BAB III
METODE PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan metode kuantitatif dan studi literatur. Metode kuantitatif yang digunakan menekankan pada pengukuran secara objektif serta analisis data yang dikumpulkan dan diolah menggunakan teknik komputasi (Babbie, 2010; Muijs, 2004), yaitu *Random Forest* yang merupakan salah satu algoritma klasifikasi *machine learning*. Data yang dikumpulkan dan diolah oleh algoritma tersebut merupakan data penelitian yang diperoleh dari studi literatur.

3.2 Objek Penelitian

Objek penelitian adalah data siswa dari dua sekolah menengah di Portugis yaitu Sekolah Menengah Gabriel Pereira dan Sekolah Menengah Mousinho da Silveira, sebanyak 395 data siswa yang diperoleh melalui situs UCI *Machine Learning Repository* (Cortez & Silva, 2008). Meskipun data siswa tersebut diperoleh dari data sekunder, namun data memiliki atribut yang diperlukan pada penelitian, serta kemiripan karakteristik dengan data siswa sekolah menengah di Indonesia, misalnya nilai, jumlah ketidakhadiran, serta kegagalan pada mata pelajaran sebelumnya.

3.3 Data Penelitian

Data penelitian yang digunakan merupakan data sekunder. Data sekunder adalah data yang tidak diperoleh secara langsung oleh peneliti, namun merupakan data yang dikumpulkan dari tangan kedua atau sumber lain yang telah tersedia sebelum penelitian dilakukan (Silalahi, 2015). Keputusan untuk menggunakan data sekunder dalam penelitian ini merupakan hasil dari berbagai pertimbangan karena situasi pandemi Covid-19 yang terjadi dan belum selesai sampai saat ini. Situasi tersebut mempengaruhi kegiatan di sekolah dan menyebabkan pengambilan data primer yang seharusnya dilakukan tidak dapat dilakukan sesuai rencana sebelumnya.

Data ini merupakan dataset yang memiliki atribut yang paling lengkap diantara data lainnya, meskipun bukan merupakan data siswa yang sekolah di

Indonesia, namun data ini memiliki atribut yang serupa dengan data akademik siswa di Indonesia seperti nilai siswa, jumlah ketidakhadiran, dan jumlah kegagalan pada mata pelajaran di sekolah.

Data diperoleh dari hasil kuesioner kepada siswa serta penilaian dari sekolah yang dikumpulkan dari tahun 2005 – 2006 oleh Paul Cortez untuk penelitiannya terkait kinerja siswa di sekolah menengah Portugis (Cortez & Silva, 2008). Data terdiri dari 33 atribut yang disajikan pada Tabel 3.1.

Tabel 3.1 Atribut Data Penelitian

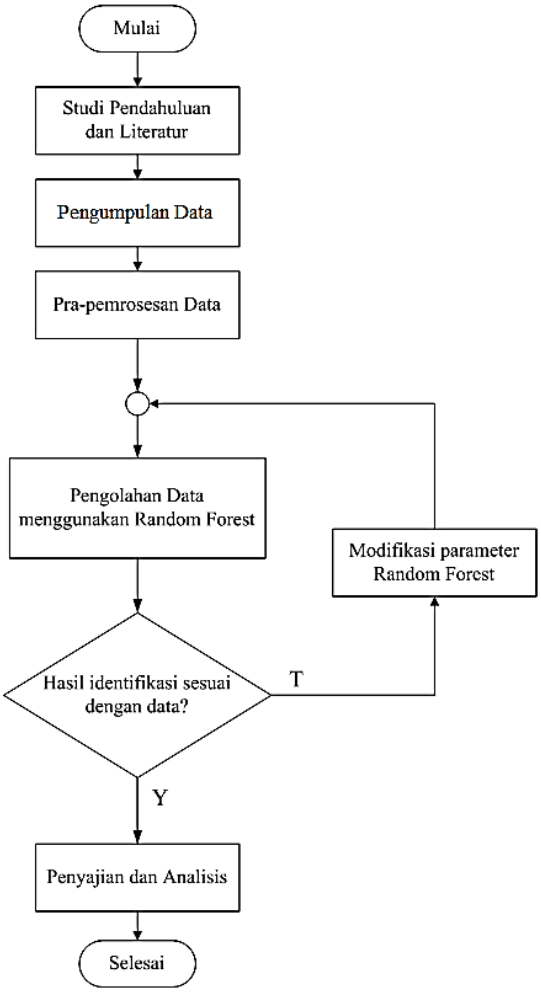
No	Atribut	Deskripsi
1	*school*	Sekolah siswa (Gabriel Pereira atau Mousinho da Silveira) Pilihan: GP atau MF
2	*sex*	Jenis kelamin siswa (laki-laki atau perempuan) Pilihan: F atau M
3	*age*	Usia siswa Numerik: rentang 15 sampai 22
4	*address*	Tipe tempat tinggal siswa (kota atau desa) Pilihan: U atau R
5	*famsize*	Banyak anggota keluarga (kurang dari/sama dengan 3, atau lebih dari 3) Pilihan: LE3 atau GT3
6	*Pstatus*	Status tinggal bersama orang tua (bersama atau terpisah) Pilihan: T atau A
7	*Medu*	Pendidikan Ibu Numerik: rentang 0 sampai 4 Ket. 0 – tidak ada 1 – sekolah dasar (sampai dengan kelas 4) 2 – sekolah dasar (dari kelas 5 sampai 9) 3 – sekolah menengah 4 – sekolah tinggi
8	*Fedu*	Pendidikan Ayah Numerik: rentang 0 sampai 4 Ket. 0 – tidak ada 1 – sekolah dasar (sampai dengan kelas 4) 2 – sekolah dasar (dari kelas 5 sampai 9) 3 – sekolah menengah 4 – sekolah tinggi
9	*Mjob*	Pekerjaan Ibu (Guru, bidang kesehatan, layanan sipil, di rumah, dan lainnya)

		Pilihan: *teacher*, *health*, *services*, *at_home*, atau *other*
10	*Fjob*	Pekerjaan Ayah (Guru, bidang kesehatan, layanan sipil, di rumah, dan lainnya) Pilihan: *teacher*, *health*, *services*, *at_home*, atau *other*
11	*reason*	Alasan memilih sekolah (dekat rumah, reputasi sekolah, preferensi sekolah, dan alasan lainnya) Pilihan: *home*, *reputation*, *course*, *other*
12	*Guardian*	Wali siswa (Ibu, Ayah, atau lainnya) Pilihan: *mother*, *father*, *other*
13	*traveltime*	Lama perjalanan ke sekolah Numerik: rentang 1 – 4 Ket. 1 – kurang dari 15 menit 2 – 15 sampai 30 menit 3 – 30 menit sampai 1 jam 4 – lebih dari 1 jam
14	*studytime*	Lama belajar selama satu minggu Numerik: rentang 1 – 4 Ket. 1 – kurang dari 2 jam 2 – 2 sampai 5 jam 3 – 5 sampai 10 jam 4 – lebih dari 10 jam
15	*failures*	Jumlah kelas sebelumnya yang gagal Numberik: rentang 1 – 4 Ket. 1 – 1 kelas 2 – 2 kelas 3 – 3 kelas 4 – lebih dari 3 kelas
16	*schoolsup*	Dukungan pendidikan tambahan (Ya atau Tidak) Pilihan: Y atau N
17	*famsup*	Dukungan pendidikan dari keluarga (Ya atau Tidak) Pilihan: Y atau N
18	*paid*	Kelas tambahan berbayar (Ya atau Tidak) Pilihan: Y atau N
19	*activities*	Aktivitas ekstrakurikuler (Ya atau Tidak) Pilihan: Y atau N
20	*nursery*	Pernah TK (Ya atau Tidak) Pilihan: Y atau N
21	*higher*	Keinginan untuk melanjutkan pendidikan (Ya atau Tidak) Pilihan: Y atau N
22	*internet*	Akses internet di rumah (Ya atau Tidak) Pilihan: Y atau N

23	*romantic*	Dalam sebuah hubungan (Ya atau Tidak) Pilihan: Y atau N
24	*famrel*	Kondisi hubungan keluarga Numerik: rentang 1 – 5 Ket. 1 – angat buruk 2 – buruk 3 – sedang 4 – baik 5 – sangat Baik
25	*freetime*	Waktu luang setelah sekolah Numerik: rentang 1 – 5 Ket. 1 – sangat sedikit 2 – sedikit 3 – sedang 4 – banyak 5 – sangat banyak
26	*goout*	Pergi bersama teman Numerik: rentang 1 – 5 Ket. 1 – sangat jarang 2 – jarang 3 – normal 4 – sering 5 – sangat sering
27	*Dalc*	Komsumsi alkohol hari kerja Numerik: rentang 1 – 5 Ket. 1 – sangat rendah 2 – rendah 3 – normal 4 – tinggi 5 – sangat tinggi
28	*Walc*	Komsumsi alkohol hari libur Numerik: rentang 1 – 5 Ket. 1 – sangat rendah 2 – rendah 3 – normal 4 – tinggi 5 – sangat tinggi
29	*health*	Kondisi kesehatan terkini Numerik: rentang 1 – 5 Ket. 1 – sangat buruk 2 – buruk

		3 – sedang 4 – baik 5 – sangat baik
30	*absences*	Jumlah ketidakhadiran di sekolah Numerik: rentang 1 – 93
31	G1	Nilai pertama Numerik: rentang 0 – 20
32	G2	Nilai kedua Numerik: rentang 0 – 20
33	G3	Nilai akhir Numerik: 0 – 20

3.4 Prosedur Penelitian



Gambar 3.1 Diagram Alir Prosedur Penelitian

Prosedur penelitian dapat dilihat pada diagram alir Gambar 3.1, yang terdiri dari beberapa tahapan. Tahapan pertama merupakan awal dari penelitian, dengan melakukan studi pendahuluan berupa diskusi atau wawancara dengan pihak yang berkaitan serta studi literatur mengenai topik penelitian. Setelah itu, dilakukan proses pengumpulan data penelitian dengan studi literatur, hal ini berkaitan dengan penggunaan data sekunder karena keterbatasan pengambilan data yang terjadi akibat pandemik Covid-19.

Tahap selanjutnya merupakan tahap pra-pemrosesan data, dimana data akan diproses terlebih dahulu sebelum diolah. Data yang didapatkan dibuat dalam satu database berbentuk .csv, kemudian transformasi data dilakukan untuk menyesuaikan data dalam format yang diperlukan saat pengolahan.

Tahap berikutnya adalah tahap utama, yaitu pengolahan data menggunakan Algoritma Random Forest untuk identifikasi dini siswa yang mengalami kendala akademik. Pada tahapan ini, implementasi dilakukan dengan menggunakan *library scikit-learn* klasifikasi Random Forest yang terdapat pada perangkat lunak Anaconda Navigator, dengan *template* yang diperoleh dari salah satu projek pembelajaran daring *Udacity Machine Learning Nanodegree* (T. Smith, 2016; Udacity, 2016). Setelah hasil pengolahan data didapatkan, hasil tersebut dibandingkan dengan penelitian-penelitian sebelumnya yang serupa untuk memberi batas-batas tertentu terhadap hasil implementasi, dan modifikasi dilakukan apabila hasil tidak sesuai.

Tahap terakhir merupakan penyajian dan analisis berdasarkan hasil penelitian yang dilakukan secara deskriptif, beserta pengambilan kesimpulan, implikasi, dan rekomendasi penelitian.

3.5 Analisis Data

Analisis data untuk memperoleh hasil identifikasi dini siswa yang mengalami kendala akademik dilakukan dengan mengimplementasikan salah satu algoritma klasifikasi *Machine Learning*, yaitu Random Forest dengan tahapan sebagai berikut.

1. Penginputan dan pra-pemrosesan dataset.

2. Pembagian dataset ke dalam data latih dan data uji.
3. Pembuatan model Random Forest, lalu dilakukan *training* dengan data latih, dan dilakukan *testing* dengan data uji. Proses *training* dan *testing* dilakukan beberapa kali menggunakan parameter *n_est* yang berbeda sebagai acuan awal.
4. Menampilkan persentase fitur/atribut data yang paling berpengaruh terhadap identifikasi siswa yang mengalami kendala akademik.
5. Pencarian parameter *n_est* berdasarkan acuan awal yang diperoleh dari tahapan ketiga, untuk mendapatkan model terbaik.
6. Evaluasi model dengan menggunakan *confusion matrix*, Kurva ROC dan AUC *Score* sebagai representasi hasil implementasi model terhadap data uji.
   a. *Confusion Matrix*

      Suatu model dari algoritma klasifikasi akan mengelompokkan data ke dalam kelas tertentu. Keputusan dari model tersebut direpresentasikan dalam bentuk tabel atau biasa disebut *confusion matrix*, yang terdiri dari 4 kategori yaitu TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negative*). Selain itu, dari *confusion matrix* ini juga akan diperoleh beberapa nilai seperti spesifisitas, NPV (*Negative Predictive Value*), serta akurasi (Davis & Goadrich, 2016).

   b. Kurva ROC dan AUC *Score*

      ROC (*Receiver Operating Characteristic*) merupakan kurva yang merepresentasikan TP *Rate* sebagai fungsi dari FP *Rate*. Sebuah model dari suatu algoritma klasifikasi yang sempurna akan direpresentasikan dengan kurva yang mendekati titik koordinat (0,1) pada ruang ROC, maka dari itu semakin baik suatu model, kurva akan mendekati titik tersebut. Sedangkan AUC (*Area Under Curve*) *Score* merupakan nilai yang diperoleh dari luas dibawah Kurva ROC, jika model tersebut sempurna maka nilai akan sama dengan 1 (Galdi & Tagliaferri, 2018).