BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian sebelumnya memiliki peran penting dalam menyediakan landasan teori untuk penelitian ini. Memahami penelitian sebelumnya memberikan pemahaman terkait metode, pendekatan dan hasil yang dicapai pada setiap konteks. Kajian ini mencakup berbagai studi terkait yang mendalami permasalahan serupa. Berikut adalah tinjauan beberapa penelitian terdahulu yang menjadi rujukan utama dalam penelitian ini.

Table 1 Penelitian Terdahulu

No	JUDUL PENELITIAN	METODE	HASIL PEMBAHASAN
1	Customer churning analysis using machine learning algorithms(2023)	SGB, LR, RF dan KNN	SGB memiliki performa terbaik dengan AUROC 0.839, diikuti oleh RF dengan AUROC 0.829.
2	Telecommunication customer churn prediction using machine learning methods(2022)	SVM, KNN, DT, RF, NB.	Support Vector Machine (SVM) memperoleh akurasi Terbaik74.54%
3	Integrated churn prediction and customer segmentation Framework for telco Business(2021)	LR, DT, RF, AdaBoost, MLP	AdaBoost memberikan F1-Score terbaik (63,11%) untuk Dataset 1, dan RF unggul untuk Dataset 2 (77,20. %).
4	Customer Churn Prediction Model using Explainable Machine learning(2023)	LR, RF, DT ,XGBOOST dan SHAP	XGBoost sebesar 0.962 dan F1-score 0.836. SHAP untuk mengidentifikasi fitur yang paling berpengaruh

5	Penerapan Seleksi Fitur Analysis of Variance Pada Algoritma Random Forest Classifier Dalam Klasifikasi Nilai Mahasiswa	Random Forest dan ANOVA	Seleksi fitur ANOVA tingkatkan akurasi dari 85,65% ke 87,31%.
6	Analisis Optimasi Forward Selection pada Klasifikasi Nilai Mahasiswa dengan Algoritma Naïve Bayes	Naïve Bayes Classifier dan Forward Selection	Akurasi meningkat dari 86% ke 93% setelah seleksi fitur.

Penelitian oleh B. Prabadevi et al. (2023) menggunakan empat algoritma, yaitu Stochastic Gradient Booster (SGB), Random Forest (RF), Logistic Regression (LR), dan K-Nearest Neighbors (KNN). Hasilnya, SGB memberikan performa terbaik dengan AUROC 0,839 diikuti RF (0,829), menunjukkan bahwa metode ensemble mampu meningkatkan akurasi prediksi churn dibandingkan model klasik.[4]

Penelitian oleh S. Wu et al. (2021) mengintegrasikan prediksi churn dan segmentasi pelanggan menggunakan SMOTE untuk mengatasi ketidakseimbangan data. Dari beberapa model yang diuji, AdaBoost dan Random Forest menunjukkan F1-Score terbaik (63,11% dan 77,20%), dengan segmentasi berbasis Bayesian Logistic Regression yang memberikan rekomendasi strategis untuk retensi pelanggan[5]. Pada penelitian Jitendra dan Harsh Maan (2023), digunakan model XGBoost dan SHAP untuk menjelaskan kontribusi fitur terhadap churn pelanggan. Model XGBoost mencapai akurasi 0,962 dengan F1-Score 0,836, menunjukkan bahwa kombinasi model boosting dan interpretasi SHAP dapat meningkatkan transparansi serta keandalan hasil prediksi[7].

Sementara itu, Muhammad Fath Thoriq et al. (2022) menunjukkan bahwa penerapan seleksi fitur ANOVA pada algoritma Random Forest Classifier dapat

meningkatkan akurasi model dari 85,65% menjadi 87,31%. Hasil ini memperkuat bahwa seleksi fitur mampu memperbaiki performa model[10].

2.2 Loyalitas Pelanggan

Loyalitas pelanggan adalah ukuran sejauh mana pelanggan tetap berkomitmen untuk menggunakan produk atau layanan dari suatu perusahaan dalam jangka waktu panjang. Loyalitas pelanggan bukan hanya mencerminkan kesetiaan pelanggan, tetapi juga menjadi indikator keberhasilan perusahaan dalam membangun hubungan yang kuat dengan konsumennya. Loyalitas pelanggan biasanya diidentifikasi melalui beberapa indikator, seperti durasi pelanggan menggunakan layanan (tenure), frekuensi pembelian ulang, dan kepuasan terhadap kualitas layanan. Dalam industri telekomunikasi, loyalitas pelanggan menjadi salah satu penentu utama keberhasilan perusahaan di tengah persaingan yang semakin ketat. Faktor-faktor seperti kualitas layanan, jenis kontrak, harga layanan, dan pengalaman pelanggan berperan penting dalam membangun loyalitas pelanggan. Oleh karena itu, memahami faktor-faktor yang memengaruhi loyalitas pelanggan menjadi sangat penting untuk menyusun strategi retensi yang efektif [6], [11]

2.3 Pemilihan Fitur dan Teknik Interpretasi

2.3.1 SFS

Metode Sequential Forward Selection (SFS) merupakan metode seleksi fitur dari kategori Wrapper yang berarti membutuhkan sebuah Learning Algorithm dalam mencari subset terbaik dari keseluruhan fitur yang berisi kombinasi fitur dengan jumlah tertentu. Teknik seleksi fitur SFS ini berjalan dengan strategi pencarian sekuensial untuk menemukan subset fitur terbaik. Pada setiap iterasi, SFS memilih fitur yang jika ditambahkan ke dalam subset yang sudah ada, akan menghasilkan peningkatan terbesar terhadap metrik evaluasi model. Proses ini terus berlangsung hingga tercapai jumlah fitur tertentu (ditentukan oleh parameter kkk) atau hingga penambahan fitur berikutnya tidak lagi meningkatkan performa secara signifikan.[9]

Keunggulan utama dari SFS adalah meningkatkan performa model dengan memilih hanya fitur yang paling relevan, sehingga membantu mengurangi overfitting dan meningkatkan interpretabilitas model. Selain itu, metode ini juga mempercepat proses pelatihan karena mengurangi jumlah fitur yang digunakan. Meskipun SFS dapat memakan waktu komputasi yang cukup lama, terutama dengan dataset besar, keuntungannya dalam menghasilkan model yang lebih sederhana, efisien, dan lebih mudah dipahami membuatnya sangat berguna dalam berbagai aplikasi pembelajaran mesin, termasuk klasifikasi dan regresi. Adapun formula untuk SFS dapat dilihat pada persamaan(2.7):

$$SFS(X) = arg\ max\ Metric(model(X_s \cup F_i), y) - Metric(model(X_s), y) \quad (2.7)$$

Dimana:

F= semua himpunan fitur

f_i = $fitur\ yang\ dipertimbangkan\ ke\ subset\ fitur\ x_s$

Metric(model(x_s),y) adalah metrik kinerja model pada fitur subset x_s

2.3.2 SHAP

SHAP (*SHapley Additive exPlanations*) adalah metode interpretabilitas untuk model menggunakan prinsip teori permainan untuk mengukur kontribusi masing-masing fitur dalam prediksi model. SHAP bekerja dengan cara menghitung nilai kontribusi marjinal dari setiap fitur terhadap prediksi model, kemudian membaginya secara adil berdasarkan semua kemungkinan kombinasi fitur yang ada. Keunggulan utama SHAP adalah sifatnya yang fleksibel, sehingga dapat digunakan untuk berbagai algoritma machine learning, termasuk XGBoost, LightGBM, dan Random Forest. Metode ini memungkinkan pengguna untuk memahami bagaimana perubahan nilai pada suatu fitur memengaruhi hasil prediksi model, baik dalam arah positif maupun negatif terhadap variabel target.[12].

Penelitian yang dilakukan oleh Kim et al(2025)memanfaatkan XGBoost-SHAP untuk menganalisis faktor risiko hipertensi pada wanita pascamenopause dan menemukan bahwa variabel usia serta lingkar pinggang memiliki kontribusi SHAP tertinggi terhadap prediksi tekanan darah[13]. Penelitian lain oleh Zhang dan Zhao menerapkan XGBoost-SHAP untuk memprediksi Environmental, Social, and Governance (ESG) rating perusahaan dan menemukan bahwa metode ini meningkatkan interpretabilitas serta akurasi model[14]. Berdasarkan hal tersebut, dalam penelitian ini SHAP digunakan untuk menginterpretasikan model XGBoost serta mengidentifikasi fitur yang paling berpengaruh terhadap prediksi churn pelanggan.

2.3.3 ANOVA

Analysis of Variance merupakan teknik standar untuk mengukur signifikansi statistik dari suatu set variabel independen dalam memprediksi variabel dependen. Teknik ini bekerja dengan membandingkan variasi antar kelompok (between groups) dan variasi di dalam kelompok (within groups) untuk menghasilkan nilai F, yang digunakan untuk menilai signifikansi hubungan antara fitur dan target.

Dalam penelitian ini, ANOVA digunakan bukan sebagai metode seleksi fitur utama, melainkan sebagai analisis pendukung untuk mengidentifikasi fitur-fitur yang berpengaruh signifikan terhadap variabel churn. Hasil analisis ANOVA kemudian digunakan sebagai dasar untuk membandingkan konsistensi dengan hasil seleksi fitur utama menggunakan Sequential Forward Selection (SFS) dan interpretasi model dengan SHAP. Dengan demikian, ANOVA berperan membantu memvalidasi apakah fitur yang signifikan secara statistik juga memiliki kontribusi besar terhadap performa dan interpretasi model. [10]

Berikut adalah formula untuk menghitung ANOVA dapat dilihat pada persamaan(2.1.1),(2.2),(2.3):

$$F = \frac{variance\ between\ groups}{variance\ within\ groups} \qquad (0)$$

$$variance\ between\ groups = \frac{\sum_i^n n_i(\bar{Y}_i - \bar{Y}_i)^2}{(K-1)} \quad (2.2)$$

$$variance\ within\ groups\ = \frac{\sum_{i=1}^{k}\sum_{j=1}^{ni}(\bar{Y}_{ij} - \bar{Y}_i)}{(n-k)} \quad (2.1)$$

Keterangan :
ni = Jumlah sampel di grup ke-i
$\bar{Y}i$ = Rata-rata sampel di grup ke-i
$\bar{Y}$ = Rata-rata total sampe
k = Jumlah grup
n = Jumlah total sampel

2.4 XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma boosting yang sangat populer karena efisiensi tinggi dan kemampuannya menghasilkan prediksi yang akurat. Algoritma ini menggunakan pendekatan regulasi kuat, termasuk L1 dan L2 regularization, untuk mengurangi risiko overfitting. Dengan desain paralel dan pemanfaatan memori yang optimal, XGBoost sangat cocok untuk melatih model dengan data besar dan kompleks dalam waktu singkat[1] .

Keunggulan utama XGBoost terletak pada fleksibilitas dan kecepatannya. Dalam kompetisi pembelajaran mesin, XGBoost sering kali digunakan karena kemampuannya menghasilkan model dengan performa tinggi. Setelah pelatihan, model yang telah terlatih dievaluasi pada test set untuk memastikan kemampuan generalisasi terhadap data baru. Pendekatan ini menjadikan XGBoost pilihan utama di berbagai bidang, termasuk analisis keuangan, prediksi kesehatan, dan analisis data pelanggan ormula dasar untuk XGBoost dalam regresi dapat dijelaskan dengan persamaan(2.8):

$$F(x) = \sum_{K=1}^{K} f_k(x) \quad (2.8)$$

Di mana:

F(x)= prediksi model

K= jumlah iterasi

$f_k(x) = \ prediksi\ pojon\ keputusan$

2.5 Random Forest

Random Forest adalah algoritma klasifikasi yang terdiri dari banyak pohon keputusan. Semakin banyak pohon dalam forest, semakin kuat prediksi dengan akurasi yang lebih tinggi. Algoritma ini menggunakan metode *bagging* dan keacakan fitur ketika membangun setiap pohon individu untuk menciptakan forest yang tidak berkorelasi. Dalam algoritma ini, setiap pohon akan menghasilkan keluaran tersendiri dari dataset yang disediakan, sehingga keluaran akhir ditentukan berdasarkan suara mayoritas dari pohon-pohon tersebut[15].

Random Forest sering digunakan untuk data yang memiliki banyak fitur, termasuk yang saling berinteraksi. Dengan menggunakan suara mayoritas untuk menentukan hasil akhir dalam klasifikasi, algoritma ini memastikan prediksi yang lebih andal dibandingkan model tunggal. Selain itu, Random Forest juga mampu memberikan informasi penting mengenai kontribusi setiap fitur dalam prediksi, yang berguna untuk analisis data lebih lanjut.[11] Adapun formula untuk Random Forest dapat dilihat pada persamaan(2.9):

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X) \tag{2.9}$$

Di mana:

$\hat{Y} = \ prediksi\ final$

N= jumlah pohon

$f_k(x) = \ prediksi\ pojon\ keputusan$

2.6 Bahasa Python

Python adalah bahasa pemrograman yang sangat populer dan serbaguna, ideal untuk berbagai aplikasi. Filosofi desainnya menekankan kemudahan dalam menulis, memelihara, dan memahami kode. Bahasa pemrograman Python merupakan bahasa pemrograman populer yang memiliki keunggulan yaitu mudah untuk

digunakan dalam mengembangkan sebuah produk perangkat lunak, perangkat keras, Internet of Things, aplikasi web, maupun video game. Selain memiliki keterbacaan kode yang tinggi, sehingga kode mudah dipahami, bahasa pemrograman ini memiliki library yang sangat banyak dan luas. Merupakan bahasa yang mendukung ekosistem Internet of Things dengan sangat baik[16]. [17]