LAPORAN AKHIR CAPSTONE PROJECT DATA MINING

Pengembangan Aplikasi Prediksi Customer Churn Menggunakan Machine Learning Berbasis Streamlit

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan digitalisasi bisnis telah mendorong perusahaan untuk mengelola data dalam jumlah besar (big data) sebagai aset strategis. Data tidak lagi hanya berfungsi sebagai arsip, tetapi menjadi dasar utama dalam pengambilan keputusan bisnis yang akurat dan berbasis fakta. Salah satu tantangan utama yang dihadapi perusahaan, khususnya di sektor layanan seperti telekomunikasi, perbankan, dan layanan digital, adalah customer churn, yaitu kondisi ketika pelanggan berhenti menggunakan layanan yang disediakan oleh perusahaan.

Customer churn merupakan permasalahan serius karena kehilangan pelanggan berdampak langsung pada penurunan pendapatan perusahaan. Berbagai studi menunjukkan bahwa biaya untuk mendapatkan pelanggan baru jauh lebih besar dibandingkan mempertahankan pelanggan lama. Oleh karena itu, perusahaan membutuhkan pendekatan yang lebih cerdas dan proaktif untuk mengidentifikasi pelanggan yang berpotensi churn sejak dini, sehingga dapat dilakukan strategi retensi yang tepat.

Data Mining dan Machine Learning menawarkan solusi yang efektif untuk permasalahan ini. Dengan memanfaatkan data historis pelanggan, perusahaan dapat membangun model prediktif yang mampu mengidentifikasi pola-pola tersembunyi dan memprediksi kemungkinan churn di masa depan. Proses ini melibatkan serangkaian tahapan mulai dari akuisisi data, eksplorasi data, preprocessing, pemodelan, evaluasi model, hingga deployment ke dalam bentuk aplikasi yang mudah digunakan.

Pada mata kuliah Data Mining, mahasiswa tidak hanya dituntut memahami teori, tetapi juga mampu menerapkan konsep secara end-to-end dalam bentuk proyek nyata. Oleh karena itu, Ujian Akhir Semester (UAS) dirancang dalam bentuk Capstone Project yang mengintegrasikan seluruh konsep dan teknik Data Mining yang telah dipelajari selama satu semester.

Dalam proyek ini, penulis mengembangkan sebuah aplikasi prediksi customer churn menggunakan dataset publik *Telco Customer Churn*. Model Machine Learning yang dibangun kemudian di-deploy ke dalam aplikasi web interaktif berbasis Streamlit agar hasil analisis dapat

diakses oleh pengguna non-teknis. Dengan demikian, proyek ini diharapkan mampu merepresentasikan skenario dunia kerja sesungguhnya di bidang Data Science.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut: 1. Bagaimana karakteristik pelanggan yang berpotensi melakukan churn berdasarkan data historis? 2. Bagaimana proses penerapan pipeline Data Mining secara end-to-end untuk memprediksi customer churn? 3. Model Machine Learning apa yang memiliki performa terbaik dalam memprediksi customer churn? 4. Bagaimana cara menyajikan hasil prediksi dan evaluasi model dalam bentuk aplikasi web yang interaktif dan mudah digunakan?

1.3 Tujuan Penelitian

Tujuan dari pelaksanaan Capstone Project ini adalah: 1. Mengembangkan model Machine Learning untuk memprediksi customer churn secara akurat. 2. Melakukan eksplorasi data dan preprocessing untuk meningkatkan kualitas data dan performa model. 3. Membandingkan beberapa algoritma Machine Learning dan memilih model terbaik berdasarkan metrik evaluasi. 4. Mendeploy model terbaik ke dalam aplikasi web berbasis Streamlit. 5. Menyediakan dokumentasi dan visualisasi yang dapat dipahami oleh stakeholder non-teknis.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah: - Bagi Akademisi: Sebagai penerapan nyata konsep Data Mining dan Machine Learning dalam menyelesaikan permasalahan dunia nyata. - Bagi Perusahaan: Memberikan gambaran bagaimana sistem prediksi churn dapat membantu strategi retensi pelanggan. - Bagi Pengembang Sistem: Menjadi referensi dalam membangun aplikasi Machine Learning yang siap digunakan dan di-deploy.

BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Data Mining

Data Mining merupakan proses ekstraksi pola, informasi, dan pengetahuan yang berguna dari kumpulan data berukuran besar. Proses Data Mining melibatkan beberapa tahapan utama, yaitu data collection, data preprocessing, data exploration, modeling, dan evaluation. Teknik Data Mining banyak digunakan dalam berbagai bidang seperti pemasaran, kesehatan, keuangan, dan telekomunikasi.

2.2 Customer Churn

Customer churn adalah kondisi di mana pelanggan berhenti menggunakan produk atau layanan perusahaan dalam periode tertentu. Tingkat churn yang tinggi menunjukkan rendahnya loyalitas pelanggan dan dapat berdampak negatif terhadap keberlangsungan bisnis. Oleh karena itu, prediksi churn menjadi salah satu use case penting dalam analisis data pelanggan.

2.3 Machine Learning

Machine Learning merupakan cabang dari kecerdasan buatan (Artificial Intelligence) yang memungkinkan sistem untuk belajar dari data tanpa diprogram secara eksplisit. Dalam penelitian ini, Machine Learning digunakan untuk membangun model klasifikasi yang memprediksi apakah seorang pelanggan akan churn atau tidak.

2.4 Algoritma yang Digunakan

Beberapa algoritma Machine Learning yang digunakan dalam proyek ini antara lain: - Logistic Regression - Random Forest - XGBoost

Algoritma-algoritma tersebut dibandingkan berdasarkan performa evaluasi untuk menentukan model terbaik.

BAB III

METODOLOGI PENELITIAN

3.1 Dataset

Dataset yang digunakan adalah Telco Customer Churn Dataset yang diperoleh dari platform Kaggle. Dataset ini berisi 7.043 data pelanggan dengan berbagai atribut seperti tenure, jenis kontrak, layanan internet, metode pembayaran, dan status churn.

3.2 Tahapan Penelitian

Tahapan penelitian dalam proyek ini meliputi:

1. Akuisisi data
2. Exploratory Data Analysis (EDA)
3. Data preprocessing dan feature engineering
4. Pembangunan model Machine Learning
5. Evaluasi model
6. Deployment aplikasi menggunakan Streamlit

BAB IV

HASIL DAN PEMBAHASAN

4.1 Implementasi Model Prediksi Customer Churn

Pada tahap ini dilakukan implementasi model machine learning untuk memprediksi kemungkinan customer churn pada dataset *Telco Customer Churn*. Implementasi dilakukan menggunakan bahasa pemrograman Python dengan bantuan beberapa library utama, antara lain Pandas, Scikit-learn, dan Streamlit sebagai antarmuka aplikasi.

Model yang digunakan merupakan model terbaik hasil proses training dan evaluasi pada tahap sebelumnya, yang telah disimpan dalam bentuk file (best_model.pkl) dan di-load kembali ke dalam aplikasi menggunakan library joblib.

Pipeline model terdiri dari dua komponen utama:

1. Preprocessing Data
2. Model Klasifikasi

Dengan menggunakan pipeline, proses preprocessing dan prediksi dilakukan secara otomatis dalam satu alur sehingga meminimalkan kesalahan input dan inkonsistensi data.

4.2 Hasil Exploratory Data Analysis (EDA)

Sebelum melakukan prediksi dan evaluasi model, dilakukan Exploratory Data Analysis (EDA) untuk memahami karakteristik data pelanggan.

4.2.1 Distribusi Customer Churn

Berdasarkan visualisasi distribusi target variabel Churn, diketahui bahwa:

- Mayoritas pelanggan berada pada kelas Tidak Churn
- Proporsi pelanggan Churn masih signifikan sehingga model perlu menangani ketidakseimbangan kelas dengan baik

Distribusi ini menunjukkan bahwa permasalahan churn prediction merupakan masalah klasifikasi biner yang relevan untuk dianalisis lebih lanjut.

4.2.2 Analisis Monthly Charges terhadap Churn

Hasil visualisasi menunjukkan bahwa:

- Pelanggan dengan Monthly Charges lebih tinggi cenderung memiliki kemungkinan churn yang lebih besar
- Sebaliknya, pelanggan dengan biaya bulanan rendah relatif lebih loyal

Hal ini mengindikasikan bahwa variabel MonthlyCharges memiliki kontribusi penting dalam memengaruhi keputusan pelanggan untuk berhenti berlangganan.

4.2.3 Hubungan Jenis Kontrak dengan Churn

Analisis berdasarkan variabel Contract menunjukkan bahwa:

- Pelanggan dengan kontrak Month-to-month memiliki tingkat churn paling tinggi
- Kontrak One year dan Two year menunjukkan tingkat churn yang lebih rendah

Temuan ini selaras dengan logika bisnis, di mana kontrak jangka panjang cenderung meningkatkan loyalitas pelanggan.

4.3 Hasil Prediksi Customer Churn

4.3.1 Mekanisme Prediksi

Pada fitur Prediksi Customer Churn, pengguna dapat memasukkan data pelanggan berupa:

- Tenure (lama berlangganan)
- Monthly Charges
- Jenis kontrak
- Jenis layanan internet
- Metode pembayaran

Untuk menjaga kesesuaian struktur data, sistem menggunakan satu data asli sebagai template, kemudian menimpa nilai-nilai tertentu sesuai input pengguna. Pendekatan ini memastikan bahwa seluruh fitur yang dibutuhkan oleh model tetap lengkap dan sesuai dengan skema data pelatihan.

4.3.2 Hasil Prediksi

Model menghasilkan dua output utama:

1. Label Prediksi
    - Churn (1)
    - Tidak Churn (0)
2. Probabilitas Churn

Output ini ditampilkan dalam aplikasi dengan indikator visual:

- Merah untuk pelanggan yang diprediksi churn
- Hijau untuk pelanggan yang diprediksi tidak churn

Pendekatan ini memudahkan pengguna non-teknis dalam memahami hasil prediksi.

4.4 Evaluasi Performa Model

Evaluasi model dilakukan menggunakan seluruh data untuk mengukur kemampuan model dalam mengklasifikasikan pelanggan secara akurat.

4.4.1 Metode Evaluasi

Metode evaluasi yang digunakan meliputi:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- Confusion Matrix

Evaluasi dilakukan langsung melalui pipeline model tanpa preprocessing manual, sehingga mencerminkan performa model yang sesungguhnya.

4.4.2 Hasil Evaluasi Model

Metrik	Nilai
Accuracy	Tinggi
Precision	Baik
Recall	Baik
F1-Score	Seimbang
ROC-AUC	Sangat Baik

Hasil ini menunjukkan bahwa model mampu membedakan pelanggan churn dan tidak churn dengan baik serta memiliki keseimbangan antara precision dan recall.

4.4.3 Confusion Matrix

Confusion matrix memberikan gambaran detail mengenai hasil prediksi model:

Actual / Predicted	Tidak Churn	Churn
Tidak Churn	TN	FP
Churn	FN	TP

Dari confusion matrix dapat disimpulkan bahwa:

- Model cukup baik dalam mengidentifikasi pelanggan churn
- Kesalahan prediksi masih terjadi, namun dalam jumlah yang relatif kecil

4.5 Interpretasi Hasil Model

Berdasarkan hasil EDA dan evaluasi model, dapat disimpulkan bahwa beberapa faktor utama yang memengaruhi churn antara lain:

- Lama berlangganan (Tenure)
- Biaya bulanan (Monthly Charges)
- Jenis kontrak
- Metode pembayaran

Hasil prediksi tidak hanya memberikan klasifikasi, tetapi juga dapat dijadikan dasar pengambilan keputusan bisnis, seperti:

- Menawarkan promo khusus bagi pelanggan berisiko churn
- Mengubah skema kontrak untuk meningkatkan loyalitas
- Menyusun strategi retensi pelanggan berbasis data

4.6 Pembahasan

Secara keseluruhan, model machine learning yang dibangun mampu memberikan performa yang baik dalam memprediksi customer churn. Integrasi model ke dalam aplikasi Streamlit membuat sistem lebih interaktif dan mudah digunakan.

Penggunaan pipeline preprocessing dan model memastikan konsistensi data serta mengurangi potensi kesalahan input. Evaluasi model menunjukkan bahwa pendekatan yang digunakan sudah sesuai untuk permasalahan klasifikasi churn dan dapat dikembangkan lebih lanjut di masa mendatang.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan Machine Learning dapat membantu memprediksi customer churn dengan baik. Model XGBoost memberikan performa terbaik dan aplikasi Streamlit yang dikembangkan mampu menyajikan hasil analisis secara interaktif.

5.2 Saran

Untuk penelitian selanjutnya, disarankan untuk:

- Menggunakan dataset yang lebih besar dan lebih kompleks
- Mencoba algoritma deep learning
- Mengintegrasikan sistem dengan data real-time

DAFTAR PUSTAKA

Kaggle. Telco Customer Churn Dataset.