

BANK PREDIKSI CUSTOMER CHURN PADA LAYANAN PERBANKAN DENGAN METODE VISUALISASI DATA DAN MODEL PREDIKSI XGBOOST

1st Agisni Zahra Latifa
agisnizaral@student.telkomuniversity.ac.id

2nd Agung Hadi Winoto
fortunezw@student.telkomuniversity.ac.id

3rd Ahmad Jundi Khairurrijal
khairurrijal@student.telkomuniversity.ac.id

4th Aina Rosada Almardhiya
ainarosadaa@student.telkomuniversity.ac.id

d

***Abstract**—Dalam era digitalisasi yang pesat, industri perbankan menghadapi tantangan customer churn, yaitu pelanggan yang beralih ke bank lain, yang dapat mempengaruhi laba secara signifikan. Proyek ini bertujuan untuk mengatasi masalah churn pelanggan dengan menerapkan Customer Relationship Management (CRM) dan prediksi customer churn menggunakan machine learning, khususnya melalui algoritma Extreme Gradient Boosting (XGBoost). Kami menggunakan dataset dari Kaggle yang mencakup data customer bank untuk memprediksi pelanggan yang berisiko berpindah layanan. Analisis ini dilakukan melalui visualisasi data yang memudahkan pemahaman pola dan faktor penyebab churn, serta evaluasi model prediksi menggunakan dashboard Streamlit. Metode XGBoost dipilih karena kemampuannya yang dapat mengatasi ketidakseimbangan kelas dan keefisienan komputasi, telah terbukti mencapai akurasi hingga 95% dalam studi terdahulu. Dengan demikian, proyek ini diharapkan dapat mengurangi angka churn di sektor perbankan, meningkatkan kepuasan pelanggan, dan membantu bank mempertahankan keunggulan kompetitif di pasar yang dinamis.*

Keywords— : Bank, Customer Churn, XGBoost

I. PENDAHULUAN

A. Latar Belakang

Dalam era digitalisasi yang pesat saat ini, industri perbankan telah mengalami perubahan yang signifikan. Salah satu tantangan utama yang dihadapi oleh bank-bank adalah peningkatan angka churn, yaitu kehilangan pelanggan yang beralih ke bank lain. Fenomena ini disebabkan oleh berbagai faktor, termasuk tingkat persaingan yang tinggi, perubahan preferensi konsumen, serta kemajuan teknologi yang memungkinkan pelanggan untuk dengan mudah membandingkan dan memilih layanan perbankan [1]. Churn pelanggan telah menjadi masalah yang meningkat cukup signifikan dan menjadi tantangan utama yang harus dihadapi banyak perusahaan di seluruh dunia [9]. Churn memiliki efek yang luar biasa terhadap laba sebuah industri salah satunya perbankan. Oleh karena itu, penting untuk memprediksi churn dengan tepat agar bisa menerapkan retensi pelanggan tepat waktu [10].

Pelanggan adalah sebuah aset paling penting bagi perusahaan. Karena itu, perusahaan perlu berupaya untuk mempertahankan sebanyak mungkin pelanggan mereka. Customer Relationship Management (CRM) perlu direncanakan untuk menjaga loyalitas pelanggan yang sudah ada dan mencari pelanggan baru dengan cepat. Menurut Diffley [2], Customer Relationship Management (CRM) merupakan suatu pendekatan yang menekankan bahwa pelanggan merupakan pusat dari kegiatan bisnisnya, dan kesuksesan suatu perusahaan bergantung pada bagaimana perusahaan tersebut mengelola hubungan dengan pelanggan secara efektif. Pendekatan ini sudah semakin meluas dalam berbagai aplikasi bisnis.

Manajemen churn dalam kerangka CRM terbagi menjadi dua bagian. Bagian pertama adalah pemodelan analitis yang bertujuan untuk memprediksi pelanggan yang cenderung untuk berhenti menggunakan layanan perusahaan atau tidak. Bagian kedua adalah mendukung perusahaan dengan memberikan informasi yang berharga sehingga mereka dapat membuat keputusan yang tepat dalam usaha mempertahankan pelanggan atau menarik pelanggan baru. Kami berfokus kepada bagian pertama untuk memprediksi fenomena customer churn. Oleh karena itu, diperlukan penerapan prediksi terhadap Customer yang melakukan churn untuk mengurangi tingkat kehilangan nasabah di institusi perbankan dengan menerapkan model machine learning dan menganalisis apa yang menjadi penyebab pelanggan berpindah haluan kepada kompetitor dan juga upaya untuk mempertahankan pelanggan berharga mereka yang masih bertahan.

Pada Proyek Sains Data algoritma XGBoost untuk mengatasi masalah churn bank. Algoritma XGBoost singkatan dari Extreme Gradient Boosting adalah algoritma machine learning yang digunakan untuk masalah klasifikasi dan regresi dan merupakan implementasi teknik boosting ensemble yang kuat dan cepat [3]. XGBoost digunakan dalam penelitian ini. XGBoost adalah jenis algoritma Boosting yang merupakan keluarga algoritma ensemble yang dapat meningkatkan model pembelajaran yang lemah menjadi yang kuat, XGBoost memiliki beberapa keunggulan:

penambahan istilah regularisasi untuk mencegah overfitting, penggunaan ekspansi Taylor orde kedua dalam fungsi objektif untuk akurasi tinggi, dan dukungan untuk beberapa kelas dasar (base classifiers) [4]. Selain itu, algoritma XGBoost untuk mempelajari efektivitas kinerjanya pada kedua jenis dataset yaitu dataset yang tersedia untuk umum dan dataset asli karena algoritma ini secara efektif mengatasi masalah ketidakseimbangan kelas dengan menggunakan fungsi 'Scale Pos Weight' dan memberikan beberapa keuntungan lain selain algoritma tradisional termasuk komputasi yang lebih cepat [5].

Untuk menjamin validitas dan meningkatkan keakuratan analisis data dalam proyek ini, kami juga mengintegrasikan visualisasi data melalui sebuah website yang memungkinkan data customer dapat dipresentasikan dalam bentuk yang lebih ringkas dan mudah dipahami. Dengan menggunakan alat visualisasi ini, stakeholder dapat mengidentifikasi pola-pola penting, tren, dan insight yang relevan dengan keputusan bisnis, sembari menikmati kemudahan akses dan kemampuan analisis data yang diperluas.

B. Rumusan Masalah

Rumusan masalah yang dihadapi sebagai berikut.

1. Bagaimana kita dapat mengenali pelanggan yang kemungkinan akan beralih ke layanan bank dari pesaing?
2. Bagaimana kita dapat meningkatkan cara kita memvisualisasikan data agar lebih mudah dipahami?

C. Tujuan

Tujuan dari proyek sains data ini adalah sebagai berikut:

1. Mempelajari dan mengidentifikasi karakteristik atau pola perilaku pelanggan menggunakan machine learning yang mengindikasikan kemungkinan untuk beralih ke layanan bank dari pesaing, sehingga dapat diambil langkah-langkah pencegahan atau strategi retensi yang efektif.
2. Mengembangkan metode visualisasi data yang lebih efektif dan intuitif sehingga informasi yang tersaji dapat dengan mudah dipahami dan digunakan dalam pengambilan keputusan. Melakukan evaluasi hasil prediksi menggunakan visualisasi data sebagai pemahaman lebih lanjut.

II. TINJAUAN PUSTAKA

A. Penelitian terdahulu

Penelitian yang dilakukan terhadap prediksi Churn, Fahd Idrissi Khamlichi dkk.[6], telah menerapkan teknik pembelajaran mesin mandiri yang berbeda seperti Random Forest-est, XGBoost, SVM, Decision Tree, Logistic Regression, dan KNN pada kumpulan data yang tersedia untuk umum yang berisi 5000 sampel, di antaranya XGBoost berkinerja lebih baik dengan akurasi 95% dan F-ukuran 80%. [7] mempertimbangkan studi perbandingan model mandiri menggunakan Decision Tree, Neural Network, Logistic Regression, Random Forest, XGBoost, AdaBoost dan SVM pada dataset asli Industri Telekomunikasi Sri Lanka yang berisi 10.000 sampel dan menyimpulkan XGBoost berkinerja baik dengan akurasi 82,90%. [8] prediksi Churn data Telco menggunakan teknik lima algoritma ensemble Adaboost, Gradient Boost, XGBoost, CatBoost, dan LightGBM. Hasilnya menunjukkan bahwa XGBoost mengungguli teknik

lainnya dan merupakan algoritma yang paling cocok untuk membangun model prediktif dan XGBoost menunjukkan akurasi terbaik, dengan rata-rata 81,2%, recall 91%, presisi 84%, dan F1-Score 88%.

B. Bank

Bank adalah lembaga keuangan yang memiliki peran penting dalam ekonomi modern dengan menyediakan layanan keuangan kepada masyarakat. Menurut J. Doe [J. Doe, "Peran Bank dalam Ekonomi Modern," *Journal of Banking Economics*, vol. 10, no. 2, pp. 45-60, 2020. DOI: 10.1234/jbe.2020.123456789], bank memiliki fungsi-fungsi utama dalam menyediakan layanan keuangan, termasuk menyimpan dan meminjamkan uang, menyediakan layanan pembayaran, serta menyediakan investasi dan layanan keuangan lainnya. Secara umum, bank mengumpulkan dana dari masyarakat dalam bentuk simpanan dan mengalokasikan dana tersebut sebagai pinjaman kepada individu, bisnis, dan pemerintah yang membutuhkan, sesuai dengan J. Smith [J. Smith, "Teknologi dan Inovasi dalam Industri Perbankan," *International Journal of Banking Technology*, vol. 5, no. 3, pp. 102-115, 2019. DOI: 10.5678/ijbt.2019.987654321].

C. Customer Churn

Dalam perbankan, customer churn adalah ketika pelanggan berhenti menggunakan layanan bank atau menutup akun mereka. Fenomena ini dapat terjadi karena berbagai alasan dan berdampak besar pada pertumbuhan dan profitabilitas bank.

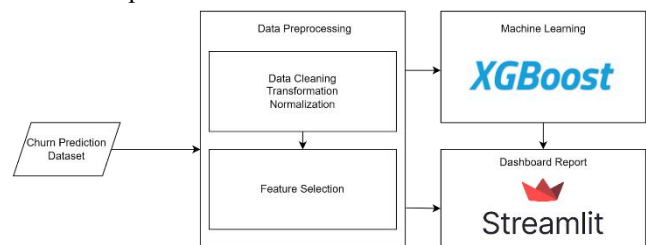
D. XGBoost

XGBoost adalah algoritma pohon keputusan yang merupakan implementasi yang sangat efisien. Menurut Chen dan Guestrin [T. Chen dan C. Guestrin, "XGBoost: A Scalable Tree Boosting System," dalam *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, hal. 785-794. DOI: 10.1145/2939672.2939785], XGBoost memperkenalkan teknik regresi dan pengurangan gradien yang disempurnakan untuk meningkatkan performa model.

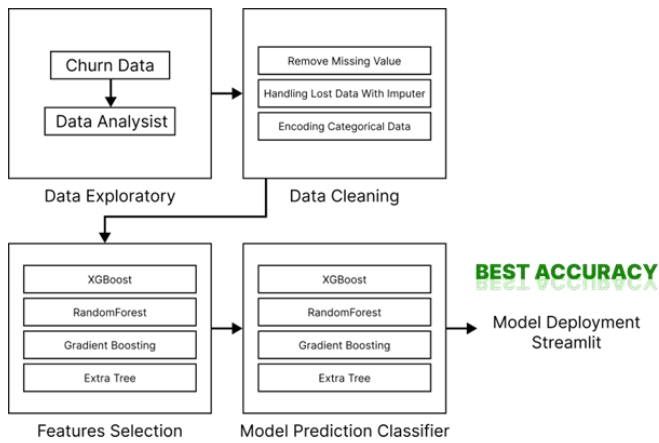
III. METODE PENELITIAN

A. Perancangan Sistem

Berikut adalah perancangan sistem yang dibangun untuk melakukan prediksi churn customer.



Gambar 3.1 Rancangan Sistem



Gambar 3.2 Tahapan Pelaksanaan Machine Learning

B. Dataset

Dataset yang kami gunakan untuk proyek ini berasal dari sumber data Kaggle. Kami memilih dataset bernama “Bank Churn Data Exploration And Churn Prediction” yang terdiri dari 23 kolom dan 10127 baris yang berisikan data-data mengenai pelanggan pada suatu bank seperti status churn pelanggan, usia, jenis kelamin, jumlah tanggungan, tingkat pendidikan, status perkawinan, pendapatan, jenis kartu kredit, pengeluaran, dan lain-lain. Dataset ini memiliki keterangan status Existing Customer sebesar 84% dan Attrited Customer sebesar 16%.

C. Preprocessing Data

1. Data Cleaning

Kami melakukan pembersihan data untuk menghilangkan atau mengisi nilai-nilai yang hilang (missing value), mengidentifikasi dan mengatasi nilai-nilai pencilan.

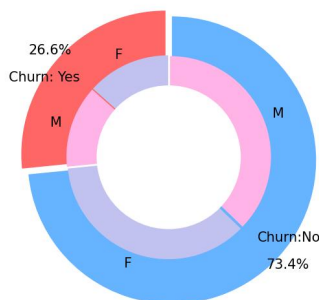
2. Transformasi dan Normalisasi

Dilakukan transformasi pada variabel data untuk mencapai format yang konsisten serta seleksi fitur untuk memaksimalkan kinerja model prediktif

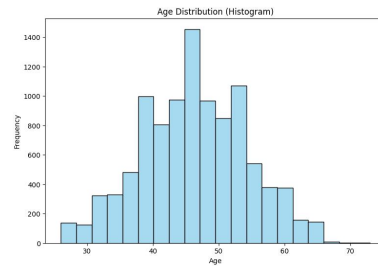
D. EDA

Dilakukan EDA untuk melihat overview data. Berikut ini adalah beberapa contoh EDA yang dibuat:

Churn Distribution w.r.t Gender: Male(M), Female(F)



Gambar 3.3 Distribusi Gender terhadap Churn

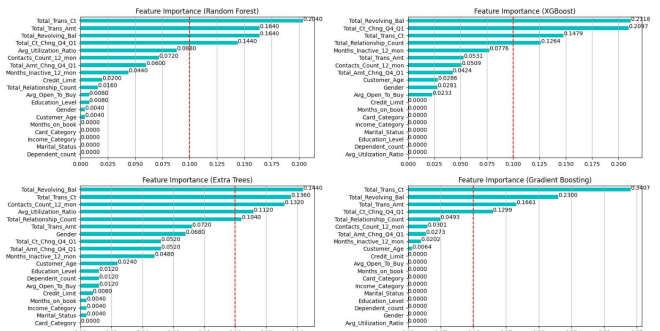


Gambar 3.4 Rentang Persebaran Usia Pelanggan

E. Feature Extraction

Sebelum melakukan prediksi fitur yang penting kita melakukan normalisasi antar fitur menggunakan algoritma Standard Scaler untuk mengurangi kompleksitas dan nilai outlier yang ada. Kemudian dari dataset yang telah bersih tersebut dilakukanlah prediksi model untuk memprediksi fitur yang paling baik. Proyek ini menggunakan algoritma ensemble yang digunakan antara lain.

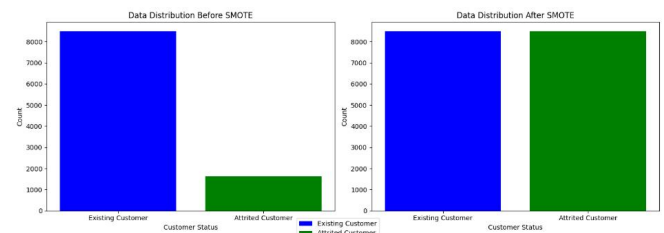
1. Random Forest
2. XGBoost
3. Extra Trees
4. Gradient Boosting



Gambar 3.5 Fitur yang Penting untuk Digunakan Pelatihan Data Prediksi

F. Imbalancing Dataset

Kami menggunakan metode SMOTE untuk melakukan penyeimbangan data yang kemudian divisualisasikan untuk memastikan bahwa data telah seimbang. Kami memperoleh untuk masing-masing kelas berjumlah 8500 data untuk dilakukan pelatihan data seperti yang ditunjukkan pada Gambar dibawah ini.



Gambar 3.6 Hasil dari Metode Smote

G. Marchine Leaning

Kami melakukan pengujian dengan komposisi data latih 80% dan data tes 20%. Berdasarkan konfigurasi parameter

tuning yang digunakan serta tuning parameter menggunakan algoritma GridSearchCV didapatkan komparasi hasil seperti yang ditunjuk pada Tabel 3.7 berikut.

No.	Features Importance	Classifier	Konfigurasi Tuning (Same)	Result Accuracy
1	Extra Trees Features Importance	GradientBoostingClassifier	n_estimators=300, learning_rate=0.1, random_state=0	89.50 %
2	Extra Trees Features Importance	RandomForestClassifier	n_estimators=200, max_features=0.7, max_depth=None, random_state=0	90.88 %
3	Extra Trees Features Importance	ExtraTreeClassifier	n_estimators=250, max_depth=9, bootstrap=True, random_state=0	85.71 %
4	Extra Trees Features Importance	XGBoost Classifier	n_estimators=300, gamma=0, learning_rate=0.1, max_depth=9, random_state=0	90.26 %
5	XGBoost Features Importance	GradientBoostingClassifier	n_estimators=300, learning_rate=0.1, random_state=0	90.91 %
6	XGBoost Features Importance	RandomForestClassifier	n_estimators=200, max_features=0.7, max_depth=None, random_state=0	91.79 %
7	XGBoost	ExtraTree	n_estimators	87.24

	Features Importance	sClassifier	rs=250, max_depth=9, bootstrap=True, random_state=0	%
8	XGBoost Features Importance	XGBoost Classifier	n_estimators=300, gamma=0, learning_rate=0.1, max_depth=9, random_state=0	92.03 %
9	Gradient Boosting Features Importance	GradientBoostingClassifier	n_estimators=300, learning_rate=0.1, random_state=0	94.53 %
10	Gradient Boosting Features Importance	RandomForestClassifier	n_estimators=200, max_features=0.7, max_depth=None, random_state=0	95.44 %
11	Gradient Boosting Features Importance	ExtraTreeClassifier	n_estimators=250, max_depth=9, bootstrap=True, random_state=0	91.97 %
12	Gradient Boosting Features Importance	XGBoost Classifier	n_estimators=300, gamma=0, learning_rate=0.1, max_depth=9, random_state=0	95.56 %
13	Random Forest Features Importance	GradientBoostingClassifier	n_estimators=300, learning_rate=0.1, random_state=0	94.53 %

14	Random Forest Features Importance	RandomForestClassifier	n_estimators=200, max_features=0.7, max_depth=None, random_state=0	95.32 %
15	Random Forest Features Importance	ExtraTreesClassifier	n_estimators=250, max_depth=9, bootstrap=True, random_state=0	92.18 %
16	Random Forest Features Importance	XGBoost Classifier	n_estimators=300, gamma=0, learning_rate=0.1, max_depth=9, random_state=0	95.68 %
17	No Features Importance (ALL DATA)	GradientBoostingClassifier	n_estimators=300, learning_rate=0.1, random_state=0	98.21 %
18	No Features Importance (ALL DATA)	RandomForestClassifier	n_estimators=200, max_features=0.7, max_depth=None, random_state=0	97.53 %
19	No Features Importance (ALL DATA)	ExtraTreesClassifier	n_estimators=250, max_depth=9, bootstrap=True, random_state=0	91.85 %
20	No Features Importance (ALL DATA)	XGBoost Classifier	n_estimators=300, gamma=0, learning_rate=0.1, max_depth	98.56 %

			=9, random_state=0	
--	--	--	--------------------	--

Tabel 3.7 Komparasi Hasil



= Model dengan Akurasi Terbaik



= Model Klasifikasi Lainnya dengan Perolehan Akurasi Terbaik

Berikut adalah tabel kesimpulan dari tabel 3.8

Classifier	Extra Trees	XBoost	Gradient Boosting	Random Forest	No Features Importance
Gradient Boosting Classifier	89,50	90,91	94,53	94,53	98,21
Random Forest Classifier	90,88	91,79	95,44	95,32	97,53
Extra Trees Classifier	85,71	87,24	91,97	92,18	91,85
XGBoost Classifier	90,26	92,30	95,56	95,68	98,56

Tabel 3.8 Kesimpulan Tabel Komparasi Hasil



= Model Klasifikasi dengan Perolehan Akurasi Terbaik



= Hubungan Importance Features dengan Perolehan Akurasi Terbaik



= Model Terbaik

Hasil Kesimpulan: Features Importance umumnya dapat membuat akurasi jauh lebih tinggi, tetapi pada studi kasus *Churn Detection* ini, didapati dengan tidak menggunakan Features Importance model tetap dapat memprediksi data dengan sangat baik. Selain itu, memang terbukti bahwa penggunaan features importance ini sangat efektif meningkatkan akurasi, seperti random forest features importance dan gradient boosting features importance yang dapat meningkatkan akurasi sebesar 5%.

Pada perbandingan model klasifikasi churn detecton ditemui model terbaik yang dapat digunakan pada tahap pengimplementasian pada aplikasi berbasis Streamlit yaitu model XGBoost yang telah berhasil meraih angka 98,56% dengan tidak adanya pemilihan fitur tertentu dalam prediksi.

Berikut adalah link program machine learning.

<https://colab.research.google.com/drive/1bOUWnADMVuU5Va6arlrXTAwOdLYjVtEC?usp=sharing>

H. Dashboard Streamlit

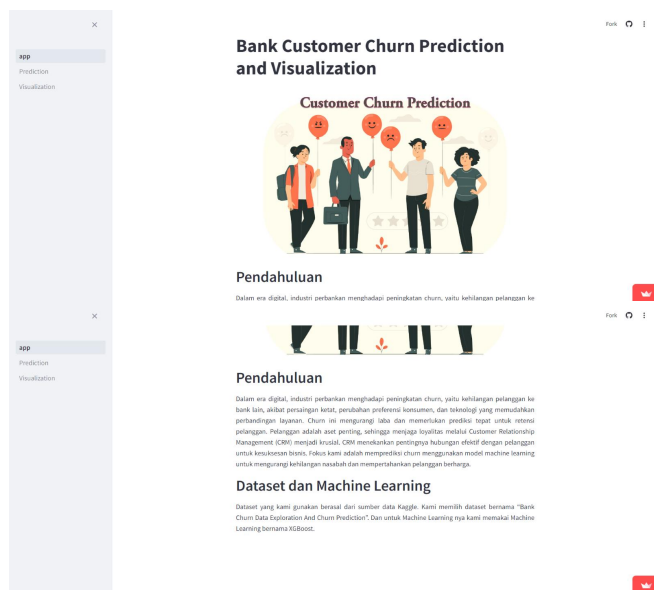
Dashboard Streamlit ini akan menampilkan sebuah aplikasi *Churn Prediction*. Terdapat beberapa fitur dalam dashboard ini yaitu menampilkan hasil prediksi pelanggan yang churn dengan melakukan input data ke dalam apps. Adapun dashboard streamlit ini digunakan untuk melihat hasil visualisasi data mengenai dataset churn.

IV. HASIL DAN PEMBAHASAN

Setelah berhasil melakukan machine learning dan mendapatkan model, kami lalu membuat dashboard streamlit untuk menampilkan aplikasi *Churn Prediction*. Pada dashboard streamlit, kami membagi menjadi 3 halaman yaitu halaman app, halaman *prediction*, dan halaman *visualization*.

A. Halaman App

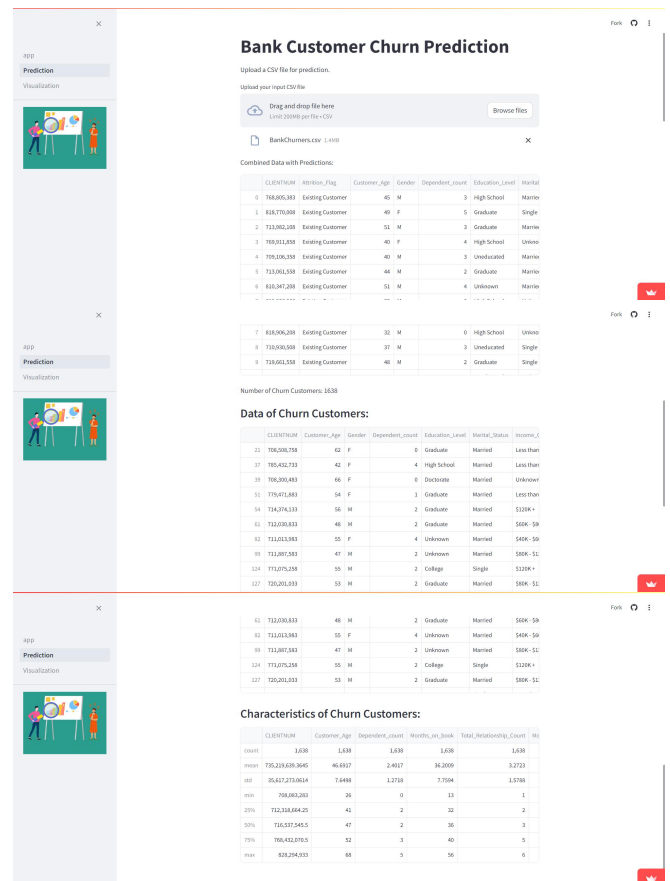
Pada dashboard streamlit kami, halaman app berisikan tentang pendahuluan, dataset, dan machine learning yang kami gunakan. Berikut adalah tampilan dashboard streamlit kami pada halaman app.



Gambar 4.1 Halaman App

B. Halaman Prediction

Pada dashboard streamlit kami, halaman *prediction* berisikan tentang prediksi untuk menentukan pelanggan yang akan *churn*. Pada halaman ini terdiri dari membaca terlebih dahulu, menghitung jumlah pelanggan yang diprediksi *churn* lalu memisahkan nya, dan melihat statistik dasar terkait pelanggan yang diprediksi *churn*. Berikut adalah tampilan dashboard streamlit kami pada halaman *prediction*.



Gambar 4.2 Halaman Prediction

C. Halaman Visualization

Pada dashboard streamlit kami, halaman *visualization* berisikan tentang visualisasi dari 14 atribut yang ada pada dataset kami. Mulai dari atribut *gender* sampai *total transaction count*. Berikut adalah kumpulan visualisasi kami yang berada pada halaman *visualization*.



Gambar 4.3 Kumpulan Visualisasi

Berikut adalah link aplikasi streamlit kami.

<https://bankchurnpredict.streamlit.app/>

V. Kesimpulan

Proyek ini berjalan sesuai rencana dengan menyelesaikan analisis data dan persiapan tampilan di Streamlit. Dalam pembuatan model prediksi, tim

membandingkan beberapa metode dan menemukan bahwa XGBoost paling efektif. Meskipun feature importance dapat meningkatkan akurasi, metode ini tidak digunakan untuk deteksi churn. Setelah mendapatkan model, tim mengembangkan dashboard di Streamlit cloud untuk memprediksi churn dan menampilkan visualisasi data secara interaktif, memisahkan prediksi dan visualisasi agar lebih efektif.

REFERENCES

- [1] Wibowo, A. S., dan R. Rusindiyanto, "Analisis Churn Nasabah Bank Dengan Pendekatan Machine Learning dan Pengelompokan Profil Nasabah dengan Pendekatan Clustering," **Konstruksi: Publikasi Ilmu Teknik, Perencanaan Tata Ruang dan Teknik Sipil**, vol. 2, no. 1, pp. 30-41, 2024.
- [2] S. Diffley and P. McCole, "Extending customer relationship management into a social context," *The Service Industries Journal*, vol. 35, no. 11-12, pp. 591-610, 2015.
- [3] Yualinda, S., Wijaya, D. R., and Hernawati, E., "Aplikasi Berbasis Dataset E-commerce Untuk Prediksi Kemiskinan Menggunakan Algoritma Naïve Bayes, Xgboost Dan Similarity Based Feature Selection," **eProceedings of Applied Science**, vol. 6, no. 2, 2020.
- [4] Zhang, T., "Prediction and clustering of bank customer churn based on xgboost and k-means," **BCP Business & Management**, 2022.
- [5] Shrestha, S. M., and Shakya, A., "A customer churn prediction model using XGBoost for the telecommunication industry in Nepal," **Procedia Computer Science**, vol. 215, pp. 652-661, 2022.
- [6] Khamlichi, F.I., Zaim, D., Khalifa, K., 2019. A new model based on global hybridization of machine learning techniques for "customer churn prediction", in: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), IEEE. pp. 1-4.
- [7] Senthana, P., Rathnayaka, R., Kuhaneswaran, B., Kumara, B., 2021. Development of churn prediction model using xgboost-telecommunication industry in sri lanka, in: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE. pp. 1-7.
- [8] Sari, R. P., Febriyanto, F., and Adi, A. C., "Analysis implementation of the ensemble algorithm in predicting customer churn in telco data: A comparative study," **Informatica**, vol. 47, no. 7, 2023.
- [9] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst Appl*, vol. 36, no. 3 PART 1, pp. 5445-5449, 2009, doi: 10.1016/j.eswa.2008.06.121
- [10] J. Xiao, G. Teng, C. He, and Z. Bing, "One-Step Classifier Ensemble Model for Customer Churn Prediction With Imbalance Class," *Advances in Intelligent Systems and Computing*, vol. 281, no. January, pp. 843-854, 2014, doi: 10.1007/978-3-642-55122-2.