BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini dilakukan untuk meneliti objek yang telah ditentukan, yaitu intensitas penagihan nasabah dan preferensi atau kecenderungan perilaku nasabah perusahaan *P2P Lending* dalam berinteraksi dengan *customer service P2P Lending* terhadap berbagai *communication channel*. Perusahaan *P2P Lending* menyediakan produk atau layanan yang mendukung kelancaran proses *collection* pada perusahaan *P2P Lending* berupa layanan komunikasi yang dilakukan oleh *customer service* untuk melakukan penagihan terkait pinjaman nasabah melalui berbagai *communication channel*, yakni *email*, *robocall*, *SMS*, dan *telephony*. Intensitas penagihan dapat ditentukan dari keberhasilan dari komunikasi antara 2 pihak yang dilakukan melalui *communication channel*, yaitu pihak *customer service* dengan nasabah, serta data pinjaman nasabah. Oleh karena itu, perusahaan *P2P Lending* memerlukan model *machine learning* dibutuhkan dalam mengklasifikasi intensitas penagihan berdasarkan penilaian dari data pinjaman dan interaksi nasabah. Selain itu, model prediksi dibutuhkan untuk mendukung perusahaan dalam memperoleh perkiraan jumlah interaksi dari nasabah pada *communication channel* yang direkomendasikan untuk memberikan pendekatan yang tepat terhadap setiap nasabah. Penelitian ini menggunakan data yang memuat data nasabah perusahaan *P2P Lending*, seperti *customer data*, *loan data*, dan *interaction data*.

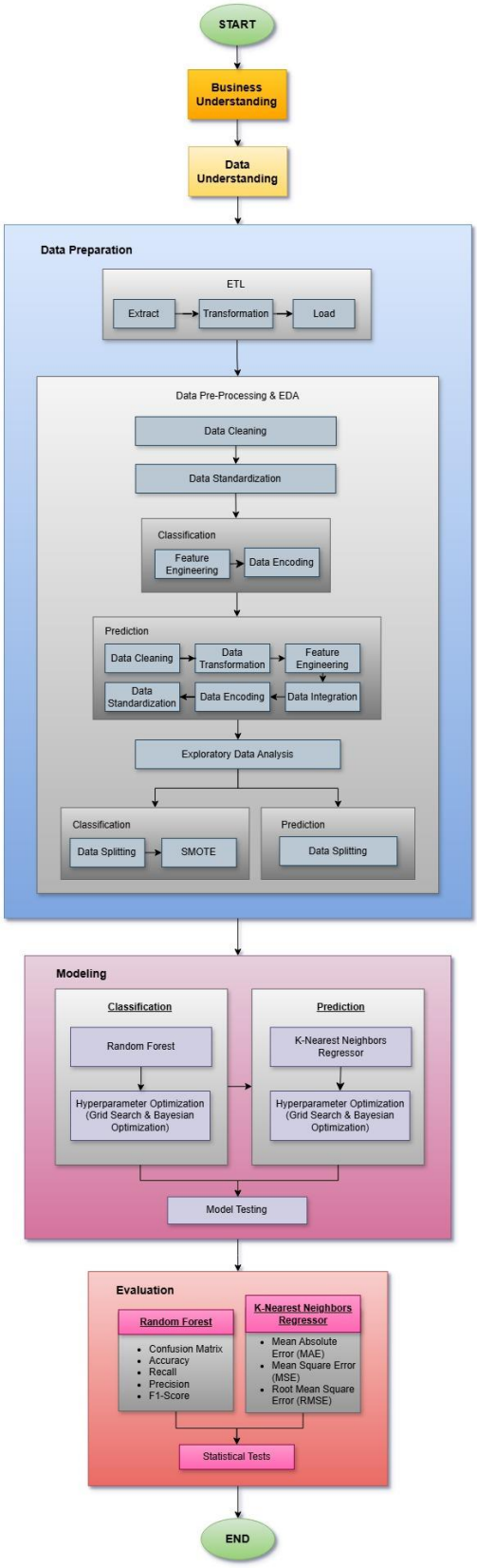
3.2 Metode Penelitian

Pendekatan yang digunakan dalam penelitian dibagi menjadi dua jenis, yaitu pendekatan kuantitatif dan pendekatan kualitatif. Pendekatan kuantitatif merujuk pada suatu metode penelitian yang melibatkan proses penyelidikan, pembuatan hipotesis atau prediksi hasil, pengumpulan data empiris, serta penarikan kesimpulan berdasarkan analisis data dengan pengukuran, penghitungan, rumus, dan data numerik, termasuk penggunaan statistik. Pendekatan kualitatif merujuk pada suatu metode penelitian yang melibatkan proses pengumpulan data empiris, analisis data, dan penarikan kesimpulan tanpa penghitungan numerik, bersifat deskriptif dan

melibatkan beberapa teknik, seperti pengamatan, wawancara secara mendalam, analisis isi cerita, serta jurnal dan kuesioner terbuka [90]. Penelitian ini menggunakan pendekatan kuantitatif karena melibatkan penggunaan data numerik untuk memperoleh klasifikasi intensitas penagihan dalam *Collection Intensity Scoring* (*CIS*) beserta prediksi jumlah interaksi nasabah dalam *channel recommendation*. Pada penelitian ini, model akan dibuat dengan algoritma *Random Forest* dan *K-Nearest Neighbors Regressor* menggunakan *dataset* yang telah disediakan.

3.2.1 Alur Penelitian

Penelitian ini memiliki tahapan yang telah digambarkan sesuai dengan *CRISP-DM Framework*. Namun, terdapat tahapan yang tidak digunakan berupa tahapan *deployment* karena penelitian ini hanya melakukan perbandingan metode optimasi pada model yang dibuat. Selain itu, penelitian ini menggunakan proses *ETL* (*Extract*, *Transformation*, *Load*) dari [42] untuk tahap *Data Preparation* dalam *CRISP-DM Framework*. Proses *ETL* ini digunakan untuk *data integration*, yaitu menggabungkan data dari berbagai sumber menjadi satu tampilan, terutama untuk kebutuhan *reporting*, analisis, dan *Business Intelligence* [91]. Berikut terdapat alur penelitian *Collection Intensity Scoring* (*CIS*) dan *channel recommendation* yang ditunjukkan pada Gambar 3.1.



Gambar 3.1 Alur Penelitian

3.2.1.1 *Business Understanding*

Business understanding merupakan tahap awal dengan peranan penting dalam *CRISP-DM*. Pada tahap *business understanding*, suatu organisasi atau perusahaan harus mengetahui tujuan dari bisnis yang didirikan sehingga mengerti akan kebutuhan dan proses yang harus diselenggarakan. Hal ini bertujuan agar proses data mining dapat dilakukan secara efektif dalam menyelesaikan permasalahan yang ada.

Dalam tahap *business understanding* ini, adapun tujuan yang ingin dicapai oleh perusahaan *P2P Lending*, yaitu mengklasifikasi intensitas penagihan nasabah yang diperoleh *Collection Intensity Scoring* (*CIS*) dan memberikan jenis *treatment* yang tepat berdasarkan *communication channel* yang direkomendasikan, serta mendapatkan kembali pembayaran pinjaman yang belum lunas dari nasabah. Perusahaan *P2P Lending* menggunakan sistem *Collection Intensity Scoring* guna memberikan penilaian intensitas penagihan nasabah. Akan tetapi, terdapat permasalahan berupa proses klasifikasi intensitas penagihan yang dilakukan secara manual tanpa mengetahui informasi terkait *communication channel* yang digunakan oleh nasabah. Selain itu, penyelenggaraan komunikasi pada *communication channel* yang jarang digunakan oleh nasabah dapat mengakibatkan peningkatan biaya komunikasi bisnis. Oleh sebab itu, klasifikasi intensitas penagihan nasabah dan prediksi jumlah interaksi nasabah dibutuhkan untuk menghasilkan *insight* guna penerapan dan evaluasi terkait jenis *treatment* penagihan terhadap nasabah melalui *communication channel* yang tepat.

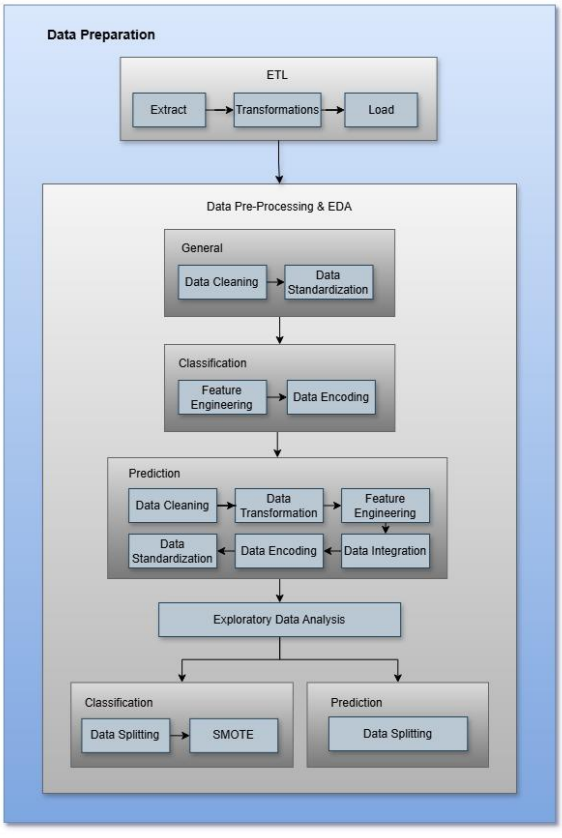
3.2.1.2 *Data Understanding*

Data understanding merupakan suatu tahap dimana organisasi atau perusahaan menyelidiki data yang akan dipakai untuk tahap atau penelitian selanjutnya. Pada tahap ini, seorang ahli *data mining* harus mengetahui dan memahami apakah data tersebut cocok untuk digunakan untuk penyelesaian suatu permasalahan bisnisnya. Tahap *data understanding* dalam penelitian ini menggunakan data sekunder dari perusahaan *P2P Lending*. Data tersebut

sesuai untuk digunakan dalam penelitian ini oleh karena adanya informasi terkait nasabah (*customer*), pinjaman (*loan*), interaksi (*interaction*), dan variabel lainnya sebagai variabel independen. Data yang diperoleh memiliki periode dari tanggal 01 Juli 2023 hingga 31 Desember 2023. Jumlah data yang diperoleh sebesar 23.880 data yang disimpan dalam *database* PostgreSQL di DBeaver.

3.2.1.3 *Data Preparation*

Data Preparation merupakan tahap dimana data akan disiapkan terlebih dahulu sebelum memasuki tahap pembuatan model. *Data preparation* pada penelitian ini melibatkan proses *ETL* (*Extract*, *Transformation*, *Load*) untuk kebutuhan integrasi data dengan bahasa PostgreSQL (PL/pgSQL), serta *Data Pre-processing and Exploratory Data Analysis* (*EDA*) yang dilakukan menggunakan Python sebagaimana ditampilkan pada Gambar 3.2.



Gambar 3.2 *Data Preparation*

Pada awalnya, proses *ETL* dimulai dengan *extract* data dengan mengimpor data ke dalam *database* PostgreSQL di DBeaver dan diikuti oleh *transformation*. *Transformation* mencakup data *combination* dari berbagai sumber untuk menghasilkan beberapa tabel sesuai dengan jenisnya. Adapun *data combination* dalam bentuk penggabungan *ID unique* berupa *user_id* dengan data nasabah, dan penggabungan *user_id* ke dalam tabel *loan* dan *interaction* yang seiringan dengan *data de-duplication*. Proses selanjutnya adalah *data filtering* berdasarkan kondisi yang ditentukan, diikuti dengan *data aggregation* untuk menghasilkan tabel yang bersifat *final*. Kemudian, tabel *final* tersebut akan di*export* dalam bentuk CSV (*Comma Separated Values*) dan dilakukan *load* untuk kebutuhan *data pre-processing* dan *EDA* (*Exploratory Data Analysis*) menggunakan bahasa pemrograman Python dalam *IDE* Jupyter *Notebook*.

Data pre-processing dan *EDA* mencakup *data cleaning* pada data hasil *ETL*, seperti pengecekan *missing* data, perubahan tipe data, serta *data standardization*. Terdapat *data pre-processing* untuk pendekatan klasifikasi yang terdiri atas *feature engineering* dan *data encoding*. Kemudian, dilakukan *data pre-processing* untuk pendekatan prediksi yang mencakup *data cleaning*, *data transformation*, *feature engineering*, *data integration*, *data encoding*, dan *data standardization*. Selanjutnya, *Exploratory Data Analysis* (*EDA*) dilakukan untuk memperoleh *data analysis* dalam bentuk *descriptive statistics* dan visualisasi data guna memperoleh pola maupun tren dari *dataset* yang digunakan. Penelitian ini melibatkan penerapan *Synthetic Minority Oversampling Technique* (*SMOTE*) dari [17] setelah *data splitting*. Teknik *SMOTE oversampling* digunakan untuk menggeneralisasi distribusi data [92]. Hal ini dilakukan untuk memperoleh *dataset* dengan jumlah data yang seimbang untuk kebutuhan klasifikasi. *Data splitting* pada penelitian ini menggunakan 80% untuk data *training* dan 20% untuk data *testing*. Rasio pada *data splitting* tersebut dipilih berdasarkan penggunaan rasio pada [18], [20] dimana rasio 80% *training data* dan 20% *testing data* dapat memberikan nilai *accuracy* dan *MAE*

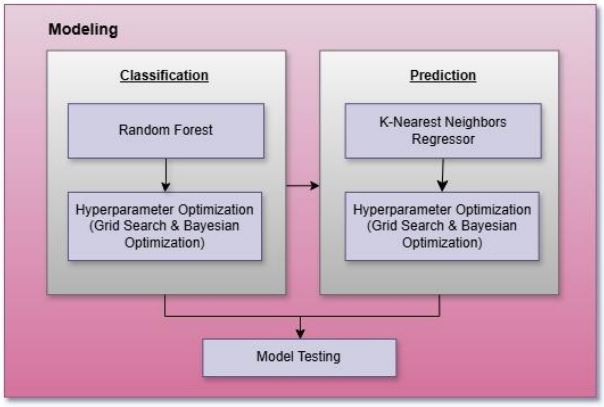
sejumlah 81% dan 0.0142 secara berturut-turut. Berikut terdapat tabel rincian untuk *data splitting* untuk penelitian ini yang terdapat pada Tabel 3.1.

Tabel 3.1 Rincian *Data Splitting*

No.	Persentase Pembagian Data	Kegunaan Data
1.	80%	*Training Data*
2.	20%	*Testing Data*

3.2.1.4 *Modeling*

Modeling merupakan tahap dimana dibutuhkan pembuatan model menggunakan algoritma yang sesuai untuk kebutuhan klasifikasi dan prediksi. Tahap *modeling* dalam penelitian ini akan melibatkan bahasa pemrograman Python dengan *Integrated Development Environment* (*IDE*). Berikut terdapat tahap *modeling* dalam penelitian ini yang ditampilkan pada Gambar 3.3.



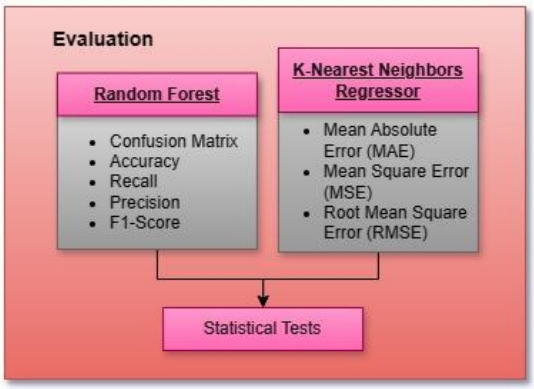
Gambar 3.3 *Modeling*

Tahap *modeling* pada penelitian ini dilakukan dengan dua jenis pendekatan, dimulai dari pendekatan klasifikasi dengan pembuatan model dengan algoritma *Random Forest* untuk klasifikasi intensitas penagihan nasabah dalam *Collection Intensity Scoring* (*CIS*). Selain itu, terdapat pendekatan prediksi yang melibatkan pembuatan model dengan algoritma *K-Nearest Neighbors Regressor* untuk memprediksi jumlah interaksi nasabah dalam *channel recommendation*. Penelitian ini menerapkan

hyperparameter optimization berupa *Grid Search* dan *Bayesian Optimization* pada kedua pendekatan. Dalam *hyperparameter optimization*, terdapat penggunaan *cross-validation* guna menentukan *hyperparameter training* terbaik. Metode yang dikenal sebagai *k-fold cross validation* dilakukan untuk memecah set pelatihan menjadi *k* set lebih kecil [93]. Model konvensional dan model yang dioptimasi pada tahapan ini akan digunakan untuk kebutuhan evaluasi performa model. Selain itu, dilakukan pengujian pada model atau *model testing* menggunakan data *dummy* guna memperoleh hasil pemodelan dalam pendekatan klasifikasi *CIS* dan prediksi *channel recommendation*.

3.2.1.5 *Evaluation*

Evaluation merupakan tahap yang melibatkan interpretasi dan melakukan pengujian model yang telah dibuat dengan data *testing*. Penelitian ini mencakup evaluasi pada model *Random Forest*, *Random Forest* x *Grid Search* dan *Random Forest* x *Bayesian Optimization* berdasarkan nilai *accuracy* untuk pendekatan klasifikasi. Selain itu, terdapat evaluasi pada model *K-Nearest Neighbors Regressor*, *K-Nearest Neighbors Regressor* x *Grid Search* dan *K-Nearest Neighbors Regressor* x *Bayesian Optimization* berdasarkan nilai *Mean Absolute Error* (*MAE*) untuk pendekatan prediksi. Berikut terdapat tahap *evaluation* beserta *evaluation metrics* yang digunakan dalam penelitian ini sebagaimana ditunjukkan pada Gambar 3.4.



Gambar 3.4 *Evaluation*

Pada tahap evaluasi ini, terdapat penggunaan *confusion matrix* dan nilai *accuracy* yang menunjukkan nilai klasifikasi yang benar oleh model *Collection Intensity Scoring* (*CIS*) terhadap kategori intensitas penagihan. *Accuracy* memegang peranan penting dalam kegiatan *scoring* untuk memperoleh kebenaran dan ketepatan dalam klasifikasi intensitas penagihan nasabah. Selain itu, nilai *precision*, *recall*, dan *f*1-*score* digunakan untuk evaluasi lebih lanjut pada model klasifikasi. Pada model *channel recommendation*, terdapat nilai *Mean Absolute Error* (*MAE*) mengukur keakuratan prediksi model terhadap jumlah interaksi pada *communication channel* yang direkomendasikan. Semakin rendah nilai *MAE* yang diperoleh, maka semakin baik model dalam memprediksi jumlah interaksi. Selain itu, nilai *Mean Square Error* (*MSE*) dan *Root Square Mean Error* (*RMSE*) digunakan untuk evaluasi lebih lanjut pada model prediksi. Seluruh model *Random Forest* dan *K-Nearest Neighbors Regressor* akan dibandingkan untuk menentukan model terbaik berdasarkan metrik *accuracy* dan *MAE*.

Evaluasi performa pada model-model yang dihasilkan dilakukan melalui *cross-validation* dengan jenis *k-fold* pada metrik *accuracy* dan *Mean Absolute Error* (*MAE*) dari *training data* dan *testing*. Pada *k-fold cross validation*, dilakukan *cross validation* sebanyak *k* kali, dengan setiap grup *k* memiliki kesempatan berperan sebagai *testing data*, dimana *k* kelompok tersisa digunakan sebagai *training data*. Teknik ini menghasilkan *k* estimasi kesalahan prediksi yang berbeda. Kinerja prediksi kemudian diukur dengan mengambil rata-rata dari estimasi kesalahan prediksi. *Cross-validation* mengevaluasi seberapa baik model statistik *machine learning* dalam menggeneralisasi data baru yang tidak digunakan dalam pelatihan model [94]. Hasil dari *cross-validation* digunakan untuk menentukan apakah model yang dihasilkan mengalami *overfitting*, *underfitting*, atau memiliki *good fit*. Kemudian, dilakukan penerapan *statistical test* seperti Shapiro-Wilk *test* untuk menguji normalitas distribusi data, sedangkan *t-test*

untuk membandingkan kinerja antara model terbaik dan model-model lainnya berdasarkan metrik *accuracy* dan *MAE*.

3.2.2 Metode *Data Mining*

Penelitian ini menggunakan *framework Cross Industry Standard Process – Data Mining* (*CRISP-DM*) guna membuat model klasifikasi terhadap data yang telah dikumpulkan. *Cross Industry Standard Process – Data Mining* (*CRISP-DM)* merupakan suatu model yang menjelaskan berbagai proses independen untuk tujuan *data mining* [62]. *Framework* ini merupakan salah satu *framework* yang terkenal dan secara umum digunakan dalam perusahaan. Adapun perbandingan antara *CRISP-DM framework* dengan *framework* lainnya, seperti *KDD* (*Knowledge Discovery in Databases*) dan *SEMMA* (*Sample, Explore, Modify, Model and Assess*) yang disajikan dalam Tabel 3.2.

Tabel 3.2 Perbandingan *Framework KDD, SEMMA,* dan *CRISP-DM*

Aspek Pembanding	***KDD***	***SEMMA***	***CRISP-DM***
Jumlah Tahapan	5	5	6
Tujuan	Mengekstrak pengetahuan tersembunyi dari data [95]	Memberikan panduan dalam penerapan aplikasi *Data Mining* [95]	Menciptakan proses yang dapat diandalkan dan bertahap untuk menghasilkan suatu nilai [95]
Tahapan	• *Selection* • *Pre-Processing* • *Data Transformation* • *Data Mining* • *Evaluation* [96]	• *Sample* • *Explore* • *Modify* • *Model* • *Assess* [97]	• *Business Understanding* • *Data Understanding* • *Data Preparation* • *Modeling* • *Evaluation* • *Deployment* [96]
Kelebihan	Bersifat *data-centric* yang menekankan sifat iteratif dan interaktif dari tugas analisis data [97]	• Menerapkan proses iteratif [97] • Memiliki instruksi dalam penggunaan aplikasi *DM* [98]	• Bersifat linier dan berurutan, memperhatikan proses iteratif dan *loop* umpan balik • Memiliki setiap fase atas tugas umum dengan *input* dan *output* yang diidentifikasi dengan jelas yang mencakup seluruh proses dan aplikasi *DM*, serta tetap valid mengikuti perkembangan teknologi yang tidak terduga [97]

Aspek Pembanding	*KDD*	*SEMMA*	*CRISP-DM*
Kelebihan			• Memiliki pendekatan yang sistematis dan terdefinisi dengan baik, serta tidak bergantung pada alat *data mining*. • Memiliki dokumentasi dengan studi kasus yang tercatat dengan baik [96]
Kekurangan	Kurangnya fokus pada aspek bisnis [97]	Proses model yang bersifat *vendor-specific* dari *SAS Enterprise Miner* membatasi penerapan *data mining* pada lingkungan yang berbeda [99]	Sifat siklus hidup yang sebagian besar berurutan yang menyangkut kurangnya iterasi dan interaksi yang jelas antar tugas [97]

Berdasarkan perbandingan yang telah dirinci pada Tabel 3.2, maka penelitian ini memilih *CRISP-DM framework* oleh karena mencakup tahapan sistematis dengan dokumentasi yang jelas dalam melakukan analisis data. Selain itu, penelitian ini memungkinkan untuk melakukan perubahan berupa pengulangan pada proses analisis datanya, serta tidak bergantung pada alat *data mining* tertentu, sehingga diperlukan *CRISP-DM framework* yang fleksibel terhadap tahapan iteratif. Oleh sebab itu, kelebihan tersebut dapat berujung pada kesesuaian *framework* untuk digunakan dalam penelitian ini.

3.3 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini berupa data sekunder yang diperoleh secara langsung dari PT. XYZ. Data tersebut menjelaskan sebagian besar mengenai informasi, pinjaman, interaksi nasabah yang dilakukan pada salah satu perusahaan *P2P Lending* yang menggunakan layanan PT. XYZ. Selain itu, data tersebut berisi jumlah interaksi yang diberikan oleh *customer service* terhadap nasabah melalui setiap *communication channel* yang digunakan.

3.3.1 Populasi dan Sampel

Pengambilan sampel (*sampling*) merupakan pemilihan subjek tertentu dari populasi yang ditentukan sebagai perwakilan dari populasi tersebut [100]. Metode pengambilan sampel (*sampling*) dilakukan untuk mendukung peneliti dalam pemilihan sampel yang representatif dan memberikan panduan terkait

seberapa besar sampel yang dibutuhkan untuk memastikan tingkat kepercayaan yang diinginkan untuk kesimpulan dan generalisasi [101]. Populasi yang digunakan dalam penelitian ini adalah nasabah dari perusahaan *P2P Lending*. Penelitian ini menggunakan metode pengambilan sampel non-probabilitas (*non-probability sampling*) dengan jenis *convenience sampling*. *Convenience sampling* melibatkan penggunaan aksesbilitas dan kenyamanan untuk penentuan sampel penelitian [102]. *Convenience sampling* digunakan karena ketersediaan data yang dimiliki oleh perusahaan dengan pada periode yang ditetapkan.

3.3.2 Periode Pengambilan Data

Penelitian ini melibatkan pengambilan data yang dilakukan oleh *P2P Lending* dengan periode yang ditentukan. Dalam hal ini, data nasabah yang diambil berdasarkan data pinjaman nasabah yang berada pada periode 01 Juli 2023 hingga 31 Desember 2023. Tanggal 31 Desember 2023 merupakan batas akhir data yang diambil sebelum data tersebut digunakan lebih lanjut. Jumlah data nasabah perusahaan *P2P Lending* sebesar 23.880 data.

3.4 Variabel Penelitian

Variabel penelitian yang digunakan terbagi menjadi dua jenis, yaitu variabel dependen dan variabel independen. Setiap variabel memiliki peran tersendiri dalam melakukan klasifikasi dan prediksi. Adapun penjelasan dari variabel dependen dan variabel independen yang digunakan pada penelitian ini, yaitu sebagai berikut.

3.4.1 Variabel Dependen

Variabel dependen merupakan variabel yang dipengaruhi atau yang menjadi akibat oleh karena keberadaan variabel independen [103]. Penelitian ini menggunakan variabel dependen berupa kategori intensitas penagihan yang bernama *intensity_category* untuk klasifikasi *Collection Intensity Scoring* dan *interaction_count* untuk prediksi *channel recommendation*. Adapun struktur dari variabel dependen yang digunakan dalam penelitian ini yang direpresentasikan pada Tabel 3.3.

Tabel 3.3 Struktur Data pada Variabel Dependen

No.	Variabel	Deskripsi	Tipe Data	Nilai Valid
Collection Intensity Scoring Classification				
1.	*intensity_category*	Kategori intensitas penagihan nasabah. Nilai '0' menandakan intensitas penagihan rendah (*low*). Nilai '1' menandakan intensitas penagihan sedang (*medium*). Nilai '2' menandakan intensitas penagihan tinggi (*high*).	Numerik	0, 1, 2
Channel Recommendation Prediction				
2.	*interaction_count*	Jumlah interaksi yang berhasil direspon nasabah pada *communication channel* yang direkomendasikan	Numerik	Numerik

3.4.2 Variabel Independen

Variabel independen merupakan merupakan variabel yang menjadi penyebab atau memungkinkan adanya pengaruh terhadap variabel lain [103]. Penelitian ini melibatkan penggunaan variabel independen yang digunakan untuk pendekatan klasifikasi *Collection Intensity Scoring* dan prediksi *channel recommendation*. Berikut struktur data pada variabel independen untuk pendekatan klasifikasi *Collection Intensity Scoring* dan prediksi *channel recommendation* yang ditunjukkan pada Tabel 3.4 dan Tabel 3.5 secara berturut-turut.

Tabel 3.4 Struktur Data pada Variabel Independen untuk Klasifikasi *Collection Intensity Scoring*

Collection Intensity Scoring Classification				
No.	**Variabel**	**Deskripsi**	**Tipe Data**	**Nilai Valid**
Customer Data				
1.	*gender*	Jenis Kelamin Nasabah	Kategorikal	*Male, Female*
2.	*age*	Usia Nasabah	Numerik	Numerik
Interaction Data				
1.	*email_count*	Jumlah interaksi *email* yang berhasil	Numerik	Numerik
2.	*robocall_count*	Jumlah interaksi *robocall* yang berhasil	Numerik	Numerik
3.	*sms_count*	Jumlah interaksi SMS yang berhasil	Numerik	Numerik

No.	Variabel	Deskripsi	Tipe Data	Nilai Valid
		Interaction Data		
4.	*telephony_count*	Jumlah interaksi *telephony* yang berhasil	Numerik	Numerik
5.	*total_interaction*	Total interaksi yang berhasil dari semua *communication channel*	Numerik	Numerik
		Loan Data		
1.	*total_loan*	Jumlah *loan_id* nasabah	Numerik	Numerik
2.	*paid_loans*	Jumlah pinjaman yang telah dibayarkan	Numerik	Numerik
3.	*avg_interacted_to_paid*	Rata-rata durasi dari interaksi hingga pembayaran	Numerik	Numerik
4.	*total_paid_amount*	Total pinjaman yang dibayarkan	Numerik	Numerik

Tabel 3.5 Struktur Data pada Variabel Independen untuk Prediksi *Channel Recommendation*

Recommendation Channel Prediction				
No.	**Variabel**	**Deskripsi**	**Tipe Data**	**Nilai Valid**
		Customer Data		
1.	*age*	Usia Nasabah	Numerik	Numerik
2.	*gender*	Jenis Kelamin Nasabah	Kategorikal	*Male, Female*
		Interaction Data		
1.	*total_interaction*	Total interaksi yang berhasil dari semua *communication channel*	Numerik	Numerik
2.	*last_interaction_type*	Jenis interaksi terakhir yang diterima nasabah	Kategorikal	*Email, Robocall, SMS, Telephony, None*
3.	*recommended_channel*	Jenis *communication channel* yang direkomendasikan	Kategorikal	*Email, robocall, SMS, telephony*
		Loan Data		
1.	*avg_interacted_to_paid*	Rata-rata durasi dari interaksi hingga pembayaran	Numerik	Numerik
2.	*intensity_category*	Kategori intensitas penagihan naabah	Numerik	0, 1, 2

3.5 Teknik Analisis Data

Teknik analisis data pada penelitian ini melibatkan proses *Extract, Transformation, Load* (*ETL*), *data pre-processing*, *modeling*, dan *statistical tests*. Proses *ETL* melibatkan penggunaan bahasa pemrograman *PostgreSQL* (PL/pgSQL) karena kemampuan dalam melakukan komputasi yang kompleks. Proses *ETL* diterapkan dalam DBeaver karena *software* ini yang bersifat *open source* dan gratis yang mendukung *database* PostgreSQL. Selanjutnya, *data pre-processing* dan *modeling* dijalankan menggunakan Python karena memiliki *library* analisis data yang beragam. Penggunaan Python ini dilakukan melalui Jupyter Notebook sebagai *Integrated Development Environment* (*IDE*) berbasis *web* yang bersifat *open-source* dan gratis. Jupyter *Notebook* digunakan oleh karena kemampuannya dalam meintegrasikan teks, persamaan matematika, visualisasi data, dan kode dalam satu dokumen, sehingga memudahkan dalam berinteraksi langsung dengan kode dan melihat hasil eksekusinya secara *real time*.

Penelitian ini melakukan pemodelan menggunakan algoritma *Random Forest* dan *K-Nearest Neighbors Regressor*. *Random Forest* digunakan untuk klasifikasi intensitas penagihan nasabah, sedangkan *K-Nearest Neighbors Regressor* digunakan untuk prediksi jumlah interaksi yang akan direspon nasabah pada *channel* yang direkomendasikan. Untuk meningkatkan performa model, diterapkan metode optimasi *Grid Search* dan *Bayesian Optimization*. Hasil *modeling* ini akan dievaluasi dan dibandingkan dengan *evaluation metrics* tersendiri. Adapun proses *statistical tests* yang dilakukan dengan R *Programming Language* dengan fokus utama pada teknik statistik dengan RStudio sebagai *IDE*. RStudio dipilih oleh karena antarmuka yang terintegrasi yang memungkinkan untuk membuat *script* dan melihat perintah, data, dan hasil secara *real time*. Pemilihan algoritma didasarkan kinerja melalui metrik utama berupa nilai *accuracy* yang tinggi pada *Random Forest* dan *Mean Absolute Error* (*MAE*) yang rendah pada *K-Nearest Neighbors Regressor*. Kinerja pada algoritma *Random Forest* dan *K-Nearest Neighbors Regressor* telah terbukti melalui komparasi dengan algoritma lainnya pada sejumlah penelitian terdahulu terkait *credit scoring* dan *recommendation system* yang ditunjukkan secara berturut-turut dalam Tabel 3.6 dan Tabel 3.7.

Tabel 3.6 Komparasi Algoritma pada Penelitian Terdahulu terkait *Credit Scoring*

Credit Scoring				
Algoritma	**Jurnal**			
	A study on credit scoring modeling with different feature selection and machine learning approaches [16]	*Loan default prediction of Chinese P2P market: a machine learning methodology* [17]	*A Comparative Performance Assessment of Ensemble Learning for Credit Scoring* [18]	*Random Forest-Bayesian Optimization for Product Quality Prediction with Large Scale Dimensions in Process Industrial Cyber-Physical Systems* [22]
	Nilai *Accuracy*			
NN	-	97.20	79.14	-
SVM	77.02	-	79.19	
LR	-	-	79.15	71.22
DT	92.21	-	79.00	72.03
NB	70.12	-	68.28	-
Bayesian	71.01	-	-	-
XGBoost	-	98.20	79.47	-
AdaBoost	-	-	78.61	-
LightGBM	-	-	79.47	-
Stacking	-	-	79.19	-
GBM	-	97.90	-	-
GLM	-	96.90	-	-
SVC	-	-	-	71.38
BPNN	-	-	-	72.29
RF	**93.12**	**98.40**	-	83.19
RF x Grid Search	-	-	**81.05**	-
RF x BO	-	-	-	**90.33**

Tabel 3.7 Komparasi Algoritma pada Penelitian Terdahulu terkait *Recommendation System*

Recommendation System				
Algoritma	**Jurnal**			
	A Hybrid Action-Related K-Nearest Neighbour (HAR-KNN) Approach for Recommendation Systems [19]	*E-Learning Course Recommender System Using Collaborative Filtering Models* [20]	*Water Quality Prediction Using Machine Learning Models based on Grid Search Method* [21]	*Estimation of rainfall erosivity factor in Italy and Switzerland using Bayesian optimization based machine learning models* [23]
	Nilai *Mean Absolute Error* (*MAE*)			
IBCF	1.2919	-	-	-
CDIBCF	1.2890	-	-	-
TWIBCF	1.2276	-	-	-
TRIBCF	1.1154	-	-	-
TCIBCF	1.1137	-	-	-
CF	0.9765	-	-	-
CBF	0.7685	-	-	-
SVD	-	-	-	-
NCF	-	-	-	-
SVD	-	0.0237	-	-
NCF	-	0.0156	-	-
DT	-	-	0.005	15.071
SVR	-	-	0.004	-
MLP	-	-	0.003	-
RF	-	-	-	13.901
GB	-	-	-	14.821
XGB	-	-	-	14.287
KNN	**0.7165**	**0.0142**	-	-
KNN x Grid Search	-	-	**0.009**	-
KNN x BO	-	-	-	**20.576**