Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Pemilihan Program Studi Sarjana

Tesis

Rio Rizki Aryanto
19917033

Tesis diajukan sebagai syarat untuk meraih gelar Magister Komputer
Konsentrasi Sains Data
Program Studi Informatika Program Magister
Fakultas Teknologi Industri
Universitas Islam Indonesia
2021

Lembar Pengesahan Pembimbing

Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Pemilihan Program Studi Sarjana

Rio Rizki Aryanto
19917033

Yogyakarta, Desember 2021

Pembimbing

Ahmad M. Raf'ie Pratama, S.T., M.I.T., Ph.D.

Lembar Pengesahan Penguji

Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Pemilihan Program Studi Sarjana

Rio Rizki Aryanto
19917033

Yogyakarta, Desember 2021

Tim Penguji,

Ahmad M. Raf'ie Pratama, S.T., M.I.T., Ph.D.
Ketua Penguji

Dr. R Teduh Dirgahayu, S.T., M.Sc., Ph.D.
Penguji 1

Dhomas Hatta F, S.T., M.Eng., Ph.D.
Penguji 2

Mengetahui,

Ketua Program Studi Informatika Program Magister
Universitas Islam Indonesia

Izzati Muhimmah, S.T., M.Sc., Ph.D.

Abstrak

Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Pemilihan Program Studi Sarjana

Pemilihan program studi jenjang sarjana menjadi salah satu tantangan bagi calon mahasiswa, yaitu siswa tingkat akhir Sekolah Menengah Atas (SMA) dan sederajat. Keputusan fase tersebut dapat berimbas tidak hanya pada kelancaran studi namun juga terhadap jalur karir setamat studi. Adanya pandemi Covid-19 dan pembatasan yang dilakukan pemerintah Indonesia menjadikan tantangan tersebut menjadi semakin sulit dan menuntut pihak perguruan tinggi bertindak lebih inovatif untuk menjangkau calon mahasiswanya. Salah satu inovasi yang dapat dilakukan adalah inisiasi sistem rekomendasi program studi sarjana. Sistem tersebut dapat membantu calon mahasiswa untuk mengetahui program studi yang cocok dengan karakteristik mereka. Sistem rekomendasi serupa telah banyak ditemukan di penelitian sebelumnya, akan tetapi kebanyakan masih menggunakan model berbasis aturan (*rule-based*) dan fuzy. Masih belum banyak ditemukan sistem rekomendasi yang mengimplementasikan model berbasis pembelajaran mesin (*machine learning*). Penelitian ini bertujuan untuk mengeksplorasi implementasi sains data khususnya terkait model *machine learning* pada sistem rekomendasi program studi. Implementasi tersebut diwujudkan dalam bentuk model klastering dan model klasifikasi. Model klastering digunakan untuk menyeleksi kelompok mahasiswa yang akan digunakan sebagai data latih pada sistem sedangkan model klasifikasi digunakan sebagai model yang memberikan hasil rekomendasi kepada pengguna. Studi komparasi penelitian akan melibatkan beberapa model klastering (KMeans, Agglomerative, Birch dan DBSCAN), model klasifikasi dengan pendekatan *single-stages* dan *multi-stages*, metode (*multinomial logictic regressions*, *random forest* dan *support vector machine*), dan skema preparasi *dataset* (dengan atau tanpa label berbasis IPK). Penelitian ini menemukan bahwa model KMeans merupakan model klastering untuk digunakan sebagai alat bantu seleksi kelompok mahasiswa, sedangkan model terbaik pada sistem rekomendasi adalah model klasifikasi dengan pendekatan *single-stage* dan metode *random forest*.

Kata kunci

studi komparasi, sistem rekomendasi program studi, pembelajaran berbasis mesin, *single-stage*, *multi-stages*

Abstract

Comparative Study of Machine Learning Classification Model in Recommendation System for Undergraduate Study Program Selection

Selecting college major is one of many challenges for prospective students. The choice may not only affect on the study itself but even more to student's career path afterward. It is seems to be to be harder due to Covid-19 pandemic, especially here in Indonesia with the social distancing and other restrictions. That makes universities should take an innovative way to be able to reach its prospective students. One kind of innovation that can be pursued is initialize a recommendation system for undergraduate study program. This system can assist prospective students determine what majors most suitable for them. Similar recommendation systems have been found in previous research, however mostly used either rule-based or fuzzy model. Yet found recommendation systems using machine learning approach. This research aims to explore how data science, specifically machine learning models such clustering and classification can be implemented in the recommendation system. Active undergraduate students' data from Universitas Islam Indonesia was used to train the recommendation system. The comparative study was conducted to compare clustering models (i.e., KMeans, agglomerative, birch and DBSCAN), classification model using either single-stage or multi-stages approach, algorithms (i.e., multinomial logistic regression, random forest and support vector machine), as well as the preparation scenarios (i.e., with or without GPA-based). This research found that the KMeans model work best among other clustering models and that classification with single-stage approach using random forest perform best across any scenarios.

Kata kunci

comparative study, recommendation system, college major selection, machine learning, single-stage, multi-stages

Pernyataan Keaslian Tulisan

Dengan ini saya menyatakan bahwa dokumen tesis ini merupakan tulisan asli dari penulis, dan tidak berisi material yang telah diterbitkan sebelumnya, juga tulisan dari penulis lain terkecuali referensi yang telah disebutkan dalam dokumen. Apabila terdapat kontribusi dari penulis lain, maka nama penulis tersebut telah disebutkan secara eksplisit di dalam dokumen tesis.

Saya menyatakan bahwa segala kontribusi dari pihak lain termasuk di dalamnya penyedia sumber data penelitian, bantuan analisis data, dan prosedur teknik lainnya yang terdapat pada penelitian telah disebutkan secara eksplisit dalam dokumen.

Segala bentuk hak cipta yang terdapat dalam dokumen tesis ini berada dalam kepemilikan pemilik hak cipta masing-masing. Apabila dibutuhkan, penulis telah mendapatkan izin dari pemilik hak cipta untuk menggunakan ulang materialnya di dalam penelitian tesis.

Yogyakarta, Desember 2021

Rio Rizki Aryanto

Daftar Publikasi

<u>Publikasi yang menjadi bagian dari tesis:</u>

Pratama, A. R., Rio Rizki Aryanto, & Lizda Iswari. (2021). Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Program Studi . *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(5), 853 - 862.

Kontributor	Jenis Kontribusi
Rio Rizki Aryanto, S.Si.	Mendesain eksperimen (65%) Menulis dan mengedit *paper* (60%)
Ahmad M. Raf'ie Pratama, , S.T., M.I.T., Ph.D.	Mendesain eksperimen (20%) Menulis dan mengedit *paper* (30%)
Lizda Iswari, S.T., M.Sc.	Mendesain eksperimen (15%) Menulis dan mengedit *paper* (10%)

Halaman Kontribusi

Penulis berterima kasih kepada tim Badan Sistem Informasi (BSI) dan tim Penerimaan Mahasiswa Baru (PMB) Universitas Islam Indonesia yang telah menyediakan data mahasiswa jenjang sarjana untuk digunakan dalam penelitian tesis.

Halaman Persembahan

Karya ini penulis persembahkan untuk semua pihak baik yang terlibat langsung maupun tidak langsung dalam semua aspek penyusunan penelitian tesis ini. Kepada istri penulis, keluarga, juga rekan-rekan mahasiswa sains data dan dosen-dosen pengajar di Universitas Islam Indonesia. Tidak lupa, karya ini juga penulis persembahkan untuk semua pihak dalam ruang lingkup Universitas Islam Indonesia. Semoga, karya tesis penulis dapat memberikan kontribusi ilmu yang nantinya mampu memacu perkembangan ilmu sains data ke depannya.

Kata Pengantar

Assalamualaikum, Wr. Wb.

Alhamdulillah, puji syukur penulis panjatkan kepada Allah SWT atas segala rahmat dan hidayahnya sehingga penulis mampu menyelesaikan penelitian tesis ini. Penulis mengucapkan banyak terima kasih kepada semua pihak yang berkontribusi selama masa studi penulis sampai dengan proses penyusunan tesis ini. Untuk itu, penulis hendak mengucapkan terima kasih kepada:

1. Agustina Anggraeni, selaku istri yang telah menemani dan tanpa lelah memberikan dukungan moral kepada penulis selama masa studi magister
2. Ibunda Muji Rahayu dan Sutri, ayahanda Sujadi dan alm. Samsudi, yang selalu mendukung setiap keputusan penulis
3. Bapak Ahmad M. Raf'ie Pratama dan Ibu Lizda Iswari selaku dosen pembimbing, yang telah membimbing penulis baik pada tesis ini maupun penelitian lainnya
4. Bapak Dhomas Hatta Fudholi dan Bapak R. Teduh Dirgahayu selaku dosen penguji tesis, yang telah memberikan kritikan dan saran
5. Dosen Magister Informatika khususnya konsentrasi sains data, yang telah membagi ilmu kepada penulis selama masa studi
6. Rekan-rekan Magister Informatika konsentrasi sains data angkatan ke-4 yaitu Mbak Siwi, Mas Eko dan Mas Wahyu, selaku teman seperjuangan
7. Departemen Badan Sistem Informasi (BSI) dan Penerimaan Mahasiswa Baru (PMB) Universitas Islam Indonesia yang telah menyediakan data penelitian
8. Teman-teman divisi analitik & data perusahaan Thrive yaitu Rifki Anisa dan Samsudin membantu penulis membagi waktu selama bekerja dan masa studi

Semoga semua bantuan dan dukungan yang diberikan mendapatkan balasan yang setimpal dari Allah SWT. Terakhir, semoga penelitian ini dapat memberikan manfaat kepada khalayak luas dan mampu memberikan kontribusi ilmu yang baru.

Wassalamualaikum, Wr. Wb.

Yogyakarta, Desember 2021

Rio Rizki Aryanto

Daftar Isi

Daftar Tabel

Daftar Gambar

Glosarium

Preliminary study	- penelitian sebelumnya yang menjadi dasar dan pondasi dari penelitian ini
Streaming	- pengiriman data konten ke perangkat elektronik melalui transmisi internet secara konstan

BAB 1
Pendahuluan

1.1 Latar Belakang

Studi jenjang perguruan tinggi merupakan salah satu tujuan bagi siswa tingkat akhir Sekolah Menengah Atas (SMA) dan sederajat. Namun, sebelum melanjutkan studi jenjang tersebut calon mahasiswa dihadapkan oleh tantangan terkait proses pemilihan progam studi atau jurusan kuliah. Kesalahan terkait pemilihan program studi sendiri adalah hal umum yang sering terjadi, tidak terkecuali pada mahasiswa Indonesia. Hasil survey Indonesia Career Centre Network (ICNN) tahun 2017 yang dilansir Jawapos (Safutra, 2019) menunjukkan bahwa 87% mahasiswa di Indonesia telah merasakan hal tersebut. Kesalahan terkait pemilihan program studi tidak hanya berimbas pada kelancaran proses studi, akan tetapi juga dapat berimbas pada jenjang karir selanjutnya. Pemerhati pendidikan Yohana Elizabeth Hardjadinata, MBA melalui pernyataannya yang dikutip pada situs JPNN (Mesya, 2019) mengatakan bahwa sebanyak 71.7% pekerja di Indonesia memiliki profesi yang tidak sesuai dengan latar pendidikan jenjang perguruan tinggi. Temuan tersebut sedikit banyak menjadi bukti bahwa proses pemilihan program studi masih menjadi tantangan yang belum terselesaikan bagi mahasiswa Indonesia.

Kondisi menjadi semakin sulit akibat pandemi Covid-19. Adanya pembatasan yang diberlakukan sedikit banyak menjadikan aktivitas calon mahasiswa dalam mengumpulkan informasi terkait program studi pilihan menjadi lebih terbatas. Pihak perguruan tinggi perlu bertindak lebih inovatif untuk dapat menjangkau calon mahasiswanya. Salah satu inovasi yang dapat dilakukan adalah menyediakan sistem rekomendasi pemilihan progam studi. Sistem yang dapat digunakan oleh calon mahasiswa untuk melihat program studi apa yang cocok dengan kemampuan mereka. Penelitian terkait sistem rekomendasi serupa telah banyak ditemukan, namun kebanyakan masih menggunakan model berbasis aturan (*rule-based*) seperti pada (Marbun & Hansun, 2019), (Pare, 2013) dan (Okaviana & Susanto, 2014) atau menggunakan model fuzzy seperti pada (Permatasari, Kridalaksana, & Suyatno, 2015).

Meskipun telah banyak digunakan pada penelitian sebelumnya, akan tetapi terdapat kelemahan dari kedua model tersebut. Baik *rule-based* dan fuzzy merupakan model yang berbasis pada aturan. Pada model fuzzy aturan tersebut dipetakan pada lingkup yang lebih

luas sehingga berjalan tidak sekaku model *rule-based*. Akan tetapi, karena berbasis pada aturan, maka kedua model sangat bergantung kepada kemampuan dan tingkat pengetahuan manusia dalam mendefinisikan aturan-aturan yang diimplementasikan. Mengingat kemampuan dan pengetahuan manusia yang tidak dapat disamakan, maka performa model juga menjadi sulit untuk divalidasi. Selain itu, model berbasis aturan juga memiliki kelemahan ketika menemui data baru. Model tidak akan mampu mengenali data baru jika aturan pada model tidak diperbarui terlebih dahulu. Artinya, tingkat kepintaran model menjadi sangat bergantung kepada frekuensi aturan model diperbarui oleh manusia.

Mengingat kelemahan yang dimiliki kedua model tersebut maka penelitian ini bertujuan untuk menginisiasi sistem rekomendasi menggunakan model berbasis pembelajaran mesin (*machine learning*). Model *machine learning* sendiri adalah model yang mempelajari dan memetakan pola hubungan antar variabel pada data. Dapat dikatakan bahwa model lebih bergantung kepada ketersediaan data dibandingkan kemampuan dan pengetahuan manusia. Oleh karenanya, performa model *machine learning* menjadi lebih mudah untuk dijustifikasi. Keunggulan lain adalah model *machine learning* dapat digunakan bahkan pada data baru yang tidak pernah ditemui sebelumnya. Hal yang tidak dapat dilakukan oleh model *rule-based* dan fuzzy. Model *machine learning* secara otomatis akan mempelajari pola pada data baru dan memberikan hasil luaran yang sesuai dengan kesesuaian data. Keunggulan tersebut yang menjadi dasar penggunaan model *machine learning* pada sistem rekomendasi penelitian ini.

Pada dasarnya penelitian bertujuan untuk melihat bagaimana implementasi sains data khususnya model *machine learning* dalam implementasi sistem rekomendasi pemilihan program studi sarjana. Studi komparasi di akhir penelitian juga dilakukan untuk mendapatkan model klasifikasi terbaik. Penelitian merupakan lanjutan dari penelitian sebelumnya (*preliminary study*) oleh (Pratama, Aryanto, & Pratama, 2021) terkait sistem rekomendasi pemilihan program studi menggunakan model klasifikasi *single-stage* yang membandingkan tiga algoritma yaitu *multinomial logistic regression*, *random forest* dan *support vector machine*. Temuan pada *preliminary study* menunjukkan bahwa model klasfiikasi *single-stage* menggunakan algoritma *random forest* berhasil memberikan performa terbaik dengan akurasi sebesar 86%. Akan tetapi peneliti berasumsi terdapat beberapa aspek yang dapat diperbaiki dari *preliminary study* tersebut. Beberapa perbaikan yang diinisiasikan pada penelitian adalah dengan mengujicoba pendekatan *multi-stages* untuk dibandingkan dengan *single-stage*. Selain itu, juga menambahkan *dataset* baru yang sebelumnya tidak tersedia pada *preliminary study*. Penambahan *dataset* baru tersebut

bertujuan untuk memperdalam sisi analisis pada pengembangan model klasifikasi sistem rekomendasi.

Adapun data yang digunakan pada penelitian adalah data mahasiswa sarjana Universitas Islam Indonesia (UII). Meskipun sistem rekomendasi memiliki target pengguna adalah calon mahasiswa atau siswa tingkat SMA, model klasifikasi pada sistem rekomendasi akan dilatih menggunakan data mahasiswa. Hal tersebut untuk memastikan agar sistem rekomendasi mampu memberikan rekomendasi yang tepat. Sederhananya, model klasifikasi pada sistem rekomendasi akan dilatih menggunakan data mahasiswa sarjana UII untuk melihat kecocokan dari karakteristik dan kemampuan masing-masing individu terhadap jenis program studi yang ditempuh. Untuk mengakomodasi tujuan tersebut maka model klasifikasi sistem rekomendasi akan dilatih hanya dengan menggunakan data kelompok mahasiswa yang berhasil atau terbukti sukses beradaptasi pada program studi yang diambil. Seleksi kelompok mahasiswa yang diasumsikan sukses beradaptasi dengan program studi yang diambil tersebut akan dilakukan dengan menggunakan bantuan model *machine learning* tidak tersupervisi yaitu model klastering dan dilanjutkan dengan implementasi tahapan *semi-supervised learning*.

Terdapat dua jenis *dataset* yang digunakan pada penelitian. Pertama adalah *dataset* mahasiswa dari seluruh program studi sarjana UII, sedangkan *dataset* yang kedua berisikan data mahasiswa khusus program studi Informatika. Kedua *dataset* memiliki jumlah data dan variabilitas data poin yang berbeda. *Dataset* pertama berisikan informasi umum sekaligus data per individu semasa jenjang SMA, sedangkan *dataset* kedua berisikan data akademik per mata kuliah dari mahasiswa Informatika selama menempuh studi di UII. Adanya perbedaan dan keterbatasan data menjadi dasar bagi peneliti untuk melakukan beberapa pendekatan yang berbeda dalam proses pengembangan model klasifikasi sistem rekomendasi. Terdapat dua pendekatan yang dilakukan yaitu implementasi model *machine learning* tidak tersupervisi dan penerapan teknik *semi-supervised learning*. Kedua tahap akan digunakan pada proses penentuan kelompok mahasiswa sebagai data latih model klasifikasi sistem rekomendasi. Hal tersebut juga menjadi bentuk pembaruan dari *preliminary study*.

Model *machine learning* tidak tersupervisi yang digunakan adalah model klastering. Model akan digunakan untuk proses pengelompokan mahasiswa program studi Informatika. Variabel prediktor yang digunakan pada model adalah variabel terkait nilai akademik per mata kuliah masing-masing mahasiswa. Hal serupa telah dilakukan pada penelitian sebelumnya oleh (Ezz, 2019) yang mengembangkan sistem rekomendasi

konsentrasi program studi berdasarkan data capaian akademik mahasiswa pada tahun pertama. Model klastering pada penelitian akan memberikan hasil luaran berupa label kelas bagi masing-masing mahasiswa Informatika.

Teknik *semi-supervised learning* akan digunakan untuk memprediksi label kelas pada mahasiswa non-Informatika. Karena keterbatasan data pada *dataset*, diketahui bahwa data terkait capaian akademik selama menempuh studi perguruan tinggi tidak ditemukan pada mahasiswa non-Informatika. Artinya, proses pengelompokkan menggunakan model klastering menjadi tidak mungkin untuk dilakukan. Oleh karenanya, akan digunakan teknik *semi-supervised learning* dengan implementasi model klasifikasi *single-stage* yang selanjutnya akan disebut dengan model klasifikasi *semi-supervised learning*. Model klasifikasi tersebut akan dilatih menggunakan data mahasiswa Informatika dengan variabel target adalah label kelas yang didapatkan dari model klastering sebelumnya. Sedangkan variabel prediktor yang digunakan adalah data karakteristik dan capaian akademik jenjang SMA dari masing-masing mahasiswa. Pemilihan variabel prediktor tersebut disesuaikan dengan ketersediaan data pada *dataset*. Selain itu, peneliti juga menemukan kemiripan dari sisi capaian akademik jenjang SMA antara mahasiswa program studi Informatika dan non-Informatika. Temuan tersebut menjadi dasar peneliti dalam pemilihan variabel prediktor model klasifikasi *semi-supervised learning*.

Tahap akhir penelitian adalah melakukan studi komparasi guna menemukan model klasfikasi dengan performa terbaik. Komparasi melibatkan beberapa model klasifikasi baik *single-stage* maupun *multi-stages* dan juga algoritma seperti *multinomial logistic regression*, *random forest*, *support vector machine*. Studi komparasi juga melibatkan dua skenario preparasi *dataset* yang dengan atau tanpa menerapkan seleksi data berdasarkan jumlah Satuan Kredit Semester (SKS), nilai Indeks Prestasi Kumulatif (IPK), dan status mahasiswa. Total terdapat 9 model klasifikasi yang akan dibandingkan. 6 diantaranya merupakan representasi dari model *single-stage,* 2 model representasi model *multi-stages*, dan satu model representasi *preliminary study*. Melalui studi komparasi tersebut dapat diketahui model klasifikasi apa yang memberikan performa terbaik, sekaligus untuk mengetahui apakah terdapat improvisasi dibandingkan model dari penelitian sebelumnya.

1.2 Rumusan Masalah

Berikut adalah rumusan masalah yang akan dijawab melalui penelitian:

1. Bagaimana implementasi sains data dalam pengembangan sistem rekomendasi program studi sarjana?

2. Model klasifikasi apakah yang terbaik berdasarkan hasil studi komparasi?
3. Bagaimana performa model klasifikasi penelitian jika dibandingkan dengan model klasifikasi *preliminary study*?
4. Bagaimana hasil teknik *semi-supervised learning* dengan data capaian studi mahasiswa Informatika dalam memprediksi label kelas mahasiswa program studi lainnya?

1.3 Batasan Penelitian

Penelitian mempunyai beberapa batasan masalah mempertimbangkan ketersediaan data, diantaranya adalah sebagai berikut:

1. Ruang penelitian secara spesifik merepresentasikan kondisi Universitas Islam Indonesia
2. Jenis program studi yang digunakan pada penelitian ini adalah program studi jenjang sarjana pada Universitas Islam Indonesia
3. Pengembangan sistem rekomendasi menyesuaikan ketersediaan data mahasiswa yang dikelola oleh Badan Sistem Informasi (BSI) dan tim Penerimaan Mahasiswa Baru (PMB) Universitas Islam Indonesia
4. Beberapa kondisi yang didesain pada penelitian tidak sepenuhnya merepresentasikan kondisi yang sebenarnya ada di lapangan
5. Penelitian hanya menggunakan metode klasifikasi berbasis pembelajaran mesin

BAB 2
Tinjauan Pustaka

2.1 Landasan Teori

2.1.1 Sistem Rekomendasi

Sistem rekomendasi adalah sistem yang berfungsi memberikan rekomendasi atau preferensi dari suatu benda kepada pengguna. Menurut (Konstan & Riedl, 2012), sistem rekomendasi adalah sistem berbasis penyaringan informasi yang dikembangkan untuk mengatasi permasalahan terkait banyaknya informasi yang harus diolah. Contoh dari sistem rekomendasi dapat ditemukan pada beberapa aplikasi *streaming* seperti Spotify, Netflix dan YouTube.

Implementasi dari sistem rekomendasi tidak hanya ditemukan pada industri hiburan, namun juga pada dunia pendidikan. Sistem rekomendasi juga familiar dikenal dengan nama Sistem Pendukung Keputusan (SPK). (Grewal & Kaur, 2016) pada penelitiannya mengembangkan sistem rekomendasi untuk membantu mahasiswa dalam menentukan mata kuliah yang akan diambil dengan memanfaatkan model *machine learning* tidak tersupervisi yaitu model klastering, dan tersupervisi seperti model klasifikasi. Penelitian serupa juga dilakukan oleh (Parameswaran, Venetis, & Garcia-Molina, 2011), dimana sistem mampu memberikan rekomendasi mata kuliah dengan menyesuaikan kebutuhan kurikulum dari masing-masing mahasiswa. Hal tersebut dilakukan mengingat setiap mahasiswa memiliki kebutuhan yang berbeda guna menyelesaikan program studi mereka. Implementasi model *machine learning* pada sistem rekomendasi juga ditemukan pada penelitian (Andriani, 2013) dan (Kumalasari & Susanto, 2019). (Andriani, 2013) mengembangkan sistem rekomendasi yang digunakan pihak perguruan tinggi dalam menentukan calon penerima beasiswa. Sistem menggunakan model *ensemble* yaitu *decision tree* dengan attribut tunggal yaitu nilai Indeks Prestasi Kumulatif (IPK). Model tersebut berhasil memberikan nilai akurasi sebesar 71.43% dengan nilai kurva *Receiver Operating Characteristic* (ROC) sebesar 0.660. (Kumalasari & Susanto, 2019) pada penelitiannya mengembangkan sistem rekomendasi karir atau profesi bagi mahasiswa jurusan teknik Informatika. Model pada sistem rekomendasi tersebut melihat kecocokan antara kemampuan mahasiswa dengan kualifikasi yang dibutuhkan pada masing-masing jenis pekerjaan. Sistem rekomendasi menggunakan model *collaborative filtering* dengan metode *K-Nearest Neighbour* (KNN).

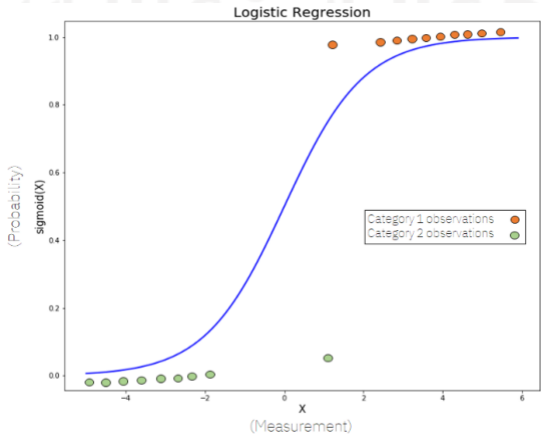
Dari penelitian-penelitian tersebut dapat diketahui bahwa sudah cukup banyak ditemukan implementasi sistem rekomendasi pada dunia pendidikan dengan ruang lingkup yang cukup beragam. Sistem rekomendasi tidak hanya mempunyai target murid, siswa atau mahasiswa namun juga terdapat sistem yang oleh pihak penyedia jasa pendidikan seperti perguruan tinggi. Beberapa dari sistem tersebut juga telah mengimplementasikan model *machine learning* dan bahkan memberikan performa yang cukup baik seperti ditunjukkan pada penelitian (Andriani, 2013).

2.1.2 *Supervised Machine Learning*

Supervised machine learning (SML) adalah metode atau algoritma *machine learning* dimana data latih memiliki variabel target. Dalam konteks klasifikasi, algoritma SML digunakan untuk memprediksi kelas target dari data masukan yang diberikan. Menurut (Kotsiantis, 2007) proses klasifikasi menggunakan SML merupakan salah satu teknik yang paling banyak digunakan pada sistem cerdas (*intelligent system*). Beberapa contoh metode SML yang cukup populer digunakan pada klasifikasi antara lain adalah *logistic regression, support vector machine*, dan *random forest.*

2.1.3 *Multinomial Logistic Regression*

Logistic regression adalah metode klasifikasi yang pada dasarnya menggunakan fungsi logistik berupa kurva sigmoid untuk memodelkan probabilitas antar kelas pada variabel target. Metode ini bekerja dengan mencari batas (*boundary*) yang mampu memisahkan antar kelas pada variabel target. Gambar 2-1 menunjukkan ilustrasi kurva sigmoid pada model *logistic regression*.

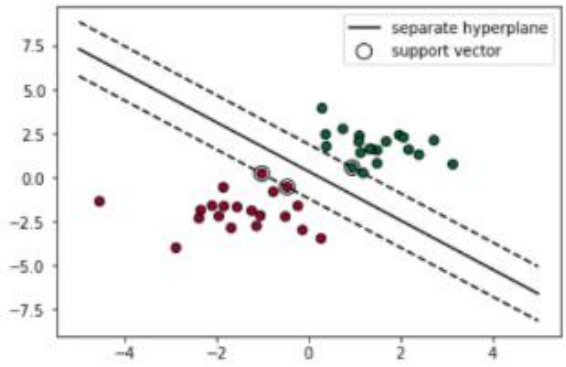


Gambar 2-1 Ilustrasi Kurva Sigmoid *Logistic Regression*

Logistic regression biasanya digunakan pada kasus *binary classification* dimana variabel target hanya mempunyai 2 kelas. Akan tetapi metode ini tetap dapat digunakan pada kasus *multi-class* dengan mengimplementasikan pendekatan tambahan seperti *one-versus-one* (OVO) atau *one-versus-all* (OVA). *Logistic regression* pada kasus *multi-class* dikenal dengan nama *multinomial logistic regression* (MLR). Metode MLR dapat digunakan ketika variabel target memiliki banyak kelas yang jenisnya bukan ordinal, artinya tidak ada unsur urutan dalam kelas tersebut (Kwak & Clayton-Matthews, 2002).

2.1.4 *Support Vector Machine*

Support vector machine (SVM) adalah metode klasifikasi *machine learning* yang cukup populer digunakan untuk kasus klasifikasi (Cortes & Vapnik, 1995). Ide dari metode SVM adalah menemukan pemisah (*hyperplane*) yang memaksimalkan jarak antar kelas (Kancherla, Bodapati, & Veeranjaneyulu, 2019). Ilustrasi konsep *hyperplane* dapat dilihat pada Gambar 2-2.

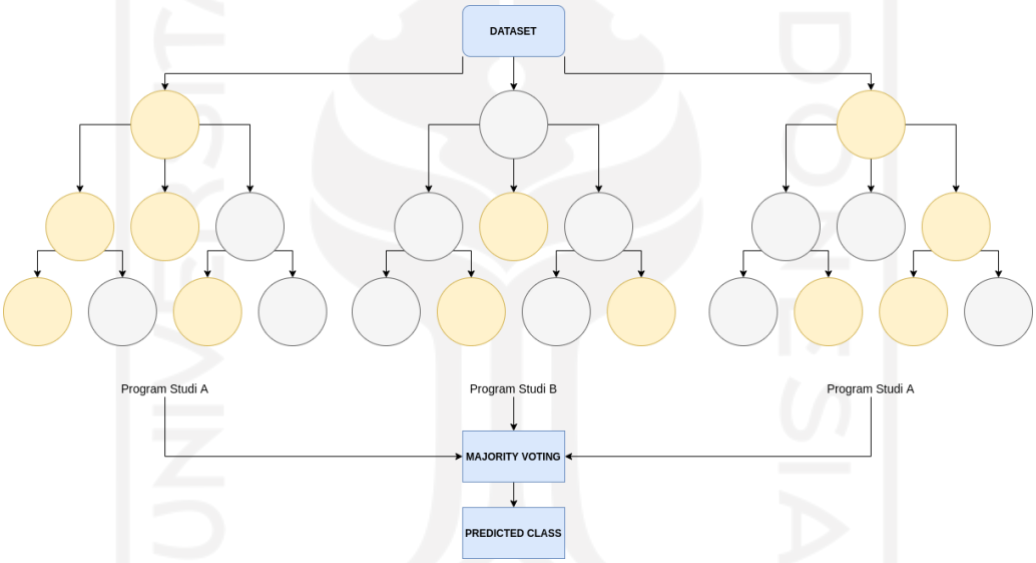


Gambar 2-2 Konsep *Hyperplane Support Vector Machine*

Hyperplane adalah sebuah pembatas yang memisahkan dan mengklasifikasikan data menjadi kelas berbeda, sedangkan *support vector* adalah titik dari masing-masing kelas yang memiliki jarak terdekat dengan *hyperplane*. Pada kasus *non-linear, hyperplane* model SVM tidak berbentuk garis linier, namun menyesuaikan distribusi data. SVM biasanya digunakan pada kasus *binary classification* sama seperti *logistic regression*. Pendekatan OVO atau OVA dapat digunakan ketika hendak menggunakan SVM pada kasus *multi-class*.

2.1.5 *Random Forest*

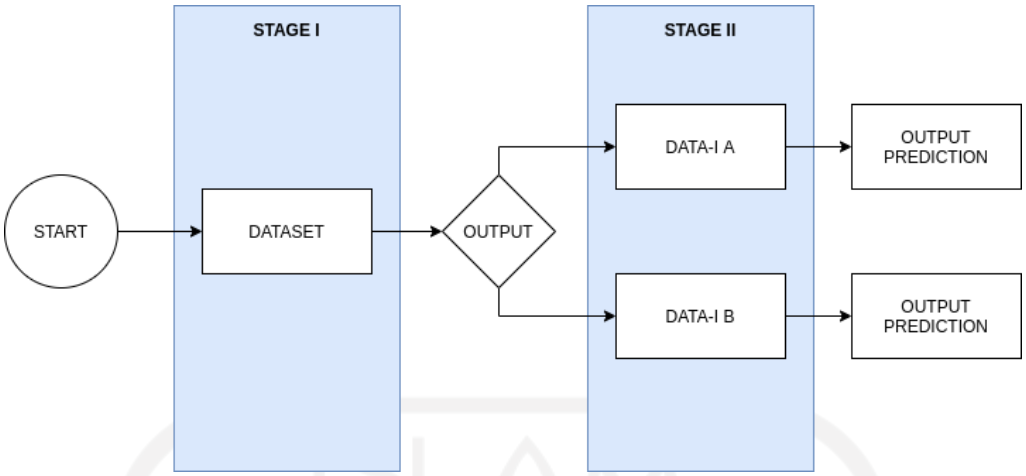
Random forest adalah metode klasifikasi yang merupakan pengembangan dari *decision tree*. *Decision tree* sendiri adalah suatu teknik non-parametrik yang dapat digunakan untuk regresi maupun klasifikasi. Metode ini menggunakan keseluruhan data dan seluruh atribut untuk mendesain 1 pohon keputusan, sedangkan *random forest* akan membentuk lebih dari 1 pohon keputusan seperti terlihat pada Gambar 2-3. Masing-masing pohon pada metode *random forest* tidak akan menggunakan keseluruhan atribut, namun secara acak (*random*) mengambil beberapa atribut untuk kemudian digunakan dalam pembentukan pohon keputusan. Untuk mendapatkan hasil prediksi, metode *random forest* melakukan pengambilan suara (*voting*) yang mengikutsertakan keseluruhan pohon keputusan yang dimiliki (Pal, 2005).



Gambar 2-3 Ilustrasi *Random Forest*

2.1.6 *Multi-Stages Classification*

Multi-stages classification adalah pendekatan pada model klasifikasi dimana dalam prosesnya data dipecah menjadi kelompok berbeda pada beberapa tahapan atau dikenal dengan *stage*. Dalam pembagian tersebut biasanya tidak ada data yang tumpang tindih. Artinya setiap kelompok pada masing-masing *stage* memiliki anggota yang berbeda. Ilustrasi model klasifikasi *multi-stages* dapat dilihat pada Gambar 2-4.



Gambar 2-4 Contoh Diagram Alir Model *Multi-Stage*

Pendekatan *multi-stages* biasanya digunakan ketika *dataset* memiliki jumlah kelas yang banyak (*multi class*). Pada kasus *multi class*, sering kali ditemukan ketidakseimbangan jumlah data pada masing-masing kelas. Artinya, terdapat suatu kelas dengan jumlah data yang jauh lebih banyak atau jauh lebih sedikit dibandingkan kelas lainnya. Dengan pendekatan *multi-stages*, maka model klasifikasi mampu mengenali secara spesifik karakteristik data pada kelas atau grup yang diasumsikan sama. Selain itu pendekatan *multi-stages* juga dapat digunakan untuk mereduksi jumlah atribut model klasifikasi, seperti yang dilakukan oleh (Patil & Atique, 2014) dan (Poorna & Nair, 2019).

Arsitektur dari pendekatan *multi-stages* sendiri cukup beragam. Tidak aturan baku atau *rule of thumbs* terkait berapa jumlah *stage* yang harus digunakan, atau metode atau algoritma apa yang harus diimplementasikan. (Salgado, et al., 2017) dalam penelitiannya mengklasifikasikan data klinis pada suatu rumah sakit di Portugal dengan menggunakan model klasifikasi *multi-stages*. Terdapat dua *stages* pada model yang menggunakan model *logistic regression* pada *stage* pertama dan model fuzzy Takeno-Sugeno pada *stage* akhir. (Isler, Narin, Ozer, & Perc, 2018) menggunakan beberapa kombinasi model seperti *K-Nearest Neighbour* (K-NN), *latent discriminant analysis* (LDA) dan *support vector machine* (SVM) dalam membangun model *multi-stages* untuk mendiagnosis penyakit gagal jantung kongestif.

Metode klasifikasi yang diimplementasikan pada model *multi-stages* juga tidak harus berbasis *machine learning*. (Mendes, Togelius, & Coelho, 2020) pada penelitiannya menggunakan metode *deep learning* dengan teknik *transfer learning* untuk mengatasi jumlah data latih yang terlalu sedikit. Namun sebagai catatan, Mendes juga menambahkan bahwa untuk *dataset* dengan ukuran yang tidak terlalu besar, metode *non-deep learning*

seperti *logistic regression*, *support vector machine*, dan *random forest* terbukti memberikan performa yang lebih baik.

Penelitian sebelumnya juga menyertakan temuan menarik terkait performa dari model *multi-stages*. (Salgado, et al., 2017) menyatakan bahwa model *multi-stages* terbukti memberikan akurasi prediksi yang lebih baik dibandingkan model *single-stage*, terutama dalam aspek sensitivitas (*sensitivity*). Model *multi-stages* pada penelitian (Isler, Narin, Ozer, & Perc, 2018) bahkan mendapatkan nilai akurasi tertinggi sebesar 98.8% ketika menggunakan total 8 kombinasi pada model klasifikasinya. Tidak hanya akurasi keseluruhan, namun peningkatan akurasi juga terlihat pada masing-masing *stage*. Hal tersebut terlihat pada model *multi-stages* penelitian (Poorna & Nair, 2019). Hal tersebut dapat terjadi dikarenakan data pada masing-masing *stage* sudah lebih spesifik dan seragam. Selain dari sisi akurasi, model *multi-stages* juga dinilai lebih baik dalam mendeteksi kejadian (*event*) dengan jumlah kemunculan sedikit (Senator, 2005). Sederhananya, model *multi-stages* bekerja dengan cara merampingkan proses klasifikasi dengan menggunakan data yang lebih terkelompok pada setiap *stage*-nya, sehingga berpotensi memberikan performa yang lebih baik dibandingkan model *single-stage*.

2.2 Kajian Pustaka

Kajian pustaka digunakan sebagai acuan dari penelitian yang dilakukan. Kajian pustaka dilakukan dengan memaparkan hasil penelitian dan ulasan kritis dari penulis terkait topik implementasi model klasifikasi pada sistem rekomendasi. Hasil kajian pustaka dapat dilihat pada Tabel 2-1.

Tabel 2-1 Ulasan Kritis Sistem Rekomendasi dan *Multi-Stages Classification*.

No.	Sub Tema	Keywords	Ulasan Kritis	Pustaka
1	Sistem rekomendasi	*Recommender System Survey*	Sistem rekomendasi berkembang berawal dari pendekatan tradisional seperti *content-based data* hingga diprediksi akan menggunakan data *Internet of Things* (IoT) di masa depan	(Bobadilla, Ortega, Hernando, & Gutierrez, 2013)
2	Sistem rekomendasi pilihan karir	*Recommendation System of Information Technology Jobs using Collaborative Filtering Method Based on LinkedIn Skills Endorsement*	Metode berbasis pembelajaran mesin digunakan untuk memberikan rekomendasi pilihan karir	(Kumalasari & Susanto, 2019)
3	Sistem rekomendasi berbasis *decision tree*	Sistem Pendukung Keputusan Berbasis *Decision Tree* dalam Pemberian Beasiswa	*Single-stages classification* menggunakan metode *decision tree* dan satu atribut yaitu nilai IPK. Sistem rekomendasi memberikan nilai akurasi sebesar 71.43% dan nilai kurva ROC 0.660	(Andriani, 2013)
4	Desain *Multi-stages classification*	*Multi-Stage Classification*	Arsitektur *multi-stages classification* mempunyai performa akurasi lebih baik dibandingkan *single-stage classification*, selain itu juga bekerja lebih baik dalam mengenali kejadian atau *event* dengan jumlah kemunculan sedikit.	(Senator, 2005)
5	*Multi-stages classification* menggunakan metode *deep learning*	*Multi-Stages Transfer Learning with an Application to Selection Process*	Penelitian menunjukkan bahwa metode *non-deep learning* seperti *logistic regression, support vector machine* dan *random forest* bekerja lebih baik pada dataset dengan ukuran tidak terlalu besar.	(Mendes, Togelius, & Coelho, 2020)

2.3 *Preliminary Study*

Penelitian terkait sistem rekomendasi pemilihan program studi sarjana pada ruang lingkup Universitas Islam Indonesia (UII) sudah pernah dilakukan oleh (Pratama, Aryanto, & Pratama, 2021). Penelitian menggunakan model klasifikasi dengan pendekatan *single-stage* dan membandingkan beberapa algoritma klasifikasi *machine learning* seperti *multinomial logistic regression* (MLR), *random forest* (RF) dan *support vector machine* (SVM). Perbandingan dari ketiga alrogitma klasifikasi pada *preliminary study* dapat dilihat pada Tabel 2-2.

Tabel 2-2 Perbandingan Performa Model Klasifikasi *Preliminary Study*

Model	*Avg. Accuracy*	*Avg. F1-score*	*Avg.* AUC ROC	*Avg. Log loss*
MLR	0.21	0.19	0.78	2.58
RF	0.86	0.84	0.97	0.66
SVM	0.22	0.17	0.79	2.58

Berdasarkan beberapa metrik evaluator terlihat bahwa model klasifikasi menggunakan algoritma *random forest* memiliki performa terbaik. Namun performa yang buruk ditunjukkan pada model klasifikasi menggunakan algoritma *multinomial logistic regression* dan *support vector machine*. Dari investigasi yang dilakukan ditemukan bahwa model klasifikasi menggunakan algoritma tersebut memiliki akurasi prediksi yang cukup baik hanya pada beberapa program studi tertentu. Akibatnya, performa keselurhan model dapat dikatakan masih jauh dari kata memuaskan. Temuan ini sekaligus menunjukkan bahwa model klasifikasi *preliminary study* masih kesulitan dalam mengatasi kasus *multi-class* yang diangkat pada penelitian.

Temuan lain yang menjadi sorotan adalah terkait proses seleksi kelompok mahasiswa sebagai data latih model klasifikasi sistem rekomendasi. *Preliminary study* melakukan seleksi data hanya berdasarkan beberapa data poin seperti status mahasiswa, jumlah Satuan Kredit Semester (SKS) dan nilai Indeks Prestasi Kumulatif (IPK). Peneliti berasumsi bahwa proses tersebut masih dapat dikembangan dengan memanfaatkan data poin lain misal data capaian akademik mahasiswa pada mata kuliah tertentu seperti yang dilakukan oleh (Ezz, 2019). Oleh karenanya, penelitian ini akan menambahkan data poin terkait capaian akademik mahasiswa yang sebelumnya tidak tersedia pada *preliminary study*.

BAB 3
Metodologi

3.1 Data

Data yang digunakan pada model klasifikasi sistem rekomendasi adalah data mahasiswa jenjang sarjana Universitas Islam Indonesia (UII). Data didapatkan dari tim Badan Sistem Informasi (BSI) dan Penerimaan Mahasiswa Baru (PMB). Kedua sumber tersebut menyediakan data dengan karakteristik yang berbeda. Data dari BSI berisikan informasi mahasiswa seperti capaian akademik, jumlah Satuan Kredit Semester (SKS), nilai Indeks Prestasi Kumulatif (IPK), status mahasiswa dan masih banyak lagi. Sedangkan data PMB berisikan informasi dari mahasiswa selama mereka duduk di bangku Sekolah Menengah Atas (SMA) atau sederajat. Informasi tersebut antara lain jenis sekolah, jurusan yang diambil, nilai per mata pelajaran SMA. Dari kedua sumber data tersebut, didapatkan dua jenis *dataset* yang berbeda. Kedua *dataset* tersebut akan disebut dengan nama **DB1** dan **DB2**.

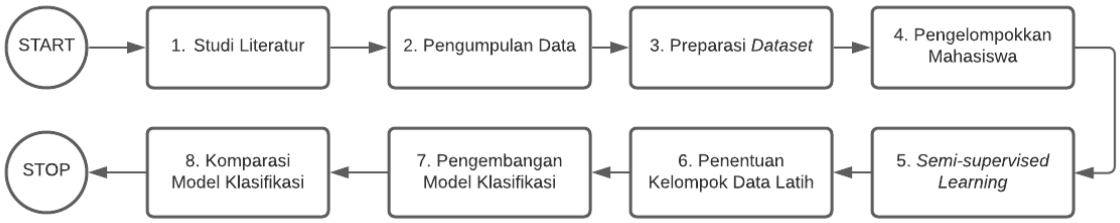
Dataset DB1 merupakan hasil gabungan dari data BSI dan PMB. Pada DB1 terdapat beberapa informasi mahasiswa baik semasa studi di perguruan tinggi maupun semasa studi pada jenjang SMA. Selain itu DB1 juga memiliki data mahasiswa dari seluruh program studi sarjana UII. Berbeda dengan DB2 yang memiliki data mahasiswa hanya dari program studi Informatika saja. *Dataset* DB2 sendiri secara spesifik menyediakan informasi terkait capaian akademik mahasiswa selama menempuh program studi Informatika di UII. Beberapa data poin pada kedua *dataset* tersebut dituliskan secara detail pada Tabel 3-1 berikut. Data poin tersebut nantinya akan banyak digunakan selama proses pengembangan sistem rekomendasi program studi sarjana.

Tabel 3-1 Data Poin per *Dataset*

Dataset	Jenis data poin
DB1	• Jenis kelamin • Jenis SMA • Jurusan SMA • Hobi • Nilai rerata mata pelajaran SMA seperti matematika, bahasa Indonesia, bahasa Inggris, fisika, kimia, biologi, geografi, sejarah, ekonomi, agama dan ketrampilan/keahlian kejuruan (11 variabel numerik) • Nomor Induk Mahasiswa (NIM) • Nilai Indeks Prestasi Kumulatif (IPK) • Jumlah Satuan Kredit Semester (SKS) • Status mahasiswa • Jenis program studi sarjana
DB2	• NIM • Nilai akademik per mata kuliah program studi Informatika • Kode semester • Kode kurikulum

3.2 Langkah-langkah Penelitian

Penelitian dimulai dari studi literatur, pengumpulan data, preparasi dataset, pengelompokkan mahasiswa, *semi-supervised learning*, penentuan kelompok data latih, pengembangan model klasifikasi dan diakhiri dengan melakukan studi komparasi. Diagram alir penelitian dapat dilihat pada Gambar 3-1.



Gambar 3-1 Diagram Alir Langkah-langkah Penelitian

3.2.1 Studi Literatur

Studi literatur dilakukan untuk mengetahui bagaimana sains data terutama model *machine learning* dapat diimplementasikan pada sistem rekomendasi program studi. Secara detail, studi literatur difokuskan terhadap beberapa tema seperti model *machine learning* tidak tersupervisi (model klastering), *semi-supervised learning*, dan model tersupervisi yaitu klasifikasi dengan beragam pendekatan (*single-stage* dan *multi-stages*). Selain itu studi literatur juga mencakup jenis algoritma klastering dan klasifikasi yang nantinya akan digunakan pada penelitian. Studi literatur tidak terbatas hanya pada ruang publikasi ilmiah, namun juga mengambil referensi dari *website*, artikel, video dan sumber-sumber lainnya.

3.2.2 Pengumpulan Data

Pengumpulan data dilakukan melalui prosedur resmi yang ditujukan kepada departemen BSI dan PMB UII selaku pengelola data mahasiswa. Data yang diminta telah disesuaikan dengan kebutuhan penelitian dan diseleksi berdasarkan hasil studi literatur sebelumnya. Demi menjaga privasi mahasiswa, data pada penelitian akan dianonimkan sehingga tidak akan diketahui pemilik dari data tersebut.

3.2.3 Preparasi *Dataset*

Preparasi *dataset* bertujuan untuk menyiapkan data sebelum digunakan pada tahap pengembangan model klasifikasi sistem rekomendasi. Mengingat penelitian memiliki dua jenis *dataset* dengan karakteristik yang berbeda, maka proses preparasi tidak dapat disamakan.

A. Preparasi *dataset* DB1

Preparasi DB1 dilakukan dengan menerapkan teknik agregasi dan seleksi data. Agregasi dilakukan untuk mendapatkan nilai rerata per mata pelajaran SMA. Selain itu juga peneliti juga melakukan agregasi terkait variabel jenis dan jurusan SMA. Teknik seleksi data dilakukan untuk mengeliminasi data yang sekiranya tidak akan digunakan pada tahap penelitian selanjutnya. Penelitian menyiapkan dua jenis seleksi yang dinamakan **Skenario-A** dan **Skenario-B**. Skenario pertama merupakan skenario yang juga diterapkan pada *preliminary study*. Skenario ini mempertahankan mahasiswa yang memiliki jumlah SKS di atas 80, nilai IPK di atas 3.00 dan memiliki status antara "Aktif" atau "Lulus". Sebaliknya, Skenario-B tidak melakukan seleksi data apapun.

Artinya pada skenario kedua keseluruhan data mahasiswa akan digunakan. Detail skema per skenario dapat dilihat pada Tabel 3-2.

Tabel 3-2 Seleksi Data per Skenario

Skenario-A	Skenario-B
• Nilai IPK >= 3.00 • SKS >= 80 • Status mahasiswa "Aktif" atau "Lulus"	Tidak ada seleksi data berdasarkan nilai IPK, jumlah SKS dan status mahasiswa

B. Preparasi *dataset* DB2

Nantinya DB2 akan digunakan pada tahap pengelompokkan mahasiswa Informatika menggunakan model klastering. Pada proses tersebut hanya akan digunakan teknik seleksi data. Akan tetapi seleksi data akan dilakukan berdasarkan tiga aspek mulai dari jenis kurikulum, kode semeter hingga kelengkapan data capaian akademik per mata kuliah. Preparasi pada DB2 juga sedikit banyak mempertimbangkan ketersediaan data pada DB1. Hal ini tetap perlu diperhatikan mengingat pada tahap penelitian selanjutnya, kedua *dataset* tersebut akan digabungkan untuk kemudian digunakan pada tahap *semi-supervised learning*.

- Seleksi berdasarkan jenis kurikulum

  Peneliti memperhatikan ketersediaan data mahasiswa khusus program studi Informatika pada DB1. Pada *dataset* tersebut didapatkan bahwa data mahasiswa Informatika yang tersedia adalah mahasiswa rentang waktu tahun 2015 sampai dengan 2020. Hal tersebut menjadi pertimbangan peneliti dalam melakukan seleksi data pada DB2 berdasarkan aspek jenis kurikulum. Pada periode waktu tersebut diketahui terdapat tiga jenis kurikulum yang berbeda yaitu KD-2010, KD-2016 dan KD-2020. Berdasarkan informasi yang dikumpulkan, diketahui bahwa diantara ketiga kurikulum tersebut, KD-2016 dan KD-2020 dapat dikatakan masih memiliki banyak kemiripan. Berbeda dengan KD-2010 yang dinilai cukup berbeda secara signifikan dibandingkan dua kurikulum lainnya. Oleh karenanya, pada preparasi DB2 diputuskan untuk dipertahankan data mahasiswa Informatika yang mendapatkan satu diantara KD-2016 dan KD-2020. Pemilihan tersebut bertujuan agar data pada DB2 menjadi lebih seragam.

- Seleksi berdasarkan semester

  Pada penelitiannya, (Ezz, 2019) menggunakan data capaian akademik mahasiswa pada tahun pertama untuk digunakan pada sistem rekomendasi konsentrasi jurusan mahasiswa. Hal tersebut menjadi referensi peneliti untuk melakukan pendekatan serupa. Selain itu peneliti juga berasumsi bahwa capaian akademik mahasiswa pada tahun pertama dapat digunakan sebagai tolak ukur seberapa sukses mahasiswa beradaptasi dengan studi jenjang perguruan tinggi. Untuk itu, preparasi DB2 akan mempertahankan data capaian akademik mahasiswa pada tahun pertama saja.
- Seleksi berdasarkan kelengkapan nilai mata kuliah

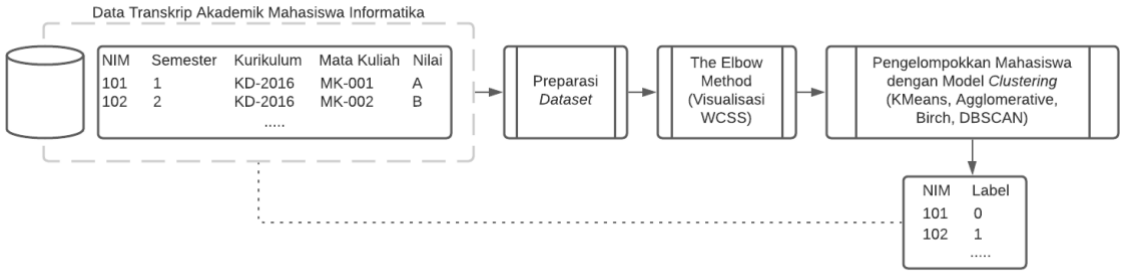
  Peneliti menemukan terdapat banyak data yang hilang (*missing value*) pada data poin atau variabel nilai mata kuliah. Banyaknya *missing value* pada *dataset* tentunya dapat memengaruhi performa dari model nantinya. Peneliti memutuskan hanya akan menggunakan variabel nilai mata kuliah yang ketersediaan datanya cukup lengkap. Angka 80% digunakan peneliti sebagai batas atau *benchmark*. Artinya, penelitian hanya akan menggunakan variabel nilai mata kuliah yang mana 80% datanya tersedia. Seleksi pada aspek ini berhasil mereduksi jumlah variabel nilai mata kuliah dari yang sebelumnya berjumlah 491 mata kuliah menjadi hanya 6 mata kuliah saja. Mata kuliah yang berhasil dipertahankan pada penelitian memiliki kode SIF101, SIF102, SIF103, SIF104, UNI600, dan UNI603.

3.2.4 Pengelompokkan Mahasiswa

Seperti dijelaskan sebelumnya, penting untuk melatih model klasifikasi sistem rekomendasi program studi menggunakan kelompok mahasiswa yang tepat. Hal ini untuk memastikan bahwa sistem mampu memberikan rekomendasi kompeten dan tidak menyesatkan. Proses seleksi kelompok mahasiswa kemudian menjadi tahap yang cukup krusial. Pada penelitian, proses seleksi mahasiswa sebagai data latih model klasifikasi sistem rekomendasi akan dibagi menjadi dua tahap yang masing-masing merepresentasikan kelompok mahasiswa Informatika dan non-Informatika. Mengingat ketersediaan data peneltian, pengelompokkan mahasiswa Informatika akan dilakukan dengan bantuan model klastering, sedangkan untuk kelompok mahasiswa non-Informatika akan menggunakan teknik *semi-supervised learning*.

Pengelompok mahasiswa Informatika menggunakan model klastering akan membandingkan beberapa algoritma klastering seperti KMeans, *agglomerative*, Birch dan

density-based spatial clustering of application with noise (DBSCAN). Penerapan beberapa algoritma tersebut bertujuan untuk membantu peneliti dalam memutuskan algoritma apa yang paling sesuai dengan kebutuhan peneliti. Peneliti akan melihat nilai koefisien Silhouette dari masing-masing algoritma untuk membantu pengambilan keputusan. Sebagai variabel prediktor model klastering, akan digunakan variabel nilai mata kuliah yang tersedia pada DB2. Ilustrasi pengelompokkan mahasiswa Informatika dapat dilihat pada Gambar 3-2 berikut.

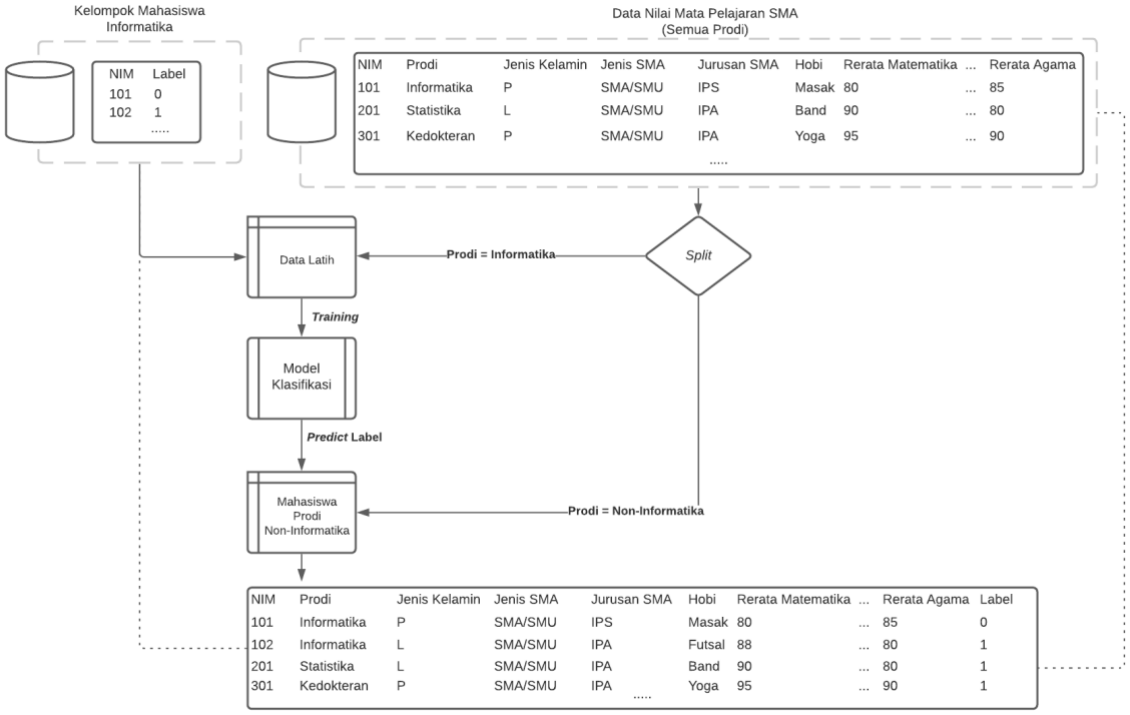


Gambar 3-2. Ilustrasi Pengelompokkan Mahasiswa Informatika Model Klastering

3.2.5 *Semi-supervised Learning*

Keterbatasan data penelitian dimana kelompok mahasiswa non-Informatika tidak memiliki data capaian akademik jenjang perguruan tinggi menyebabkan proses pengelompokkan menggunakan model klastering tidak dapat digunakan. Sebagai alternatif, proses pengelompokkan pada mahasiswa non-Informatika dilakukan dengan menerapkan teknik *semi-supervised learning*. Teknik tersebut akan mengimplementasikan model klasfikasi *single-stage* (selanjutnya akan disebut dengan model klasifikasi *semi-supervised learning*) yang dilatih menggunakan data kelompok mahasiswa Informatika. Model klasfikasi tersebut selanjutnya akan digunakan untuk memprediksi label kelompok pada mahasiswa non-Informatika. Model klasfikasi *semi-supervised learning* akan menggunakan variabel target yaitu hasil pengelompokkan model klastering pada tahap sebelumnya, sedangkan untuk varibel prediktornya akan digunakan beberapa data poin pada DB1 seperti jenis kelamin, jenis dan jurusan SMA, hobi, dan nilai rerata mata pelajaran SMA. Variabel prediktor yang digunakan tersebut sama persis dengan variabel prediktor yang nantinya akan digunakan pada model klasfikasi sistem rekomendasi program studi sarjana. Ilustrasi dari teknik *semi-supervised learning* dapat dilihat pada Gambar 3-3 di bawah. Dari ilustrasi tersebut terlihat bahwa penelitian kembali mengimplementasikan model klasifikasi. Model klasifikasi tersebut adalah model *single-stage* dan menggunakan

sekaligus membandingkan metode *multinomial logistic regression*, *random forest* dan *support vector machine*.



Gambar 3-3. Ilustrasi Teknik *Semi-supervised Learning*

Mengingat teknik *semi-supervised learning* menggunakan mahasiswa Informatika untuk memprediksi data mahasiswa non-Informatika, tentunya perlu adanya landasan yang cukup kuat untuk melakukan pendekatan tersebut. Untuk itu, sebelum menerapkan teknik *semi-supervised learning*, peneliti akan melihat terlebih dahulu apakah terdapat kemiripan antara kelompok mahasiswa Informatika dengan kelompok lainnya. Kemiripan akan dilihat berdasarkan nilai rerata mata pelajaran SMA, mengingat variabel tersebut merupakan variabel yang banyak digunakan pada model klasfikasi sistem rekomendasi program studi sarjana.

Selain itu, teknik *semi-supervised learning* juga berpotensi menimbulkan bias. Secara logika, penggunaan mahasiswa Informatika sebagai data latih pasti akan dirasa kurang cocok ketika kemudian digunakan untuk memprediksi kelompok mahasiswa yang berasal dari rumpun ilmu lain misalkan medis dan sosial. Di bagian akhir, penelitian akan memaparkan hasil analisis tambahan yang membahas hal tersebut. Analisis tersebut bertujuan untuk melihat seberapa besar bias yang terjadi, dan pada kelompok rumpun ilmu yang paling banyak terdampak.

3.2.6 Penentuan Kelompok Data Latih

Setelah melalui proses pengelompokkan mahasiswa, selanjutnya yang perlu dilakukan akan menyeleksi kelompok mahasiswa yang lebih tepat untuk digunakan sebagai data latih model klasifikasi sistem rekomendasi program studi sarjana. Untuk mengakomodasi hal tersebut maka perlu dilakukan perbandingkan antar kelompok yang berhasil didapatkan. Proses perbandingkan akan dilakukan dengan melihat sebaran data antar kelompok. Interpretasi tersebut bertujuan untuk melihat bagaimana karakteristik dari masing-masing kelompok.

Karena menggunakan dua teknik yang berbeda, maka interpretasi yang dilakukan juga dapat dipisahkan antara mahasiswa Informatika dan non-Informatika. Pada mahasiswa Informatika, perbandingan dapat dilakukan dengan melihat sebaran data berdasarkan nilai mata kuliah. Sedangkan pada mahasiswa non-Informatika, perbandingan dapat dilakukan dengan melakukan interpretasi berdasarkan nilai per mata pelajaran SMA.

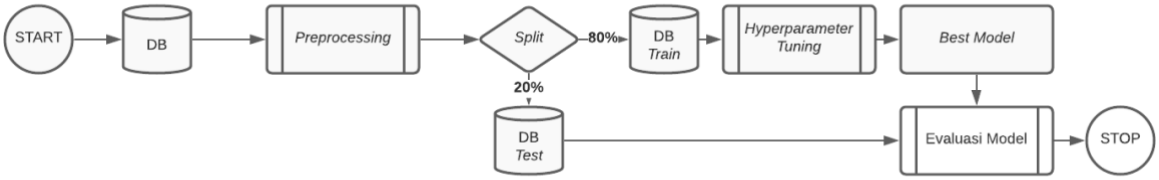
3.2.7 Pengembangan Model Klasifikasi

Penelitian menggunakan model klasifikasi dengan pendekatan (*single-stage* dan *multi-stages*), metode klasifikasi (*multinomial logistic regression*, *random forest*, dan *support vector machine*) juga skenario preparasi *dataset* (Skenario-A dan Skenario-B). Pada model *multi-stages*, prediksi kelas dilakukan secara berulang berdasarkan fase atau *stage* yang sudah didesain. Namun baik model *single-stage* dan *multi-stages* menggunakan variabel target dan variabel prediktor yang sama. Detail variabel yang digunakan pada model klasifikasi dapat dilihat pada Tabel 3-3 berikut.

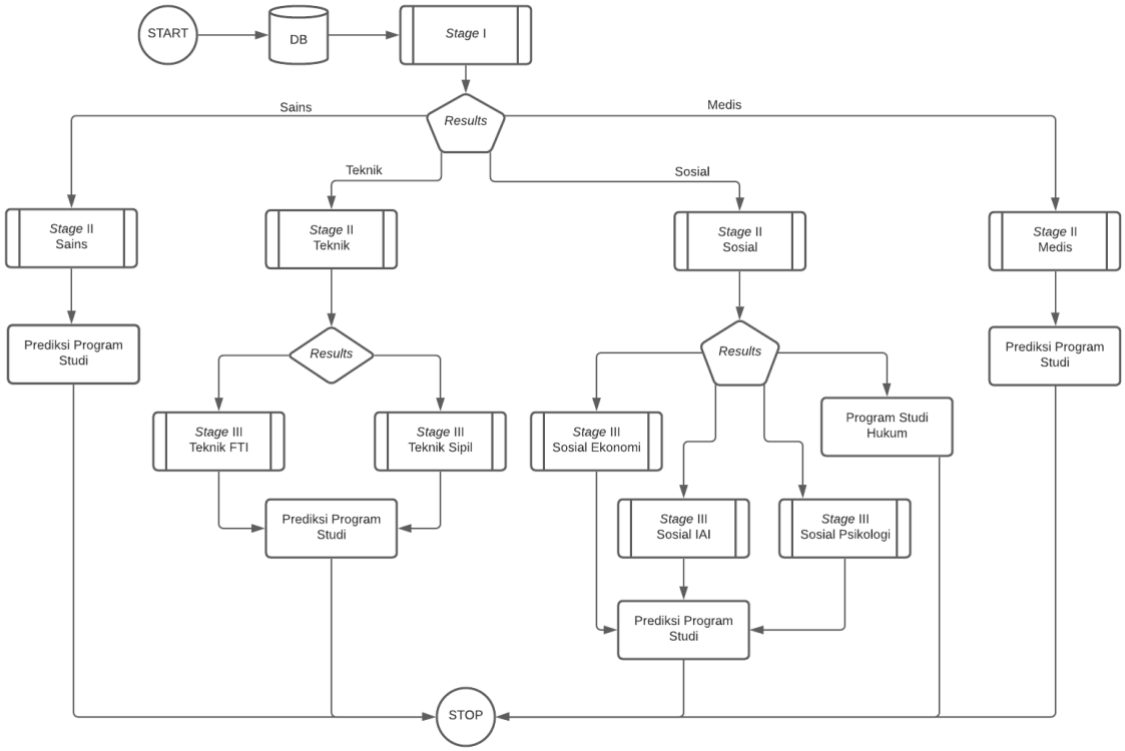
Tabel 3-3 Variabel Target dan Prediktor Model Klasifikasi Sistem Rekomendasi

Variabel Dependen (y)	Variabel Prediktor (X)
Program studi sarjana (25 prodi)	• Jenis kelamin • Jenis SMA • Jurusan SMA • Hobi • Nilai mata pelajaran matematika, bahasa Indonesia, bahasa Inggris, agama, fisika, kimia, biologi, geografi, sejarah, ekonomi, ketrampilan/keahlian kejuruan

Total terdapat 15 variabel prediktor bertipe kategorik maupun numerik yang digunakan pada model klasifikasi sistem rekomendasi. Pemilihan variabel tersebut disesuaikan dengan hasil studi literatur yang dilakukan pada awal penelitian. Diagram alir model *single-stage* dan *multi-stages* dapat dilihat pada Gambar 3-4 dan Gambar 3-5 berikut.



Gambar 3-4. Diagram Alir Model *Single-stage*



Gambar 3-5. Diagram Alir Model *Multi-stages*

Alir pada model *single-stage* lebih sederhana karena prediksi hanya dilakukan sekali jalan sedangkan pada model *multi-stages* terdapat beberapa *stage* yang harus dilalui sebelum akhirnya mendapatkan hasil prediksi program studi. Secara garis besar terdapat tiga *stage* pada model *multi-stages*. Pada *stage* I prediksi difokuskan lebih kepada rumpun ilmu progam studi. Terdapat 4 rumpun ilmu yang digunakan pada penelitian yaitu sains, medis, teknik dan sosial. *Stage* II melakukan prediksi pada spesifik rumpun ilmu

sedangkan *stage* III difokuskpan pada prediksi level fakultas. Artinya, semakin tinggi level *stage*-nya, maka data yang digunakan pun semakin detail segmennya. Pembagian rumpun ilmu dan fakultas progam studi pada penelitian ini dilakukan secara manual dan berdasarkan interpretasi pribadi peneliti. Peneliti tidak mengikuti pembagian rumpun ilmu maupun fakultas yang dilakukan oleh Universitas Islam Indonesia (UII). Hal tersebut dikarenakan peneliti memiliki interpretasi berbeda dengan apa yang diterapkan di UII. Pembagian program studi ke dalam rumpun ilmu dan fakultasnya dapat dilihat pada Tabel 3-4 berikut.

Tabel 3-4. Pembagian Rumpun Ilmu dan Fakultas

Rumpun	Fakultas	Progam studi
Sains	-	Kimia, Statistika, Pendidikan Kimia
Medis	-	Farmasi, Kedokteran
Teknik	Teknologi Industri (TI)	Informatika, Teknik Elektro, Teknik Industri, Teknik Kimia, Teknik Mesin
	Teknik Sipil dan Perencanaan (SP)	Arsitektur, Teknik Lingkungan, Teknik Sipil
Sosial	Bisnis Ekonomika (BE)	Akuntansi, Ekonomi Pembangunan, Hubungan Internasional, Perbankan dan Keuangan, Manajemen
	Hukum	Hukum
	Ilmu Agama Islam (IAI)	Ahwal Al-Syakhshiyah, Ekonomi Islam, Pendidikan Agama Islam
	Psikologi dan Sosial Budaya (PSB)	Ilmu Komunikasi, Psikologi, Pendidikan Bahasa Inggris

Model *single-stage* dan *multi-stages* akan membandingkan tiga metode klasifikasi yaitu *multinomial logistic regression*, *random forest*, dan *support vector machine*. Pemilihan tiga metode tersebut berdasarkan temuan pada penelitian (Mendes, Togelius, & Coelho, 2020) yang menyatakan bahwa model klasifikasi *machine learning* seperti *logistic regression*, *random forest* dan *support vector machine* memiliki performa yang lebih baik ketika digunakan pada *dataset* dengan ukuran kecil. Tahapan pada pengembangan model klasifikasi sistem rekomendasi sendiri dapat dibagi menjadi tiga yaitu preparasi, *training* & *validation* dan uji (*testing*).

- Preparasi

  Menggunakan dua jenis skenario preparasi data, maka penelitian juga memiliki dua jenis data latih yang akan dicoba pada model klasifikasi. Data latih menggunakan Skenario-A akan disebut dengan **DB-A** sedangkan data latih menggunakan Seknario-B akan disebut dengan **DB-B**. Pada kedua data latih tersebut peneliti menemukan ketidakseimbangan (*imbalanced*) pada jumlah program studi. Beberapa program studi memiliki jumlah data *record* yang jauh lebih sedikit dibandingkan program studi lainnya. Untuk mengatasi hal tersebut, peneliti menggunakan teknik *synthethic minority over-sampling method* (SMOTE) pada tahap preparasi untuk menyeimbangkan jumlah data masing-masing program studi. Teknik SMOTE bekerja dengan cara mensintesis data baru yang disesuaikan dengan karakteristik data asli pada masing-masing program studi. Artinya, data sintesis baru tersebut akan memiliki karakteristik yang serupa dengan data yang sebelumnya sudah tersedia pada *dataset*.

  Selain teknik SMOTE, juga digunakan teknik standardisasi untuk menyamakan skala numerik pada variabel numerik *dataset*. Tahap akhir preparasi adalah membagi data latih dengan rasio 8:2 dimana rasio yang lebih besar akan digunakan sebagai data latih sedangkan sisanya akan digunakan sebagai data uji.

- *Training* & *Validation*

  Tahap ini bertujuan untuk mengoptimasi performa dari model klasifikasi. Teknik *hyperparameter tuning* dilakukan untuk mendapatkan parameter pada model terbaik. *Hyperparameter tuning* dilakukan dengan metode *random search* menggunakan pembatasan parameter (*grid parameter*). Performa model klasifikasi diukur menggunakan skema *cross validation* dimana data validasi diambil sebanyak 20% dari data latih yang tersedia. Validasi dilakukan sebanyak 5 kali (*5 fold*) dengan menggunakan kurang lebih 100 kombinasi paramater model. Artinya, *hyperparameter tuning* akan melibatkan setidaknya 500 model klasifikasi. Dari 500 model tersebut akan dipilih satu model terbaik dan dilihat bagaimana parameter modelnya.

- *Testing*

  Tahap akhir dalam pengembangan model klasifikasi adalah proses uji (*testing*). Proses ini menggunakan data uji yang sebelumnya tidak digunakan pada tahap *training* & *validation*. Evaluasi diperlukan untuk mengukur kekuatan model

klasifikasi terhadap data baru yang sebelumnya belum pernah dikenali. Proses pengujian menggunakan model klasifikasi dengan paramater model yang didapatkan melalui proses *hyperparameter tuning*. Evaluasi model dilakukan dengan menggunakan beberapa metrik evaluator seperti *confussion matrix*, skor ROC-AUC dan *log-loss*. *Confussion matrix* digunakan untuk menilai kekuatan akurasi dan keseimbangan antara presisi dan sensitivitas model. Skor ROC-AUC digunakan untuk mengevaluasi kekuatan model klasifikasi dalam membedakan kelas atau label, sedangkan nilai *log-loss* digunakan untuk menilai kekuatan prediksi model dibandingkan dengan kelas atau label aslinya. Metrik evaluator tersebut biasanya digunakan pada kasus klasifikasi biner (*binary classification*) sedangkan penelitian ini mengangkat kasus *multi-class*. Untuk mengatasi hal tersebut akan digunakan nilai rerata dari masing-masing metrik sebelumnya pada proses evaluasi model klasifikasi sistem rekomendasi.

3.2.8 Komparasi Model Klasifikasi

Penelitian diakhiri dengan melakukan studi komparasi terkait model klasifikasi yang dikembangkan. Komparasi melibatkan pendekatan model klasifikasi (*single-stage* dan *multi-stages*), algoritma (*multinomial logistic regression, random forest* dan *support vector machine*) dan skenario preparasi (Skenario-A dan Skenario-B). Studi komparasi dilakukan untuk mendapatkan model klasifikasi terbaik, dan skenario preparasi yang mendukung model tersebut. Studi komparasi dilakukan dengan membandingkan metrik evaluator yang sama.

Total terdapat 9 model klasifikasi yang akan dikomparasi. 6 diantaranya merupakan model *single-stage*, 2 model *multi-stages* dan model *preliminary study*. Detail algoritma dan skenario preparasi pada model klasifikasi penelitian dapat dilihat pada Tabel 3-5 berikut.

Tabel 3-5 Model Klasifikasi Penelitian pada Tahap Studi Komparasi

	Model klasifikasi *single-stage*			Model klasifikasi *multi-stages*
	MLR	RF	SVM	MLR & RF & SVM
Skenario				
A	✓	✓	✓	✓
B	✓	✓	✓	✓

BAB 4
Hasil dan Pembahasan

Bab ini menjelaskan hasil dari metodologi yang diusulkan mulai dari preparasi *dataset*, pengelompokkan mahasiswa & *semi-supervised learning*, pemodelan dan pengujian model, sampai dengan studi komparasi.

4.1 Hasil dan Pembahasan Preparasi *Dataset*

Penelitian menggunakan dua *dataset* yaitu DB1 dan DB2. Kedua *dataset* mendapatkan preparasi data berbeda. Pada DB1 disiapkan dua skenario preparasi yaitu Skenario-A dan Skenario-B. Tabel 4-1 menunjukkan sebaran data pada DB1 per skenario.

Tabel 4-1 Sebaran Data *Dataset* DB1 per Skenario Preparasi

Skenario	Total Mahasiswa	Total Mahasiswa Informatika	Total Mahasiwa Non-Informatika
A	1,980	76	1,904
B	2,908	116	2,792

Dataset menggunakan Skenario-A selanjutnya akan disebut dengan DB-A dan *dataset* Skenario-B akan disebut dengan DB-B. Terlihat bahwa DB-B memiliki jumlah mahasiswa yang lebih banyak dibandingkan DB-A. Hal yang wajar mengingat Skenario-B tidak melakukan tahapan seleksi mahasiswa berdasarkan status mahasiswa, jumlah SKS dan nilai IPK. Proporsi mahasiswa Informatika pada kedua *dataset* juga termasuk sedikit. Kelompok mahasiswa Informatika tersebut nantinya akan digunakan pada tahap pengelompokan mahasiswa dan *semi-supervised learning*.

Preparasi DB2 dilakukan hanya dengan menggunakan teknik seleksi berdasarkan tiga kategori yaitu kode kurikulum, jenis semester dan kelengkapan nilai mata kuliah. Sebaran data pada DB2 sebelum dan setelah preparasi disajikan pada Tabel 4-2.

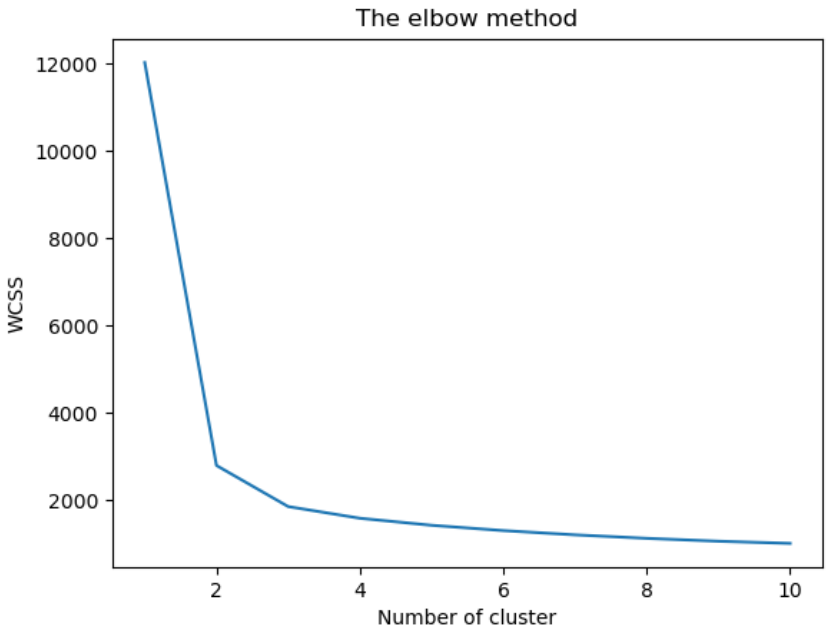
Tabel 4-2 Sebaran Data *Dataset* DB2 Sebelum dan Sesudah Tahap Preparasi

Sebelum Preparasi	Sesudah Preparasi
• 6.896 mahasiswa Informatika • 491 jenis mata kuliah • Semester 1 sampai dengan 9 • Tahun akademi 1995-2021 • 6 jenis kurikulum (1999-2020)	• 1.007 mahasiswa Informatika • 6 jenis mata kuliah • Semester 1 dan 2 • Tahun akademi 2016-2020 • 2 jenis kurikulum (KD-2016, KD-2020)

DB2 memiliki jumlah mahasiswa Informatika yang jauh lebih banyak dibandingkan DB-A dan DB-B, bahkan setelah melalui tahapan preparasi. Artinya kelengkapan data mahasiswa Informatika pada kedua *dataset* memang jauh berbeda. Ketimpangan tersebut nantinya akan berpengaruh terhadap jumlah data mahasiswa Informatika yang tersedia yang dapat digunakan pada tahap *semi-supervised learning*.

4.2 Hasil dan Pembahasan Tahap Pengelompokkan Mahasiswa

Pertama, pengelompokan mahasiswa akan dilakukan pada kelompok mahasiswa Informatika. Pengelompokkan akan dilakukan dengan menggunakan model *machine learning* tidak tersupervisi yaitu model klastering. Model tersebut akan menggunakan variabel prediktor berupa nilai mata kuliah mahasiswa. Sebelum menggunakan model klastering, terlebih dahulu perlu diketahui berapa jumlah klaster atau kelompok ideal yang dapat dibentuk dari data mahasiswa yang tersedia. Penentuan jumlah kelompok tersebut akan dilakukan dengan memvisualisasikan grafik *within cluster sum square* (WCSS). Teknik tersebut juga dikenal dengan nama metode siku (*elbow method*). Nilai WCSS dihitung dengan mengukur jarak dari masing-masing data terhadap titik tengah klaster yang didapatkan dari model. Gambar 4-1 menyajikan grafik WCSS dengan menggunakan metode KMeans.



Gambar 4-1 Visualisasi Nilai *Within Cluster Sum Square*

Jumlah kelompok ideal dapat ditentukan dengan melihat dimana grafik garis mulai melandai. Terlihat bahwa garis mulai melandai pada jumlah klaster dua. Peneliti kemudian memutuskan untuk menggunakan jumlah klaster sama dengan 2 pada model klastering. Pada penelitian ini peneliti mencoba beberapa metode klastering untuk kemudian dikomparasi nilai koefisien Silhoutte. Hal tersebut bertujuan untuk membantu peneliti dalam memutuskan model klastering mana yang sesuai dengan kebutuhan peneliti. Tabel 4-3 menunjukkan nilai koefisien Silhoutte dari model klastering menggunakan beberapa algoritma klastering.

Tabel 4-3 Performa Model *Clustering* Menggunakan Koefisien Silhoutte

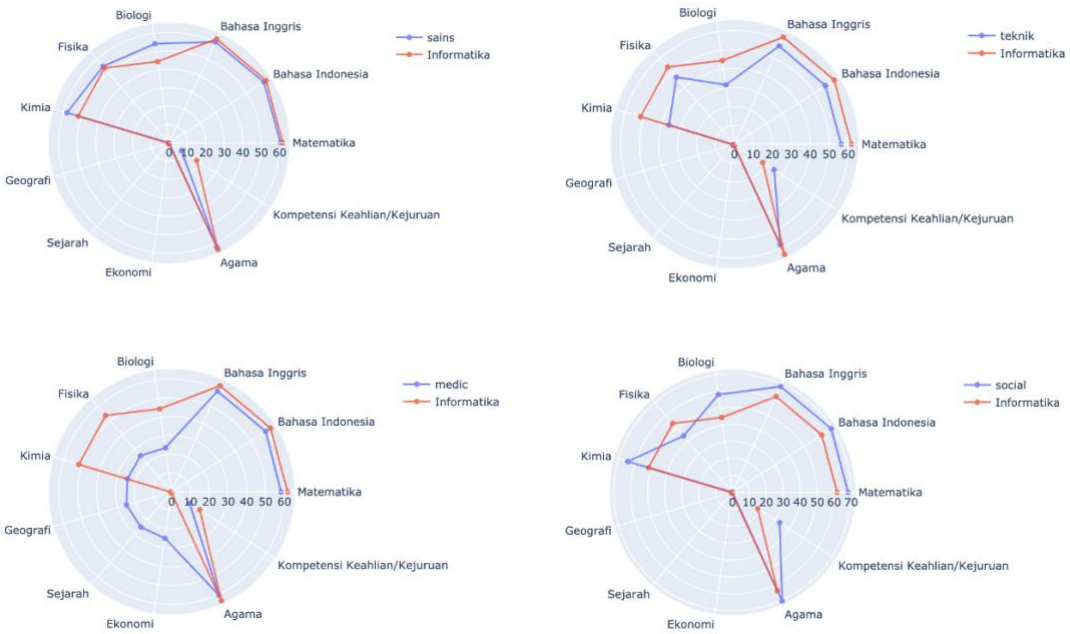
KMeans	**Agglomerative**	**Birch**	**DBSCAN**
0.729	0.727	0.621	0.013

Nilai koefisien semakin mendekati satu menunjukkan bahwa model bekerja semakin baik dan berhasil memisahkan data ke dalam kelompok yang tepat. Berdasarkan tabel di atas peneliti memutuskan untuk menggunakan model klastering KMeans sebagai alat bantu dalam pengelompokkan mahasiswa Informatika.

4.3 Hasil dan Pembahasan Tahap *Semi-supervised Learning*

Keterbatasan data penelitian mengakibatkan proses pengelompokkan mahasiswa pada penelitian harus dipisahkan antara kelompok mahasiswa Informatika dan program studi lainnya. Tidak tersedianya data capaian akademik jenjang perguruan tinggi pada kelompok mahasiswa non-Informatika menyebabkan proses pengelompokkan tidak dapat dilakukan dengan bantuan model klastering. Sebagai alternatif, penelitian menggunakan pendekatan teknik *semi-supervised learning* untuk mendapatkan label kelas pada kelompok mahasiswa program studi non-Informatika. Teknik tersebut memanfaatkan model klasifikasi *single-stage*, yang selanjutnya akan disebut dengan model klasifikasi *semi-supervised learning*, untuk memprediksi label kelas pada masing-masing mahasiswa. Model klasifikasi *semi-supervised learning* akan memanfaatkan data mahasiswa Informatika sebagai data latihnya. Variabel target model tersebut adalah label kelas yang didapatkan dari model klastering sebelumnya, sedangkan untuk variabel prediktor akan digunakan beberapa data poin yang sama dengan yang akan digunakan pada model klasifikasi sistem rekomendasi nantinya.

Mengingat teknik *semi-supervised learning* menggunakan data mahasiswa Informatika untuk memprediksi label mahasiswa non-Informatika, maka terlebih dahulu akan dilihat apakah terdapat kesamaan karakteristik dari kelompok mahasiswa tersebut. Perbandingkan antar kelompok akan dilakukan dengan melihat sebaran nilai rerata mata pelajaran SMA. Mengingat banyaknya program studi non-Informatika, maka peneliti berinisiasi untuk mengelompokkan beberapa program studi menjadi 4 rumpun ilmu yaitu sains, teknik, medis dan sosial. Rumpun ilmu tersebut adalah kelompok yang sama yang nantinya akan digunakan pada model klasifikasi *multi-stages*. Perbandingan antar kelompok mahasiswa Informatika dan rumpun ilmu dapat dilihat pada Gambar 4-2 berikut.



Gambar 4-2 Perbandingan Mahasiswa Informatika dengan Non-Informatika

Grafik radar atau *spyder charts* di atas menunjukkan perbandingkan sebaran nilai rerata per mata pelajaran SMA dari masing-masing kelompok mahasiswa. Terlihat bahwa sebaran data kelompok mahasiswa Informatika menyerupai dengan sebaran data rumpun ilmu sains, teknik dan sosial. Kemiripan terlihat paling jelas pada rumpun sains, dimana titik-titik pada grafik di setiap mata pelajaran sangat berdekatan. Perbedaan paling mencolok terlihat pada rumpun ilmu medis, dimana kelompok mahasiswa pada rumpun tersebut terlihat memiliki nilai rerata yang lebih baik pada mata pelajaran IPS. Berdasarkan perbandingan yang dilakukan menggunakan grafik radar, peneliti berasumsi bahwa memang terdapat kemiripan karakteristik antara mahasiswa program studi Informatika dengan mahasiswa program studi lainnya. Kemiripan terlihat terutama pada sebaran data nilai mata pelajaran umum seperti matematika, bahasa Indonesia dan bahasa Inggris. Hasil tersebut menjadi dasar peneliti untuk tetap menggunakan teknik *semi-supervised learning* pada penelitian. Peneliti juga menyadari bahwa teknik atau pendekatan ini berpotensi menimbulkan bias. Oleh karenanya, di bagian akhir peneliti akan memaparkan performa dari model klasifikasi sistem rekomendasi pada masing-masing rumpun ilmu tersebut, sehingga dapat diketahui pada rumpun ilmu apa teknik *semi-supervised learning* banyak mengakibatkan bias.

Dari pembahasan sebelumnya diketahui bahwa jumlah mahasiswa program studi Informatika pada DB1 dan DB2 berbeda. Karena tahap *semi-supervised learning* menggunakan data gabungan dari kedua *dataset*, maka jumlah data mahasiswa setelah digabungkan akan berkurang jauh. Tabel 4-4 menunjukkan jumlah mahasiswa Informatika setelah kedua *dataset* penelitian digabungkan. Terlihat bahwa jumlah mahasiswa Informatika pada *dataset* DB1 jauh berkurang setelah digabungkan dengan data pada *dataset* DB2.

Tabel 4-4 Jumlah Mahasiswa Informatika Gabungan *Dataset*

Skenario-A	**Skenario-B**
35	50

Kedua *dataset* penelitian juga memiliki data poin yang berbeda, oleh karenanya pada tahapan ini peneliti akan mempertahankan beberapa data poin atau variabel yang memang dibutuhkan untuk mengimplementasikan model klasifikasi *semi-supervised learning*. Data poin tersebut antara lain adalah jenis kelamin, hobi, jenis dan jurusan SMA, nilai rerata per mata pelajaran SMA dan label kelas hasil model klastering sebelumnya. Label kelas hasil model klastering akan digunakan sebagai variabel target model klasifikasi sedangkan data poin sisanya akan digunakan sebagai variabel prediktor.

Meskipun jumlah mahasiswa hasil penggabungan dua *dataset* dapat dikatakan hanya tersisa sedikit, akan tetapi peneliti memutuskan untuk tetap melanjutkan tahap *semi-supervised learning*. Model klasifikasi *semi-supervised learning*menguji dan membandingkan tiga algoritma yaitu *multinomial logistic regression* (MLR), *random forest* (RF) dan *support vector machine* (SVM). Tabel 4-5 menunjukkan performa dari model klasifikasi *semi-supervised learning*.

Tabel 4-5 Performa Model Klasifikasi *Semi-supervised Learning*

Skenario	**Model**	***Avg.Accuracy***	***Avg.F1-Score***	***Avg.ROC AUC score***	***Avg.Log Loss***
A	MLR	0.80	0.80	0.84	3.17
	RF	0.90	0.90	0.96	0.27
	SVM	0.90	0.90	1.00	0.25
B	MLR	0.87	0.87	0.85	0.49
	RF	0.87	0.87	0.87	0.48
	SVM	0.87	0.87	0.87	0.39

Pada Skenario-A terlihat bahwa model klasifikasi RF dan SVM memiliki performa yang paling baik, sedangkan pada Skenario-B ketiga model memiliki performa yang dapat dikatakan tidak jauh berbeda. Namun, jika diperhatikan lebih detail terlihat bahwa model SVM memiliki performa yang sedikit lebih unggul dibandingkan kedua model lainnya. Hal tersebut terlihat dari nilai rerata *log loss*-nya yang sedikit lebih rendah. Artinya, model klasifikasi SVM mampu memberikan hasil prediksi yang lebih mendekati nilai asli data. Peneliti memutuskan untuk menggunakan model klasifikasi SVM unutk digunakan sebagai alat bantu prediksi kelas atau label mahasiswa pada program studi non-Informatika.

4.4 Hasil dan Pembahasan Tahap Penentuan Kelompok Data Latih

Setelah mendapatkan label kelas dari masing-masing mahasiswa di semua program studi yang tersedia, selanjutnya dilakukan seleksi kelompok mahasiswa. Proses ini bertujuan untuk menentukan kelompok yang lebih tepat untuk digunakan sebagai data latih model klasifikasi sistem rekomendasi. Proses penentuan kelompok data latih dilakukan dengan melihat sebaran data dari masing-masing kelompok.

Pada mahasiswa program studi Informatika, interpretasi kelompok dilakukan dengan melihat sebaran data nilai mata kuliah dan hasil model klastering. Tabel 4-6 menunjukkan sebaran data terkait nilai mata kuliah pada mahasiswa Informatika per skenario preparasi *dataset*.

Tabel 4-6 Sebaran Data Nilai Mata Kuliah Mahasiswa Informatika

Skenario	**Kelas**	**Total Mahasiswa**	**Rerata Jumlah Mata Kuliah Lulus**	**Rerata Nilai per Mata Kuliah**
A	0	11	0.00	0.00
	1	23	6.00	3.42
B	0	14	0.42	1.41
	1	36	6.00	3.34

Berdasarkan tabel di atas terlihat bahwa kelompok mahasiswa dengan label 1 dapat dikatakan memiliki prestasi yang lebih baik. Mahasiswa kelompok tersebut rata-rata berhasil lulus di keenam mata kuliah dengan nilai 3.00-3.50 atau setara B+. Sedangkan mahasiswa dengan label kelompok 0 sepertinya berisikan mahasiswa yang tidak berhasil lulus dalam 6 mata kuliah tersebut, atau mendapatkan nilai yang sangat rendah. Terlihat

mahasiswa kelompok tersebut memiliki nilai rerata per mata kuliah di bawah 1. Interpretasi kelompok mahasiswa non-Informatika dilakukan dengan melihat sebaran nilai mata pelajaran SMA per rumpun ilmu program studi seperti yang digunakan pada tahap *semi-supervised learning*. Tabel 4-7 menunjukkan sebaran data nilai mata pelajaran.

Tabel 4-7 Sebaran Data Nilai Mata Pelajaran Prodi Informatika dan Rumpun Ilmu

Skenario-A	**INFORMATIKA**		**SAINS**		**MEDIS**		**TEKNIK**		**SOSIAL**	
	Kelas 0	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**
Matematika	41.32	70.15	55.80	65.55	57.39	76.25	56.85	66.29	51.64	58.24
Bahasa Indonesia	41.41	70.74	56.09	69.77	58.26	79.05	57.46	65.07	52.38	59.43
Bahasa Inggris	40.84	69.94	55.50	69.36	57.76	72.74	56.68	62.83	51.71	58.52
Fisika	41.66	66.60	55.33	72.24	54.60	76.63	56.27	64.21	53.93	63.28
Biologi	42.30	64.95	56.53	71.84	57.78	80.41	56.37	65.48	54.75	66.07
Kimia	41.52	73.20	56.11	71.43	57.73	66.77	56.75	66.88	53.96	61.28
Geografi	0.00	34.44	0.00	36.00	0.00	0.00	11.98	39.06	13.92	57.05
Sejarah	0.00	34.74	0.00	34.00	0.00	0.00	14.04	30.53	13.62	57.75
Ekonomi	0.00	0.00	0.00	35.00	0.00	0.00	13.80	38.30	13.93	58.37
Agama	43.01	71.27	56.96	70.90	59.41	78.83	57.64	67.94	54.11	60.17
Kompetensi Keahlian/Kejuruan	30.00	75.32	46.70	72.10	0.00	82.37	79.47	66.59	49.59	70.44
Skenario-B	**INFORMATIKA**		**SAINS**		**MEDIS**		**TEKNIK**		**SOSIAL**	
	Kelas 0	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**	**Kelas 0**	**Kelas 1**
Matematika	51.18	73.37	55.95	71.54	56.87	76.74	56.91	71.48	50.86	61.32
Bahasa Indonesia	51.04	74.36	56.16	73.66	56.75	78.78	57.05	73.45	51.87	62.49
Bahasa Inggris	50.92	73.58	55.72	73.57	57.24	75.74	57.02	71.24	51.48	61.32
Fisika	45.34	66.22	55.54	71.00	54.57	77.48	56.21	70.28	53.64	62.12
Biologi	49.63	72.28	56.40	75.08	57.13	81.09	56.39	66.30	54.12	63.35
Kimia	49.07	77.27	55.78	72.30	57.29	75.58	56.78	70.10	53.68	62.85
Geografi	0.00	34.44	0.00	36.00	0.00	0.00	11.98	59.11	14.36	56.97
Sejarah	0.00	34.74	0.00	34.00	0.00	0.00	14.04	46.50	14.02	57.71
Ekonomi	0.00	0.00	0.00	35.00	0.00	0.00	13.80	58.70	14.34	58.26
Agama	52.92	75.30	57.07	75.97	58.91	82.26	57.48	75.79	53.65	63.27
Kompetensi Keahlian/Kejuruan	56.78	77.60	52.65	74.27	0.00	84.67	76.54	75.84	50.81	77.99

Berdasarkan sebaran data terlihat bahwa kelompok mahasiswa dengan label kelompok 1 memiliki nilai rerata yang lebih tinggi. Hal tersebut sama seperti hasil interpretasi kelompok mahasiswa Informatika. Oleh karena itu, peneliti memutuskan menggunakan kelompok dengan label tersebut sebagai data latih model klasifikasi sistem rekomendasi. Sebagai tambahan akan disajikan sebaran data per program studi dari kelompok mahasiswa dengan label 1 pada Tabel 4-8.

Tabel 4-8 Sebaran Data Mahasiswa per Program Studi Sarjana

Program Studi	Skenario A	Skenario B
Ahwal Al-Syakhshiyah	20	32
Akuntansi	110	185
Arsitektur	10	29
Ekonomi Islam	16	38
Ekonomi Pembangunan	78	124
Farmasi	25	71
Hubungan Internasional	10	29
Hukum	64	97
Ilmu Komunikasi	47	75
Informatika	23	36
Kedokteran	2	4
Kimia	13	26
Manajemen	130	215
Pendidikan Agama Islam	13	26
Pendidikan Bahasa Inggris	7	10
Pendidikan Kimia	6	9
Perbankan dan Keuangan	25	37
Psikologi	43	63
Statistika	6	8
Teknik Elektro	3	32
Teknik Industri	2	17
Teknik Kimia	9	24
Teknik Lingkungan	7	23
Teknik Mesin	6	25
Teknik Sipil	15	54
Total	**690**	**1289**

4.5 Hasil dan Pembahasan Tahap Pengembangan Model Klasifikasi

Studi komparasi pada penelitian melibatkan model klasifikasi dengan beberapa pendekatan (*single-stage* dan *multi-stages*), metode (*multinomial logistic regressions*, *random forest* dan *support vector machine*) dan skenario (Skenario-A dan Skenario-B). Hasil dari pengembangan masing-masing model akan disajikan di bawah.

4.5.1 Model *Single-stage classification*

Tabel 4-9 dan 4-10 menunjukkan performa model klasifikasi *single-stage* yang dilatih menggunakan DB-A dan DB-B dengan tiga metode klasifikasi yaitu *multinomial logistic regression* (MLR), *random forest* (RF), dan *support vector machine* (SVM). Performa pada kedua tabel tersebut adalah performa dari model klasifikasi terbaik yang telah melalui proses *hyperparameter tuning* dan dievaluasi menggunakan data uji yang sama sekali tidak digunakan pada tahap *training* & *validation* model.

Tabel 4-9 Performa Model Klasifikasi *Single-stage* Skenario A

Model	***Avg.Accuracy***	***Avg.F1-Score***	***Avg.ROC AUC score***	***Avg.Log Loss***
MLR	0.44	0.43	0.90	1.76
RF	0.92	0.92	0.99	0.29
SVM	0.59	0.56	0.95	1.14

Tabel 4-10 Performa Model Klasifikasi *Single-stage* Skenario B

Model	***Avg.Accuracy***	***Avg.F1-Score***	***Avg.ROC AUC score***	***Avg.Log Loss***
MLR	0.30	0.27	0.83	2.34
RF	0.92	0.90	0.99	0.34
SVM	0.41	0.40	0.89	1.88

Dari kedua tabel di atas terlihat bahwa model klasifikasi metode RF memiliki performa yang lebih baik hampir pada semua metrik evaluator dibandingkan dengan model MLR dan SVM. Model tersebut juga bekerja sama baiknya menggunakan DB-A maupun DB-B. Jika komparasi dilakukan antar skenario, terlihat bahwa model klasifikasi yang dilatih menggunakan *dataset* Skenario-A memiliki performa lebih baik. Perbedaan jelas terlihat pada model MLR dan SVM yang memiliki kekuatan prediksi yang lebih baik dan juga memiliki nilai skor ROC-AUC dan rerata *log-loss* yang lebih baik. Dapat diambil kesimpulan bahwa pada model klasifikasi *single-stage* lebih cocok digunakan *dataset* dengan Skenario-A. Artinya seleksi data berdasarkan status mahasiswa, jumlah SKS dan

nilai IPK yang dilakukan pada Skenario-A membantu meningkatkan performa dari model klasifikasi *single-stage*.

4.5.2 Model *Multi-stages Classification*

Seperti dijelaskan sebelumnya bahwa model *multi-stages* pada penelitian memiliki total tiga *stage* yang berbeda. Oleh karenanya, sangat dimungkinkan di masing-masing *stage* diimplementasikan model klasifikasi dengan algoritma yang berbeda. Tabel 4-11 menunjukkan metode klasifikasi dengan performa terbaik yang telah diuji coba pada masing-masing *stage*.

Tabel 4-11 Model Klasifikasi Terbaik setiap *Stage*

Stage	**Model Terbaik**	
	Skenario A	**Skenario B**
Stage I	RF	RF
Stage II Medis	MLR	MLR
Stage II Sains	MLR	SVM
Stage II Sosial	RF	RF
Stage II Teknik	SVM	RF
Stage III Sosial Ekonomi	RF	RF
Stage III Sosial IAI	MLR	RF
Stage III Sosial Psikologi	RF	RF
Stage III Teknik FTI	RF	RF
Stage III Teknik Sipil	SVM	RF

Model klasifikasi dengan metode RF sekali lagi terlihat mendominasi hampir pada semua *stage*. Dominasi terlihat jelas pada model klasifikasi menggunakan DB-B. Pada model *multi-stages* menggunakan DB-A, dominasi model RF tidak terlalu terlihat dan masih terlihat kontribusi dari metode MLR dan SVM. Setelah mendapatkan model klasifikasi terbaik pada masing-masing *stage*, selanjutkan dikembangkan model *multi-stages* secara utuh untuk kemudian diukur performanya.

Performa dari model *multi-stages* dilampirkan pada Tabel 4-12. Perlu menjadi catatan pada tabel tersebut bahwa nilai rerata akurasi dan F1-*score* didapatkan dari pengujian menggunakan data uji, sedangkan pada metrik evaluator lainnya, nilai rerata didapatkan dari nilai rerata model terbaik setiap *stage*-nya. Hal tersebut dilakukan karena

metrik evaluator seperti rerata ROC-AUC dan *log-loss* tidak bisa dihitung jika menggunakan data uji.

Tabel 4-12 Performa Model Klasifikasi *Multi-stages*

Skenario	***Avg.Accuracy***	***Avg.F1-Score***	***Avg.ROC AUC score***	***Avg.Log Loss***
A	0.82	0.79	0.97	0.26
B	0.85	0.81	0.99	0.17

Hasil sedikit berbeda ditunjukkan pada model *multi-stages*. Tidak seperti model *single-stage* yang mencapai performa terbaik ketika menggunakan *dataset* DB-A, pada model *multi-stages* performa terbaik justru didapatkan ketika menggunakan *dataset* DB-B. Hal ini mungkin terjadi mengingat *dataset* DB-B memiliki jumlah mahasiswa yang lebih banyak dibandingkan DB-A, sehingga jumlah mahasiswa yang diimplementasi pada setiap *stage*-nya pun juga lebih banyak.

4.5.3 Implementasi *Hyperparameter Tuning* dan *Features Selection*

Komparasi pada pembahasan sebelumnya diketahui melibatkan model klasifikasi hasil *hyperparameter tuning*. Artinya, perbandingan hanya menggunakan model klasifikasi dengan performa terbaik. Detail hasil dari *hyperparameter tuning* sendiri tidak akan dijabarkan mengingat banyaknya model klasifikasi yang terlibat, khususnya pada model *multi-stages*. Sebagai gambaran, akan sedikit dijelaskan mengenai hasil paramater tuning pada model klasifikasi terbaik dengan pendekatan *single-stage*.

Pada *single-stage*, model klasifikasi terbaik adalah model menggunakan algoritma *random forest* dimana memiliki akurasi sebesar 92%. Model klasifikasi *random forest* tersebut menggunakan total sebanyak 400 pohon keputusan (*decision tree*). Masing-masing pohon keputusan akan menggunakan total sebanyak 4 atribut atau fitur. Jumlah atribut tersebut didapatkan dari hasil akar (*square root*) dari total atribut yang terdapat pada *dataset* yaitu sebanyak 15 atribut. Kedalaman minimal per pohon keputusan adalah 55. Selain itu, dapat pembentukannya digunakan seluruh data yang tersedia pada *dataset*. Artinya tidak terdapat seleksi data atau *bootstraping*. Menggunakan parameter tersebut, model klasifikasi *random forest* berhasil mendapatkan akurasi sebesar 90% pada proses validasi. Nilai akurasi tersebut tidak jauh berbeda jika dibandingkan dengan akurasi pada tahap pengujian. Hal yang sama juga peneliti temukan pada model klasifikasi menggunakan algoritma *multinomial logistic regression* dan *support vector machine*.

Dimana kedua model klasifikasi tersebut pada proses *hyperparameter tuning* mendapatkan akurasi terbaik yang tidak jauh berbeda dengan hasil pada tahap pengujian. Tabel 4-13 secara detail melampirkan parameter dari semua model klasifikasi terbaik pada model *single-stage* maupun *multi-stages*.

Tabel 4-13 Parameter Model Klasifikasi Terbaik

Model	Metode	Parameter
single-stage (Skenario-A)	*Random Forest*	{"n_estimators": 400, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 55, "bootstrap": false}
multi-stages (Skenario-B)	*Stage* I *Random Forest*	{"n_estimators": 100, "min_samples_split": 5, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 91, "bootstrap": false}
	Stage II Medis *Multinomial Logistic Regression*	{"solver": "lbfgs", "penalty": "l2", "multi_class": "multinomial", "max_iter": 1000, "class_weight": null, "C": 206.913808111479}
	Stage II Sains *Support Vector Machine*	{"probability": true, "kernel": "rbf", "gamma": "auto", "decision_function_shape": "ovo", "class_weight": "balanced", "C": 1.0}
	Stage II Sosial *Random Forest*	{"n_estimators": 200, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 82, "bootstrap": false}
	Stage II Teknik *Random Forest*	{"n_estimators": 900, "min_samples_split": 5, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 19, "bootstrap": false}
	Stage III Sosial IAI *Random Forest*	{"n_estimators": 1000, "min_samples_split": 5, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 46, "bootstrap": true}
	Stage III Sosial Ekonomi *Random Forest*	{"n_estimators": 500, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "auto", "max_depth": 55, "bootstrap": false}

	Stage III Sosial Psikologi *Random Forest*	{"n_estimators": 100, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 100, "bootstrap": false}
	Stage III Teknik FTI *Random Forest*	{"n_estimators": 400, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 64, "bootstrap": false}
	Stage III Teknik Sipil *Random Forest*	{"n_estimators": 400, "min_samples_split": 2, "min_samples_leaf": 2, "max_features": "sqrt", "max_depth": 64, "bootstrap": false}

Hasil *hyperparameter tuning* tabel di atas dapat digunakan sebagai panduan untuk mengembangkan model klasifikasi serupa pada penelitian di masa depan. Perlu diingat bahwa parameter model yang terlampir pada Tabel 4-13 adalah parameter berdasarkan modul Python yaitu Sklearn. Parameter model mungkin sedikit berbeda jika model dikembangkan menggunakan bahasa pemrograman lainnya. Pada model klasifikasi *random forest*, terlihat bahwa diperlukan minimal sebanyak 100 pohon keputusan untuk mendapatkan performa terbaik dari model. Selain itu kedalaman masing-masing pohon juga berkisar antara 55 sampai dengan 100. Mayoritas pohon keputusan juga tidak melakukan teknik *bootstraping*. Pada model *multinomial logistic regression* performa terbaik didapatkan ketika model menggunakan pinalti L2, pendekatan *multinomial* dan algoritma penyelesaiannya (*solver*) lbfgs. Sedangkan pada model klasifikasi *support vector machine* kernel terbaik adalah kernel rbf dan menggunakan pendekatan *multi-class one-versus-one* (OVO).

Selain memberikan panduan terkait parameter dari model klasfikasi terbaik, peneliti juga tertarik untuk melihat performa model klasifikasi dengan mengimplementasikan teknik *features selection*. Teknik *features selection* adalah teknik untuk mereduksi atribut atau variabel prediktor yang digunakan pada model klasifikasi. Analisis ini dilakukan oleh peneliti untuk melihat apakah teknik *features selection* memengaruhi performa dari model klasifikasi. Pada model klasifikasi menggunakan metode *random forest*, teknik *classification and decision tree* (CART) digunakan untuk menyeleksi atribut model, sedangkan pada metode *multinomial logistic regression* dan *support vector machine* akan digunakan koefisien atau bobot dari masing-masing atribut pada model klasifikasi. Kedua

pendekatan tersebut akan menghitung seberapa besar kontribusi dari masing-masing atribut terhadap model klasifikasi. Atribut yang akan dipertahankan pada model adalah atribut dengan nilai kontribusi di atas nilai rerata kontribusi dari semua atribut.

Setelah melalui tahap preparasi, *training* & *validation* hingga pengujian, didapatkan performa dari model klasifikasi *single-stage* dan *multi-stages* yang dilampirkan pada Tabel 4-14, sedangkan atribut yang terpilih dari teknik *features selection* dapat dilihat pada Tabel 4-15.

Tabel 4-14 Performa Model Klasifikasi dengan *Features Selection*

Model	*Single-stage* (RF)		*Multi-stages* (RF, SVM, MLR)	
Skenario	A		B	
Features selection	Tidak	Ya	Tidak	Ya
Avg.Accuracy	0.92	0.92	0.85	0.82
Avg.F1-Score	0.92	0.91	0.81	0.78
Avg.ROC AUC	0.99	0.99	0.99	0.97
Avg.Log Loss	0.29	0.31	0.17	0.24

Tabel menampilkan perbandingan dari model klasifikasi dengan dan tanpa mengimplementasikan teknik *features selection*. Terlihat bahwa performa model klasifikasi dengan teknik *features selection* tidak jauh lebih baik dibandingkan model tanpa teknik tersebut. Pada model *single-stage* terlihat bahwa kedua model memiliki kekuatan prediksi yang sama. Begitu pula dengan kemampuan membedakan antar kelas. Akan tetapi, model *single-stage* dengan teknik *features selection* memiliki nilai rerata *log loss* yang sedikit lebih besar. Artinya performa model tersebut sedikit di bawah model *single-stage* tanpa teknik *features selection*.

Perbedaan performa terlihat sedikit lebih jelas ketika membandingkan model klasifikasi *multi-stages*. Performa model dengan teknik *features selection* terlihat sedikit di bawah model tanpa teknik tersebut. Hal ini terlihat pada semua metrik evaluator yang digunakan. Dari perbandingan tersebut dapat diambil kesimpulan bahwa teknik *features seelction* yang diuji coba pada penelitian tidak terlalu memberikan efek positif pada model. Artinya, atribut yang digunakan pada model sebelumnya sudah dirasa cukup dan dapat digunakan.

Tabel 4-15 Variabel Prediktor Model Klasifikasi Teknik *Features Selection*

Model Klasifikasi	**Attribut**
Single-stage *Random forest*	Matematika, Bahasa Indonesia, Bahasa Inggris, Fisika, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage I *Random forest*	Jenis kelamin, Matematika, Bahasa Indonesia, Bahasa Inggris, Biologi, Fisika, Kimia, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage II Medis *Multinomial logistic regression*	Hobi, Bahasa Indonesia, Bahasa Inggris, Biologi, Fisika, Kimia, Geografi, Sejarah, Ekonomi, Agama
Multi-stages Stage II Sains *Support vector machine*	Matematika, Bahasa Indonesia, Bahasa Inggris, Fisika, Kimia, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage II Sosial *Random Forest*	Hobi, Matematika, Bahasa Indonesia, Bahasa Inggris, Geografi, Sejarah, Ekonomi, Agama
Multi-stages Stage II Teknik *Random Forest*	Matematika, Bahasa Indonesia, Bahasa Inggris, Fisika, Kimia, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage III Sosial IAI *Random Forest*	Jenis kelamin, Hobi, Matematika, Bahasa Indonesia, Bahasa Inggris, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage III Sosial Ekonomi *Random Forest*	Hobi, Matematika, Bahasa Indonesia, Bahasa Inggris, Geografi, Sejarah, Ekonomi, Agama
Multi-stages Stage III Sosial Psikologi *Random Forest*	Hobi, Matematika, Bahasa Indonesia, Bahasa Inggris, Geografi, Sejarah, Ekonomi, Agama
Multi-stages Stage III Teknik FTI *Random Forest*	Matematika, Bahasa Indonesia, Bahasa Inggris, Fisika, Kimia, Agama, Kompetensi Keahlian/Kejuruan
Multi-stages Stage III Teknik Sipil *Random Forest*	Matematika, Bahasa Indonesia, Bahasa Inggris, Fisika, Kimia, Agama, Kompetensi Keahlian/Kejuruan

Tabel 4-15 menunjukkan atribut yang dipertahankan hasil teknik *features selection* pada masing-masing model klasifikasi. Pada model *single-stage* terlihat bahwa teknik

tersebut mengeliminasi semua atribut atau variabel kategorik seperti jenis kelamin, hobi dan jenis dan jurusan SMA. Variabel numerik yang dipertahankan juga dapat dikatakan tidak terlalu banyak. Teknik *features selection* mengeliminasi semua mata pelajaran dari rumpun ilmu sosial dan hanya mempertahankan mata pelajaran Fisika dari rumpun ilmu sains. Sisanya terdapat mata pelajaran umum seperti matematika, bahasa Indonesia, bahasa Inggris, agama dan satu mata pelajaran kejuruan yaitu kompetensi keahlian/kejuruan.

Hasil yang lebih beragam didapatkan pada model klasifikasi *multi-stages*. Teknik *features selection* pada model beberapa kali mempertahankan variabel kategorik pada beberapa *stage* seperti pada *stage* I, *stage* II medis, *stage* II sosial, dan semua *stage* III sosial. Meski demikian, variabel kategorik yang dipertahankan hanya diantara jenis kelamin atau hobi. Tidak terlihat variabel kategorik jenis dan jurusan SMA, sehingga dapat diambil kesimpulan bahwa kedua variabel tersebut tidak berkontribusi lebih pada model di masing-masing *stage*-nya. Teknik *features selection* juga hampir selalu mempertahankan variabel numerik terkait mata pelajaran umum di setiap *stage*. Terlihat bahwa mata pelajaran seperti matematika, bahasa Indonesia dan bahasa Inggris hampir selalu dipertahankan di setiap model. Selain itu juga terlihat bahwa mata pelajaran IPA hampir selalu dipertahankan pada model klasifikasi terkait *stage* rumpun ilmu sains. Sebaliknya, mata pelajaran IPS juga selalu dipertahankan pada model klasifikasi terkait rumpun ilmu sosial.

4.6 Hasil dan Pembahasan Komparasi Model Klasifikasi

Tahap akhir penelitian adalah melakukan studi komparasi dari model *single-stage* dan *multi-stages* terhadap model klasifikasi *preliminary study*. Hal ini dilakukan untuk mengetahui apakah pembaruan dan teknik yang diterapkan pada penelitian berhasil memperbaiki performa dari model klasifikasi sistem rekomendasi. Melalui komparasi ini juga dapat dilihat apakah penambahan *dataset* baru terkait nilai mata kuliah dan implementasi model klastering dan teknik *semi-supervised learning* berhasil memberikan kontribusi terhadap model klasifikasi sistem rekomendasi itu sendiri. Komparasi antar model klasifikasi disajikan pada Tabel 4-16 berikut.

Tabel 4-16 Komparasi Model Klasifikasi dengan *Preliminary Study*

Model	*single-stage*	*multi-stages*	*preliminary study*
Algoritma	RF	RF, SVM, MLR	RF
Skenario	A	B	*preliminary study*
Avg.Accuracy	0.92	0.85	0.86
Avg.F1-Score	0.92	0.81	0.84
Avg.ROC AUC	0.99	0.99	0.97
Avg.Log Loss	0.29	0.17	0.66

Dari komparasi di atas terlihat bahwa model *single-stage* menggunakan metode *random forest* dan *dataset* dengan Skenario-A memiliki performa yang lebih baik dibandingkan dengan model klasifikasi *preliminary study*. Model *single-stage* berhasil unggul di semua metrik evaluator yang digunakan. Model *multi-stages* juga menunjukkan hal yang serupa. Meskipun dari sisi akurasi dan F1-*score* model *multi-stages* tidak jauh berbeda dibandingkan model *preliminary study*, akan tetapi model tersebut sedikit lebih baik jika dilihat dari sisi skor ROC-AUC dan rerata *log-loss*. Artinya model *multi-stages* memiliki kemampuan membedakan antar kelas yang lebih baik dibandingkan model klasifikasi *preliminary study*. Selain itu model tersebut juga memiliki probabilitas prediksi yang lebih mendekati data asli.

4.7 Analisis Bias Teknik *Semi-supervised Learning*

Pembahasan berikut akan menjelaskan bagaimana efek dari teknik *semi-supervised learning* yang diterapkan pada penelitian. Seperti diketahui teknik *semi-supervised learning* menggunakan data mahasiswa Informatika sebagai data latih model klasifikasi *semi-supervised learning* untuk kemudian digunakan sebagai alat bantu prediksi kelas atau label pada kelompok mahasiswa program studi non-Informatika. Secara logika, penggunakan data mahasiswa program studi Informatika akan menimbulkan bias ketika digunakan untuk memprediksi program studi pada rumpun ilmu sosial dan medis. Hal ini dikarenakan secara logika, progam studi Informatika akan lebih dekat dengan rumpun ilmu sains atau teknik dibandingkan dengan rumpun ilmu medis dan sosial.

Efek bias dari teknik *semi-supervised learning* dilihat berdasarkan performa model klasifikasi yang dilatih menggunakan data yang merepresentasikan masing-masing rumpun ilmu. Dengan kata lain, akan dilihat performa dari model klasifikasi pada *stage* II pada model *multi-stages*. *Stage* II model *multi-stages* mengembangkan model klasifikasi secara

spesifik pada rumpun ilmu sains, teknik, medis dan sosial. Akan digunakan model *multi-stages* yang dilatih menggunakan *dataset* Skenario-B, mengingat model memberikan performa terbaik ketika menggunakan skenario tersebut. Performa model klasifikasi pada *stage* II model *multi-stage* terlampir pada Tabel 4-17 berikut.

Tabel 4-17 Performa Model Klasifikasi per Rumpun Program Studi

Rumpun	**Model**	***Avg. Accuracy***	***Avg.F1-Score***	***Avg. ROC AUC score***	***Avg. Log Loss***
Sains	MLR	0.88	0.87	0.98	0.27
	RF	1.00	1.00	1.00	0.11
	SVM	0.94	0.94	1.00	0.14
Teknik	MLR	0.64	0.64	0.68	0.65
	RF	0.97	0.97	1.00	0.05
	SVM	0.69	0.68	0.78	0.58
Medis	MLR	1.00	1.00	1.00	0.001
	RF	1.00	1.00	1.00	0.003
	SVM	1.00	1.00	1.00	0.02
Sosial	MLR	0.46	0.46	0.75	1.17
	RF	0.93	0.93	0.99	0.28
	SVM	0.58	0.57	0.82	1.01

Sebelum melakukan komparasi antar rumpun ilmu, pembahasan akan dimulai dengan melihat performa model klasifikasi pada rumpun ilmu medis. Pada rumpun ilmu tersebut model klasifikasi baik menggunakan metode MLR, RF atau SVM memiliki performa yang dapat dikatakan hampir sempurna. Hal tersebut cukup aneh mengingat dari studi komparasi yang dilakukan sebelumnya, biasanya hanya model klasifikasi dengan metode RF yang memiliki performa bagus. Peneliti berasumsi terdapat indikasi *overfitting* pada rumpun ilmu medis. Asumsi tersebut didasari pada jumlah data program studi pada rumpun ilmu medis dengan jumlah yang memang jauh lebih sedikit dibandingkan program studi lainnya. Artinya, teknik SMOTE yang digunakan untuk mengatasi ketidakseimbangan telah mensintesis data baru dengan jumlah yang terlalu banyak sehingga menutupi karakteristik data asli yang tersedia. Oleh karena itu, tidak dapat dilihat efek bias dari penggunaan teknik *semi-supervised learning* pada rumpun ilmu medis.

Analisis untuk mengetahui efek bias teknik *semi-supervised learning* akan difokuskan pada ketiga rumpun ilmu yang tersisa yaitu sains, teknik dan sosial. Selain itu

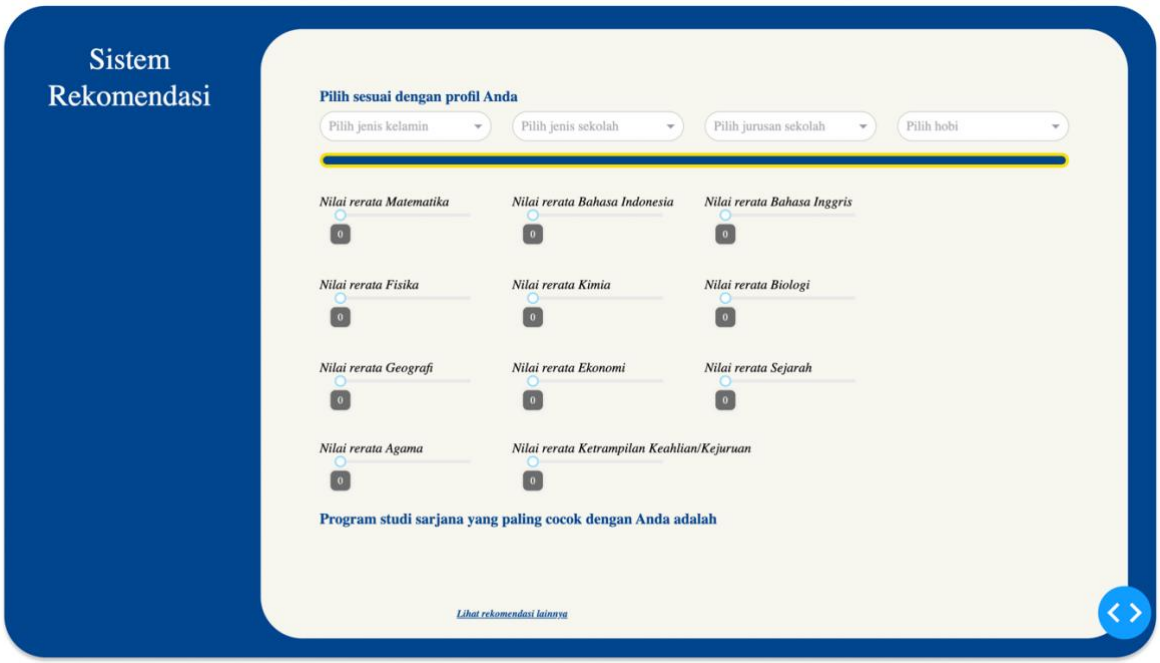
analisis juga akan difokuskan hanya pada model klasifikasi dengan metode MLR dan SVM. Metode RF dikeluarkan dari analisis karena model tersebut terlihat memiliki performa yang bagus dan cenderung tidak jauh berbeda pada ketiga rumpun ilmu.

Diantara ketiga rumpun ilmu yang tersisa, terlihat bahwa model klasifikasi MLR dan SVM memiliki performa yang paling baik ketika digunakan pada data mahasiswa progam studi sains. Model klasifikasi tersebut berhasil unggul hampir pada semua metrik evaluator. Sepertinya data mahasiswa program studi Informatika yang digunakan pada teknik *semi-supervised learning* lebih memiliki kemiripan dengan data mahasiswa pada program studi rumpun ilmu sains. Selain itu, model klasifikasi MLR dan SVM juga memiliki performa yang cukup pada rumpun ilmu teknik, meskipun memiliki nilai akurasi yang sedikit di bawah 70%. Angka tersebut masih lebih baik jika dibandingkan dengan performa model pada rumpun ilmu sosial. Artinya, pada penelitian yang dilakukan, program studi Informatika cenderung lebih mirip dengan progam studi pada rumpun ilmu teknik dibandingkan rumpun ilmu sosial.

Hasil komparasi yang dilakukan berdasarkan rumpun ilmu tersebut menunjukkan bahwa efek bias pada teknik *semi-supervised learning* lebih berimbas pada program studi rumpun ilmu sosial. Teknik *semi-supervised learning* bekerja paling baik ketika digunakan pada program studi rumpun ilmu sains. Sedangkan pada rumpun ilmu teknik, efek bias berimbas tidak sebanyak ketika digunakan pada rumpun ilmu sosial.

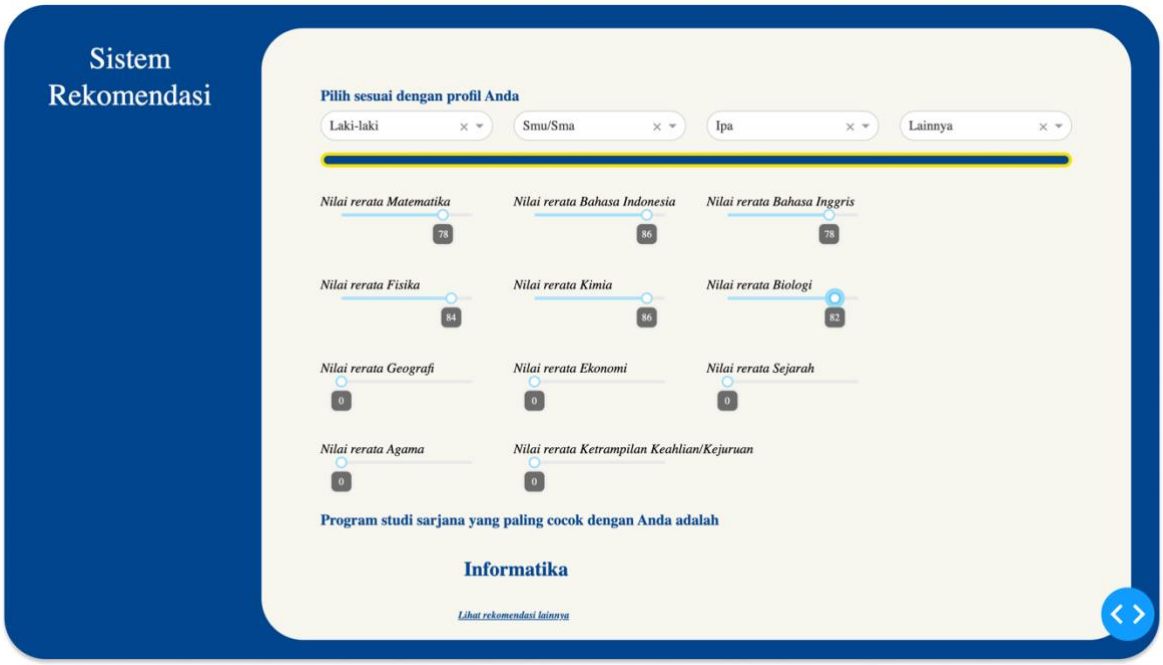
4.8 Hasil dan Pembahasan Purwarupa Sistem Rekomendasi

Sistem rekomendasi kemudian diwujudkan pada purwarupa berbasis *website* sederhana. Model yang akan diimplementasikan pada purwarupa tersebut adalah model *single-stage* menggunakan metode *random forest*. Peneliti mengembangkan purwarupa sederhana untuk memberikan gambaran sederhana kepada pembaca atau calon pengguna. Gambar 4-2 menunjukkan tampilan muka dari sistem rekomendasi.



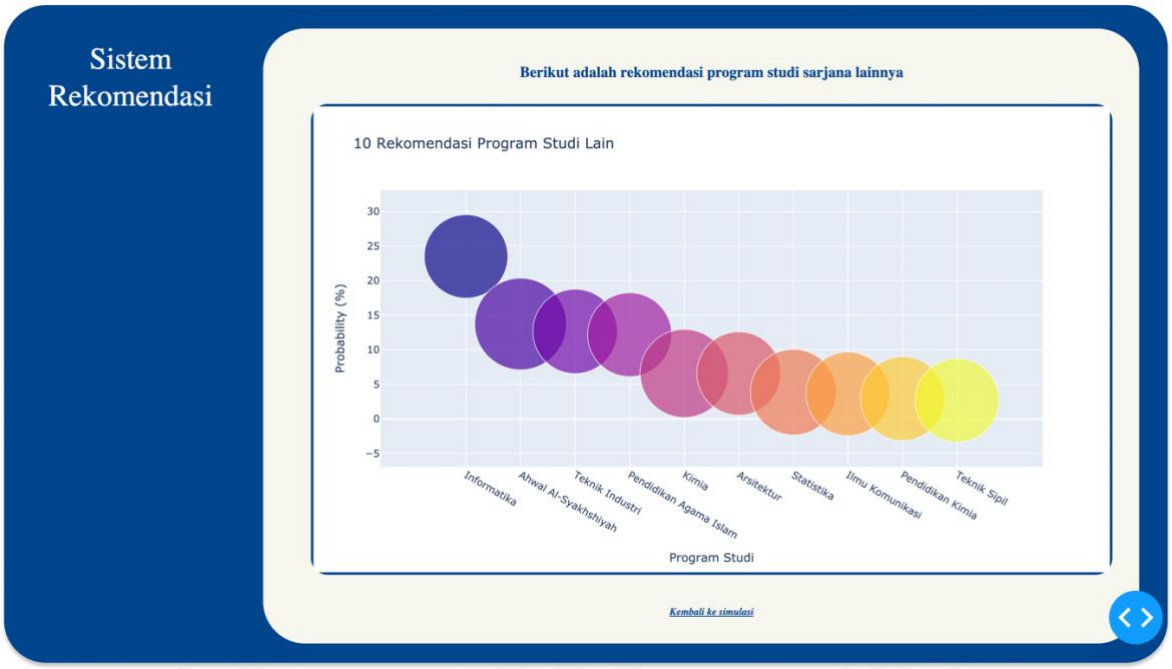
Gambar 4-3 Tampilan Halaman Muka Sistem Rekomendasi

Melalui halaman muka tersebut pengguna dapat merasakan langsung pengalaman menggunakan sistem rekomendasi pemilihan program studi sarjana. Setelah mengatur profil sesuai dengan karakteristik masing-masing, akan keluar hasil rekomendasi dari sistem. Bentuk dari hasil rekomendasi sistem dapat dilihat pada Gambar 4-3.



Gambar 4-4 Hasil Rekomendasi Sistem

Selain tampilan di atas, sistem juga memvisualisasikan hasil rekomendasi dalam bentuk *bubble* seperti pada Gambar 4-4. Hal ini dilakukan agar pengguna tetap bisa mendapatkan informasi terkait rekomendasi program studi lainnya. Selain label program studi, pada grafik tersebut juga dapat diketahui nilai probabilitas (%) dari masing-masing program studi.



Gambar 4-5 Visualisasi Hasil Rekomendasi Menggunakan *Bubble Chart*

BAB 5
Kesimpulan dan Saran

5.1 Kesimpulan

Penelitian bertujuan untuk mengetahui bagaimana sains data khususnya model *machine learning* dapat diimplementasikan pada pengembangan sistem rekomendasi program studi sarjana. Implementasi sains data kemudian diwujudkan pada beberapa tahapan seperti penggunakan model klastering, teknik *semi-supervised learning*, hingga model klasifikasi. Masing-masing model dan teknik tersebut ditempatkan sedemikian rupa pada penelitian dan bertujuan untuk menyempurnakan sistem rekomendasi program studi.

Adapun studi komparasi penelitian bertujuan untuk mengetahui model klasifikasi terbaik apa yang dapat diterapkan pada sistem rekomendasi. Studi komparasi melibatkan model klasifikasi yang didesain baik menggunakan pendekatan *single-stage* mapupun *multi-stages*. Tiga algoritma klasifikasi mulai dari *multinomial logistic regression*, *random forest* dan *support vector machine* juga dibandingkan di masing-masing model klasifikasi. Selain itu, studi komparasi juga melibatkan skenario preparasi *dataset* dengan dan tanpa menggunakan variabel status mahasiswa, jumlah SKS dan nilai IPK pada tahap awal seleksi data mahasiswa. Sebagai penutup, studi komparasi dilakukan dengan melibatkan model klasifikasi *preliminary study*.

Hasil studi komparasi menunjukkan bahwa model klasifikasi *single-stage* menggunakan algoritma *random forest* berhasil memberikan performa terbaik dibandingkan model klasifikasi *multi-stages* maupun model *preliminary study*. Model tersebut memberikan nilai akurasi sebesar 92%, yang mana 6% lebih tinggi dibandingkan akurasi dari model *multi-stages* dan *preliminary study*. Model *single-stage* memberikan performa terbaik ketika dilatih menggunakan *dataset* dengan preparasi Skenario-A. Berbeda dengan model *multi-stages* yang berhasil memberikan performa terbaik ketika menggunakan *dataset* dengan preparasi Skenario-B. Model *multi-stages* juga memberikan performa yang cukup baik jika dibandingkan dengan model klasifikasi *preliminary study*. Meskipun dari sisi akurasi kedua model tersebut dapat dikatakan tidak jauh berbeda, akan tetapi model *multi-stages* memiliki nilai rerata *log loss* yang lebih kecil.

Dari studi komparasi yang dilakukan, dapat diambil kesimpulan bahwa kedua model klasifikasi yang diinisiasikan penelitian berhasil menunjukkan perbaikan performa dibandingkan dengan model klasifikasi *preliminary study*. Pembaruan dan pendekatan

yang dilakukan pada penelitian juga terbukti memberikan kontribusi terhadap performa kedua model tersebut. Pembaruan yang dimaksud adalah adanya penambahan *dataset* baru terkait data capaian akademik mahasiswa selama studi perguruan tinggi, juga penerapan model klastering dan teknik *semi-supervised learning* yang digunakan untuk proses seleksi kelompok mahasiswa sebagai data latih model klasifikasi sistem rekomendasi.

Selain studi komparasi, penelitian juga memaparkan hasil analisis terkait implementasi teknik *semi-supervised learning* yang berpotensi menimbulkan bias. Hasil analisis menunjukkan bahwa bias teknik *semi-supervised learning* terlihat pada penerapan model klasifikasi terutama pada rumpun ilmu sosial. Pada rumpun ilmu tersebut, model klasifikasi menggunakan algoritma *multinomial logistic regression* dan *support vector machine* memberikan performa yang dapat dikatakan kurang memuaskan.

5.2 Saran

Meskipun model *single-stage* dan *multi-stages* pada penelitian berhasil memberikan performa yang lebih baik dibandingkan model klasifikasi *preliminary study*, akan tetapi masih terdapat bagian yang menurut peneliti masih dapat diperbaiki. Salah satunya terkait ketersediaan data. Peneliti merasa bahwa akan lebih baik jika penelitian dilakukan dengan jumlah data yang cukup banyak dan seimbang antar program studi. Dengan adanya data yang cukup di masing-masing program studi, maka teknik *oversampling* tidak akan terlalu banyak berkontribusi sehingga penelitian dapat dilakukan dengan data asli. Ketersediaan data terkait data capaian akademik juga perlu dilengkapi sehingga proses pengelompokkan mahasiswa dapat sepenuhnya dilakukan dengan menggunakan model klastering. Hal ini dikarenakan pada penelitian ditemukan bahwa teknik *semi-supervised learning* yang dilakukan memiliki bias terutama pada rumpun ilmu sosial.

Selain itu penelitian juga berpotensi untuk dikembangkan menjadi sistem rekomendasi dengan target pengguna yang berbeda yaitu pihak perguruan tinggi. Tahap pengelompokkan mahasiswa Informatika menggunakan model klastering dapat dikemas sedemikian rupa menjadi suatu sistem rekomendasi bagi pihak perguruan tinggi untuk menyeleksi kelompok mahasiswa mana yang mengalami kesulitan studi. Dengan mengetahui hal tersebut pihak perguruan tinggi dapat kemudian melakukan pendekatan kepada kelompok mahasiswa untuk membantu mereka dalam menemukan solusi atas permasalahan studi.

Daftar Pustaka

Andriani, A. (2013). Sistem Pendukung Keputusan Berbasis Decision Tree dalam Pemberian Beasiswa Studi Kasus: AMIK "BSI Yogyakarta". *Seminar Nasional Teknologi Informasi dan Komunikasi 2013 (SENTIKA 2013)* (pp. 163-168). Yogyakarta: SENTIKA 2013.

Bobadilla, J., Ortega, F., Hernando, A., & Gutierrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 109-132.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Springer: Mach Learn*, 273-297.

Ezz, M. E. (2019). Adaptive Recommendation System using Machine Learning Algorithm for Predict Student's Best Academic Program. *Education and Information Technologies*, 2733-2746.

Grewal, D., & Kaur, K. (2016). Developing an Intelligent Recommendation System for Course Selection by Students for Graduate Courses. *Business and Economics Journal*.

Isler, Y., Narin, A., Ozer, M., & Perc, M. (2018). Multi-stage classification of congestive heart failure based on short-term heart rate variability. *Chaos, Solitons & Fractals*, 145-151.

Kancherla, D., Bodapati, J., & Veeranjaneyulu, N. (2019). Effect of different kernels on the performance of an SVM. *Int. J. Recent Technology. Eng*, 1-6.

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *Springer Science* , 101-123.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Technique. In I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies.* IOS Press.

Kumalasari, L. D., & Susanto, A. (2019). Recommendation System of Information Technology Jobs using Collaborative Filtering Method Based on LinkedIn Skills Endorsement. *Sisforma*, 63-72.

Kwak, C., & Clayton-Matthews, A. (2002). Multinomial Logistic Regression. *Nursing Research*, 404-410.

Marbun, E., & Hansun, S. (2019). Sistem Pendukung Keputusan Pemilihan Program Studi dengan Metode SAW dan AHP. *ILKOM Jurnal Ilmiah*.

Mendes, A., Togelius, J., & Coelho, L. d. (2020). Multi-Stage Transfer Learning with an Application to Selection Process. *24th European Conference on Artificial Intelligence - ECAI 2020.* Santiago de Compostela.

Mesya. (2019, December 7). *JPNN*. Retrieved from JPNN: https://www.jpnn.com/news/hasil-survei-87-persen-mahasiswa-pilih-jurusan-tidak-sesuai-minat

Okaviana, M. R., & Susanto, R. (2014). Sistem Pendukung Keputusan Rekomendasi Pemilihan Program Studi menggunakan Metode Multifactor Evaluation Process di SMA Negeri 1 Bandung. *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*, 50-57.

Pal, M. (2005). Random forest classifier for remote sensing classificaion. *International Journal of Remote Sensing*, 217-222.

Parameswaran, A., Venetis, P., & Garcia-Molina, H. (2011). Recommendation Systems with Complex Constraint: A Course Recommendation Perspective. *ACM Transactions on Information Systems*.

Pare, S. (2013). Sistem Pendukung Keputusan Pemilihan Program Studi pada Perguruan Tinggi. *Jurnal Ilmiah Mustek Anim Ha*.

Patil, L. H., & Atique, M. (2014). A Multistage Feature Selection Model for Document Classification Using Information Gain and Rough Set. *International Journal of Advanced Research in Artificial Intelligence*.

Permatasari, H. S., Kridalaksana, A. H., & Suyatno, A. (2015). Sistem Pendukung Keputusan Pemilihan Program Studi di Universitas Mulawarman menggunakan Metode Tsukamoto (Studi Kasus : Fakultas MIPA). *Jurnal Informatika Mulawarman*, 32-37.

Poorna, S., & Nair, G. (2019). Multistage classification scheme to enhance speech emotion. *International Journal of Speech Technology*.

Pratama, A. R., Aryanto, R. R., & Pratama, A. T. (2021). Model Klasifikasi Calon Mahasiswa Baru Untuk Sistem Rekomendasi Program Studi Sarjana Berbasis Machine Learning. *Unpublished manuscript*.

Safutra, I. (2019, December 15). *JawaPos*. Retrieved from JawaPos: https://www.jawapos.com/nasional/pendidikan/15/12/2019/mahasiswa-mengaku-salah-jurusan-banyak-sarjana-yang-penting-lulus/

Salgado, C. M., Fernandes, M. P., Horta, A., Xavier, M., Sousa, J. M., & Vieira, S. M. (2017). Multistage modeling for the classification of numerical and categorical

datasets. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Naples: IEEE.

Senator, T. E. (2005). Multi-Stage Classification. *International Conference on Data Mining (ICDM'05)*. IEEE.