BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Perusahaan asuransi seringkali mengalami kerugian finansial yang signifikan akibat adanya klaim asuransi yang tidak tepat atau dipalsukan. Klaim asuransi yang dipalsukan adalah klaim yang sengaja dibuat atau dimanipulasi untuk mendapatkan pembayaran yang tidak pantas. Dampak dari klaim asuransi yang dipalsukan sangat merugikan perusahaan asuransi. Perusahaan asuransi harus menanggung biaya klaim yang tidak seharusnya dibayarkan, yang dapat mengurangi keuntungan perusahaan dan mengganggu stabilitas keuangan mereka. Selain itu, klaim asuransi yang tidak tepat atau dipalsukan juga dapat meningkatkan premi asuransi secara keseluruhan, karena perusahaan asuransi perlu mengkompensasi kerugian yang terjadi. Oleh karena itu, perusahaan asuransi perlu menggunakan metode yang tepat dalam melakukan klasifikasi untuk memprediksi klaim asuransi agar dapat mengurangi kerugian dan meningkatkan efisiensi dalam proses klaim asuransi. Penelitian ini bertujuan untuk membandingkan kinerja beberapa algoritma *machine learning* dalam memprediksi klaim asuransi.

Tujuan dari penelitian ini adalah untuk membandingkan performa beberapa algoritma *Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor* dan *Logistic Regression* menggunakan teknik sampling SMOTE dan *Undersampling* pada data klaim asuransi mobil dalam memprediksi klaim asuransi kendaraan dengan menggunakan teknik visualisasi, serta membangun model klasifikasi data klaim asuransi dengan menggunakan RapidMiner.

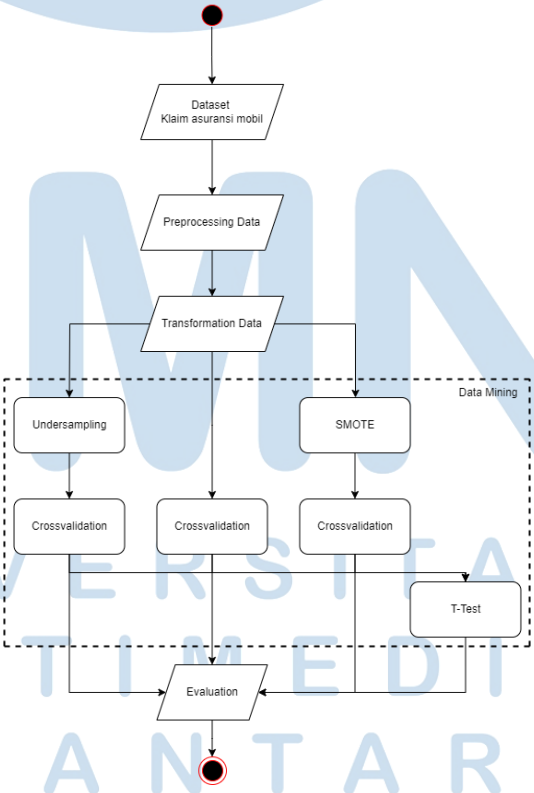
3.2 Metode Penelitian

Metode Komparatif merupakan metode dengan membandingkan suatu variabel antara dua atau lebih kelompok atau sampel. Metode ini digunakan untuk membandingkan perbedaan atau kesamaan antara kelompok atau sampel yang berbeda. Hasil evaluasi performa model dari algoritma dikumpulkan dan dianalisis, dengan menggunakan matrik-matrik yang relevan seperti *accuracy, precision* dan

recall. Kemudian, hasil analisis tersebut diinterpretasikan untuk menentukan algoritma mana yang memiliki performa lebih baik dalam menyelesaikan masalah klasifikasi data klaim asuransi mobil.

3.2.1 Alur Penelitian

Knowledge Discovery in Database (KDD) atau Penemuan Pengetahuan dalam Basis Data adalah proses ekstraksi informasi yang bermanfaat dari basis data besar atau kompleks [4]. KDD mencakup tahap pra-pemrosesan data yang melibatkan pemahaman dan persiapan data yang komprehensif sebelum tahap pemodelan dan evaluasi. Ini termasuk pemilihan data yang relevan, pembersihan data, integrasi data, transformasi data, dan langkah-langkah lain untuk memastikan kualitas data yang baik sebelum proses data mining dimulai. Proses KDD melibatkan langkah-langkah seperti pemilihan data, pemrosesan data, transformasi data, pemodelan dan evaluasi hasil untuk menghasilkan pengetahuan yang berguna bagi pengambilan keputusan.



Gambar 3.1 Alur Penelitian

KDD merupakan suatu proses yang kompleks dan melibatkan 5 tahap yaitu:

1. Pemilihan Data

   Dataset klaim asuransi mobil dipilih dari kaggle.com, sebuah situs *open-source* yang menyediakan berbagai kumpulan data. Data tersebut dikumpulkan menggunakan perangkat lunak *Angoss Knowledge Seeker* dari Januari 1994 hingga Desember 1996. Terdapat 33 atribut dalam dataset ini, dan atribut kelasnya menunjukkan apakah sebuah klaim dianggap sebagai penipuan atau tidak. Totalnya, terdapat 15.420 catatan klaim dalam format excel.

2. *Preprocessing Data*

   Pada tahap ini, Salah satu tugas utama dalam tahap ini adalah membersihkan data dari kecacatan seperti data yang hilang dan duplikat. peneliti dapat memastikan bahwa data yang digunakan untuk analisis selanjutnya bebas dari cacat dan keberatan, sehingga menghasilkan hasil yang lebih akurat dan dapat diandalkan dalam proses data mining.

3. Transformasi Data

   Pada tahap ini, Data diubah atau diubah menjadi bentuk yang lebih mudah dipahami dan diolah oleh rapid miner, seperti mengubah format data.

4. *Data mining*

   Pada tahap ini, model *machine learning* dibangun untuk memprediksi apakah klaim asuransi merupakan *fraud* atau tidak. Beberapa algoritma *machine learning* yang dapat digunakan, antara lain *Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbor* dan *Logistic Regression*.

5. Evaluasi *Data Mining*

   Model yang dibangun dievaluasi untuk mengetahui seberapa baik performanya dalam memprediksi label *fraud*. Menggunakan nilai AUC dan Matrik yang dapat digunakan, antara lain *accuracy*, *precision* dan *recall*.

3.3 Teknik Pengumpulan Data

Data sekunder adalah data yang telah dikumpulkan oleh pihak lain untuk tujuan lain dan tersedia untuk digunakan oleh pihak lain untuk keperluan penelitian atau analisis. Pada penelitian ini data klaim asuransi mobil diambil dari kaggle.com https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection, Data ini memiliki 33 atribut Pada atribut *fraudfound* memiliki 14447 No dan 923 Yes dan berisi 15.420 data.

Tabel 3.1 Atribut Data

No	Atribut	Deskripsi	Tipe
1	*Month*	Bulan-bulan terjadinya kecelakaan	*String*
2	*Weekofmonth*	Minggu di bulan kecelakaan terjadi	Int
3	*DayOfWeek*	hari-hari kecelakaan terjadi	*String*
4	*Make*	Produsen model mobil	*String*
5	*AccidentArea*	Daerah kecelakaan pedesaan atau perkotaan	*String*
6	*DayOfWeekClaimed*	Berisi hari dalam seminggu klaim diajukan	*String*
7	*MonthClaimed*	Berisi hari dari bulan pengajuan klaim	*String*
8	*WeekOfMonthClaimed*	Berisi minggu dalam bulan yang diklaim di lapangan	Int
9	*Sex*	Jenis kelamin yang membuat klaim	*String*
10	*MaritalStatus*	Status perkawinan yang membuat klaim	*String*

No	Atribut	Deskripsi	Tipe
11	*Age*	Usia individu yang membuat klaim	Int
12	*PoliceReportFiled*	Menunjukkan apakah ada laporan polisi untuk kecelakaan itu	*String*
13	*Fault*	Kategorisasi siapa yang dianggap bersalah	*String*
14	*BasePolicy*	Jenis pertanggungan asuransi	*String*
15	*NumberOfCars*	Jumlah mobil yang terlibat kecelakaan	*String*
16	*AddressChange_Claim*	Waktu dari klaim diajukan hingga saat orang tersebut pindah	*String*
17	*NumberOfSuppliments*	Total pelayanan tambahan	*String*
18	*WitnessPresent*	Menunjukkan apakah saksi hadir	*String*
19	*AgentType*	Agen yang menangani klaim	Int
20	*AgeOfPolicyHolder*	Perkiraan Usia pemegang polis	String
21	*VehicleCategory*	Kategori kendaraan	String
22	*PolicyType*	1. Jenis asuransi 2. Kategori kendaraan	String
23	*VehiclePrice*	Kisaran untuk harga kendaraan	String
24	*Deductible*	Jumlah uang yang harus dibayarkan oleh pemegang polis asuransi	Int
25	*AgeOfVehicle*	Usia kendaraan pada saat kecelakaan	String

No	Atribut	Deskripsi	Tipe
26	*PastNumberOfClaims*	Jumlah klaim sebelumnya	String
27	*Days_Policy_Claim*	Jumlah hari pengajuan pembelian dan klaim	String
28	*RepNumber*	Nomor perwakilan	Int
29	*Days_Policy_Accident*	Jumlah hari antara kecelakaan itu terjadi	String
30	*DriverRating*	Rating Pengemudi	Int
31	*FraudFound*	apakah klaim itu palsu	String
32	*PolicyNumber*	Nomor polis	Int
33	*Year*	Tahun klaim	int

3.4 Populasi dan Sampel

Populasi mengacu pada keseluruhan kumpulan data yang tersedia atau diinginkan untuk dianalisis. sampel merujuk pada subset yang diambil dari populasi yang lebih besar untuk analisis lebih lanjut.

3.4.1 Populasi

Populasi dalam penelitian ini adalah seluruh data klaim asuransi mobil yang memiliki label *fraud* dalam kurun waktu tertentu dengan total 33 atribut. Populasi ini mencakup semua klaim asuransi mobil yang dilaporkan dan telah ditandai atau diberi label *fraud* oleh perusahaan asuransi dalam periode waktu yang ditentukan.

3.4.2 Sampel

Sampel dalam penelitian ini adalah sejumlah data klaim asuransi mobil yang dipilih secara acak dari populasi yang lebih besar. Sampel tersebut digunakan untuk mewakili karakteristik dan variabilitas klaim asuransi mobil secara keseluruhan dalam populasi tersebut. Dengan menggunakan metode pemilihan acak, diharapkan bahwa sampel ini

memberikan gambaran yang representatif dan dapat digeneralisasi terhadap populasi yang lebih besar. Dengan menganalisis data klaim asuransi mobil dalam sampel ini, penelitian dapat menghasilkan temuan dan kesimpulan yang dapat diterapkan secara lebih luas pada populasi klaim asuransi mobil secara keseluruhan.

3.5 Variabel Penelitian

Variabel penelitian pada data klaim asuransi mobil dapat dibagi menjadi variabel dependen dan independen. Variabel dependen adalah variabel utama yang menjadi fokus penelitian, sedangkan variabel *independen* adalah variabel yang berpotensi mempengaruhi atau berhubungan dengan variabel *dependen* tersebut.

3.5.1 Variabel *Independen*

Variabel *independen* adalah variabel yang mempengaruhi atau memengaruhi terjadinya *fraud* pada klaim asuransi mobil. Beberapa contoh variabel independen pada penelitian ini antara lain jenis kendaraan, lokasi klaim, biaya klaim, umur pengemudi, jenis kelamin pengemudi, jenis klaim, dan sebagainya.

3.5.2 Variabel *Dependen*

Variabel *dependen* adalah variabel yang ingin diteliti atau dijelaskan, yaitu terjadinya *fraud* pada klaim asuransi mobil. Variabel dependen ini, dalam konteks yang disebutkan, diukur menggunakan data yang telah ditandai sebagai *fraud* oleh perusahaan asuransi. Variabel ini akan digunakan sebagai ukuran kinerja algoritma. Variabel ini akan digunakan sebagai tolok ukur performa algoritma dengan Kurva ROC dengan matrik AUC (*Area Under the Curve*) dan berbagai matrik *Confusion matrix*, seperti *accuracy*, Presisi, *Recall*.

3.6 Teknik Analysis

Teknik analisis komparatif adalah suatu metode yang digunakan untuk membandingkan dan menganalisis perbedaan antara dua atau lebih objek, variabel, atau kelompok. Tujuan utama dari analisis komparatif adalah untuk

mengidentifikasi perbedaan, kesamaan, keunggulan, atau kelemahan antara objek yang dibandingkan. Pada penelitian ini Teknik komparatif digunakan untuk membandingkan performa algoritma *Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbor* dan *Logistic Regresssion* dengan Teknik sampling SMOTE dan *Undersampling* menggunakan *Crossvalidation* untuk membandingkan *acuracy, precision, Recall* dan AUC kemudian menggunakan T-test untuk melihat apakah memiliki perbedaan yang signifikan. Hasil dari performa algoritma yang paling optimal dengan Teknik sampling SMOTE, tanpa Teknik sampling dan *Undersampling* akan dibandingkan lagi dengan algorima lainnya sehingga dapat menemukan algoritma yang paling optimal dengan Teknik terbaik.