BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian-penelitian yang telah dilakukan sebelumnya digunakan untuk meninjau temuan-temuan yang relevan dengan penelitian yang akan dilaksanakan dijabarkan berikut ini.

1. Bismah Hasan, Zubair, Shahrukh Ali Shaikh, Abdul Khaliq, Ghalib Nadeem (2024). Penelitian yang dilakukan oleh Bismah Hasan, Zubair, Shahrukh Ali Shaikh, Abdul Khaliq, Ghalib Nadeem dengan judul "Data-Driven Decision-Making: Accurate Customer Churn Prediction with Cat-Boost" memprediksi tingkat *churn* pelanggan di industri telekomunikasi. Dataset yang digunakan berisi riwayat pelanggan dengan jumlah 7043 data dan 21 atribut yang mencakup keseluruhan, mulai dari demografi dasar (usia, jenis kelamin, masa kerja) hingga metrik khusus layanan (biaya bulanan, jenis kontrak, dan penggunaan TV streaming. Tahap *pre-processing* yang dilakukan yaitu mengubah dataset menjadi *one-hot encoding, feature scaling*, dan data cleaning terhadap *missing value* dan outlier. Setelah dilakukan pengujian, didapatkan algoritma *CatBoost* menjadi model terbaik dibandingkan *Stochastic Gradient Boosting* dan *Extreme Gradient Boosting.* Hasil evaluasi menunjukkan bahwa *CatBoost* mencapai akurasi 94%, *recall* 97%, dan *AUC-ROC* 0.94 setelah dioptimasi. Penelitian ini menyoroti kemampuan algoritma CatBoost untuk memberikan prediksi churn yang akurat, sehingga memungkinkan perusahaan telekomunikasi melakukan pencegahan churn secara proaktif. Keunggulan algoritma ini mencakup penanganan fitur kategorikal secara efisien dan kemampuannya untuk meningkatkan strategi retensi pelanggan (Hasan et al., 2024).
2. Wahyu Nugraha dan Muhamad Syarif (2023).
   Penelitian yang dilakukan oleh Wahyu Nugraha dan Muhamad Syarif dengan judul "Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost" memprediksi *churn* di industri perbankan. Penelitian ini menggunakan dataset Churn Modelling dari Kaggle, yang berisi rincian data pelanggan bank dengan total 14 atribut, termasuk variabel target (*Exited*) yang menggambarkan apakah pelanggan menutup rekeningnya atau tetap menjadi pelanggan. Variabel lainnya mencakup CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, dan IsActiveMember. Dataset

ini menunjukkan ketidakseimbangan kelas dengan rasio churn sebesar 20.37% dan *non-churn* sebesar 79.63%. Pada tahap *pre-processing* dilakukan dengan *cleaning data* dengan menghapus data duplikat dan atribut yang tidak relevan. Penelitian ini menggunakan tiga algoritma boosting populer yaitu *CatBoost*, *XGBoost*, dan *LightGBM*. Model diuji dengan pembagian data 70% untuk training dan 30% untuk testing, dengan tuning parameter *scale pos weight* (default, 3, dan 5). Hasilnya, *CatBoost* dengan *scale pos weight* 5 memiliki nilai *recall* tertinggi (0.79). penelitian ini menunjukkan *CatBoost* terbukti unggul dengan performa *recall* terbaik, mendukung penerapannya untuk memprediksi churn di industri perbankan. Penelitian ini menyoroti pentingnya tuning parameter untuk meningkatkan kinerja model pada data yang tidak seimbang (Nugraha & Syarif, 2023).

3. Andy Hermawan, Nila Rusiardi Jayanti, Zia Tabaruk, Faizal Lutfi Yoga Triadi, Aji Saputra, M. Rahmat Hidayat Syachrudin (2024).

   Penelitian yang dilakukan oleh Andy Hermawan, Nila Rusiardi Jayanti, Zia Tabaruk, Faizal Lutfi Yoga Triadi, Aji Saputra, dan M. Rahmat Hidayat Syachrudin dengan judul "Membangun Model Prediksi Churn Pelanggan yang Akurat (Studi Kasus tentang TELCO Company)" mendeteksi pelanggan yang berisiko churn. Penelitian menggunakan dataset pelanggan dari TELCO Company, yang mencakup informasi seperti *tenure* (lama berlangganan), *monthly charges* (biaya bulanan), dan *service typ*e (jenis layanan). Dataset terdiri dari 4930 data dengan 11 atribut. Tahap *pre-prpcessing* meliputi data cleaning dengan menghapus data duplikat dan menangani outliers pada kolom numerik, juga dilakukan *feature engineering* untuk membuat fitur baru seperti *ratio monthly charges* dan *tenure group.* Hasil yang didapatkan yaitu *Gradient Boosting* memberikan performa terbaik dibandingkan *Random Forest* dan *XGBoost*. *Gradient Boosting* menghasilkan *F2 Score* sebesar 0.748 yang menunjukkan keseimbangan antara *precision* dan *recall.* Oleh karena itu, penelitian ini menyatakan bahwa *Gradient Boosting* unggul dibandingkan metode *rule-based* tradisional dalam mendeteksi pelanggan *churn* (Andy Hermawan et al., 2024).

4. Muhammad Adji Purnama, Jilang Ramadhani, Yoga Safitra Anugraha, Lusiana Efrizoni (2024).

   Penelitian yang dilakukan oleh Muhammad Adji Purnama, Jilang Ramadhani, Yoga Safitra Anugraha, Lusiana Efrizoni dengan judul "Perbandingan Performa Algoritma Random Forest Dan Gradient Boosting Dalam Mengklasifikasi Churn Telco" melakukan perbandingan algoritma *Random Forest* dan *Gradient Boosting*

dalam mengklasifikasi *churn* pelanggan Telco. Dataset yang digunakan adalah dataset publik Telco Customer Churn dari California, diambil dari Kaggle. Dataset ini memiliki 7.043 baris dan 50 atribut, mencakup informasi demografi, layanan yang digunakan pelanggan, dan status churn. *Pre-processing* dilakukan dengan *cleaning* data yaitu menghapus 30 atribut*,* transformasi data yaitu mengubah data menjadi bentuk numerik, dan *splitting* data yaitu membagi dataset menjadi 80% untuk *training* dan 20% untuk *testing*. Hasilnya, *Gradient Boosting* dengan akurasi 81% dan *ROC-AUC* 87% menunjukkan performa lebih baik dibandingkan *Random Forest.* Oleh karena itu, *Gradient Boosting* merupakan pilihan yang sangat baik untuk prediksi *churn* pelanggan karena kombinasi akurasi tinggi, kemampuan menangani hubungan data yang kompleks, dan efisiensi dalam proses pelatihan serta prediksi (Adji Purnama et al., n.d.).

5. A. Thirunirai Selvi, A. Divya Dharshini, V. Devadharshini, M. Hema, S. Aasiya Begam (2024).

   Penelitian yang dilakukan oleh A. Thirunirai Selvi, A. Divya Dharshini, V. Devadharshini, M. Hema, S. Aasiya Begam dengan judul "Client Churn Prediction Using Gradient Boosting Classifier" mengembangkan model prediksi *churn* pelanggan yang efektif dengan menggunakan algoritma *Gradient Boosting.* Dataset yang digunakan mencakup data historis pelanggan, termasuk pola penggunaan, metrik keterlibatan, dan data demografi pelanggan. Proses pengumpulan data melibatkan data terkait seperti ID pelanggan, usia, saldo, riwayat transaksi, dan interaksi pelanggan. Pada tahap *pre-processing* dilakukan pembersihan data utnuk mengisi datang yang *missing value* dengan *mean,* median, atau modus, juga dilakukan transformasi data menggunakan teknik *one-hot-encoding*, dan juga pembagian dataset. Setelah dilakukan pengujian, didapatkan akurasi 86.4%, *precision* kelas 0 (tidak *churn)* yaitu 97% dan kelas 1 (*churn)* yaitu 46%, *recall* kelas 0 yaitu 87% dan kelas 1 yaitu 79%, dan *F1-Score* kelas 0 yaitu 91% dan kelas 1 yaitu 60%. *Gradient Boosting* memiliki performa yang unggul dengan akurasi tinggi (86.4%), meskipun *precision* untuk memprediksi *churn* (kelas 1) masih relatif rendah (46%). Namun, *recall* untuk *churn* cukup tinggi (79%), menunjukkan kemampuan model untuk menangkap pelanggan yang benar-benar *churn*. Model *Gradient Boosting* disarankan untuk digunakan, terutama jika *recall churn* menjadi prioritas utama perusahaan (Selvi et al., n.d.).

2.2 Landasan Teori

2.2.1 Machine Learning

Pembelajaran mesin adalah bagian dari kecerdasan buatan yang memungkinkan komputer untuk "belajar secara mandiri" dari data pelatihan dan terus meningkat kinerjanya seiring waktu tanpa perlu diprogram secara langsung. Algoritme pembelajaran mesin dapat mengenali pola dalam data dan mempelajarinya untuk membuat prediksi secara mandiri. Secara sederhana, pembelajaran mesin memungkinkan model dan algoritme untuk belajar melalui pengalaman. Berbeda dengan pemrograman tradisional, di mana seorang insinyur komputer menulis serangkaian instruksi untuk mengubah data masukan menjadi keluaran berdasarkan aturan IF-THEN, pembelajaran mesin berfokus pada kemampuan komputer untuk memahami data dan menghasilkan hasil yang relevan. Dengan machine learning, komputer dapat mengambil keputusan atau melakukan tindakan berdasarkan analisis data tanpa memerlukan pemrograman eksplisit. Secara sederhana, mesin "mengambil pelajaran" dari pengalaman sebelumnya dan menggunakan pengetahuan yang diperoleh untuk menghadapi situasi baru (Chyan et al., n.d.).

2.2.2 Python

Python adalah bahasa pemrograman dengan sintaksis yang sederhana dan penerapannya yang luas, serta yang paling penting, bersifat open source. Karena faktor ini, Python dianggap sebagai salah satu bahasa terbaik untuk pemula. Bagi seorang akademisi atau praktisi komputer, mempelajari setidaknya satu bahasa pemrograman sangat penting, karena segala inovasi dan teknologi didasarkan pada pemahaman mendalam tentang komputer, sistem operasi, antarmuka pemrograman aplikasi perangkat lunak, atau perangkat keras tertentu. Semua itu dibuat oleh programmer yang mengikuti pola pikir tertentu. Untuk mengembangkan pola pikir seperti itu, seseorang perlu membiasakan diri dengan bahasa pemrograman tertentu dan menjadi ahli dalam pengembangan perangkat lunak (Sigmon, 2023).

2.2.3 Google Colab

Ada berbagai macam alat bantu pengajaran berbasis komputer yang digunakan di tingkat sarjana dan pascasarjana. Baru-baru ini, alat baru seperti Google Colaboratory, atau disingkat "Colab", menyediakan layanan Jupyter notebook online yang dapat diakses secara bebas untuk kolaborasi dalam kegiatan pendidikan dan penelitian. Colab banyak digunakan untuk mengajarkan pembelajaran mesin (ML) dengan menulis dan menjalankan kode

Python serta alat ML (seperti Tensorflow, Keras, dll.) melalui browser. Colab memungkinkan kita untuk berbagi eksperimen yang dapat direproduksi di Web, dan selain itu, alat ini juga sangat cocok digunakan di ruang kelas. Selain itu, siswa menjadi lebih termotivasi untuk mengerjakan tugas laboratorium tanpa perlu mengunduh atau mengonfigurasi paket perangkat lunak dan ketergantungan di komputer mereka, hanya dengan mengikuti instruksi sesi Colab menggunakan browser. Lebih lanjut, hampir semua universitas terpaksa tutup akibat pandemi COVID-19, yang memaksa kita untuk beradaptasi dengan skenario pembelajaran virtual. Colab memberikan portabilitas dan aksesibilitas, karena dapat dijalankan bahkan di smartphone (Canesche et al., 2021).

2.2.4 SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) adalah teknik oversampling yang digunakan untuk menyeimbangkan distribusi dataset dengan meningkatkan jumlah data pada kelas minoritas, sehingga jumlah data pada kelas minoritas menjadi setara dengan jumlah data pada kelas mayoritas. Penggunaan teknik oversampling dapat menyebabkan overfitting, sehingga SMOTE diusulkan sebagai solusi untuk mengatasi overfitting, dengan memanfaatkan konsep ketetanggaan terdekat (KNN) pada jumlah oversampling yang diinginkan. Rumusnya adalah sebagai berikut (Gumelar et al., n.d.):

$$X_{syn} = X_i + (X_{knn} - X_i).\sigma$$

Keterangan:

- X_{syn}: data sintesis yang akan diciptakan.
- X_i : data yang akan direplikasi.
- X_{knn}: data yang memiliki jarak terdekat dengan data yang akan direplikasi.
- σ: nilai acak antara 0 dan 1.

2.2.5 Gradient Boosting

Gradient Boosting adalah teknik machine learning yang efektif untuk mengklasifikasikan penyakit diabetes tipe 2, dengan cara menggabungkan beberapa model prediktif sederhana ("weak learners") menjadi satu model yang lebih kuat. Proses ini dimulai dengan pembuatan model pertama untuk memprediksi target, lalu model-model berikutnya dibangun untuk memperbaiki kesalahan model sebelumnya, dengan fokus pada data yang

sulit diprediksi. Dalam praktiknya, data yang digunakan sering kali mencakup berbagai fitur seperti kadar glukosa dan indeks massa tubuh. Setelah data dikumpulkan dan dibersihkan, Gradient Boosting digunakan untuk melatih model secara iteratif, dengan melakukan tuning hyperparameter untuk mengoptimalkan kinerja model. Dengan pendekatan ini, Gradient Boosting dapat memberikan akurasi tinggi dalam mengklasifikasikan pasien diabetes tipe 2, menjadikannya pilihan yang populer dalam analisis data medis .

2.2.6 CatBoost

CatBoost adalah algoritma baru dalam teknik *gradient boosting decision tree* (GBDT) yang mampu menangani fitur kategorikal dengan baik. Algoritma ini berbeda dari algoritma GBDT tradisional dalam beberapa aspek:

1. Penanganan Fitur Kategorikal: CatBoost menangani fitur kategorikal selama waktu pelatihan, bukan pada waktu praproses. CatBoost memungkinkan penggunaan seluruh dataset untuk pelatihan. Menurut Prokhorenkova et al. (2018), metode statistik target (TS) sangat efisien dalam menangani fitur kategorikal dengan kehilangan informasi minimal. Secara khusus, untuk setiap contoh data, CatBoost melakukan permutasi acak dari dataset dan menghitung nilai label rata-rata untuk contoh dengan nilai kategori yang sama yang ditempatkan sebelum yang diberikan dalam permutasi.

2. Kombinasi Fitur: Semua fitur kategorikal dapat digabungkan menjadi satu fitur baru. Ketika membangun pembelahan baru untuk pohon keputusan, CatBoost menggunakan pendekatan *greedy* untuk mempertimbangkan kombinasi fitur tersebut. Untuk pembelahan pertama di pohon, tidak ada kombinasi yang dipertimbangkan, namun untuk pembelahan kedua dan seterusnya, CatBoost menggabungkan semua kombinasi yang telah ditentukan sebelumnya dengan semua fitur kategorikal dalam dataset.

3. Peningkatan Tanpa Bias dengan Fitur Kategorikal: Saat menggunakan metode TS untuk mengubah fitur kategorikal menjadi nilai numerik, distribusinya akan berbeda dari distribusi asli, dan deviasi distribusi ini dapat menyebabkan deviasi solusi. Ini adalah masalah yang tak terhindarkan dalam metode GBDT tradisional. mengembangkan metode baru melalui analisis teoretis untuk mengatasi bias gradien ini, yang dinamakan *ordered boosting*.

CatBoost juga merupakan algoritma berbasis pohon keputusan. Meskipun memiliki banyak parameter, parameter utama yang memengaruhi akurasi dan stabilitas model sama seperti pada Random Forest (RF). Untuk model CatBoost, jumlah iterasi berkisar antara 200 hingga 800 dengan interval 100, kedalaman pohon maksimum antara 2 hingga 10 dengan interval 2, dan rasio subset dari semua dataset antara 0,5 hingga 1 dengan interval 0,05. Sedangkan untuk model SVM, parameter yang paling penting adalah koefisien regulasi dan parameter gamma dari fungsi kernel. Fungsi kernel Gaussian biasanya lebih akurat dibandingkan kernel linear. Oleh karena itu, fungsi Gaussian dipilih sebagai fungsi kernel basis radial dalam model SVM, dengan parameter C berkisar antara 10 hingga 100 dengan interval 10, dan parameter γ antara 5 hingga 50 dengan interval 5.

CatBoost banyak digunakan dalam berbagai aplikasi seperti model prediksi, termasuk dalam pemodelan hidrologi untuk mengestimasi evapotranspirasi harian (ET0) dengan dataset meteorologi yang berbeda. Algoritma ini memiliki keunggulan dalam akurasi, stabilitas, dan efisiensi waktu pelatihan, serta lebih mudah diterapkan dalam skenario dunia nyata dibandingkan dengan metode lain seperti SVM atau RF (Huang et al., 2019).

2.2.7 Prediksi churn

Penerapan Gradient Boosting dan CatBoost dalam prediksi churn melibatkan beberapa langkah penting. Proses dimulai dengan pengumpulan data pelanggan, yang mencakup informasi seperti demografi, perilaku pembelian, dan pola transaksi. Data ini kemudian melalui tahap pre-processing, termasuk menangani nilai yang hilang, menghapus fitur tidak relevan, dan melakukan encoding pada data kategorikal. Gradient Boosting digunakan untuk membangun model prediksi dengan menggabungkan banyak pohon keputusan secara iteratif, di mana setiap pohon berusaha memperbaiki kesalahan prediksi dari pohon sebelumnya. Sementara itu, CatBoost dirancang khusus untuk mengolah data kategorikal secara langsung, tanpa memerlukan transformasi yang rumit, sehingga lebih efisien dan akurat. Model yang dihasilkan kemudian dievaluasi menggunakan metrik seperti akurasi, F1-score, dan ROC-AUC untuk memastikan kemampuan model dalam memprediksi pelanggan yang cenderung churn. Dengan menganalisis fitur yang penting, seperti kategori produk atau pola pembelian, model ini dapat memberikan wawasan bagi perusahaan untuk mengurangi churn (Hafidatus Sholeha et al., 2024).

2.2.8 Perbandingan Gradient Boosting dan CatBoost

Gradient Boosting dan CatBoost adalah dua metode yang sering digunakan untuk prediksi churn, masing-masing dengan pendekatan dan keunggulan yang berbeda. Gradient Boosting bekerja dengan cara membangun model secara iteratif, di mana setiap pohon keputusan baru ditambahkan untuk mengurangi kesalahan dari model sebelumnya. Teknik ini dikenal fleksibel karena mampu menangkap pola non-linear yang kompleks dalam data. Namun, Gradient Boosting memerlukan preprocessing tambahan, seperti one-hot encoding, untuk menangani data kategorikal. Proses ini dapat meningkatkan waktu dan kompleksitas pengolahan data, terutama pada dataset dengan banyak fitur kategorikal. Di sisi lain, CatBoost dirancang khusus untuk mengatasi kelemahan ini. Algoritma ini dapat menangani data kategorikal secara langsung tanpa memerlukan one-hot encoding, sehingga lebih efisien dalam hal waktu dan sumber daya. CatBoost menggunakan teknik berbasis statistik untuk memproses data kategorikal, yang tidak hanya mengurangi risiko overfitting tetapi juga menghasilkan model yang lebih stabil dan akurat. Selain itu, CatBoost memiliki keunggulan dalam efisiensi komputasi, menjadikannya pilihan yang ideal untuk dataset besar dengan fitur kategorikal yang dominan. Dalam konteks prediksi churn, kedua metode ini memiliki keunggulan masing-masing. Gradient Boosting cenderung lebih cocok untuk data yang sudah diproses dengan baik dan didominasi oleh fitur numerik. Sebaliknya, CatBoost menunjukkan performa unggul pada dataset dengan banyak fitur kategorikal, karena kemampuannya menangani data tersebut secara langsung tanpa perlu transformasi tambahan. Dengan demikian, pilihan antara kedua metode ini bergantung pada karakteristik dataset dan kebutuhan spesifik analisis (Rahman et al., 2020).