nilai 0.130309 dan 0.360984, sementara itu nilai MAE terkecil berada pada validasi

kelima dengan nilai 0.248987. Hal ini menunjukkan bahwa sebaran data EDM yang digunakan sangat baik.

Tabel 4.12 Nilai validasi model

K-Validation	MSE	MAE	RMSE
1	0.130309	0.254369	0.360984
2	0.135093	0.249349	0.367549
3	0.145931	0.258957	0.38201
4	0.152121	0.25944	0.390027
5	0.140935	0.248987	0.375413

c. Interpretasi hasil evaluasi dan implikasinya

Fase penerapan adalah fase penerapan dari pengetahuan yang didapat berdasarkan evaluasi pada proses data mining. Secara ringkas hasil evaluasi proses data mining adalah sebagai berikut:

1. Data yang digunakan merupakan data akademik dan data non-akademik. Data akademik bersumber dari data LMS dan data informasi akademik, sementara data non-akademik merupakan data ekonomi, demografi dan data organisasi kampus.
2. Kumpulan data akademik dan data non-akademik disebut dengan data EDM, data ini memiliki 3 tipe dimana kolom NIM merupakan id dari data, sementara itu kolom jenis kelamin, domisili, ekonomi dan organisasi kampus bertipe kategorikal, sedangkan kolom lainya bertipe numerik. Kolom target adalah CGPA dengan tipe data numerik.
3. Model prediksi terdiri dari dua model yaitu model regresi dan model klasifikasi, model klasifikasi memiliki kekurangan yaitu hilangnya informasi penting dari sebuah data, sehingga model yang digunakan adalah model regresi.
4. Model yang digunakan untuk memprediksi performa akademik mahasiswa adalah GBT, GBT terbukti memiliki

tingkat kesalahan paling kecil jika dibandingkan dengan model lainnya.

5. Untuk meningkatkan akurasi dari model GBT metode yang digunakan adalah penyetelan hyperparameter, penyetelan hyperparameter dikerjakan dengan menggunakan metode GS, nilai terbaik dari parameter hasil penyetelan GS adalah *learning_rate* dengan nilai sebesar 0.02, *max_depth:* 16, *n_estimators:* 300, dan nilai *subsample* sebesar 0.5.
6. Berdasarkan uji coba yang telah dikerjakan dalam penggunaan model GBT tanpa optimasi dan GBT teroptimasi ditemukan bahwa model GBT teroptimasi memiliki nilai akurasi lebih baik dibanding dengan model tanpa optimasi.

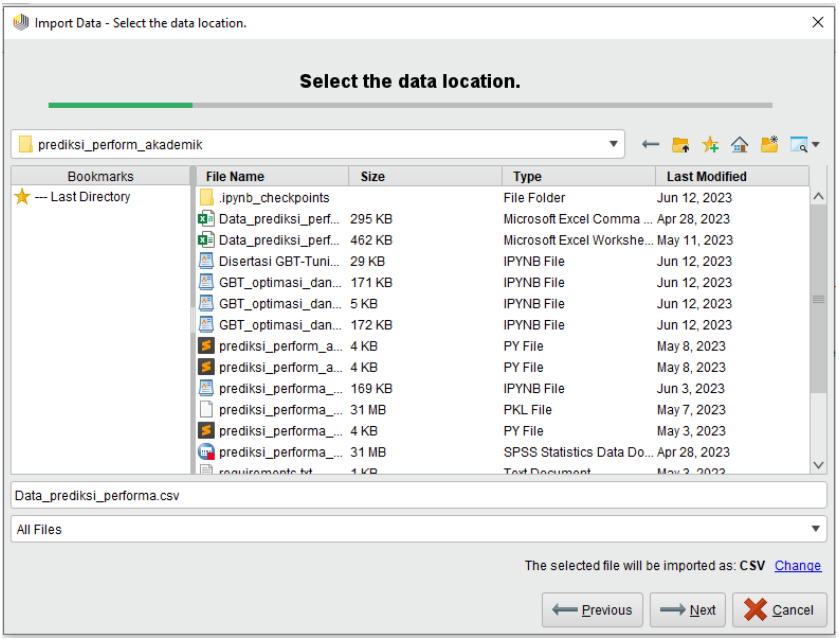
Setelah menemukan poin-poin penting dari proses evaluasi data mining, selanjutnya poin-point tersebut digunakan untuk fase penerapan. Pada fase penerapan penulis menggunakan alat bantu RapidMiner. RapidMiner merupakan salah satu tools yang digunakan dalam data mining. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. RapidMiner menempati peringkat pertama sebagai Software data mining pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011.

Penerapan data mining pada alat bantu RapidMiner terdiri dari beberapa langkah yaitu:

1. Memuat data
2. Pemilihan tipe data dari masing-masing kolom

3. Penentuan kolom id dan kolom target
4. Pemilihan model
5. Penyesuaian parameter
6. Proses prediksi
7. Evaluasi model
8. Penerapan model
9. Hasil prediksi
10. Simulasi

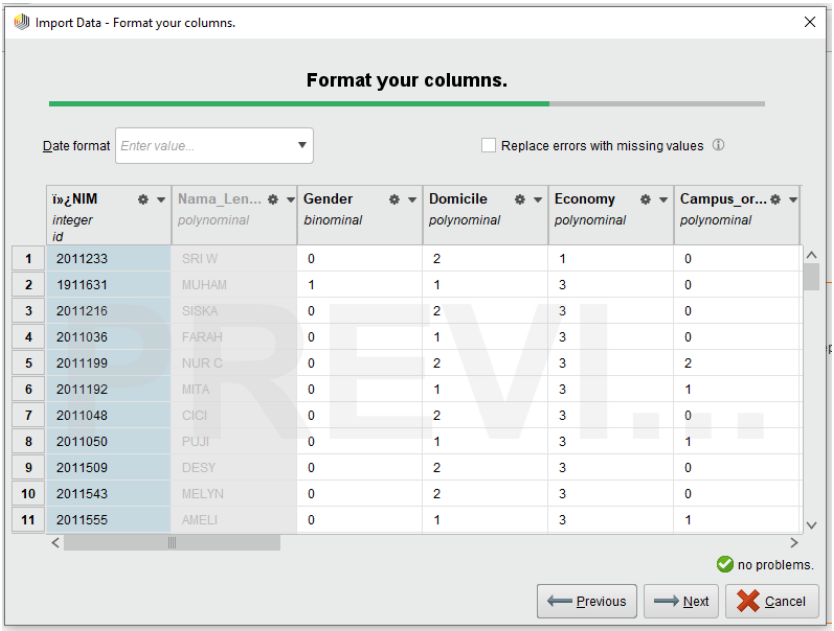
Langkah demi langkah tersebut harus dikerjakan dengan cermat dan berurutan agar mendapatkan hasil yang maksimal.



Gambar 4.50 Proses muat data EDM

Proses memuat data ditunjukan pada gambar 4.50, dimana data yang digunakan adalah data EDM berekstensi csv dengan nama Data_prediksi_performa.csv. Setelah data dimuat selanjutnya menentukan tipe data dari masing-masing kolom,

pada langkah ini kolom Nama_Lengkap tidak diikutsertakan untuk prediksi sehingga diperlukan penyesuaian dengan menerapkan perintah *Exclude column.* Masih pada halaman yang sama, untuk kolom NIM dijadikan ID, kolom Gender di ganti tipe binominal, kolom Domicile, Economy, dan kolom Campus_organization diubah menjadi tipe binominal, hal ini ditunjukkan pada gambar nomor 4.51.

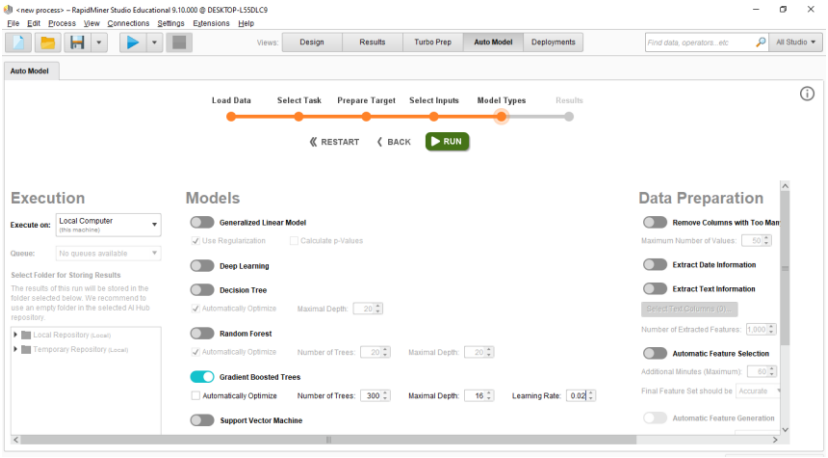


	ï¿¿NIM integer id	Nama_Len... polynominal	Gender binominal	Domicile polynominal	Economy polynominal	Campus_or... polynominal
1	2011233	SRI W	0	2	1	0
2	1911631	MUHAM	1	1	3	0
3	2011216	SISKA	0	2	3	0
4	2011036	FARAH	0	1	3	0
5	2011199	NUR C	0	2	3	2
6	2011192	MITA	0	1	3	1
7	2011048	CICI	0	2	3	0
8	2011050	PUJI	0	1	3	1
9	2011509	DESY	0	2	3	0
10	2011543	MELYN	0	2	3	0
11	2011555	AMELI	0	1	3	1

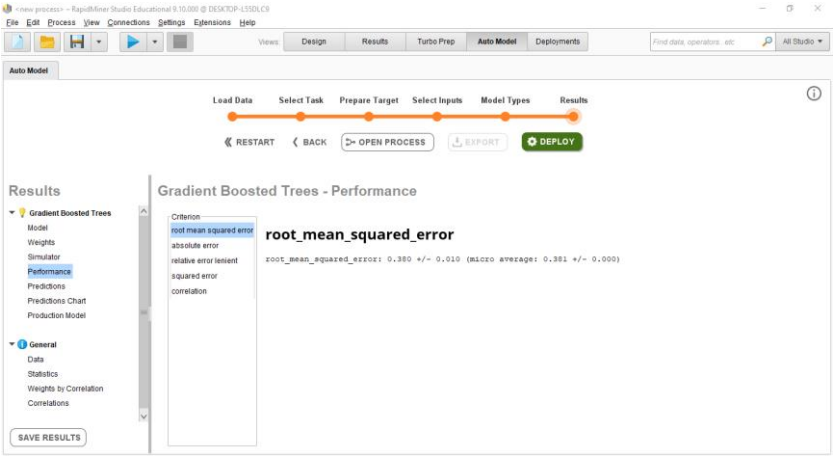
Gambar 4.51 Penentuan tipe data dan kolom id serta kolom target

Proses selanjutnya adalah pemilihan model dan penyesuaian parameter, proses ini ditunjukkan pada gambar 4.52, pada gambar tersebut ditunjukkan bahwa terdapat banyak model, namun sesuai dengan temuan pada fase evaluasi dimana model yang digunakan adalah GBT dengan parameter *learning_rate* sebesar 0.02, *max_depth:* 16, dan *n_estimators* (Number of Trees): 300.

Sementara itu gambar 4.47 merupakan visualisasi dari evaluasi hasil prediksi dengan nilai RMSE: 0.38.



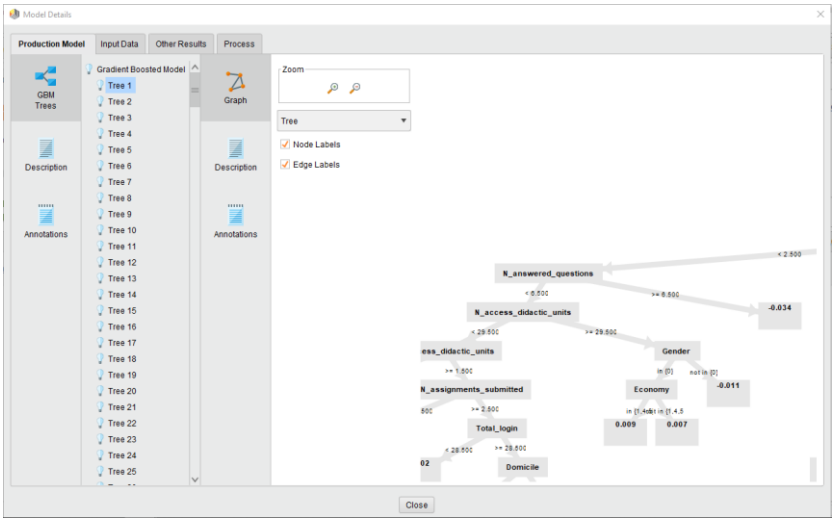
Gambar 4.52 Pemilihan model



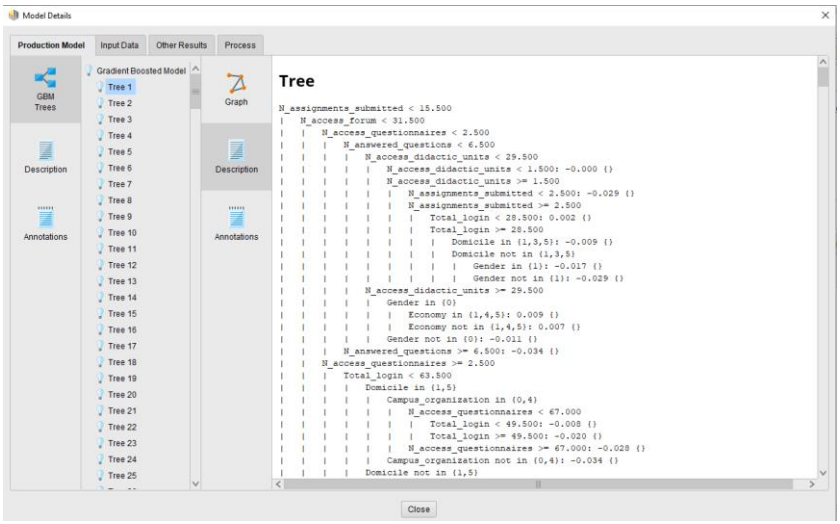
Gambar 4.53 Evaluasi Hasil Prediksi

Pada fase penerapan ini proses selanjutnya adalah penerapan model ditunjukkan pada gambar 4.53, caranya adalah dengan mengklik tombol Deploy maka akan menghasilkan model dengan dua visualisasi yaitu model berbentuk pohon dan deskripsi,

visualisasi hasil prediksi ditunjukkan pada gambar 4.54 merupakan visualisasi dalam bentuk pohon keputusan sedangkan gambar 4.55 merupakan visualisasi hasil prediksi dalam bentuk deskripsi.

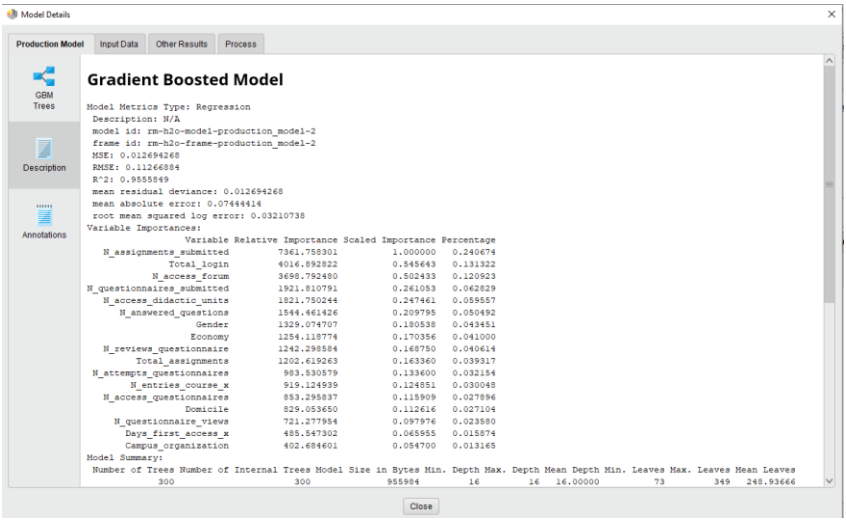


Gambar 4.54 Visualisasi hasil prediksi bentuk pohon keputusan

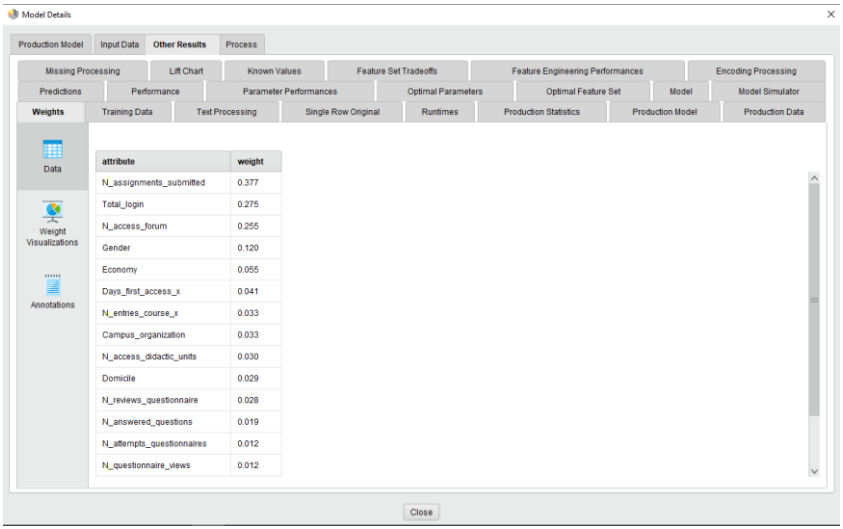


Gambar 4.55 Visualisasi hasil prediksi bentuk deskripsi

Gambar 4.56 menampilkan hasil prediksi yang berbentuk deskripsi dari model GBT untuk prediksi data EDM, pada gambar ini ditampilkan nilai evaluasi dari model serta nilai dari masing-masing fitur.



Gambar 4.56 Hasil model GBT dalam bentuk deskripsi

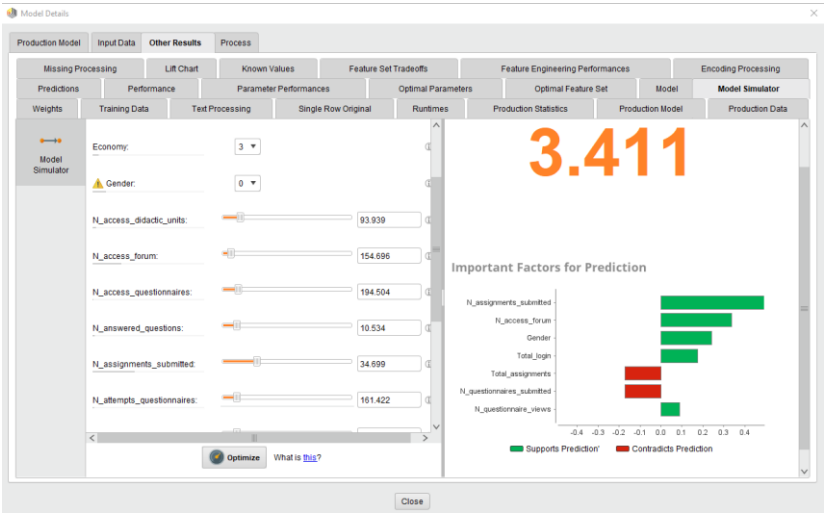


Gambar 4.57 Hasil lain yang dihasilkan oleh model GBT

Adapun gambar 4.57 merupakan hasil lain dari model, dimana pada gambar tersebut ditampilkan bobot dari masing-masing fitur. Pada gambar 4.57 ini juga menampilkan hasil-hasil lainnya dari model GBT.

4.8. Implementasi dan Penerapan Hasil Prediksi

Dalam mengimplementasikan hasil dapat dilakukan dengan cara mengintegrasi hasil prediksi ke dalam sistem pendidikan atau platform yang relevan. Platform yang relevan dapat berupa aplikasi berbasis web ataupun mobile, namun dalam buku ini penulis memanfaatkan aplikasi RapidMiner dalam menerapkan hasil prediksi melalui simulasi model ditunjukkan pada gambar 4.58. Dengan menggunakan simulasi model ini, pengguna dapat dengan mudah dalam mensimulasikan hasil prediksi performa akademik mahasiswa dengan cara memasukkan data baru.



Gambar 4.58 Model simulator

Pengelolaan dan pengawasan berkelanjutan harus dilakukan untuk memastikan keberhasilan implementasi prediksi

performa akademik siswa dalam konteks pendidikan. Berikut adalah beberapa cara untuk mengelola dan mengawasi secara berkelanjutan:

1. Tim Manajemen Proyek: Bentuk tim manajemen proyek yang bertanggung jawab untuk mengawasi dan mengelola implementasi prediksi performa akademik. Tim ini harus memiliki pemimpin proyek yang memimpin upaya tersebut, dan anggota tim yang terlibat dalam aspek teknis, administratif, dan analitis.
2. Perencanaan yang Terstruktur: Mengembangkan rencana implementasi yang terstruktur dengan tujuan, jadwal, dan tugas yang jelas. Rencana ini harus mencakup identifikasi sumber daya yang diperlukan, tahapan implementasi, dan batasan yang mungkin terjadi.
3. Koordinasi dengan Pihak Terkait: Berkomunikasi dan berkoordinasi dengan pihak terkait, seperti guru, staf administrasi, dan manajemen sekolah/universitas. Ini termasuk menjelaskan tujuan dan manfaat prediksi performa akademik, serta melibatkan mereka dalam proses implementasi dan penggunaan hasil prediksi.
4. Pelatihan dan Pendampingan: Memberikan pelatihan dan pendampingan kepada pihak terkait yang terlibat dalam penggunaan dan interpretasi hasil prediksi performa akademik. Ini termasuk memberikan pemahaman tentang metode dan teknik yang digunakan dalam prediksi, serta membantu mereka dalam mengambil keputusan berdasarkan hasil prediksi.
5. Evaluasi dan Pemantauan: Melakukan evaluasi berkala terhadap implementasi prediksi performa akademik untuk memastikan bahwa proses berjalan sesuai rencana dan mencapai tujuan yang ditetapkan. Melakukan pemantauan terhadap akurasi prediksi, kepuasan pengguna, dan dampak

penggunaan prediksi dalam pengambilan keputusan pendidikan.

6. Penanganan Masalah dan Perbaikan: Mempersiapkan rencana penanganan masalah yang terperinci untuk mengatasi kendala atau masalah yang mungkin timbul selama implementasi. Mengidentifikasi masalah secara proaktif dan melakukan perbaikan yang diperlukan untuk memastikan keberhasilan implementasi.
7. Keberlanjutan dan Peningkatan: Membuat rencana untuk menjaga keberlanjutan penggunaan prediksi performa akademik dan terus meningkatkannya seiring berjalannya waktu. Ini dapat melibatkan pengembangan model prediksi yang lebih baik, pengumpulan data yang lebih baik, dan peningkatan metode pengolahan data.

Pengelolaan dan pengawasan berkelanjutan memastikan bahwa prediksi performa akademik siswa terus diperbarui, disesuaikan dengan kebutuhan pendidikan yang berkembang, dan memberikan manfaat yang signifikan bagi pengambilan keputusan. Ini juga membantu dalam mengatasi tantangan dan hambatan yang mungkin muncul selama implementasi serta meningkatkan kualitas dan efektivitas prediksi performa akademik.

PENUTUP

1. Kesimpulan dan Tantangan di Masa Depan

a. Kesimpulan

Kesimpulan dari buku ini adalah bahwa educational data mining (EDM) merupakan teknik data mining yang digunakan untuk menggali informasi tersembunyi dari data pendidikan. Terdapat 11 area dalam EDM, dengan prediksi performa mahasiswa menjadi area yang paling populer dan berkembang pesat. Prediksi performa akademik merupakan bagian dari prediksi performa mahasiswa yang fokus pada prediksi nilai akhir mahasiswa.

Perkembangan pesat dalam bidang ini menegaskan pentingnya penelitian mengenai prediksi performa akademik mahasiswa. Untuk mencapai prediksi yang akurat, peneliti membandingkan beberapa metode dan memilih metode terbaik. Penggunaan model klasifikasi dalam prediksi performa akademik memiliki kelemahan yaitu kehilangan informasi penting saat mengklasifikasikan nilai ke dalam kelas. Oleh karena itu, beberapa peneliti menggunakan model regresi untuk menghindari kehilangan informasi.

Penyetelan hyperparameter juga dapat meningkatkan akurasi beberapa model prediksi. Metode grid search digunakan untuk melakukan penyetelan hyperparameter secara efektif. Selain mencari metode yang paling akurat, para peneliti juga berusaha menemukan faktor-faktor lain yang memengaruhi prestasi akademik mahasiswa.

Dengan demikian, buku ini mengemukakan pentingnya EDM dalam menggali informasi pendidikan, fokus pada prediksi performa akademik mahasiswa, kelemahan dan peningkatan akurasi model prediksi, serta upaya untuk menemukan faktor-faktor yang memengaruhi prestasi akademik mahasiswa.

b. Tantangan dan peluang yang dihadapi dalam prediksi performa akademik melalui Educational Data Mining

Prediksi performa akademik melalui Educational Data Mining (EDM) menghadapi tantangan dan peluang tertentu. Berikut adalah beberapa di antaranya:

1. Tantangan Data: Salah satu tantangan utama adalah ketersediaan dan kualitas data yang digunakan untuk prediksi. Data akademik sering kali kompleks, tidak terstruktur, dan beragam. Selain itu, data tersebut mungkin mengandung kecacatan, seperti kehilangan nilai atau outlier. Menghadapi tantangan ini, peneliti perlu mengatasi masalah pengumpulan data yang akurat dan memastikan bahwa data yang digunakan dalam prediksi mencerminkan kondisi yang sebenarnya.
2. Dimensi dan Kompleksitas Data: Data akademik biasanya memiliki dimensi yang tinggi dan kompleksitas yang beragam. Terdapat berbagai variabel, seperti nilai mata pelajaran, keterlibatan siswa, kehadiran, dan karakteristik personal. Mengelola dan memahami data dengan dimensi dan kompleksitas tinggi ini memerlukan teknik dan algoritma yang tepat.
3. Interpretasi Hasil: Prediksi performa akademik melalui EDM menghasilkan model yang memberikan prediksi atau klasifikasi terhadap data. Namun, interpretasi hasil tersebut mungkin menjadi tantangan. Penting untuk memahami dengan benar apa arti dari hasil prediksi tersebut, dan bagaimana hasil tersebut dapat diinterpretasikan dan diterapkan dalam konteks pendidikan.
4. Kesulitan dalam Prediksi yang Akurat: Meskipun EDM menawarkan potensi untuk meningkatkan prediksi performa akademik, masih ada kesulitan dalam mencapai prediksi yang akurat. Variabilitas yang kompleks dalam prestasi akademik,

pengaruh faktor-faktor eksternal, dan tantangan dalam memodelkan interaksi antara variabel-variabel tersebut dapat menyebabkan ketidakakuratan dalam prediksi. Meningkatkan akurasi prediksi melalui pengembangan model yang lebih canggih dan penggunaan teknik pemrosesan data yang lebih baik tetap menjadi tantangan.

Di sisi lain, prediksi performa akademik melalui EDM juga memberikan peluang yang signifikan:

1. Intervensi dan Perbaikan: Prediksi performa akademik yang akurat melalui EDM dapat membantu dalam mengidentifikasi siswa-siswa yang berisiko rendah atau tinggi. Hal ini memberikan peluang untuk melakukan intervensi dini dan memberikan bantuan yang tepat kepada siswa yang membutuhkannya. Dengan informasi yang diperoleh dari prediksi, sekolah dan pendidik dapat mengembangkan program-program yang dirancang khusus untuk meningkatkan prestasi siswa.
2. Personalisasi Pendidikan: Dengan prediksi performa akademik yang akurat, EDM dapat membantu dalam personalisasi pendidikan. Dalam konteks ini, pendidik dapat menyesuaikan pembelajaran dan instruksi untuk memenuhi kebutuhan individu siswa berdasarkan prediksi dan karakteristiknya. Hal ini dapat meningkatkan efektivitas pembelajaran dan membantu siswa mencapai potensi terbaik mereka.
3. Pengembangan Metode dan Algoritma: Prediksi performa akademik melalui EDM mendorong pengembangan metode dan algoritma baru. Para peneliti terus mencoba meningkatkan keakuratan dan kinerja model prediksi, serta mengeksplorasi pendekatan baru untuk mengatasi tantangan yang ada. Ini memberikan peluang untuk inovasi dan pengembangan ilmu baru dalam bidang EDM.

Dalam kesimpulannya, prediksi performa akademik melalui EDM menghadapi tantangan dalam hal ketersediaan dan kualitas data, dimensi dan kompleksitas data, interpretasi hasil, serta kesulitan mencapai prediksi yang akurat. Namun, melalui EDM juga terdapat peluang untuk intervensi dan perbaikan, personalisasi pendidikan, serta pengembangan metode dan algoritma yang lebih baik.

2. Pemikiran tentang arah penelitian dan pengembangan selanjutnya

Pemikiran tentang arah penelitian dan pengembangan selanjutnya untuk prediksi performa akademik mahasiswa terus berkembang seiring dengan perkembangan teknologi dan kebutuhan pendidikan. Berikut beberapa arah yang mungkin menjadi fokus penelitian dan pengembangan di masa mendatang:

1. Penggunaan Metode Pembelajaran Mesin yang Lebih Maju: Penelitian dapat difokuskan pada pengembangan dan penerapan metode pembelajaran mesin yang lebih maju dan canggih dalam prediksi performa akademik mahasiswa. Ini melibatkan eksplorasi algoritma baru, seperti jaringan saraf tiruan (neural networks), pemrosesan bahasa alami (natural language processing), dan metode pembelajaran mendalam (deep learning), untuk meningkatkan akurasi prediksi.
2. Pendekatan Multimodal: Pemikiran dapat diarahkan ke integrasi data multimodal, seperti data akademik, data sosial, data demografis, dan data emosional, dalam prediksi performa akademik. Dengan mempertimbangkan berbagai aspek ini, prediksi dapat menjadi lebih holistik dan menghasilkan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi prestasi akademik.

3. Interpretabilitas dan Transparansi: Seiring dengan peningkatan kompleksitas metode prediksi, penting untuk menjaga interpretabilitas dan transparansi hasil prediksi. Penelitian dapat difokuskan pada pengembangan metode yang menjelaskan dan memperjelas alasan di balik prediksi, sehingga pengguna (seperti guru dan siswa) dapat memahami dasar prediksi dan mengambil tindakan yang sesuai.
4. Analisis Ketidakpastian: Pengembangan metode untuk mengukur dan mengelola ketidakpastian dalam prediksi performa akademik juga menjadi arah yang menarik. Dengan mengakui ketidakpastian dalam prediksi, pengambilan keputusan dapat menjadi lebih berimbang dan dapat mengurangi risiko pengambilan tindakan yang tidak tepat.
5. Pengembangan Framework dan Instrumen Evaluasi: Perkembangan framework dan instrumen evaluasi yang lebih baik untuk mengukur kinerja dan keberhasilan model prediksi dapat menjadi fokus penelitian. Ini akan membantu dalam membandingkan dan mengevaluasi berbagai metode prediksi yang ada, serta memastikan bahwa metode yang digunakan memberikan nilai tambah yang signifikan dalam konteks pendidikan.
6. Etika dan Privasi: Perhatian terhadap isu-etika dan privasi dalam penggunaan data pendidikan juga harus menjadi fokus penelitian dan pengembangan selanjutnya. Pengembangan pedoman dan kebijakan yang mengatur penggunaan data pendidikan, perlindungan privasi siswa, dan keadilan dalam pengambilan keputusan berdasarkan prediksi adalah hal yang penting.

Dalam keseluruhan, penelitian dan pengembangan selanjutnya untuk prediksi performa akademik mahasiswa akan melibatkan penggunaan metode yang lebih maju, pendekatan

multimodal, interpretabilitas, analisis ketidakpastian, pengembangan framework evaluasi, etika dan privasi, serta mengakom

DAFTAR PUSTAKA

Abu, Amjad. 2016. "Educational Data Mining & Students' Performance Prediction." *International Journal of Advanced Computer Science and Applications* 7(5):212–20. doi: 10.14569/ijacsa.2016.070531.

Abu Saa, Amjed, Mostafa Al-Emran, dan Khaled Shaalan. 2019. *Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques*. Vol. 24. Springer Netherlands.

Al-Emran, Mostafa, Vitaliy Mezhuyev, Adzhar Kamaludin, dan Khaled Shaalan. 2018. "The impact of knowledge management processes on information systems: A systematic review." *International Journal of Information Management* 43(2):173–87. doi: 10.1016/j.ijinfomgt.2018.08.001.

Albarrak, Ahmed I., Hatim A. Aboalsamh, dan Mohamed Abouzahra. 2010. "Evaluating learning management systems for university medical education." *ICEMT 2010 - 2010 International Conference on Education and Management Technology, Proceedings* 672–77. doi: 10.1109/ICEMT.2010.5657569.

Albreiki, Balqis, Nazar Zaki, dan Hany Alashwal. 2021. "A systematic literature review of student' performance prediction using machine learning techniques." *Education Sciences* 11(9). doi: 10.3390/educsci11090552.

Alsinaidi, Aljoharah A., Dalal H. Alotaibi, Amina K. P. Zerine, Rabab A. Bukhamseen R, Farah A. Alzahrani, dan Rana M. Almazroa. 2021. "Extracurricular Activities Among Undergraduate Dental Students In Dental Colleges In Riyadh, Saudi Arabia." *Annals of Dental Specialty* 9(1):34–38. doi: 10.51847/myyl88qns3.

Alsuwaiket, Mohammad. 2018. *Measuring Academic Performance of Students in Higher Education Using Data Mining Techniques*.

Aluko, Ralph Olusola, Emmanuel Itodo Daniel, Olalekan Shamsideen Oshodi, Clinton Ohis Aigbavboa, dan Abiodun Olatunji Abisuga. 2018. "Towards reliable prediction of academic performance of architecture students using data mining techniques." *Journal of Engineering, Design and Technology* 16(3):385–97. doi: 10.1108/JEDT-08-2017-0081.

Alyahyan, Eyman, dan Dilek Düştegör. 2020. "Predicting academic success in higher education: literature review and best practices." *International Journal of Educational Technology in Higher Education* 17(1). doi: 10.1186/s41239-020-0177-7.

Amrieh, Elaf Abu, Thair Hamtini, dan Ibrahim Aljarah. 2016. "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods." *International Journal of Database Theory and Application* 9(8):119–36. doi: 10.14257/ijdta.2016.9.8.13.

Andonie, Razvan, dan Adrian Catalin Florea. 2020. "Weighted random search for CNN hyperparameter optimization." *International Journal of Computers, Communications and Control* 15(2):1–11. doi: 10.15837/IJCCC.2020.2.3868.

Arifin, Muhammad, Widowati, Farikhin, Adi Wibowo, dan Budi Warsito. 2021. "Comparative Analysis on Educational Data Mining Algorithm to Predict Academic Performance." *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021* 173–78. doi: 10.1109/iSemantic52711.2021.9573185.

Arifin, Muhammad, Widowati Widowati, dan Farikhin Farikhin. 2022. “Using Education Data Mining (EDM) and Tracer Study (TS) Data as Materials for Evaluating Higher Education Curriculum and Policies.” *KnE Social Sciences*. doi: 10.18502/kss.v7i14.11948.

Arola Anderson, Nicole, Brynn Huguenel, Amy Bohnert, dan Colleen Conley. 2021. “Organized Activity Involvement Across the Transition to College.” *Journal of College Orientation, Transition, and Retention* 28(1):1–15. doi: 10.24926/jcotr.v28i1.3597.

Asif, Raheela, Agathe Merceron, Syed Abbas Ali, dan Najmi Ghani Haider. 2017. “Analyzing undergraduate students’ performance using educational data mining.” *Computers and Education* 113:177–94. doi: 10.1016/j.compedu.2017.05.007.

Baker, Ryan S. J. D., dan Kalina Yacef. 2009. “The State of Educational Data Mining in 2009 : A Review and Future Visions.” *Journal of Educational Data Mining* 1(1):3–16.

Bakhshinategh, Behdad, Osmar R. Zaiane, Samira ElAtia, dan Donald Ipperciel. 2018. “Educational data mining applications and tasks: A survey of the last 10 years.” *Education and Information Technologies* 23(1):537–53. doi: 10.1007/s10639-017-9616-z.

Bergstra, James, dan Yoshua Bengio. 2012. “Random search for hyper-parameter optimization.” *Journal of Machine Learning Research* 13:281–305.

Bohnert, Amy, Jennifer Fredricks, dan Edin Randall. 2010. “Capturing unique dimensions of youth organized activity involvement: Theoretical and methodological considerations.” *Review of Educational Research* 80(4):576–610. doi: 10.3102/0034654310364533.

Bravo-Agapito, Javier, Sonia J. Romero, dan Sonia Pamplona.

2021. “Early prediction of undergraduate Student’s academic performance in completely online learning: A five-year study.” *Computers in Human Behavior* 115(October 2020):106595. doi: 10.1016/j.chb.2020.106595.

Casanova, Joana R., Cristiano Mauro Assis Gomes, Ana B. Bernardo, José Carlos Núñez, dan Leandro S. Almeida. 2021. “Dimensionality and reliability of a screening instrument for students at-risk of dropping out from Higher Education.” *Studies in Educational Evaluation* 68(December 2020). doi: 10.1016/j.stueduc.2020.100957.

Chai, T., dan R. R. Draxler. 2014. “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.” *Geoscientific Model Development* 7(3):1247–50. doi: 10.5194/gmd-7-1247-2014.

Chu, Larry F., Chelsea A. Young, Lynn K. Ngai, Tony Cun, Ronald G. Pearl, dan Alex MacArio. 2010. “Learning management systems and lecture capture in the medical academic environment.” *International Anesthesiology Clinics* 48(3):27–51. doi: 10.1097/AIA.0b013e3181e5c1d5.

Cohn, Elchanan, Sharon Cohn, Richard E. Hult, Donald C. Balch, dan James Bradley. 1998. “The Effects of Mathematics Background on Student Learning in Principles of Economics.” *Journal of Education for Business* 74(1):18–22. doi: 10.1080/08832329809601655.

Conijn, Rianne, Chris Snijders, Ad Kleingeld, dan Uwe Matzat. 2017. “Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS.” *IEEE Transactions on Learning Technologies* 10(1):17–29. doi: 10.1109/TLT.2016.2616312.

Costa, Francisco José da, Marcelo de Souza Bispo, dan Rita de Cássia de Faria Pereira. 2018. “Dropout and retention of

undergraduate students in management: a study at a Brazilian Federal University." *RAUSP Management Journal* 53(1):74–85. doi: 10.1016/j.rauspm.2017.12.007.

Coussement, Kristof, Minh Phan, Arno De Caigny, Dries F. Benoit, dan Annelies Raes. 2020. "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model." *Decision Support Systems* 135(May):113325. doi: 10.1016/j.dss.2020.113325.

Deda, Eriona, Sokol Pacukaj, dan Luan Vardari. 2021. "Education and its role in the economic development of the country and government policies to be undertaken to increase the quality of education, the case of albania." *Journal of Educational and Social Research* 11(1):188–99. doi: 10.36941/jesr-2021-0018.

Delen, Dursun. 2010. "A comparative analysis of machine learning techniques for student retention management." *Decision Support Systems* 49(4):498–506. doi: 10.1016/j.dss.2010.06.003.

Delen, Dursun, Kazim Topuz, dan Enes Eryarsoy. 2020. "Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition." *European Journal of Operational Research* 281(3):575–87. doi: 10.1016/j.ejor.2019.03.037.

Du, Chaoming, Honglian Liu, dan Liu Wei. 2010. "On the characteristics and application of integrated e-learning." *2010 2nd International Conference on E-Business and Information System Security, EBISS2010* 659–61. doi: 10.1109/EBISS.2010.5473280.

Duarte, Edson, dan Jacques Wainer. 2017. "Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters." *Pattern Recognition Letters* 88:6–11. doi: 10.1016/j.patrec.2017.01.007.

Elgeldawi, Enas, Awny Sayed, Ahmed R. Galal, dan Alaa M. Zaki. 2021. “Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis.” *Informatics* 8(4):1–21. doi: 10.3390/informatics8040079.

Fernandes, E. 2019. “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil.” *Journal of Business Research* 94:335–43. doi: 10.1016/j.jbusres.2018.02.012.

Florea, Adrian Cătălin, dan Razvan Andonie. 2019. “Weighted Random Search for hyperparameter optimization.” *International Journal of Computers, Communications and Control* 14(2):154–69. doi: 10.15837/ijccc.2019.2.3514.

Friedman, Jerome. 2001. “Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL : http://www.” *The Annals of Statistics* 29(5):1189–1232.

Friedman, Jerome H. 2002. “Stochastic gradient boosting.” *Computational Statistics & Data Analysis* 38(4):367–78. doi: 10.1016/S0167-9473(01)00065-2.

Gerritsen, Leon. 2017. “Predicting Student Performance with Neural Networks.” (May):1–30.

Gunawan, Alexander A. S., Tania, dan Derwin Suhartono. 2016. “Developing recommender systems for personalized email with big data.” Hal. 77–82 in *2016 International Workshop on Big Data and Information Security (IWBIS)*. IEEE.

Helal, Sumyea, Jiuyong Li, Lin Liu, Esmaeil Ebrahimie, Shane Dawson, Duncan J. Murray, dan Qi Long. 2018. “Predicting academic performance by considering student heterogeneity.” *Knowledge-Based Systems* 161(July):134–46. doi: 10.1016/j.knosys.2018.07.042.

Hellas, Arto, Petri Ihantola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, dan Soohyun Nam Liao. 2018. "Predicting academic performance: A systematic literature review." *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE* 175–99. doi: 10.1145/3293881.3295783.

Hernández-Blanco, Antonio, Boris Herrera-Flores, David Tomás, dan Borja Navarro-Colorado. 2019. "A Systematic Review of Deep Learning Approaches to Educational Data Mining." *Complexity* 2019. doi: 10.1155/2019/1306039.

Hew, Khe Foon, Xiang Hu, Chen Qiao, dan Ying Tang. 2020. "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach." *Computers & Education* 145(July 2019):103724. doi: 10.1016/j.compedu.2019.103724.

Ibitoye, Ayodeji O. J., Bunmi Borokini, dan Jesujoba O. Alabi. 2019. "Knowledge Based Performance Evaluation and Predictive Model for Undergraduate Students." *Asian Journal of Research in Computer Science* 2(3):1–7. doi: 10.9734/ajrcos/2018/v2i330074.

Jenitha, T., S. Santhi, dan J. Monisha Privthy Jeba. 2021. "Prediction of Students' Performance based on Academic, Behaviour, Extra and Co-Curricular Activities." *Webology* 18(SpecialIssue1):262–79. doi: 10.14704/WEB/V18SI01/WEB18058.

Jenkins, Martin, dan Janet Hanson. 2003. *E-learning series No. 1: A guide for senior managers*. 1 ed. diedit oleh M. Jenkins dan J. Hanson. Netherlands: LTSN Generic Centre.

Joy, Johnsymol, dan Mercy Paul Selvan. 2022. "A comprehensive study on the performance of different Multi-class Classification Algorithms and Hyperparameter Tuning

Techniques using Optuna." *Proceedings of International Conference on Computing, Communication, Security and Intelligent Systems, IC3SIS 2022*. doi: 10.1109/IC3SIS54991.2022.9885695.

Kaewkiriya, Thongchai, Nattavee Utakrit, dan Hiroshi Tsuji. 2011. "Experimental evaluation of distributed e-Learning management system." *IEEE Region 10 Annual International Conference, Proceedings/TENCON* 1193–97. doi: 10.1109/TENCON.2011.6129301.

Ketui, Nongnuch, Warawut Wisomka, dan Kanitha Homjun. 2019. "Using Classification Data Mining Techniques for Students Performance Prediction." Hal. 359–63 in *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE.

Kumar, Mukesh, dan Yass Khudheir Salal. 2019. "Systematic review of predicting student's performance in academics." *International Journal of Engineering and Advanced Technology* 8(3):54–61. doi: 10.13140/RG.2.2.26667.69923.

Kumar, Mukesh, A. J. Singh, dan Disha Handa. 2017. "Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques." *International Journal of Education and Management Engineering* 7(6):40–49. doi: 10.5815/ijeme.2017.06.05.

Kumar, Pardeep. 2012. "A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems." *International Journal of Data Mining & Knowledge Management Process* 2(5):25–42. doi: 10.5121/ijdkp.2012.2503.

Kumar, Sheo, Anil Kumar Gankotiya, dan Kamlesh Dutta. 2011. "A comparative study of moodle with other e-learning

systems." *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology* 5:414–18. doi: 10.1109/ICECTECH.2011.5942032.

Labini, Mauro Sylos. 2015. "Higher education and economic welfare." *The University and the Economy: Pathways to Growth and Economic Development* (February):28–46. doi: 10.4337/9781782549499.00010.

Lau, Linda. 2003. "Institutional Factors Affecting Student Retention." *Education-Indianapolis then Chula Vista*- 124(1):126–36.

Liashchynskyi, Petro, dan Pavlo Liashchynskyi. 2019. "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS." (2017):1–11.

Lorraine, Jonathan, Paul Vicol, dan David Duvenaud. 2019. "Optimizing Millions of Hyperparameters by Implicit Differentiation." 108.

Ma, Hongwei, Xin Yang, Jianrong Mao, dan Hao Zheng. 2018. "The Energy Efficiency Prediction Method Based on Gradient Boosting Regression Tree." *2nd IEEE Conference on Energy Internet and Energy System Integration, EI2 2018 - Proceedings* 1(4). doi: 10.1109/EI2.2018.8581904.

MacKay, Matthew, Paul Vicol, Jon Lorraine, David Duvenaud, dan Roger Grosse. 2019. "Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions." *7th International Conference on Learning Representations, ICLR 2019* 1(1):1–25.

Mais Haj Qasem, Raneem Qaddoura, Bassam Hammo. 2017. "Educational Data Mining: A Review." Hal. 149–56 in *Proceedings of the New Trends in Information Technology*. Vol. 97. Amman, Jordan: The University of Jordan.

Maphosa, Mfowabo, dan Vusumuzi Maphosa. 2020. "Educational data mining in higher education in sub-saharan africa." Hal.

1–7 in *Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications*. New York, NY, USA: ACM.

Mekonen, Tesfa, Wubalem Fekadu, Tefera Chane Mekonnen, dan Shimelash Bitew Workie. 2017. "Substance Use as a Strong Predictor of Poor Academic Achievement among University Students." *Psychiatry Journal* 2017:1–9. doi: 10.1155/2017/7517450.

Mengash, Hanan Abdullah. 2020. "Using data mining techniques to predict student performance to support decision making in university admission systems." *IEEE Access* 8:55462–70. doi: 10.1109/ACCESS.2020.2981905.

Morris, David S. 2015. "Actively Closing the Gap? Social Class, Organized Activities, and Academic Achievement in High School." *Youth and Society* 47(2):267–90. doi: 10.1177/0044118X12461159.

Morris, Libby V. 2016. "Mining Data for Student Success." *Innovative Higher Education* 41(3):183–85. doi: 10.1007/s10755-016-9367-6.

Nachouki, Mirna, dan Mahmoud Abou Naaj. 2022. "Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm." *International Journal of Distance Education Technologies* 20(1):1–17. doi: 10.4018/ijdet.296702.

Pinheiro, Rómulo, Gerald Wangenge-ouma, dan Elizabeth Balbachevsky. 2015. "Institutionalized Features in Higher Education." 225–26.

Probst, Philipp, Anne Laure Boulesteix, dan Bernd Bischl. 2019. "Tunability: Importance of hyperparameters of machine learning algorithms." *Journal of Machine Learning Research* 20:1–32.

Probst, Philipp, Marvin N. Wright, dan Anne-Laure Boulesteix.

2019. “Hyperparameters and tuning strategies for random forest.” *WIREs Data Mining and Knowledge Discovery* 9(3). doi: 10.1002/widm.1301.

Putatunda, Sayan, dan Kiran Rama. 2018. “A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost.” Hal. 6–10 in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning - SPML ’18*. New York, New York, USA: ACM Press.

Putatunda, Sayan, dan Kiran Rama. 2019. “A Modified Bayesian Optimization based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting.” *2019 15th International Conference on Information Processing: Internet of Things, ICINPRO 2019 - Proceedings*. doi: 10.1109/ICInPro47689.2019.9092025.

Ragab, Mahmoud, Ahmed M. K. Abdel Aal, Ali O. Jifri, dan Nahla F. Omran. 2021. “Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques.” *Wireless Communications and Mobile Computing* 2021. doi: 10.1155/2021/6241676.

Ramaswami, Gomathy, Teo Susnjak, Anuradha Mathrani, James Lim, dan Pablo Garcia. 2019. “Using educational data mining techniques to increase the prediction accuracy of student academic performance.” *Information and Learning Science* 120(7–8):451–67. doi: 10.1108/ILS-03-2019-0017.

Rijsdijk, Jorai, Lichao Wu, Guilherme Perin, dan Stjepan Picek. 2021. “Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis.” *IACR Transactions on Cryptographic Hardware and Embedded Systems* 2021(3):677–707. doi: 10.46586/tches.v2021.i3.677-707.

Romer, David. 1993. “Do Students Go to Class? Should They?” *Journal of Economic Perspectives* 7(3):167–74. doi: 10.1257/jep.7.3.167.

Romero, Cristbal, dan Sebastin Ventura. 2010. “Educational data mining: A review of the state of the art.” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 40(6):601–18. doi: 10.1109/TSMCC.2010.2053532.

Romero, Cristóbal, Mykola Ventura, Sebastian Pechenizkiy, dan Ryan S. J. .. Baker. 2010. *Handbook of Educational Data Mining*. 1 ed. United States of America: Springer US.

Romero, Cristobal, dan Sebastian Ventura. 2013. “Data mining in education.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3(1):12–27. doi: 10.1002/widm.1075.

Romero, Cristobal, dan Sebastian Ventura. 2020. “Educational data mining and learning analytics: An updated survey.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(3):1–21. doi: 10.1002/widm.1355.

Sawant, Tejashree U. 2019. “Student Placement Prediction Model Using Gradient Boosted Tree.” 6(5):489–99.

Schratz, Patrick, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, dan Alexander Brenning. 2019. “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data.” *Ecological Modelling* 406(April 2018):109–20. doi: 10.1016/j.ecolmodel.2019.06.002.

Seow, Poh Sun, dan Gary Pan. 2014. “A Literature Review of the Impact of Extracurricular Activities Participation on Students’ Academic Performance.” *Journal of Education for Business* 89(7):361–66. doi: 10.1080/08832323.2014.912195.

Shekar, B. H., dan Guesh Dagnew. 2019. “Grid search-based hyperparameter tuning and classification of microarray cancer data.” *2019 2nd International Conference on Advanced Computational and Communication Paradigms, ICACCP 2019*. doi: 10.1109/ICACCP.2019.8882943.

Shekhar, Shashank, Adesh Bansode, dan Asif Salim. 2021. “A Comparative study of Hyper-Parameter Optimization Tools.” *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2021*. doi: 10.1109/CSDE53843.2021.9718485.

Suthar, Ms. Falguni, Hiralben Patel, dan Bhavesh Patel. 2019. “A Study on Educational Data Mining.” *International Journal for Research in Applied Science and Engineering Technology* 7(2):1089–96. doi: 10.22214/ijraset.2019.2172.

Valli Jayanthi, S., Santhi Balakrishnan, Angela Lim Siok Ching, Noor Aaqilah Abdul Latiff, dan A. M. A. Nasirudeen. 2014. “Factors Contributing to Academic Performance of Students in a Tertiary Institution in Singapore.” *American Journal of Educational Research* 2(9):752–58. doi: 10.12691/education-2-9-8.

Victoria, A. Helen, dan G. Maragatham. 2021. “Automatic tuning of hyperparameters using Bayesian optimization.” *Evolving Systems* 12(1):217–23. doi: 10.1007/s12530-020-09345-2.

Villalobos-Arias, Leonardo, dan Christian Quesada-López. 2021. “Comparative study of random search hyper-parameter tuning for software effort estimation.” Hal. 21–29 in *Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering*. New York, NY, USA: ACM.

Villalobos-Arias, Leonardo, Christian Quesada-López, Jose Guevara-Coto, Alexandra Mart\’\inez, dan Marcelo Jenkins. 2020. “Evaluating Hyper-Parameter Tuning Using Random

Search in Support Vector Machines for Software Effort Estimation." Hal. 31–40 in *Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering, PROMISE 2020*. New York, NY, USA: Association for Computing Machinery.

Weerts, Hilde J. P., Andreas C. Mueller, dan Joaquin Vanschoren. 2020. "Importance of Tuning Hyperparameters of Machine Learning Algorithms."

Witten, Ian H., Eibe Frank, dan Mark A. Hall. 2011. *Data Mining Practical Machine Learning Tools and Techniques*. Third Edit. USA: Morgan Kaufmann Publishers.

Wolaver, Amy M. 2002. "Effects of heavy drinking in college on study effort, grade point average, and major choice." *Contemporary Economic Policy* 20(4):415–28. doi: 10.1093/cep/20.4.415.

Wong, Jenna, Travis Manderson, Michal Abrahamowicz, David L. Buckeridge, dan Robyn Tamblyn. 2019. "Can Hyperparameter Tuning Improve the Performance of a Super Learner?: A Case Study." *Epidemiology* 30(4):521–31. doi: 10.1097/EDE.0000000000001027.

Yaacob, Wan Fairos Wan, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, dan Norafefah Mohd Sobri. 2019. "Supervised data mining approach for predicting student performance." *Indonesian Journal of Electrical Engineering and Computer Science* 16(3):1584–92. doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

Zhang, Jun, Qin Wang, dan Weifeng Shen. 2022. "Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library." *Chinese Journal of Chemical Engineering* 52:115–25. doi: 10.1016/j.cjche.2022.04.004.

Zhang, Zhongheng, Yiming Zhao, Aran Canes, Dan Steinberg,

dan Olga Lyashevska. 2019. “Predictive analytics with gradient boosting in clinical medicine.” *Annals of Translational Medicine* 7(7):152–152. doi: 10.21037/atm.2019.03.29.

BIOGRAFI PENULIS

Dr (c). Muhammad Arifin, S.Kom, M.Kom, MCE adalah seorang mahasiswa akhir Doktor Sistem Informasi di Universitas Diponegoro (Undip) Semarang, selain mahasiswa dia merupakan dosen pada program studi Sistem Informasi di Universitas Muria Kudus (UMK) dari tahun 2014 sampai dengan sekarang. Area penelitian yang ditekuni adalah Data Mining, Machine Learning, Educational Data Mining, E-learning, dan Sistem Informasi. Dia menyelesaikan sarjana di program studi Sistem Informasi UMK tahun 2011. Adapun gelar magister komputer diperoleh dari program studi Teknik Informatika dari Universitas Dian Nuswantoro Semarang tahun 2013. Sedangkan gelar sertifikasi Microsoft Certified Educator (MCE) diperoleh pada tahun 2018. Dia telah menghasilkan 61 artikel yang terdiri dari 9 artikel terindek Scopus, 51 artikel terindek di GoogleScholar dan 2 artikel terindek di Web of Science (WOS). Keanggotaan ilmiah yang diikuti adalah Aptikom, Asosiasi Peneliti Data Science Indonesia, EDM dan UIKTEN Serbia.

Prof. Dr. Widowati, S.Si, M.Si. adalah Guru besar di Universitas Diponegoro (Undip) Semarang, beliau merupakan Dekan sekaligus dosen pada Fakultas Sains dan Matematika (FSM) Undip. Beliau juga seorang dosen pada Doktor Sistem Informasi di Undip. Area penelitian yang ditekuni adalah Modeling, Optimization, Systems and Control. Beliau menyelesaikan sarjana strata satu di program studi Matematika di Undip tahun 1993. Adapun gelar magister Matematika diperoleh dari Institut Teknologi Bandung (ITB), tahun 2000. Sedangkan

gelar Doktor matematika juga diperoleh dari ITB, tahun 2005. Sedangkan pengukuhan guru besar disematkan sejak tahun 2014. Beliau telah menghasilkan 145 artikel ilmiah, 10 buku, 30 paten, dan 32 projects. Keanggotaan ilmiah yang diikuti adalah IndoMS (Indonesian Mathematical Society), MIPANe dan UIKTEN Serbia.

Farikhin, M.Si., Ph.D adalah dosen di Universitas Diponegoro (Undip) Semarang, beliau merupakan Wakil Dekan I pada Fakultas Sains dan Matematika (FSM) Undip. Beliau juga seorang dosen pada Doktor Sistem Informasi Undip. Area penelitian yang ditekuni adalah Numerical Optimization, Approximation, and Mathematics for decision science. Beliau menyelesaikan sarjana strata satu di program studi Matematika di Universitas Gadjah Mada UGM) tahun 1998. Adapun gelar magister Matematika diperoleh juga dari UGM pada tahun 2002. Sedangkan gelar Ph.D diperoleh dari Universiti Malaysia Terengganu, Malaysia, pada tahun 2011. Beliau telah menghasilkan 61 artikel ilmiah, yang telah dipublikasikan baik di prosiding maupun jurnal. Keanggotaan ilmiah yang diikuti Indonesian Mathematics Society (IndoMS) and Indonesian Operation resaerch Society (IORA) dan UIKTEN Serbia.

Buku ini membahas peran Educational Data Mining (EDM) dalam mengungkap nilai-nilai tersembunyi dari data yang dihasilkan oleh proses pendidikan. Prediksi performa akademik mahasiswa merupakan bagian sangat penting. Tujuannya adalah untuk membantu lembaga pendidikan mengidentifikasi mahasiswa yang berisiko putus sekolah (DO) atau lulus terlambat.

Buku ini terbagi menjadi tiga bagian utama, yaitu:

1. Membahas definisi dan tujuan EDM, pentingnya prediksi performa akademik mahasiswa, kontribusi dan manfaat, perkembangan penelitian, faktor-faktor yang mempengaruhi, jenis data, metode serta teknik. Pengumpulan dan persiapan, analisis eksplorasi data, pemilihan fitur, penggunaan metode gradient boosting trees disertai contoh soal dan penyelesaiannya. Evaluasi MSE, MAE, dan RMSE, validasi model menggunakan cross-validation.
2. Studi kasus dan penyelesaian prediksi performa akademik mahasiswa. Data akademik dan non-akademik, model prediksi gradient boosting trees, GridsearchCV digunakan untuk penyetelan hyperparameter. Penggunaan software Python dan RapidMiner. Bagian ini juga menjelaskan secara detail ekstraksi log LMS.
3. Kesimpulan, tantangan dan peluang yang dihadapi, serta arah penelitian dan pengembangan selanjutnya.

Buku ini sangat cocok untuk mahasiswa jurusan komputer, termasuk Sistem Informasi, Teknik Informatika, Manajemen Informatika, dan jurusan sejenisnya. Namun, buku ini juga sangat bermanfaat bagi mahasiswa yang ingin memantau performa akademik mereka sendiri, serta berguna bagi perguruan tinggi dalam memantau prestasi mahasiswa mereka.