BAB III
METODOLOGI PENELITIAN

3.1. Subjek dan Object Penelitian

Subjek penelitian pemodelan klasifikasi data tidak seimbang.Objekpenelitian ini adalah penggunaan teknik *oversampling* SMOTE untuk meningkatkan akurasi model.

3.2. Alat dan Bahan

Alat:

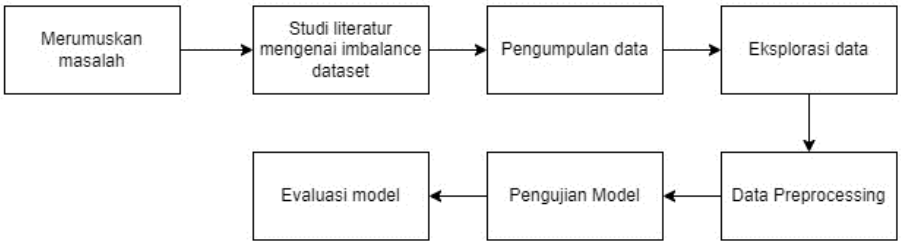
1. Laptop IdeaPad 3 14IIL05 Core i3-1005G1
2. Python 3.9

Bahan:

1. Imbalanced dataset

3.3. Alur Penelitian

Penelitian ini menggunakan metode machine learning untuk mengklasifikasikan data yang tidak seimbang. Proses pada alur 3.1 memberikan arah yang jelas mengenai tahapan-tahapan pada penelitian ini



Gambar 3.1 Alur Penelitian

3.4. Merumusan Masalah

Tahap Merumuskan Masalah memiliki peran sentral dalam membentuk landasan proyek analisis data. Dalam penelitian ini, fokus utama adalah menyelesaikan masalah terkait ketidakseimbangan dataset. Ketidakseimbangan

dataset sering kali menjadi tantangan dalam analisis data, terutama dalam konteks klasifikasi di mana kelas-kelas memiliki distribusi yang tidak merata. Membuat model memiliki kencendrungan untuk mengandalkan kelas mayoritas

3.5. Studi Literatur

Setelah tahap Merumuskan Masalah yang telah mengidentifikasi ketidakseimbangan dalam dataset sebagai isu yang perlu diselesaikan, langkah berikutnya adalah melakukan Studi Literatur yang berfokus pada penelitian dan pendekatan yang telah ada terkait penanganan dataset yang tidak seimbang.

Melalui studi literatur ini, dapat diperoleh wawasan tentang bagaimana para peneliti sebelumnya telah menghadapi masalah yang serupa dan menerapkan solusi dalam konteks dataset yang tidak seimbang. Studi literatur ini dapat memberikan panduan berharga dalam memilih pendekatan yang sesuai untuk proyek ini, serta menghindari perluasan tenaga kerja dalam mengembangkan solusi yang mungkin telah dijelajahi dalam literatur sebelumnya.

3.6. Pengumpulan Data

Data yang digunakan dalam penelitian berasal dari tujuh dataset yang dikumpulkan dari *UCI Machine Learning Repository* (https://archive-beta.ics.uci.edu/datasets) dan juga Kaggle. Ke empat dataset yang dipilih memiliki dua kelas target atau termasuk ke dalam kelas biner *(binary classification*).

3.6.1. *Haberman's Survival*

Dataset ini memuat kasus-kasus yang terkait dengan sebuah studi yang dilakukan terhadap kelangsungan hidup pasien setelah menjalani operasi untuk kanker payudara. Studi tersebut dilakukan dalam rentang waktu antara 1958 dan 1970 di Rumah Sakit Billings, Universitas Chicago. Informasi dataset ini mencakup kasus-kasus yang dapat diakses melalui tautan https://archive.ics.uci.edu/dataset/43/haberman+s+survival.

Dataset ini mencakup berbagai data seperti usia pasien, tahun operasi, jumlah nodus aksila positif yang terdeteksi, dan status kelangsungan hidup pasien

setelah operasi. Tujuan utama dataset ini adalah untuk melakukan klasifikasi apakah pasien akan hidup lebih dari 5 tahun setelah operasi atau tidak.

Tabel 3.1 Dataset Haberman's Suvival

No	Age	Op_Year	Nb_pos_detected	Surv_status
1	30	64	1	1
2	30	62	3	1
3	30	65	0	1
4	31	59	2	1
5	31	65	4	1
...	...	...	...	...
302	75	62	1	1
303	76	67	0	1
304	77	65	3	1
305	78	65	1	2
306	83	58	2	2

3.6.2. Indian Liver Patient Records

Dataset ini digunakan untuk mengevaluasi algoritma prediksi dalam upaya untuk mengurangi beban kerja bagi para dokter. Dataset ini terdiri dari 416 rekam medis pasien dengan penyakit liver dan 167 rekam medis pasien tanpa penyakit liver yang dikumpulkan dari bagian Timur Laut Andhra Pradesh, India. Kolom "Dataset" merupakan label kelas yang membagi kelompok menjadi pasien dengan penyakit liver atau tidak (tanpa penyakit). Dataset ini mencakup 441 rekam medis pasien pria dan 142 rekam medis pasien wanita. Setiap pasien yang usianya melebihi 89 tahun dicatat sebagai usia "90". Kolom-kolom dalam dataset mencakup usia pasien, jenis kelamin, total bilirubin, direct bilirubin, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, total protiens, albumin, albumin and globulin ratio, serta kolom "Dataset" yang membagi data menjadi dua kelompok (pasien dengan penyakit liver dan tanpa penyakit). Dataset ini diunduh dari *UCI ML Repository* dan dapat digunakan untuk menentukan pasien yang menderita penyakit liver dan yang tidak

Tabel 3.2 Dataset Indian Liver Patient Records

No	Age	Gender	Alkaline_Phosphotase	...	Dataset
1	65	Female	187	...	1
2	62	Male	699	...	1
3	62	Male	490	...	1
4	58	Male	182	...	1
5	72	Male	195	...	1
...	...	...	...	...	...
579	60	Male	500	...	2
580	40	Male	98	...	1
581	52	Male	245	...	1
582	31	Male	184	...	1
583	38	Male	216	...	2

3.6.3. Adult

Adult merupakan sebuah data yang digunakan untuk klasifikasi pendapatan seseorang berdasarkan umur dan beberapa variabel lainnya[37]. dengan ukuran 32.562 baris dan 16 kolom. Dataset ini berisi kumpulan rekaman data yang telah diolah dengan ketentuan berikut: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). Tujuan prediksinya adalah untuk menilai apakah seseorang memiliki penghasilan > $50K dan <=per tahun. Contoh data Adult terdapat pada Tabel 3.3.

Tabel 3.3 Dataset Adult

No	age	...	fnlwgt	native.country	income
1	90	...	77053	United-States	<=50K
2	82	...	132870	United-States	<=50K
3	66	...	186061	United-States	<=50K
4	54	...	140359	United-States	<=50K
5	41	...	264663	United-States	<=50K
...	...	...	...	...	...
32557	22	...	310152	United-States	<=50K

32558	27	...	257302	United-States	<=50K
32559	40	...	154374	United-States	>50K
32560	58	...	151910	United-States	<=50K
32561	22	...	201490	United-States	<=50K

3.6.4. Credit Risk

Dataset Credit Risk ini berisi simulasi data dari lembaga kredit yang mencakup berbagai kolom untuk mensimulasikan informasi dari lembaga kredit. Data ini memberikan wawasan tentang dunia evaluasi risiko kredit yang kompleks. Dataset ini terdiri dari berbagai kolom yang secara kolektif mencerminkan atribut-atribut yang umumnya digunakan dalam skenario penilaian kredit. Atribut-atribut ini, ketika diinterpretasikan dan dianalisis, memungkinkan lembaga keuangan dan agensi kredit untuk membuat keputusan yang berdasar pada kelayakan kredit seseorang.

Tabel 3.4 Dataset *Credit Risk*

No	person_age	person_income	...	loan_status	cb_person_cred_hist_length
1	22	59000	...	1	3
2	21	9600	...	0	2
3	25	9600	...	1	3
4	23	65500	...	1	2
5	24	54400	...	1	4
...	...	...	...		
32577	57	53000	...	0	30
32578	54	120000	...	0	19
32579	65	76000	...	1	28
32580	56	150000	...	0	26
32581	66	42000	...	0	30

3.7. Eksplorasi Data

Tahap eksplorasi data merupakan langkah yang penting dalam proses analisis. Pada tahap ini, dilakukan eksplorasi secara menyeluruh terhadap dataset

yang telah terkumpul. Tujuannya adalah untuk mendapatkan pemahaman yang lebih mendalam tentang karakteristik-karakteristik yang terdapat dalam data tersebut. Dalam konteks eksplorasi data, analis melakukan berbagai metode visualisasi seperti grafik, histogram, diagram sebar, dan plot lainnya untuk membantu menggambarkan distribusi data. Hal ini membantu mencerahkan karakteristik data, serta memberikan wawasan awal tentang bagaimana variabel-variabel saling berhubungan. Selain itu, eksplorasi data juga berfungsi untuk mengenali potensi masalah yang mungkin ada dalam dataset. Contohnya, dapat ditemukan adanya *outlier* (data yang jauh dari nilai-nilai lainnya), kecacatan atau ketidaklengkapan data, serta indikasi ketidakseimbangan antara kelas dalam kasus dataset klasifikasi. Semua informasi ini penting untuk merumuskan strategi yang lebih baik dalam tahap pra-pemrosesan dan pembangunan model selanjutnya.

Pada akhir tahap eksplorasi data, diharapkan bahwa analis mendapatkan gambaran yang lebih kaya tentang data yang akan diolah. Informasi yang ditemukan dapat membantu mengarahkan langkah-langkah berikutnya dalam proses analisis, serta memastikan bahwa solusi yang dihasilkan relevan dan efektif dalam menangani permasalahan atau pertanyaan yang ingin dijawab dalam proyek ini.

3.8. *Data Preprocessing*

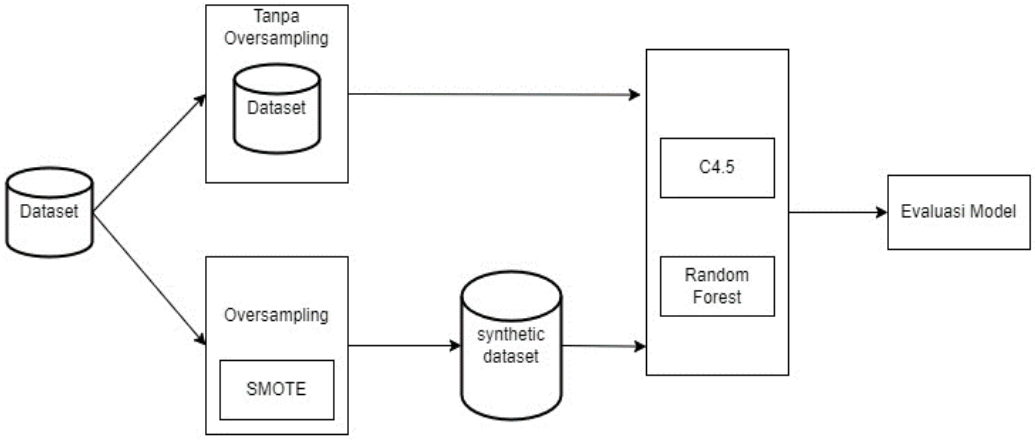
Penelitian ini melibatkan beberapa tahapan preprocessing, seperti penghapusan data yang tidak relevan, imputasi data, *labelling*, dan standarisasi data menggunakan *robust scaller*, dan *oversampling* menggunakan SMOTE. Penghapusan data yang tidak relevan menjadi langkah krusial dalam *preprocessing*, yangbertujuan untuk meningkatkan kualitas dataset dengan mengeliminasi informasi yang tidak memberikan kontribusi signifikan pada analisis atau pembangunan model. Setelah itu, dilakukan imputasi data untuk menangani nilai-nilai yang hilang dalam dataset, membantu melengkapi informasi yang kosong, serta memastikan integritas dan validitas data.

Langkah *labelling* digunakan untuk mengubah data kategorikal menjadi format yang dapat dipahami oleh model, memastikan bahwa model dapat memproses informasi dengan baik.

Lalu langkah standarisasi data dilakukan untuk menstandarisasi rentang nilai setiap fitur dalam dataset, dan terakhir SMOTE digunakan untuk meyeimbangkan antara jumlah kelas pada data yang tidak seimbang. Keseluruhan langkah-langkah *preprocessing* ini dirancang untuk menyelaraskan dataset agar siap digunakan dalam analisis lebih lanjut atau untuk melatih model dengan tingkat akurasi yang tinggi sesuai dengan tujuan penelitian ini.

3.9. Pengujian Model

Setelah tahap eksplorasi data dan *data preprocessing*, langkah berikutnya adalah membangun model menggunakan metode *Random Forest* dan *C4.5* Tujuan utama dari tahap ini adalah mengembangkan model yang dapat memahami dan menjelaskan pola dalam data serta memberikan prediksi atau solusi yang berguna. Berikut merupakan alur dari dari pengujian algoritma *Random Forest* dan *C4.5* ditunjukkan pada Gambar 3.2



Gambar 3. 2 Alur Pengujian Algoritma

1. Data yang telah dibersihkan akan melewati dua proses berbeda. Pertama, proses oversampling dengan menggunakan metode *SMOTE*. Kedua, terdapat proses tanpa *oversampling*, di mana data yang tidak melalui proses

oversampling langsung akan diarahkan ke tahap pengujian model menggunakan algoritma *Random Forest* dan *C4.5*.

2. Data yang telah melalui proses oversampling akan menghasilkan data sintetis, yang selanjutnya akan diklasifikasikan dengan menggunakan algoritma *Random Forest* dan *C4.5*.
3. Setelah proses pengujian dengan algoritma *Random Forest* dan *C4.5*.
4. selesai dilakukan, langkah selanjutnya adalah mengevaluasi kinerja hasil klasifikasinya.

3.10. Evalusi Data

Tahap evaluasi melibatkan penilaian kinerja model atau hasil analisis. Model yang dikembangkan dievaluasi berdasarkan kriteria yang ditentukan sebelumnya, yaitu akurasi model.