

**ANALISIS PREDIKSI CHURN PELANGGAN
PERUSAHAAN TELEKOMUNIKASI DENGAN
MEMANFAATKAN PROSES *DATA MINING* NAIVE
BAYES**

PRA-TUGAS AKHIR



*temp late
Service*

Oleh :

Muhammad Daffa Nur Alif

162019028

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI NASIONAL BANDUNG
2023**

INTISARI

Untuk melindungi eksistensinya, perusahaan telekomunikasi cenderung meningkatkan kegiatan promosi dan mengembangkan lebih banyak inovasi saat ini dengan demikian, mereka dapat mencegah konsumennya beralih ke pesaing(Zeniarja & Luthfiarta, 2015). Banyak perusahaan telekomunikasi yang mengalami *customer churn* terutama karena kesalahan penargetan dalam menerapkan program retensi. Secara teoritis, tujuan utama dari program retensi adalah untuk mencegah potensi churn, yang berarti program retensi seharusnya ditargetkan hanya kepada pelanggan yang berpotensi menjadi tidak loyal. Namun, dalam kehidupan nyata, program retensi sering kali diterapkan pada semua pelanggan perusahaan, termasuk pelanggan yang berpotensi loyal, sehingga program-program tersebut berjalan dengan sia-sia karena perusahaan telah menghabiskan banyak uang untuk melaksanakan program tersebut. Oleh karena itu, prediksi *customer churn* yang didasarkan pada klasifikasi karakteristik loyalitas pelanggan sangat bermanfaat dalam menentukan bagaimana program retensi akan dijalankan secara tepat sehingga perusahaan dapat menghemat sumber dayanya dan mempertahankan pelanggan semaksimal mungkin. Analisis klasifikasi *data mining* adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya, disebut juga dengan supervised learning. Metode-metode yang telah dikembangkan oleh periset untuk menyelesaikan kasus klasifikasi, antara lain: Pohon keputusan (Decision Tree), Naïve Bayes, Jaringan Syaraf Tiruan, Analisis Statistik, Algoritma Genetik, Rough Sets, kNearest Neighbour, Metode Berbasis Aturan, Memory Based Reasoning, Support Vector Machine. Pada penelitian prediksi costumer churn ini menggunakan metode klasifikasi *data mining* *Naïve bayes Classifier* yang merupakan sebuah metode klasifikasi yang berdasarkan teorema bayes. Algoritma ini digunakan untuk mengklasifikasikan data dengan menggunakan metode probabilitas dan statistik, yang bertujuan memprediksi peluang di masa depan berdasarkan pengalaman di masa lampau. Berdasarkan penjelasan tersebut, tujuan dari penelitian ini adalah untuk mengetahui bagaimana klasifikasi menggunakan algoritma *Naïve bayes* dapat memprediksi *customer churn* pada industri telekomunikasi.

DAFTAR ISI

DAFTAR ISI.....	ii
DAFTAR GAMBAR	iv
DAFTAR TABEL.....	v
BAB I	1
PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	3
1.3. Tujuan	4
1.4. Ruang Lingkup	4
1.5. Sistematika penulisan	4
BAB II.....	6
KAJIAN TEORI	6
2.1. Literatur Mapping.....	6
2.2. <i>Data mining</i>	6
2.3. Posisi <i>Data mining</i>	8
2.4. Pekerjaan Dalam <i>Data mining</i>	8
2.4.1. model prediksi	9
2.4.2. analisis kelompok	9
2.4.3. analisis asosiasi.....	10
2.4.4. deteksi anomali	10
2.5. <i>Dataset</i>	11
2.6. Klasifikasi	13
2.7. Naive Bayes Classifier.....	13
2.8. <i>Customer churn</i>	14

2.9.	Prediksi Csutomer Churn.....	15
2.10.	CRISP <i>Data mining</i>	16
2.11.	Confusion Matrix.....	19
2.12.	EDA	20
2.13.	SMOTE.....	20
2.14.	Tinjauan Pustaka.....	21
	BAB III	24
	METODOLOGI PENELITIAN	24
3.1.	Alur Penelitian.....	24
3.2.	<i>Data Collection</i>	24
3.3.	<i>Data Understanding</i>	25
3.4.	<i>Data Preparation</i>	27
3.5.	<i>Modeling</i>	28
3.6.	<i>Prediction</i>	29
3.7.	<i>Evaluation</i>	29
	DAFTAR PUSTAKA	30

DAFTAR GAMBAR

Gambar 2. 1. <i>Literatur map</i> penelitian.....	6
Gambar 2. 2. <i>Data mining</i>	7
Gambar 2. 3. Posisi <i>data mining</i> pada penelitian.....	8
Gambar 2. 4. Pekerjaan <i>data mining</i>	9
Gambar 2. 5. Tahapan pada CRISP-DM.....	17
Gambar 3. 1. Alur penelitian.....	24
Gambar 3. 2. Diagram <i>pie attribut churn</i>	27

DAFTAR TABEL

Tabel 2. 1. Tipe atribut <i>dataset</i>	12
Tabel 2. 2. <i>Confusion matrix</i>	19
Tabel 2. 3. Tinjauan pustaka	21
Tabel 3. 1. Definisi atribut <i>dataset customer churn</i>	25

DAFTAR RUMUS

DAFTAR ISTILAH

BAB I

PENDAHULUAN

1.1. Latar Belakang

Teknologi informasi saat ini sudah berkembang dengan cepat. Masyarakat kini menggunakan telekomunikasi, seperti internet, sebagai cara untuk berkomunikasi jarak jauh dan dalam waktu yang singkat (Zeniarja & Luthfiarta, 2015). Meningkatnya kebutuhan komunikasi jarak jauh ini mengakibatkan persaingan bisnis yang ketat di antara perusahaan. Untuk melindungi eksistensinya, perusahaan telekomunikasi cenderung meningkatkan kegiatan promosi dan mengembangkan lebih banyak inovasi saat ini dengan demikian, mereka dapat mencegah konsumennya beralih ke pesaing. Menurut Customer Relationship Management, mempertahankan pelanggan yang sudah ada adalah strategi pemasaran yang lebih baik daripada mencari pelanggan baru. Mencari pelanggan baru lebih tidak efisien karena membutuhkan lebih banyak waktu dan biaya(Herawati et al., 2016).

Customer churn merupakan kondisi ketika perusahaan kehilangan pelanggannya karena pelanggan berhenti berlangganan atau membeli produk setelah beberapa saat(Arina & Ulfah, 2022). Banyak perusahaan telekomunikasi yang mengalami *customer churn* terutama karena kesalahan penargetan dalam menerapkan program retensi. Secara teoritis, tujuan utama dari program retensi adalah untuk mencegah potensi churn, yang berarti program retensi seharusnya ditargetkan hanya kepada pelanggan yang berpotensi menjadi tidak loyal(Novendri et al., 2021). Tindakan ini harus dilakukan dengan harapan bahwa pelanggan tersebut akan memiliki beberapa alasan untuk mencegah niat mereka untuk meninggalkan perusahaan. Namun, dalam kehidupan nyata, program retensi sering kali diterapkan pada semua pelanggan perusahaan, termasuk pelanggan yang berpotensi loyal, sehingga program-program tersebut berjalan dengan sia-sia karena perusahaan telah menghabiskan banyak uang untuk melaksanakan program tersebut (Wardani et al., 2018).

Faktor .

Penyebab utama dari *customer churn* biasanya terdiri dari beberapa faktor utama seperti ketidak sesuaian harga yang ditawarkan kepada pelanggan, pengalaman pelanggan yang kurang memuaskan, ketertarikan pelanggan yang berkurang karena kurangnya inovasi yang ditawarkan, produk yang ditawarkan tidak sesuai kebutuhan, dan pelanggan menemukan produk lain yang lebih baik dari produk yang ditawarkan(Yulianti, 2018).

Oleh karena itu, prediksi *customer churn* yang didasarkan pada klasifikasi karakteristik loyalitas pelanggan sangat bermanfaat dalam menentukan bagaimana program retensi akan dijalankan secara tepat sehingga perusahaan dapat menghemat sumber dayanya dan mempertahankan pelanggan semaksimal mungkin(Wardani et al., 2022).

Analisis klasifikasi *data mining* adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya, disebut juga dengan supervised learning(Batubara & Windarto, 2019). Metode-metode yang telah dikembangkan oleh periset untuk menyelesaikan kasus klasifikasi, antara lain: Pohon keputusan (Decision Tree), Naïve Bayes, Jaringan Syaraf Tiruan, Analisis Statistik, Algoritma Genetik, Rough Sets, kNearest Neighbour, Metode Berbasis Aturan, Memory Based Reasoning, Support Vector Machine.

Pada penelitian prediksi costumer churn ini menggunakan metode klasifikasi *data mining* *Naïve bayes Classifier* yang merupakan sebuah metode klasifikasi yang berdasarkan teorema bayes(Wardani et al., 2022). Algoritma ini digunakan untuk mengklasifikasikan data dengan menggunakan metode probabilitas dan statistik, yang bertujuan memprediksi peluang di masa depan berdasarkan pengalaman di masa lampau.

Kelemahan pada metode *data mining* ini terletak pada masalah probabilitas. *Naive Bayes* memiliki masalah probabilitas nol, terutama saat Anda menemukan kata-kata dalam data pengujian untuk kelas tertentu yang tidak ada dalam data pelatihan. Kemungkinan besar Anda akan berakhir

dengan probabilitas “Zero Frequency”. Meski begitu, probabilitas nol ini dapat diatasi dengan *smoothing* (teknik penghalusan). Tambahkan faktor penghalusan pada pembilang dan penyebut setiap probabilitas untuk menghindari munculnya nilai nol.

*data' y tidak
balance*



balancing data



Sampling

oversampling

SMOTE



*oleh karena
itu*

*penelitian
ini*

akhirnya menerapkan

SMOKE u/

meningkatkan

performa Naive Bayes.

Sebelum melakukan proses *modeling* menggunakan metode Naive Bayes. Data training yang akan dipakai harus melalui proses *balancing* data jika data yang terdapat pada *dataset* memiliki kelas yang tidak seimbang atau *imbalance*. Proses ini menggunakan metode balancing data yaitu dengan metode *sampling* dan *oversampling*.

} *stasi*

Variable yang digunakan dalam analisis prediksi costumer churn ini meliputi indikator churn yang berisi data apakah pelanggan churn atau tidak, data tentang layanan apa saja yang pengguna gunakan, data tentang akun pengguna layanan seperti berapa lama mereka berlangganan layanan dan metode pembayaran yang digunakan, dan data tentang informasi demografis pengguna layanan.

Berdasarkan penjelasan tersebut, tujuan dari penelitian ini adalah untuk mengetahui bagaimana klasifikasi menggunakan algoritma *Naive Bayes* yang dapat memprediksi *customer churn* pada industri telekomunikasi. Hasil dari penelitian ini diharapkan dapat digunakan perusahaan untuk menentukan program strategi retensi dan akuisisi yang akan dilakukan. Jika *churn rate* yang dimiliki oleh perusahaan telekomunikasi yang diteliti memiliki nilai yang tinggi maka perusahaan harus melakukan strategi retensi pelanggan dan jika *churn rate* yang dimiliki perusahaan telekomunikasi memiliki nilai yang kecil maka perusahaan lebih baik melakukan program strategi akuisisi pelanggan. Dan juga dengan dilakukannya penelitian ini, perusahaan telekomunikasi dapat melakukan program retensi dengan tepat sasaran kepada pelanggan yang berpotensi untuk meninggalkan perusahaan.

1.2. Rumusan Masalah

Dari analisis prediksi churn pelanggan menggunakan klasifikasi decision tree ini ditemukan beberapa rumusan masalah yaitu sebagai berikut:

1. Bagaimana cara ~~mengetahui prediksi~~ memprediksi costumer churn ~~dari perusahaan~~ berdasarkan ~~data~~ dan, telekomunikasi.
 2. Bagaimana cara analisis costumer churn memanfaatkan proses *data mining Naïve bayes Classifier*.
 3. Bagaimana ~~datanya~~ meningkatkan akurasi dengan mengatasi masalah ~~balancing data~~ penerapan smote pada langkah *preprocessing*. *us* meningkatkan.
- 1.3. Tujuan**

Berikut adalah beberapa tujuan dari pembuatan analisis prediksi churn pelanggan memanfaatkan proses *data mining Naïve bayes Classifier*:

1. Mengetahui prediksi costumer churn dari perusahaan telekomunikasi.
2. Mengetahui cara analisis costumer churn memanfaatkan proses *data mining Naïve bayes Classifier*.
3. Mengetahui cara meningkatkan akurasi dengan mengatasi *balancing data*

1.4. Ruang Lingkup

Pembuatan laporan ini hanya berisi pemaparan kajian pustaka dan metodologi dari penelitian yang akan dilakukan. Proses implementasi belum dilakukan pada penelitian ini.

1.5. Sistematika penulisan

Secara sistematis isi dari laporan ini disusun sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini akan berisi latar belakang, rumusan masalah, tujuan penelitian, ruang lingkup, dan sistematika penulisan laporan yang terkait dengan penelitian ini.

BAB II TINJAUAN PUSTAKA

Berisikan Tinjauan Pustaka dari penelitian Terkait, serta materi-materi yang terkait dengan topik yang dipilih.

BAB III METODOLOGI PENELITIAN

Berisikan tahapan – tahapan dalam penelitian untuk mencapai tujuan penelitian.

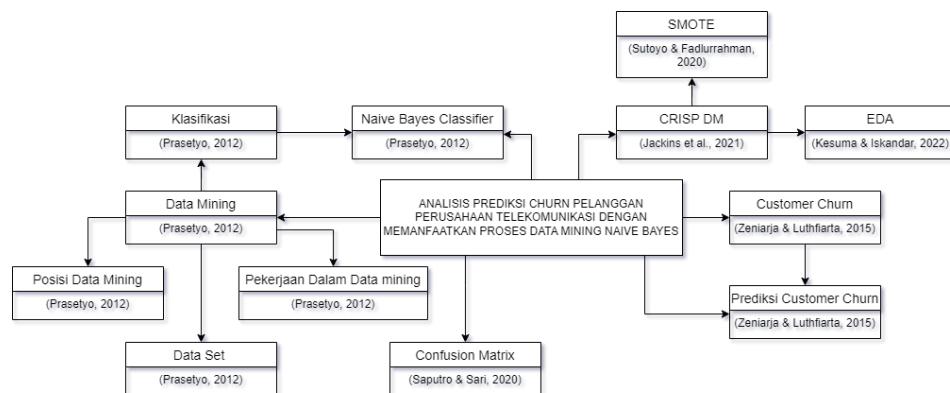
BAB II

KAJIAN TEORI

Pada bab ini berisi kajian pustaka yang diambil dari berbagai sumber tertulis seperti literatur, jurnal-jurnal dan penelitian-penelitian terdahulu yang relevan dengan topik penelitian ini

2.1. Literatur Mapping

Literatur map pada penelitian ini digunakan untuk mengarahkan penelitian agar analisis dan gagasan yang dihasilkan menjadi tepat. Setiap proses pada penelitian membutuhkan dasar analisis sehingga dapat menghasilkan keputusan yang sesuai dengan tujuan penelitian. Hal ini perlu mempertimbangkan teori-teori yang telah dibahas dan sumber-sumber teori lainnya, literatur map pada penelitian ini ditunjukkan oleh Gambar 2. 1.

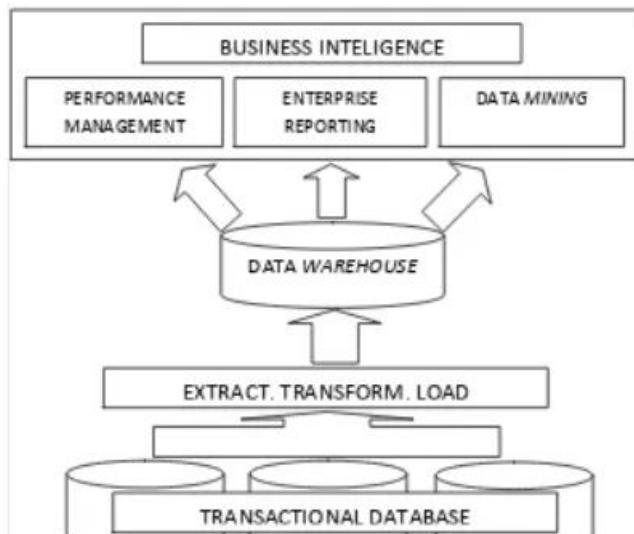


Gambar 2. 1. *Literatur map* penelitian

2.2. Data mining

Data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar (Prasetyo, 2012). *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga knowledge discovery.

Salah satu teknik yang dibuat dalam *data mining* adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam *data mining*, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai.n (Prasetyo, 2012)



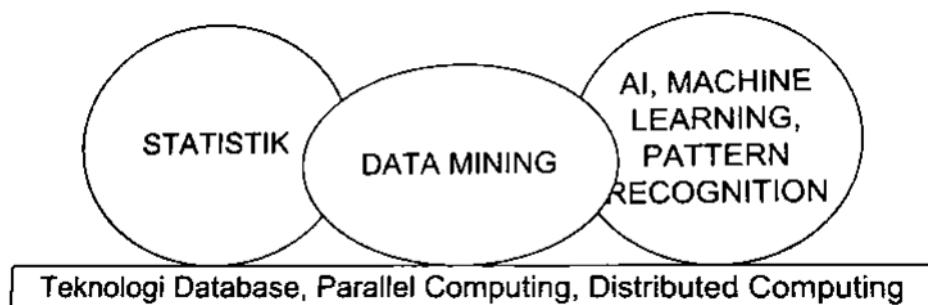
Gambar 2. 2. *Data mining*

Gambar 2. 2 mengilustrasikan posisi antara *data mining* dan data warehouse. Dari gambar ter- sebut, dapat dilihat bahwa *data mining* adalah bidang yang sepenuhnya menggunakan apa yang dihasilkan oleh data warehouse, bersama dengan bidang yang menangani masalah pelaporan dan manajemen data. Sementara, data warehouse sendiri bertugas untuk menarik/meng-query data dari basis data mentah untuk memberikan hasil data yang nantinya digunakan oleh bidang yang menangani manajemen, pelaporan, dan *data mining*. Dengan *data mining* inilah, penggalian Informasi baru dapat dilakukan dengan bekal data mentah yang diberikan oleh data

warehouse. Hasil yang diberikan oleh ketiga bidang tersebut berguna untuk mendukung aktivitas bisnis Cerdas (*business intelligence*).

2.3. Posisi *Data mining*

Algoritma pencarian, teknik pemodelan, dan teori pembelajaran, seperti yang ditunjukkan pada Gambar 2. 3

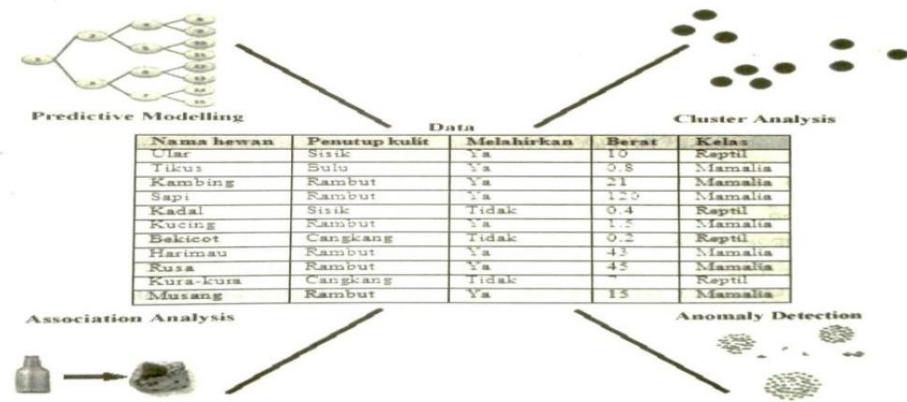


Gambar 2. 3. Posisi *data mining* pada penelitian

Bidang lain yang juga memengaruhi *data mining* adalah teknologi basis data, yang mendukung penyediaan penyimpanan yang efisien, pengindeksan, dan pemrosesan query (Prasetyo, 2012). Teknik komputasi paralel sering digunakan untuk memberikan kinerja yang tinggi untuk ukuran set data yang besar, sedangkan komputasi terdistribusi dapat digunakan untuk menangani masalah ketika data tidak dapat disimpan di satu tempat.

2.4. Pekerjaan Dalam *Data mining*

Pekerjaan yang berkaitan dengan *data mining* dapat dibagi menjadi empat kelompok, yaitu model prediksi (prediction modelling), analisis kelompok (cluster analysis), analisis asosiasi (association analysis), dan deteksi anomali (anomaly detection) (Prasetyo, 2012). Ilustrasi empat pekerjaan tersebut dapat dilihat pada Gambar 2. 4.

Gambar 2. 4. Pekerjaan *data mining*

2.4.1. model prediksi

Model prediksi berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Ada dua jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan untuk variabel target diskret, sedangkan regresi untuk variabel target kontinu.

Misalnya, pekerjaan untuk melakukan deteksi jenis penyakit pasien berdasarkan sejumlah nilai parameter penyakit yang diderita masuk dalam jenis klasifikasi karena di sini target yang diharapkan adalah diskret, hanya beberapa jenis kemungkinan nilai target yang didapatkan, tidak ada nilai deret waktu (time series) yang harus didapatkan untuk mendapat target nilai akhir. Sementara, pekerjaan prediksi jumlah penjualan yang didapatkan pada tiga bulan ke depan termasuk regresi karena untuk mendapatkan nilai penjualan bulan ketiga, nilai penjualan bulan kedua harus didapatkan dan untuk mendapatkan nilai penjualan bulan kedua, nilai penjualan bulan pertama harus didapatkan. Di sini ada nilai deret waktu yang harus dihitung untuk sampai pada target akhir yang diinginkan, ada nilai kontinu yang harus dihitung untuk mendapatkan nilai target akhir yang diinginkan.

2.4.2. analisis kelompok

Contoh pekerjaan yang berkaitan dengan analisis kelompok (cluster analysis) adalah bagaimana caranya mengetahui pola pembelian barang oleh para konsumen pada waktu-waktu tertentu. Dengan mengetahui pola kelompok pembelian tersebut, perusahaan/pengecer dapat menentukan jadwal promosi yang dapat diberikan sehingga omzet penjualan bisa ditingkatkan.

Analisis kelompok melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dalam kelompok yang berbeda jika keluar dari batas kesamaan dengan kelompok tersebut.

2.4.3. analisis asosiasi

Analisis asosiasi (association analysis) digunakan untuk menemukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Pola yang ditemukan biasanya merepresentasikan bentuk aturan implikasi atau subset fitur. Tujuannya adalah untuk menemukan pola yang menarik dengan cara yang efisien.

Penerapan yang paling dekat dengan kehidupan sehari-hari adalah analisis data keranjang belanja. Sebagai contoh, pembeli adalah ibu rumah tangga yang akan membeli barang kebutuhan rumah tangga di sebuah supermarket. Jika ibu tersebut membeli beras, sangat besar kemungkinannya bahwa itu juga akan membeli barang lain, misalnya minyak, telur, dan tidak mungkin atau jarang membeli barang lain seperti topi atau buku. Dengan mengetahui hubungan yang lebih kuat antara beras dengan telur daripada beras dengan topi, pengecer dapat menentukan barang-barang yang sebaiknya disediakan dalam jumlah yang cukup banyak.

2.4.4. deteksi anomali

Pekerjaan deteksi anomali (anomaly detection) berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain. Data-data yang karakteristiknya menyimpang (berbeda) dari data yang lain disebut outlier. Algoritma deteksi anomali yang baik harus mempunyai laju deteksi yang tinggi dan laju eror yang rendah. Deteksi anomali dapat diterapkan pada sistem jaringan untuk mengetahui pola data yang memasuki jaringan sehingga penyusupan bisa ditemukan jika pola kerja data yang datang berbeda. Perilaku kondisi cuaca yang mengalami anomali juga dapat dideteksi dengan algoritma ini.

2.5. *Dataset*

Set data (*data set*) dapat dipandang sebagai kumpulan objek data. Nama lain yang sering digunakan adalah record, point, vector, pattern, event, observation, case, atau bahkan data. Sementara objek data digambarkan dengan sejumlah atribut yang menangkap (*capture*) karakter dasar objek data (Prasetyo, 2012). Contohnya tinggi badan yang memberikan nilai kuantitatif tinggi badan seseorang, waktu yang menangkap saat sebuah peristiwa terjadi. Atribut terkadang juga disebut variabel, karakteristik, medan (*field*), fitur, atau dimensi.

Atribut adalah sifat atau properti atau karakteristik objek data yang nilainya bisa bermacam-macam dari satu objek ke objek yang lain, dari satu waktu ke waktu yang lain. Misalnya, warna kulit seseorang bisa berbeda dengan warna kulit orang lain, berat badan seseorang juga bisa berubah dari waktu ke waktu. Warna kulit bisa mempunyai nilai simbolik [hitam, putih, kuning langsat, cokelat, sawo matang], sedangkan berat badan bisa berupa nilai angka numerik, misalnya 35, 50, 70, 85, dan sebagainya.

Atribut yang menjadi elemen setiap data mempunyai jenis yang beragam. Berat badan, pada contoh sebelumnya, mempunyai nilai numerik sehingga dapat dibandingkan satu sama lain, sedangkan warna kulit tidak bisa

dibandingkan karena menggunakan nilai yang sifatnya kualitatif. Umumnya, tipe atribut ada dua, yaitu kategoris (kualitatif) dan numerik (kuantitatif).

Ada empat sifat penting yang dimiliki atribut secara umum, yaitu

1. distinctness, = dan #
2. order, <, <, >, dan >
3. addition, + dan -
4. multiplication, * dan /

Dari keempat sifat tersebut dapat diturunkan empat tipe atribut, yaitu nominal, ordinal, interval dan rasio. Tabel 2. 1 menginformasikan penjelasan ketiga tipe atribut dan kaitannya dengan sifat-sifat di atas.

Tabel 2. 1. Tipe atribut *dataset*

Tipe Atribut		Penjelasan	Contoh
Kategori (kualitatif)	Nominal	Nilai atribut bertipe nominal membedakan nilai berupa nama. Dengan nama inilah sebuah atribut membedakan dirinya pada data yang satu dengan yang lain ($=, \neq$).	Kode pos, nomor KTP, NRP, jenis kelamin.
	Ordinal	Nilai atribut bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi terurut ($<, \leq, \geq, >$).	Suhu{dingin, normal, panas}.
Numerik (kuantitatif)	Interval	Nilai atribut di mana perbedaan di antara dua nilai mempunyai makna yang berarti (+,-)	Tanggal, suhu (dalam celcius dan fahrenheit)
	Rasio	Nilai atribut di mana perbedaan di antara dua nilai dan rasio dua nilai mempunyai makna yang berarti (*,/).	Suhu (dalam kelvin), umur, panjang, tinggi.

2.6.Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012). Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu:

1. Pembangunan model sebagai prototipe untuk disimpan sebagai memori dan
2. Penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya.

Contoh aplikasi yang sering ditemui adalah pengklasifikasian jenis hewan, yang mempunyai sejumlah atribut. Dengan atribut tersebut, jika ada hewan baru, kelas hewannya bisa langsung diketahui. Contoh lain adalah bagaimana melakukan diagnosis penyakit kulit kanker melanoma, yaitu dengan melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebut untuk mengidentifikasi penyakit pasien baru sehingga diketahui apakah pasien tersebut menderita kanker atau tidak.

2.7.Naive Bayes Classifier

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naive Bayes, model yang digunakan adalah "model fitur independen" (Prasetyo, 2012).

Dalam Bayes (terutama Naive Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya, pada kasus

klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat, dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dipastikan juga menyusui. Di sini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan yang bertelur biasanya tidak menyusui. Dalam Bayes, hal tersebut tidak dipandang sehingga masing-masing fitur seolah tidak memiliki hubungan apa pun.

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut:

$$P(H|E) = \frac{P(E|H)xP(H)}{P(E)}$$

Keterangan:

1. $P(H|E)$ = probabilitas akhir bersyarat (Condition probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
2. $P(E|H)$ = Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H
3. $P(H)$ = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun
4. $P(E)$ = Probabilitas awal (priori) nukti E terjadi tanpa memandang hipotesis/bukti yang lain.

2.8. Customer churn

Pelanggan adalah aset yang paling penting dari semua jenis bisnis. Prospek usaha hanya mungkin dapat dilakukan dengan kehadiran pelanggan yang puas yang selalu setia dan membangun hubungan mereka dengan perusahaan. Untuk alasan ini, perusahaan harus merencanakan dan menerapkan strategi untuk menciptakan pelanggan, umumnya dikenal sebagai Customer Relationship Management (CRM). K. Tsiptsis dan A. Chorianopoulos mendefinisikan CRM sebagai strategi yang terkait dengan

mempertahankan, mengelola, dan meningkatkan hubungan pelanggan setia dan langgeng. Merujuk ke perspektif bisnis intelijen, proses manajemen churn dalam kerangka CRM terdiri dari dua bagian utama pemodelan analitis yang memprediksi bagi mereka yang cenderung churn atau tidak dan mendukung operator penyedia untuk membuat keputusan yang berharga dalam mempertahankan atau meningkatkan pelanggan baru. Oleh karena itu, artikel ini difokuskan pada pertimbangan dalam prediksi pelanggan churn(Zeniarja & Luthfiarta, 2015).

2.8.1. Churn Rate

Churn rate adalah persentase banyaknya konsumen yang berhenti berlangganan atau tak lagi membeli produk.

Menghitung churn rate sangat penting karena bisa menjadi tolok ukur apakah sebuah bisnis mampu mempertahankan pelanggan dengan baik. Terutama, untuk jenis bisnis yang menjual produk dengan sistem berlangganan.

Dengan model bisnis tersebut, pendapatan bisnis berasal dari setiap konsumen yang melakukan pembelian berulang. Jika konsumen berhenti membeli, di sanalah terjadi churn yang membahayakan bisnis.

Rumus menghitung churn rate

$$\text{Churn Rate} = \left(\frac{\text{Jumlah Pelanggan yang Hilang}}{\text{Jumlah Pelanggan di Awal Periode}} \right) \times 100$$

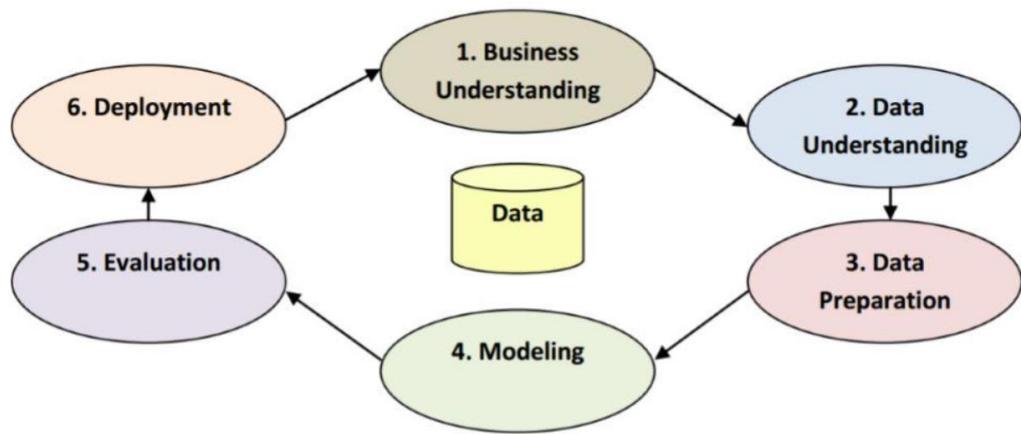
2.9.Prediksi Csutomer Churn

Prediksi churn pelanggan adalah bagian dari manajemen churn, yang memprediksi perilaku pelanggan dengan klasifikasi sebagai pelanggan setia dan mana yang cenderung untuk pindah ke kompetitor lain. “Pelanggan churn” berarti kehilangan klien. Ini memiliki arti yang sama seperti gesekan pelanggan, pembelotan pelanggan, dan perputaran pelanggan. Churn pelanggan juga didefinisikan oleh Hung et al. di mana layanan nirkabel

industri telekomunikasi yang umum digunakan dalam jangka gerakan pelanggan dari satu operator ke yang lain. Keakuratan prediksi ini mutlak diperlukan karena tingginya tingkat migrasi pelanggan untuk perusahaan pesaing. Manajemen churn merupakan tugas penting bagi perusahaan untuk mempertahankan pelanggan yang berharga. Riset pemasaran menunjukkan bahwa rata-rata nilai pelanggan yang churn atau pindah ke pesaing lain dari perusahaan operator seluler adalah sekitar 2,2% per bulan. Hung et al. menyebutkan bahwa ada sekitar 27% dari pelanggan hilang setiap tahun. Berdasarkan riset pasar, keadaan ini mendorong perusahaan untuk menyediakan biaya untuk dukungan penjualan, pemasaran, iklan, dan komisi untuk mendapatkan pelanggan layanan mobile dengan pelanggan baru adalah sekitar \$ 300 sampai \$ 600. Dengan demikian, biaya untuk mendapatkan pelanggan baru jauh lebih tinggi daripada mempertahankan yang baru dan karenanya, kemampuan untuk memprediksi churn pelanggan adalah suatu keharusan(Zeniarja & Luthfiarta, 2015).

2.10. CRISP *Data mining*

Metodologi CRISP-DM telah dimanfaatkan dalam dunia industri. Dunia industri yang beragam bidangnya memerlukan proses yang standard yang mampu mendukung penggunaan *data mining* untuk menyelesaikan masalah bisnis. Berdasarkan ‘*best practice*’(Jackins et al., 2021). Praktisi-praktisi dan peneliti *Data mining* mengusulkan beberapa proses seperti *workflow* atau pendekatan dengan tahapan-tahapan yang sederhana untuk memperbesar peluang keberhasilan dalam melaksanakan berbagai projek dalam *Data mining*. Usaha-usaha tersebut menghasilkan beberapa proses yang dijadikan sebagai standard, salah satu yang sudah terstandarisasi dan termasuk terpopuler yakni Metodologi *Cross-Industry Standard Process for Data mining* (CRISP-DM).



Gambar 2. 5. Tahapan pada CRISP-DM

Pada Gambar 2. 1 menggambarkan proses dalam metodologi CRISP-DM. Ada enam tahap berurutan yang dimulai dengan *Business Understanding* (Pemahaman Terhadap Bisnis), *Data Understanding* (Pemahaman Terhadap Data), *Data Preparation* (Persiapan Data), *Modeling* (Pemodelan), *Evaluation* (Evaluasi) dan *Deployment* (Penyebaran).

Dari setiap tahapan dalam CRISP DM dapat dijelaskan sebagai berikut:

1. *Bussines Understanding.* Situasi bisnis harus dinilai untuk mendapatkan gambaran umum tentang sumber daya yang tersedia dan yang diperlukan. Penentuan tujuan Penentuan tujuan *data mining* adalah salah satu aspek terpenting dalam fase ini. Pertama, jenis *data mining* harus dijelaskan (mis. klasifikasi) dan kriteria keberhasilan *data mining* (seperti presisi). Rencana proyek wajib harus dibuat.
2. *Data understanding.* Mengumpulkan data dari sumber data, mengeksplorasi dan mendeskripsikannya dan memeriksa kualitas data adalah tugas penting dalam fase ini. Untuk membuatnya lebih konkret, panduan pengguna menggambarkan tugas deskripsi data dengan menggunakan analisis statistik dan menentukan atribut dan kolasinya.

3. atribut dan kolasi mereka. Seleksi data harus dilakukan dengan menentukan kriteria inklusi dan eksklusi. Kualitas data yang buruk dapat ditangani dengan membersihkan data. Tergantung pada model yang digunakan (didefinisikan pada tahap pertama) atribut turunan harus dibangun. Untuk semua langkah ini, berbagai metode yang berbeda dapat dilakukan dan tergantung pada model.
4. Data Preparation. Fase pemodelan data terdiri dari pemilihan teknik pemodelan, membangun kasus uji dan model. Semua teknik *data mining* dapat digunakan. Secara umum, pilihannya tergantung pada masalah bisnis dan data. Yang lebih penting adalah, bagaimana menjelaskan pilihan tersebut. Untuk membangun model, parameter spesifik harus ditetapkan. Untuk menilai model, adalah tepat untuk mengevaluasi model terhadap kriteria evaluasi dan memilih yang terbaik.
5. Modeling. Fase pemodelan data terdiri dari pemilihan teknik pemodelan, membangun kasus uji dan model. Semua teknik *data mining* dapat digunakan. Secara umum, pilihannya tergantung pada masalah bisnis dan data. Yang lebih penting adalah, bagaimana menjelaskan pilihan tersebut. Untuk membangun model, parameter spesifik harus ditetapkan. Untuk menilai model, adalah tepat untuk mengevaluasi model terhadap kriteria evaluasi dan memilih yang terbaik.
6. Evaluation. Dalam fase evaluasi, hasilnya diperiksa terhadap tujuan bisnis yang telah ditetapkan. Oleh karena itu, hasilnya harus diinterpretasikan dan tindakan lebih lanjut harus didefinisikan. Hal lain adalah, bahwa proses harus ditinjau secara umum
7. Deployment. Fase penyebaran dijelaskan secara umum dalam panduan pengguna. Ini bisa berupa laporan akhir atau komponen perangkat lunak. Panduan pengguna panduan pengguna menjelaskan bahwa fase penyebaran terdiri dari perencanaan penyebaran, pemantauan dan pemeliharaan.

2.11. Confusion Matrix

Merupakan tabel yang menggambarkan performa dari sebuah model atau algoritma secara spesifik. Setiap baris dari matrix tersebut, merepresentasikan kelas aktual dari data, dan setiap kolom merepresentasikan kelas prediksi dari data (atau sebaliknya) (Saputro & Sari, 2020). Matrix tersebut dijelaskan pada Tabel.

Tabel 2. 2. *Confusion matrix*

	<i>Predicted Negative</i>	<i>Predicted Positive</i>
<i>Predicted Negative</i>	<i>True Negative (TN)</i>	<i>False Positive (FP)</i>
<i>Predicted Positive</i>	<i>False Negative (FN)</i>	<i>True Positive (TP)</i>

Keterangan:

1. True Positive = Berarti seberapa banyak data yang aktual kelasnya positif, dan model juga memprediksi positif.
2. True Negative = Berarti seberapa banyak data yang aktual kelasnya negatif, dan model memprediksi negatif.
3. False Positive = Berarti seberapa banyak data yang aktual kelasnya negatif, namun model memprediksi positif.
4. False Negative = Berarti seberapa banyak data yang aktual kelasnya positif, namun model memprediksi negatif

Melalui 4 data tersebut, dapat diperoleh data lain yang sangat berguna untuk mengukur perfoma sebuah model, diantaranya:

1. Accuracy = Total keseluruhan seberapa sering model benar mengklasifikasi. Formula accuracy dapat ditulis menggunakan persamaan:

$$\frac{TP + TN}{\text{Total}}$$

2. Precision = Ketika model memprediksi positif, seberapa sering prediksi itu benar. Formula precision dapat ditulis menggunakan persamaan:

$$\frac{TP}{FP + TP}$$

3. Recall (Sensitivity / True Positive Rate) = Ketika kelas aktualnya positif, seberapa sering model memprediksi positif. Formula recall dapat ditulis menggunakan persamaan:

$$\frac{TP}{FN + TP}$$

4. F1-Score = Merupakan rata-rata harmonik dari Precision dan Recall. Formula f1-score dapat ditulis menggunakan persamaan:

$$2 * \frac{precision * call}{precision + call}$$

2.12. EDA

Analisis Data Eksploratif atau biasa disebut dengan Exploratory Data Analysis merupakan sebuah cara (metode) penjelajahan lebih banyak data (eksplorasi) dengan teknik aritmatika dan visual grafis dalam meringkas data yang diamati (Kesuma & Iskandar, 2022). Metode ini digunakan untuk mencari tahu pattern pola sebaran, meringkas serta memvisualisasikan data dalam berbagai bentuk grafik, plot dan table dengan tujuan disajikan secara menyeluruh ringkasan statistik secara visual.

Dengan tersedianya data sedemikian besar, biasanya terdapat informasi yang terpendam dan dapat digali, teknik yang dapat digunakan bisa menggunakan statistic deksriptif biasa hingga teknik *Data mining* dan *Exploratory Data Analysis* (EDA).

2.13. SMOTE

Over-sampling SMOTE digunakan untuk meningkatkan jumlah *dataset* untuk mencapai *dataset* yang seimbang (Sutoyo & Fadlurrahman, 2020).

Metode *Synthetic Minority Over-sampling Technique* (SMOTE) merupakan metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan *dataset* dengan cara sampling ulang sampel kelas minoritas.

2.14. Tinjauan Pustaka

Tabel 2. 3. Tinjauan pustaka

No	Peneliti	Tahun	Judul	Ringkasan
1	Novendri, Risky Andreswari, Rachmadita Pratiwi, Oktariani Nurul	2021	Implementasi <i>Data mining</i> Untuk Memprediksi <i>Customer churn</i> Menggunakan Algoritma Naive Bayes Implementation of <i>Data mining</i> To Predict <i>Customer churn</i> s Using Naive Bayes Algorithm	Naive Bayes, Data customer chur PT. Telekomunikasi. hasil penelitian ini dapat disimpulkan bahwa implementasi <i>data mining</i> untuk memprediksi <i>customer churn</i> menggunakan algoritma naive bayes, dapat dilakukan dengan sangat baik. Dari hasil pengujian akurasi, hasil prediksi yang dihasilkan tidak terlalu berpengaruh terhadap perbedaan rasio data yang digunakan. Sehingga berapa pun rasio data yang dipakai, tidak akan berpengaruh terhadap model yang akan dilatih menggunakan algoritma naive bayes. Hal ini dikarenakan, algoritma naive bayes merupakan algoritma yang menggunakan prediksi dengan perhitungan probabilitas.
2	Wardani, Ni Wayan Arnidya, Diah Juniari	2022	Prediksi Pelanggan Loyal Menggunakan	Naive Bayes, transaksi penjualan, Berdasarkan dari hasil penelitian dapat ditarik beberapa

No	Peneliti	Tahun	Judul	Ringkasan
	Agus, I Nyoman Putra, Suarya Made, Ni Rosa, Mila		Metode Naïve Bayes Berdasarkan Segmentasi Pelanggan dengan Pemodelan RFM	kesimpulan, dengan implementasi <i>data mining</i> untuk memprediksi pelanggan loyal mempunyai frekuensi rata-rata. Sedangkan segmen dengan jumlah pelanggan paling sedikit adalah segmen At Risk dengan pelanggan yang karakteristiknya menghabiskan banyak uang dan sering membeli, tetapi sudah lama sekali.
3	Saputro, Irkham Widhi Sari, Bety Wulan	2020	Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa	Naive Bayes, data kelulusan mahasiswa, Penelitian terhadap 300 data alumni menggunakan Algoritma Naïve Bayes yang digunakan untuk klasifikasi waktu kelulusan mahasiswa menghasilkan model klasifikasi dengan rata-rata nilai akurasi, precision, recall, dan f1-score sebesar 68%, 61.3%, 65.3%, dan 61% yang dihitung menggunakan metode 10-Fold Cross Validaiton, dan Confusion Matrix.
4	Wardani, Ni Wayan Dantes, Gede Rasben Indrawan, Gede	2018	Prediksi <i>Customer churn</i> Dengan Algoritma Decision Tree C4 . 5	Decision Tree C4.5, data UD. Mawar Sari, Hasil kinerja Algoritma Decision Tree C4.5 pada kelas pelanggan Dormant yaitu recall 97.51%, precision 75.18%, dan accuracy 76.18%. Hasil kinerja pada kelas pelanggan everyday yaitu recall 100%, precision 99.04%, dan accuracy 99.04%. Hasil kinerja pada kelas pelanggan golden yaitu

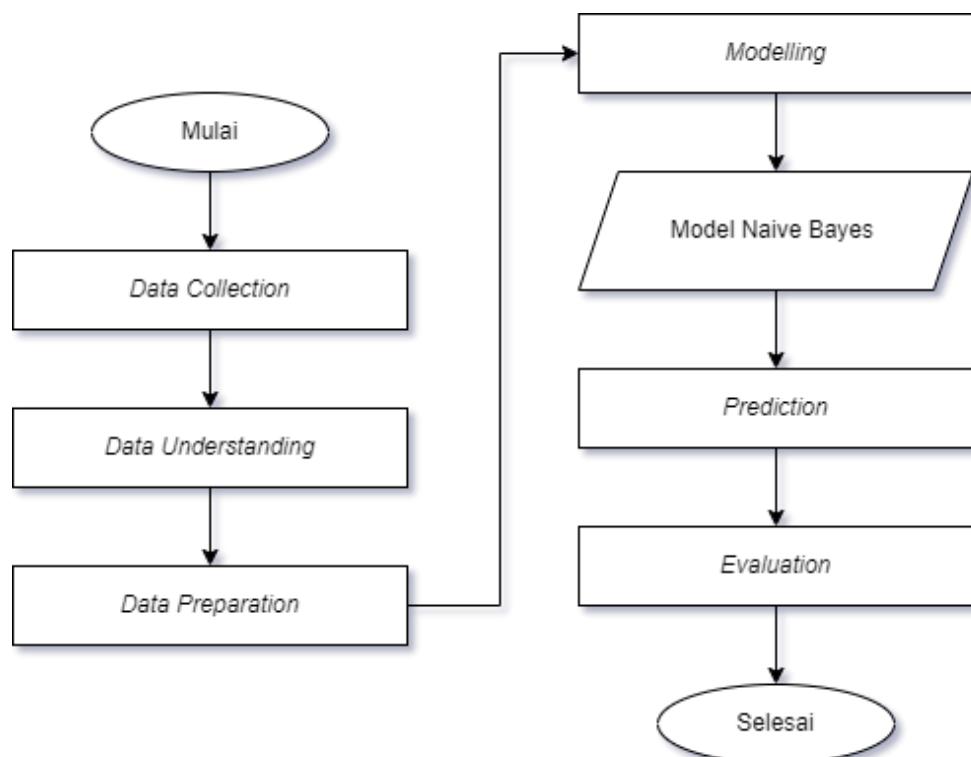
No	Peneliti	Tahun	Judul	Ringkasan
				recall 100%, precision 98.84%, dan accuracy 98.84%. Hasil kinerja pada kelas pelanggan superstar yaitu recall 96.15%, precision 99.43%, dan accuracy 95.63%.
5	Utami, yohana Tri Shofiana, Dewi Asiah Heningtyas, Yunda	2020	Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi	Decision Tree C4.5, Telco, Berdasarkan penelitian yang telah dilakukan, diperoleh beberapa kesimpulan bahwa algoritma decision tree C4.5 untuk melakukan klasifikasi pada set data pengguna jasa telekomunikasi menghasilkan nilai akurasi sebesar 87%, precision sebesar 87,5% dan recall sebesar 96% yang dianggap memperoleh hasil yang cukup baik. Atribut yang dianggap paling menarik dari hasil pengujian ini adalah atribut harga yang selanjutnya dapat diidentifikasi sebagai prediksi pola perpindahan pelanggan (<i>customer churn</i>).

BAB III

METODOLOGI PENELITIAN

3.1. Alur Penelitian

Bagian ini menjelaskan mengenai detail tahapan penelitian yang dilakukan dalam penelitian yang dapat dilihat pada Gambar 3. 1.



Gambar 3. 1. Alur penelitian

3.2. Data Collection

Data collection adalah proses pengumpulan, pengukuran, dan analisis berbagai tipe informasi menggunakan teknik berstandar. Tujuan utama *data collection* adalah untuk mengumpulkan informasi dan data terpercaya sebanyak-banyaknya, yang kemudian dianalisis untuk membuat sebuah keputusan bisnis yang krusial. Ketika sudah berhasil dikumpulkan, data ini kemudian melalui sejumlah proses meliputi pembersihan dan pemrosesan data agar dapat digunakan oleh perusahaan.

Pada tahap ini metode yang digunakan untuk pengumpulan data adalah dengan menggunakan metode sekunder yang berarti data yang didapat dan digunakan dalam penelitian customer churn ini berasal dari data yang sudah ada sebelumnya yang berada pada situs internet yaitu Kaggle.

3.3. Data Understanding

Proses *data understanding* yang digunakan adalah *Exploratory Data Analysis* (EDA). *Exploratory Data Analysis* (EDA) adalah bagian dari proses *data science*. EDA menjadi sangat penting sebelum melakukan *feature engineering* dan modeling karena dalam tahap ini kita harus memahami datanya terlebih dahulu.

Pada tahap ini dilakukan pendefinisian setiap atribut pada data yang digunakan dalam penelitian yang bertujuan untuk memahami data yang akan digunakan. Berikut adalah definisi dari setiap atribut dari data *customer churn* perusahaan telekomunikasi.

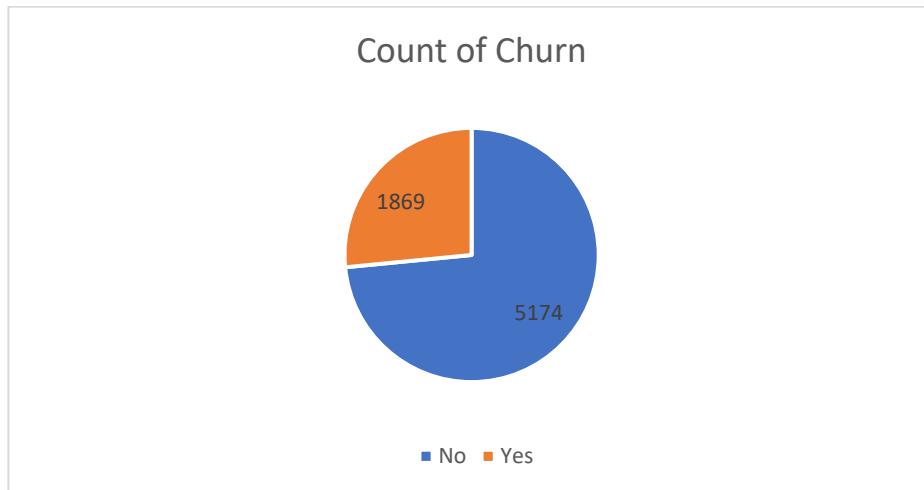
Tabel 3. 1. Definisi atribut *dataset customer churn*

No	Atribut	Deskripsi	Kriteria Jawaban
1	Customer ID	berisi data customer id dari pelanggan yang berlangganan	Kode pelanggan
2	Gender	Berisi data jenis kelamin dari pelanggan yang berlangganan	1= Male 0= Female
3	Senior Citizen	berisi data apakah pelanggan adalah orang lanjut usia atau tidak	1= Yes 0= No
4	Partner	Berisi data apakah pelanggan memiliki pasangan atau tidak	1= Yes 0= No
5	Dependents	berisi data apakah pelanggan memiliki tanggungan atau tidak	1= Yes 0= No
6	Tenure	berisi data berapa lama custumer berlangganan layanan dalam bulan	0 – 72
7	Phone Service	berisi data apakah pelanggan memiliki layanan telepon atau tidak	1= yes 0= no
8	Multiple Lines	berisi data apakah pelanggan memiliki layanan telpon lain atau tidak.	1= yes 0= no 2= Other

No	Atribut	Deskripsi	Kriteria Jawaban
9	Internet Service	berisi data jenis layanan internet yang digunakan	1= yes 0= no 2= Other
10	Online Security	berisi data apakah pelanggan memiliki keamanan online atau tidak	1= yes 0= no 2= Other
11	Online Backup	berisi data apakah pelanggan memiliki pencadangan online atau tidak	1= yes 0= no 2= Other
12	Device Protection	berisi data apakah pelanggan memiliki perangkat perlindungan atau tidak.	1= yes 0= no 2= Other
13	Tech Support	berisi data apakah pelanggan memiliki dukungan teknis atau tidak	1= yes 0= no 2= Other
14	Streaming TV	berisi data apakah pelanggan memiliki TV streaming atau tidak	1= yes 0= no 2= Other
15	Streaming Movies	berisi data apakah pelanggan memiliki movie streaming atau tidak	1= yes 0= no 2= Other
16	Contract	berisi data jangka waktu kontrak pelanggan	1= Month to Month 0= Two year 2= Other
17	Paperless Billing	berisi data apakah pelanggan memiliki paperless billing atau tidak	1= yes 0= no
18	Payment Method	berisi data jenis pembayaran yang dilakukan pelanggan.	1= Electronic check 0= Mailed check 2= Other
19	Montly Charges	berisi data besaran biaya yang dikeluarkan pelanggan per bulan	18- 119
20	Total Charge	berisi data besaran biaya yang dikeluarkan pelanggan selama berlangganan.	18- 8.68k
21	Churn	berisi data apakah pelanggan berhenti berlangganan atau tidak	1= yes 0= no

Kemudian pada tahap ini juga dilakukan penentuan karakteristik data. Cara mementukan karakteristik data yaitu dengan melihat seimbang atau tidaknya kelas yang ada pada dataset.

Dari data *customer churn* yang didapat dihasilkan presentase sebagai berikut:



Gambar 3. 2. Diagram pie atribut *churn*

Dari diagram yang tampilakan pada Gambar 3. 2, dapat diketahui bahwa kelas yang ada pada *dataset* yang dimiliki tidak seimbang.

3.4. Data Preparation

Pada tahap ketiga dilakukan penyiapan data awal yang akan digunakan pada fase berikutnya. Pilih kasus dan variable yang ingin dan yang sesuai untuk dianalisis. Kemudian dilakukan pembersihan, reduksi, transformasi, integrasi data, balancing data sehingga siap digunakan pada tahap pemodelan.

1. Tahap pembersihan - *Dataset* yang digunakan pada penelitian ini dilakukan pembersihan yaitu melakukan pengisian data yang hilang, menghaluskan noisy, mengenali atau mengilangkan outlier, dan memecahkan ketidak konsistenan pada data.
2. Tahap reduksi – *Dataset* yang digunakan dilakukan pengurangan volume yang bertujuan pemrosesan data berlangsung lebih cepat dengan hasil analitikal yang sama atau mirip.
3. Tahap transformasi – Nilai seluruh rangkaian atribut pada *dataset* yang digunakan kemudian dipetakan lalu diganti ke nilai pengganti yang

baru contoh seperti pada atribut gender data laki-laki diubah menjadi ‘1’ dan perempuan menjadi ‘0’.

4. Tahap integrasi - Pada *dataset* yang digunakan dilakukan penggabungan data yang bertujuan untuk mengurangi redundansi data.
5. Balancing data – Proses ini merupakan pengecekan terhadap data yang akan digunakan, apakah data tersebut sudah balance atau belum, karena berdasarkan penelitian sebelumnya hal tersebut mempengaruhi terhadap proses *data mining* yang akan dilakukan. Jika *dataset* yang digunakan memiliki kelas yang tidak seimbang maka dilakukanlah proses balancing data ini. Ada dua jenis metode yang dilakukan untuk menyelesaikan permasalahan *imbalancing* data yaitu *sampling* dan *oversampling*. Metode *sampling* atau yang lebih dikenal dengan *resample* adalah metode umum yang digunakan dalam menyelesaikan permasalahan *imbalance* data. Dengan adanya penerapan *sampling* pada data yang *imbalance*, tingkat *imbalance* semakin kecil dan klasifikasi dapat dilakukan dengan tepat. Sedangkan metode *Oversampling* dilakukan dengan menyeimbangkan jumlah distribusi data dengan meningkatkan jumlah data kelas minor. Metode *oversampling* yang biasa digunakan dalam menyelesaikan masalah *imbalance* data adalah *random oversampling*, SMOTE, dan *borderline* SMOTE

3.5. Modeling

Pada tahap modeling ini dilakukan menentukan dan menerapkan teknik pemodelan yang sesuai, mengkalibrasi aturan model untuk mengoptimalkan hasil, dan dapat kembali ke fase pengolahan data ke dalam bentuk kebutuhan tertentu. Modeling data yang digunakan pada penelitian ini adalah Naive Bayes.

Data training nantinya akan digunakan untuk melatih algoritma dalam mencari model yang sesuai, sedangkan data testing akan dipakai untuk

menguji dan mengetahui performa model yang didapatkan pada tahapan testing.

3.6. *Prediction*

Hasil dari metode modelling menggunakan Niave Bayes ini adalah berupa prediksi *customer churn* yang nantinya akan digunakan dalam proses pengetesan.

3.7. *Evaluation*

Setelah hasil pengukuran diketahui, maka langkah selanjutnya adalah evaluasi menggunakan *Confusion Matrix*. *Confusion Matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*.

DAFTAR PUSTAKA

- Arina, F., & Ulfah, M. (2022). Analisa survival untuk mengurangi *customer churn* pada perusahaan telekomunikasi. *Journal Industrial Servicess*, 8(1), 59. <https://doi.org/10.36055/jiss.v8i1.14313>
- Batubara, D. N., & Windarto, A. P. (2019). Analisa Klasifikasi *Data mining* Pada Tingkat Kepuasan Pengunjung Taman Hewan Pematang Siantar Dengan Algoritma. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 3(1), 588–592. <https://doi.org/10.30865/komik.v3i1.1664>
- Herawati, M., Wibowo, I. L., & Mukhlash, I. (2016). Prediksi *Customer churn* Menggunakan Algoritma Fuzzy Iterative Dichotomiser 3. *Limits: Journal of Mathematics and Its Applications*, 13(1), 23. <https://doi.org/10.12962/j1829605x.v13i1.1913>
- Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *Journal of Supercomputing*, 77(5), 5198–5219. <https://doi.org/10.1007/s11227-020-03481-x>
- Kesuma, M. E.-K., & Iskandar, R. (2022). Analisis Toko dan Asal Toko Fashion Pria di Shopee Menggunakan Data Scrapping dan Exploratory Data Analysis. *Majalah Ilmiah Teknologi Elektro*, 21(1), 127. <https://doi.org/10.24843/mite.2022.v21i01.p17>
- Novendri, R., Andreswari, R., & Pratiwi, O. N. (2021). *Implementasi Data mining Untuk Memprediksi Customer churn Menggunakan Algoritma Naive Bayes Implementation of Data mining To Predict Customer churn s Using Naive Bayes Algorithm*. 8(2), 2762–2773.
- Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, 6(1), 1.

<https://doi.org/10.24076/citec.2019v6i1.178>

Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, 6(3), 379. <https://doi.org/10.26418/jp.v6i3.42896>

Wardani, N. W., Arnidya, D. J., Agus, I. N., Putra, S., Made, N., & Rosa, M. (2022). *Prediksi Pelanggan Loyal Menggunakan Metode Naïve Bayes Berdasarkan Segmentasi Pelanggan dengan Pemodelan RFM*. 12(2), 113–124.

Wardani, N. W., Dantes, G. R., & Indrawan, G. (2018). Prediksi Customer churn Dengan Algoritma Decision Tree C4 . 5. *Jurnal Resistor*, 1(1), 16–24.

Yulianti. (2018). Metode *Data mining* untuk Prediksi Churn Pelanggan. *Jurnal ICT Akademi Telkom Jakarta*, 17(May 2018), 46–52.

Zeniarja, J., & Luthfiarta, A. (2015). *Prediksi Churn Dan Segmentasi Pelanggan Menggunakan Backpropagation Neural Network Berbasis Evolution Strategies*. 14(1), 49–54.

Prasetyo, E. (2012). *Data mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI