DigitalSkola

Week 12

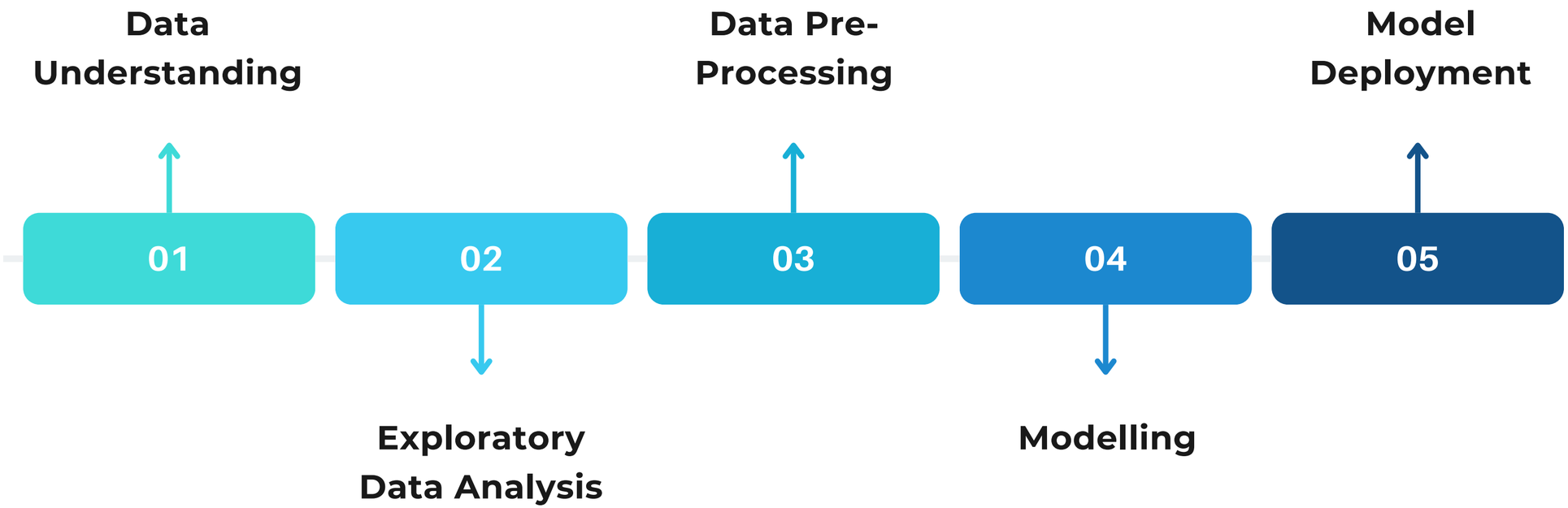
FINAL PROJECT

Data Science Bootcamp Batch 32

Materi:

Customer Churn Prediction

CONTENTS


Data Understanding
01
Exploratory Data Analysis
02
Data Pre-Processing
03
Modelling
04
Model Deployment
05

Deskripsi Dataset

Dataset Churn Risk

Data behaviour dari masing-masing customer yang churn/tidak churn, yang direpresentasikan dalam bentuk rate.
Semakin besar nilai rate churn maka semakin besar kemungkinan customer tersebut akan meninggalkan menggunakan layanan.

Tujuan Project:

Membuat model prediksi machine learning yang paling efektif untuk mengetahui apakah customer akan churn/no churn berdasarkan data behaviour customer.

Cerita Dataset:

- Dataset ini terdiri dari total 10.000 baris (entries) dengan rentang indeks baris dari 0 hingga 9999, dan terdapat 7 kolom.
- Variabel independen melibatkan Gender, Age, CreditScore, EstimatedSalary, dan HasCrCard, yang berisi informasi tentang pelanggan.
- Variabel dependen adalah Exited, yang menunjukkan apakah pelanggan tersebut telah meninggalkan layanan.

Fitur

- CustomerId: ID unik untuk setiap pelanggan.
- Gender: Jenis kelamin pelanggan (Female / Male).
- Age: Usia pelanggan.
- CreditScore: Skor kredit pelanggan.
- EstimatedSalary: Perkiraan gaji pelanggan.
- HasCrCard: Menunjukkan apakah pelanggan memiliki kartu kredit (1 untuk ya, 0 untuk tidak).
- Exited: Variabel target yang menunjukkan apakah pelanggan telah keluar dari layanan (1 untuk ya, 0 untuk tidak).

DigitalSkola

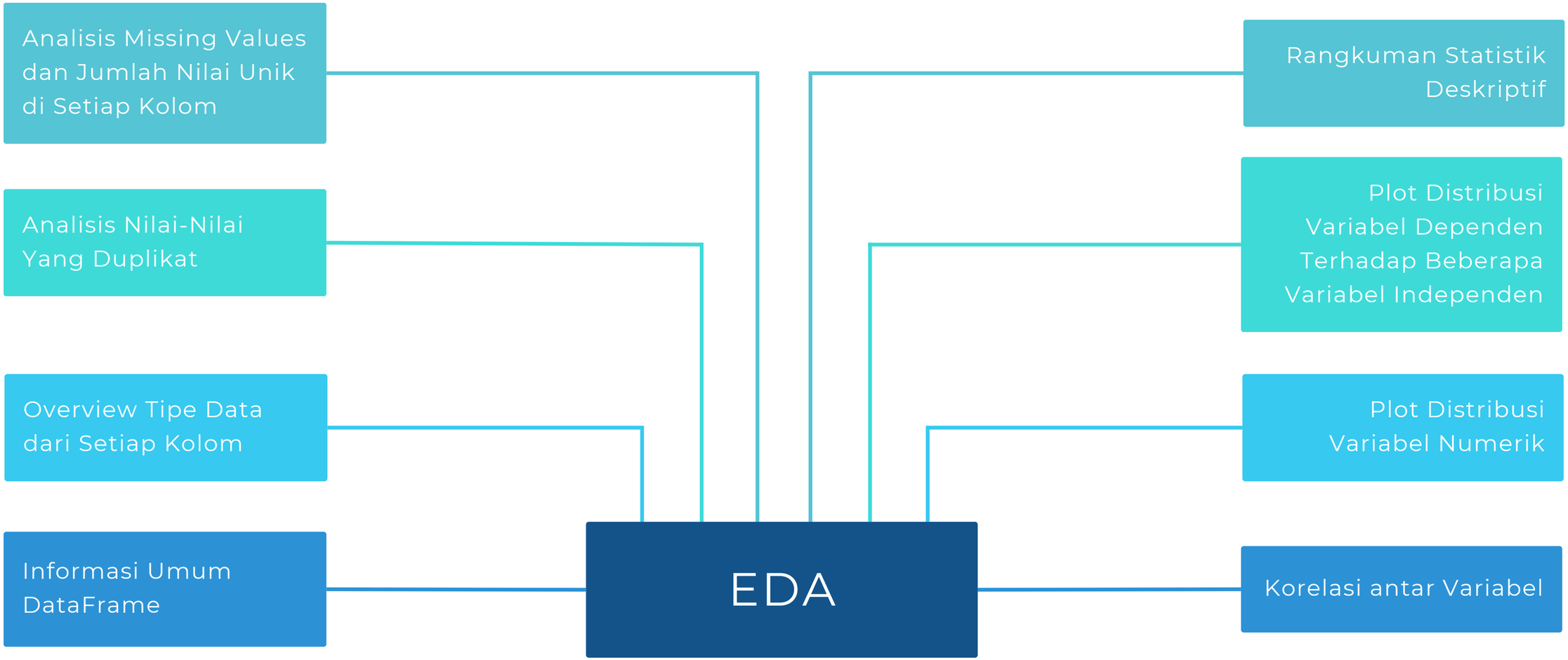
Dataset Overview

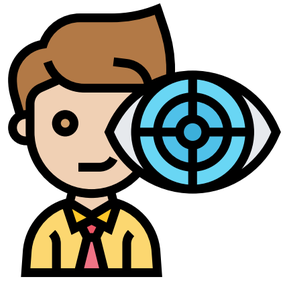
	CustomerId	Gender	Age	CreditScore	EstimatedSalary	HasCrCard	Exited
0	15634602	Female	42	619	101348.88	1	1
1	15647311	Female	41	608	112542.58	0	0
2	15619304	Female	42	502	113931.57	1	1
3	15701354	Female	39	699	93826.63	0	0
4	15737888	Female	43	850	79084.10	1	0

Exploration Data Analysis

Final Project - Algowizard Team

Exploration Data Analysis (EDA)





Exploratory Data Analysis

Info Data Frame

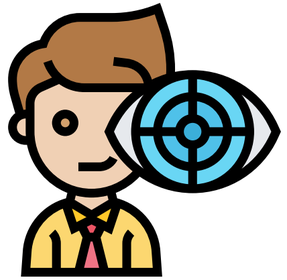
```
Informasi umum mengenai DataFrame:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CustomerId       10000 non-null  int64
 1   Gender           10000 non-null  object
 2   Age              10000 non-null  int64
 3   CreditScore      10000 non-null  int64
 4   EstimatedSalary  10000 non-null  float64
 5   HasCrCard        10000 non-null  int64
 6   Exited           10000 non-null  int64
dtypes: float64(1), int64(5), object(1)
```

Missing Value

```
Jumlah missing values:
CustomerId         0
Gender             0
Age                0
CreditScore        0
EstimatedSalary    0
HasCrCard          0
Exited             0
dtype: int64
```

Unique Value

```
Jumlah nilai unik:
CustomerId         10000
Gender                 2
Age                   70
CreditScore          460
EstimatedSalary     9999
HasCrCard              2
Exited                 2
dtype: int64
```



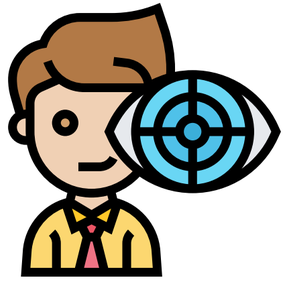
Exploratory Data Analysis

Duplikat

```
Duplikat dalam DataFrame:
Empty DataFrame
Columns: [CustomerId, Gender, Age, CreditScore, EstimatedSalary, HasCrCard, Exited]
Index: []
```

Tipe Data

```
Tipe data dari setiap kolom:
CustomerId           int64
Gender              object
Age                  int64
CreditScore          int64
EstimatedSalary    float64
HasCrCard            int64
Exited               int64
dtype: object
```

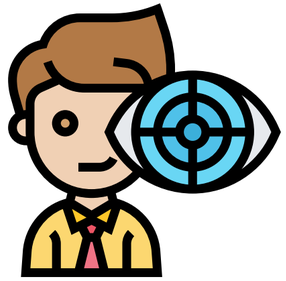


Exploratory Data Analysis

Statistik Deskriptif

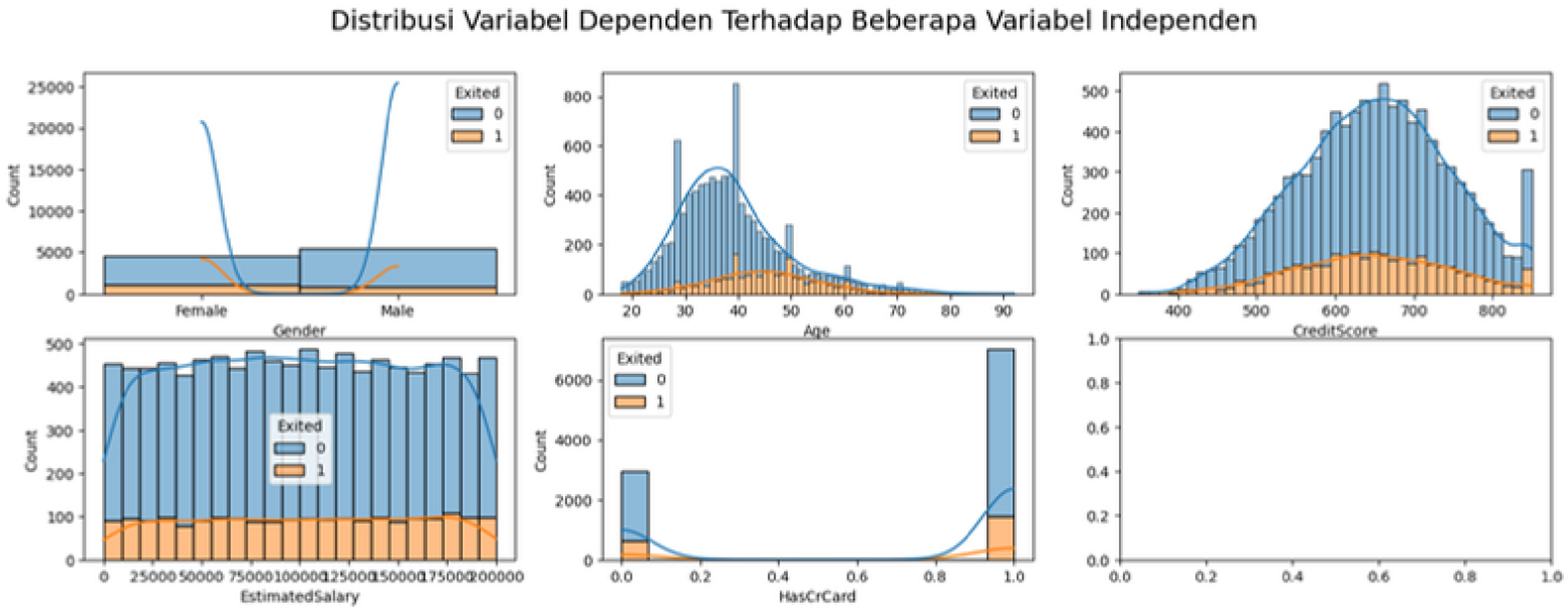
Statistik deskriptif dari kolom numerik DataFrame:

	CustomerId	Age	CreditScore	EstimatedSalary	HasCrCard	Exited
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000
mean	1.569094e+07	38.921800	650.528800	100090.239881	0.70550	0.203700
std	7.193619e+04	10.487806	96.653299	57510.492818	0.45584	0.402769
min	1.556570e+07	18.000000	350.000000	11.580000	0.00000	0.000000
25%	1.562853e+07	32.000000	584.000000	51002.110000	0.00000	0.000000
50%	1.569074e+07	37.000000	652.000000	100193.915000	1.00000	0.000000
75%	1.575323e+07	44.000000	718.000000	149388.247500	1.00000	0.000000
max	1.581569e+07	92.000000	850.000000	199992.480000	1.00000	1.000000



Exploratory Data Analysis

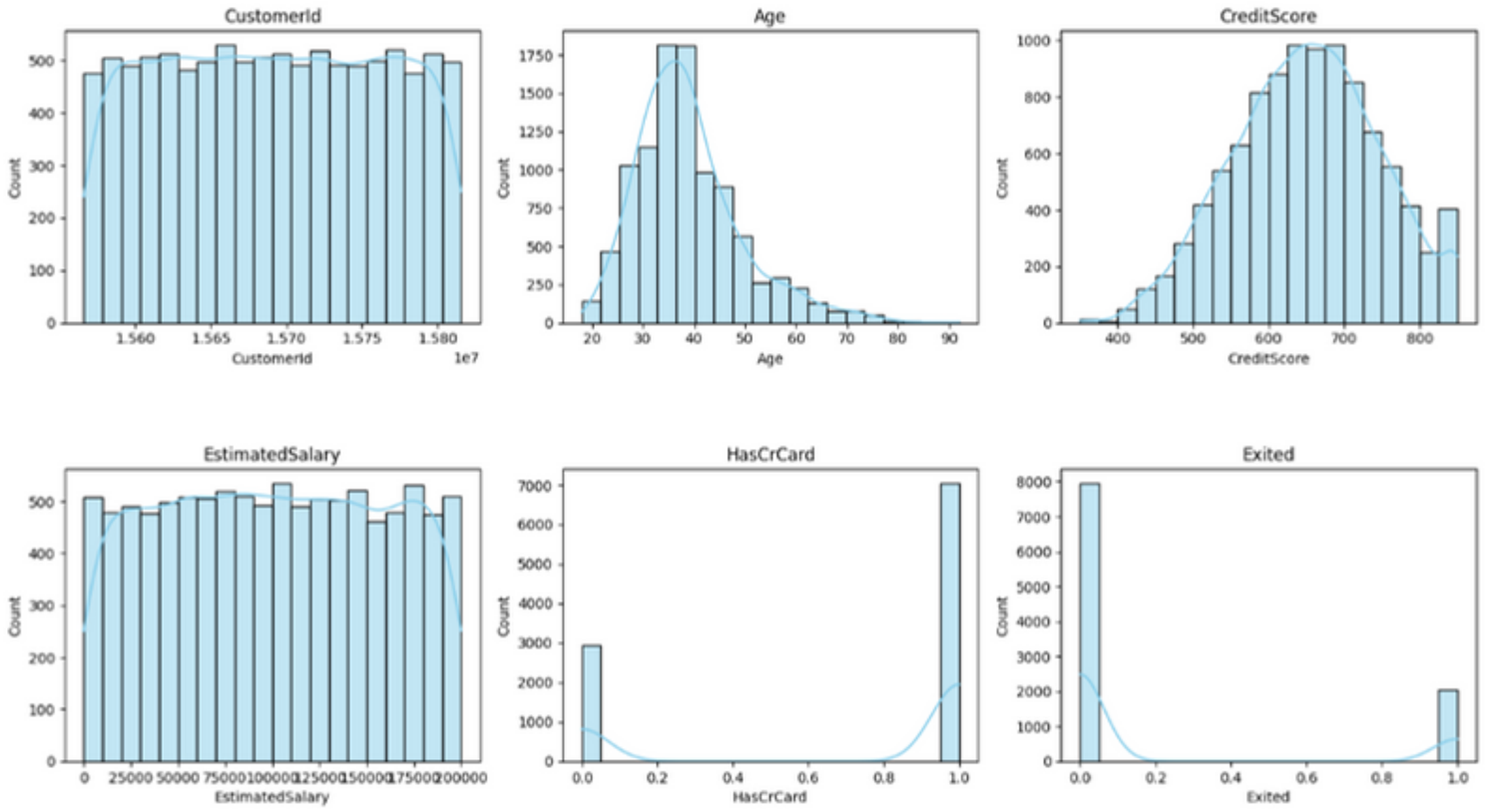
Plot Distribusi Variabel Dependen Terhadap Beberapa Variabel Independen



1. Distribusi variabel dependen 'Exited' terhadap variabel independen 'Gender':
2. Distribusi variabel dependen 'Exited' terhadap variabel independen 'Age'
3. Distribusi variabel dependen 'Exited' terhadap variabel independen 'CreditScore'
4. Distribusi variabel dependen 'Exited' terhadap variabel independen 'EstimatedSalary'
5. Distribusi variabel dependen 'Exited' terhadap variabel independen 'HasCrCard'

Exploratory Data Analysis

Distribusi Variabel Numerik



CustomerId : tipe data numerik selanjutnya akan mejadi tipe data objek

Age : Rentang usia dari '30-50' memiliki jumlah frekuensi "Exited" yang lebih tinggi dibandingkan dengan '20-30', tetapi menurun untuk '50++'.

CreditScore : memiliki bentuk yang mirip dengan distribusi normal, tetapi terdapat peningkatan yang tiba-tiba melonjak di atas nilai 800.

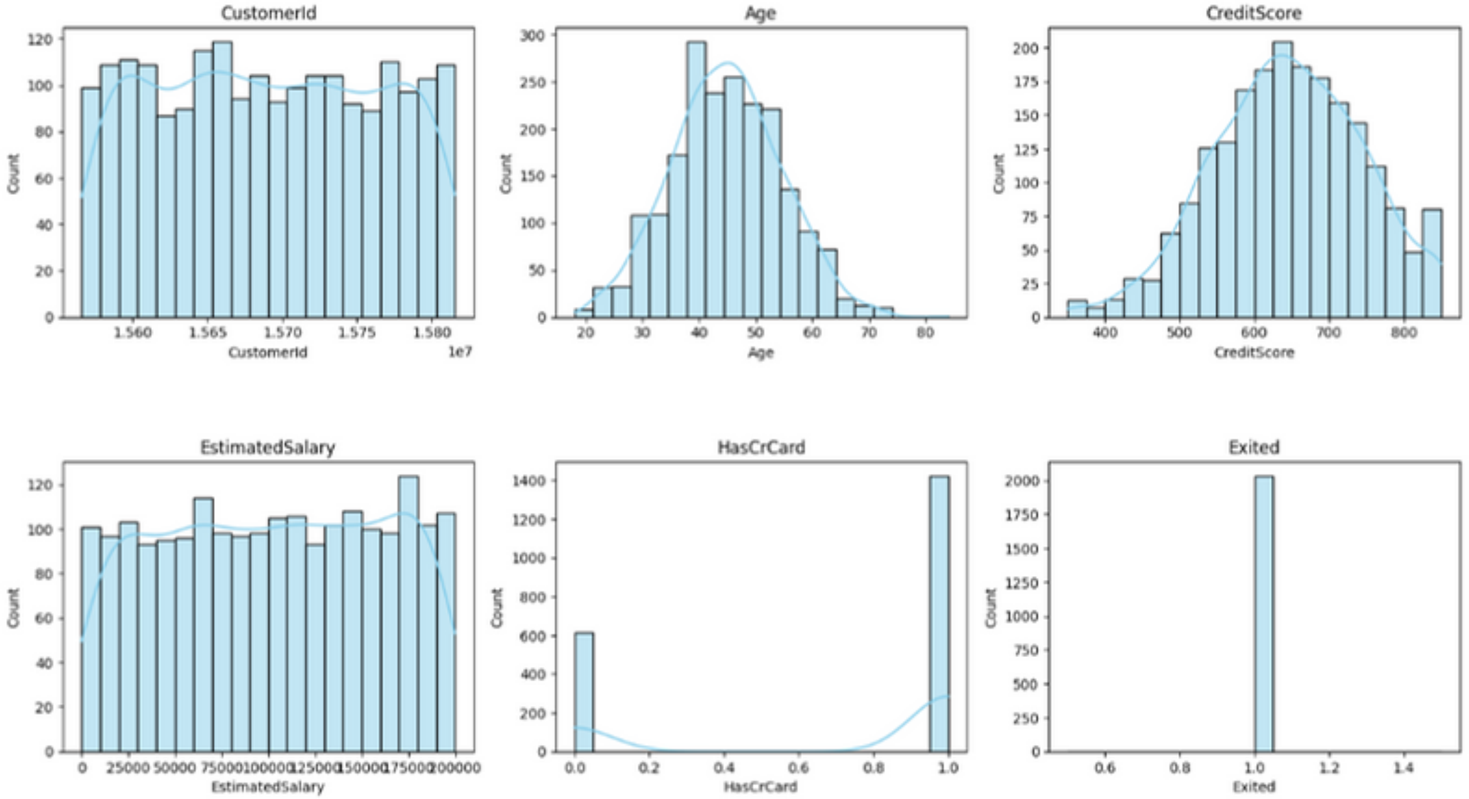
EstimatedSalary : cenderung seragam, dengan jumlah frekuensi yang relatif stabil di setiap bin edge.

HasCrCard : pemegang kartu kredit (HasCrCard=1) memiliki jumlah frekuensi "Exited" yang lebih tinggi dibandingkan dengan yang bukan pemegang kartu kredit (HasCrCard=0).

Exited : mayoritas data adalah 0 sementara label 1 hanya muncul di akhir

Exploratory Data Analysis

Distribusi Variabel Dependen Terhadap Variabel Lainnya



CustomerId : tipe data numerik selanjutnya akan mejadi tipe data objek

Age : bentuk plot yang normal dengan sedikit right-skewed. Mayoritas data berada dalam rentang usia '36-50'.

CreditScore : bentuk plot yang normal, namun sedikit left-skewed.

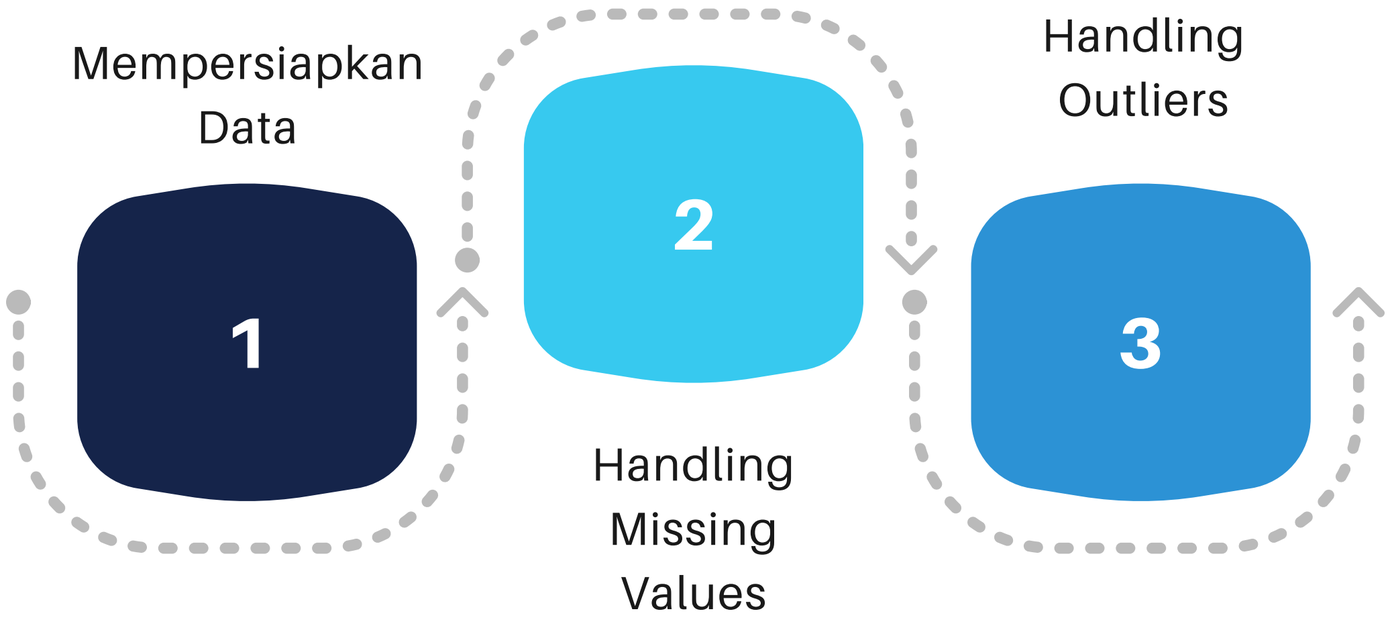
EstimatedSalary : bentuk plot yang seragam, dengan frekuensi yang relatif stabil di setiap bin edge.

HasCrCard : menunjukkan bahwa mayoritas data berada pada kategori pemegang kartu kredit (HasCrCard=1). Pemegang kartu kredit (HasCrCard=1) memiliki jumlah frekuensi "Exited" yang lebih tinggi dibandingkan dengan yang bukan pemegang kartu kredit (HasCrCard=0).

Exited : label 0 tidak ditampilkan tapi fokus ada pada label 1

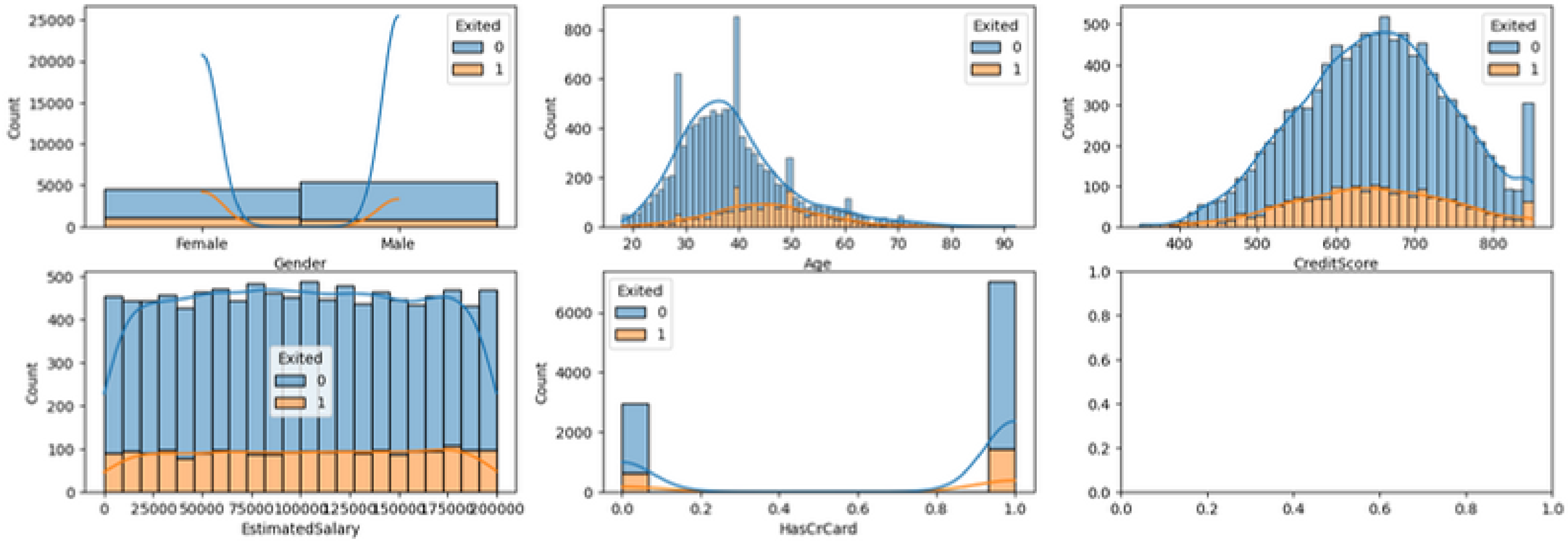
Pra-Pemrosesan Data

Pra-Pemrosesan Data



Mempersiapkan Data

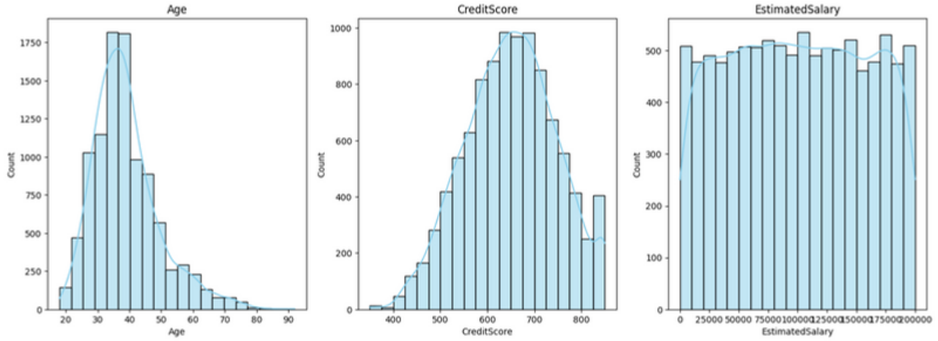
Mengubah Tipe Data Kolom ‘CustomerId’ Menjadi ‘object’

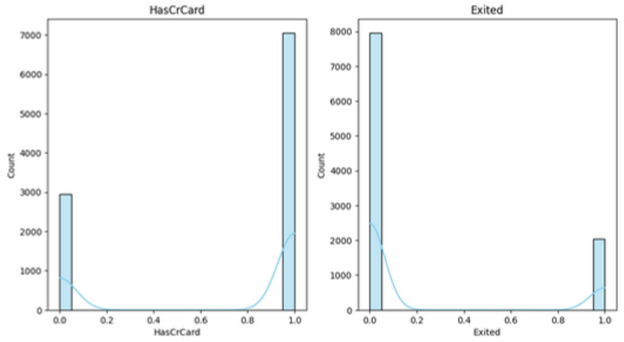


Distribusi Variabel Dependen Terhadap Beberapa Variabel Independen

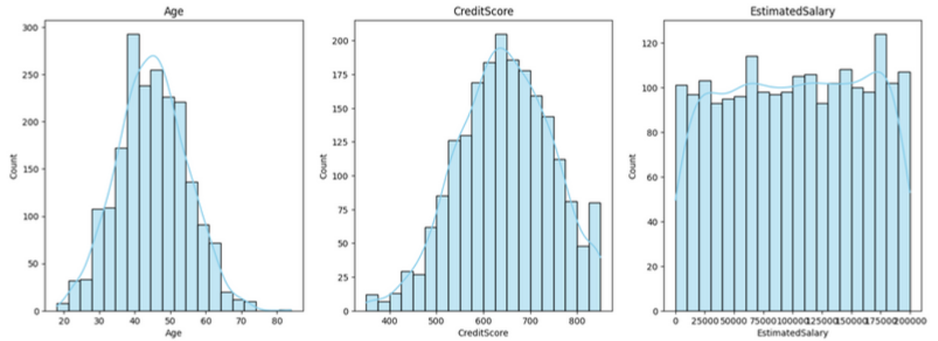
Mempersiapkan Data

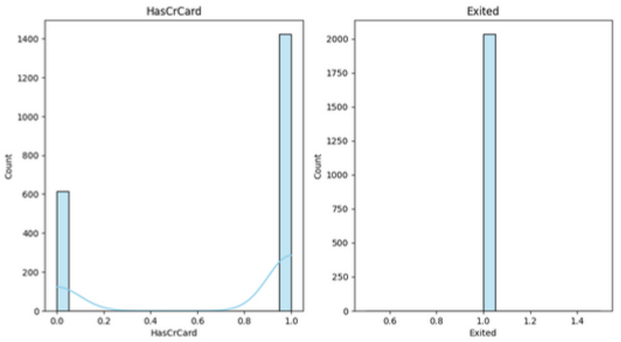
Mengubah Tipe Data Kolom ‘CustomerId’ Menjadi ‘object’





Distribusi Variabel Numerik

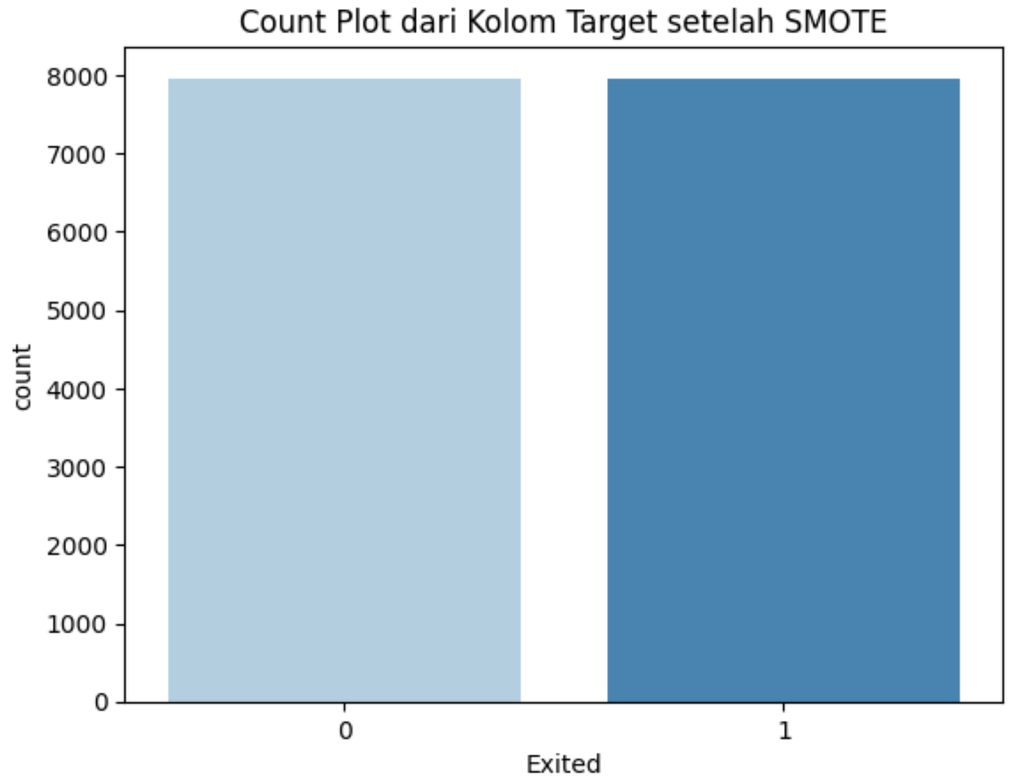




Distribusi Variabel Dependen Terhadap Variabel Lainnya

Mempersiapkan Data

Plot tersebut menunjukkan bahwa setelah menggunakan teknik SMOTE, **distribusi data menjadi lebih seimbang**. Sebelum menggunakan teknik SMOTE, jumlah data untuk nilai 1 lebih sedikit daripada jumlah data untuk nilai 0. **Setelah menggunakan teknik SMOTE, jumlah data untuk kedua nilai tersebut menjadi lebih merata.**



- Nilai 0
  - Sebelum menggunakan teknik SMOTE, jumlah data untuk nilai 0 adalah sekitar 7.000.
  - Setelah menggunakan teknik SMOTE, jumlah data untuk nilai 0 adalah sekitar 8.000.
- Nilai 1
  - Sebelum menggunakan teknik SMOTE, jumlah data untuk nilai 1 adalah sekitar 3.000.
  - Setelah menggunakan teknik SMOTE, jumlah data untuk nilai 1 adalah sekitar 4.000.

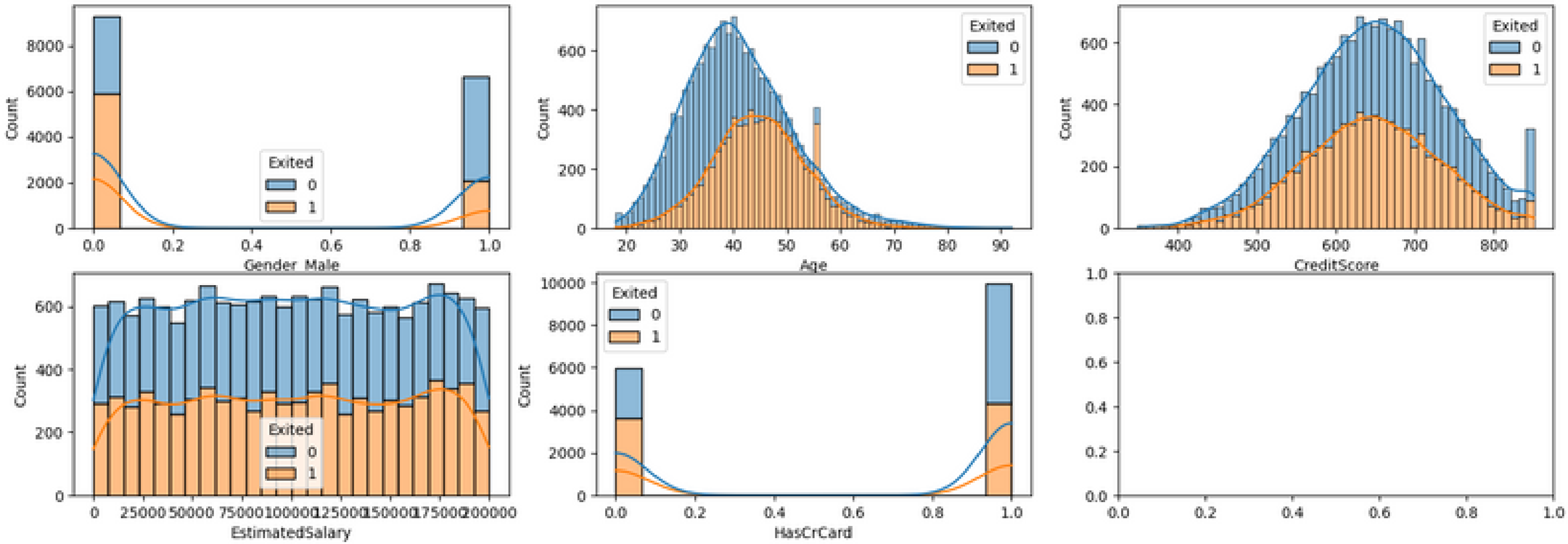
Note:

Label 0 : pelanggan tidak keluar,
Label 1 : pelanggan keluar,

Sumbu x : nilai kolom target
Sumbu y : jumlah data.

Mempersiapkan Data

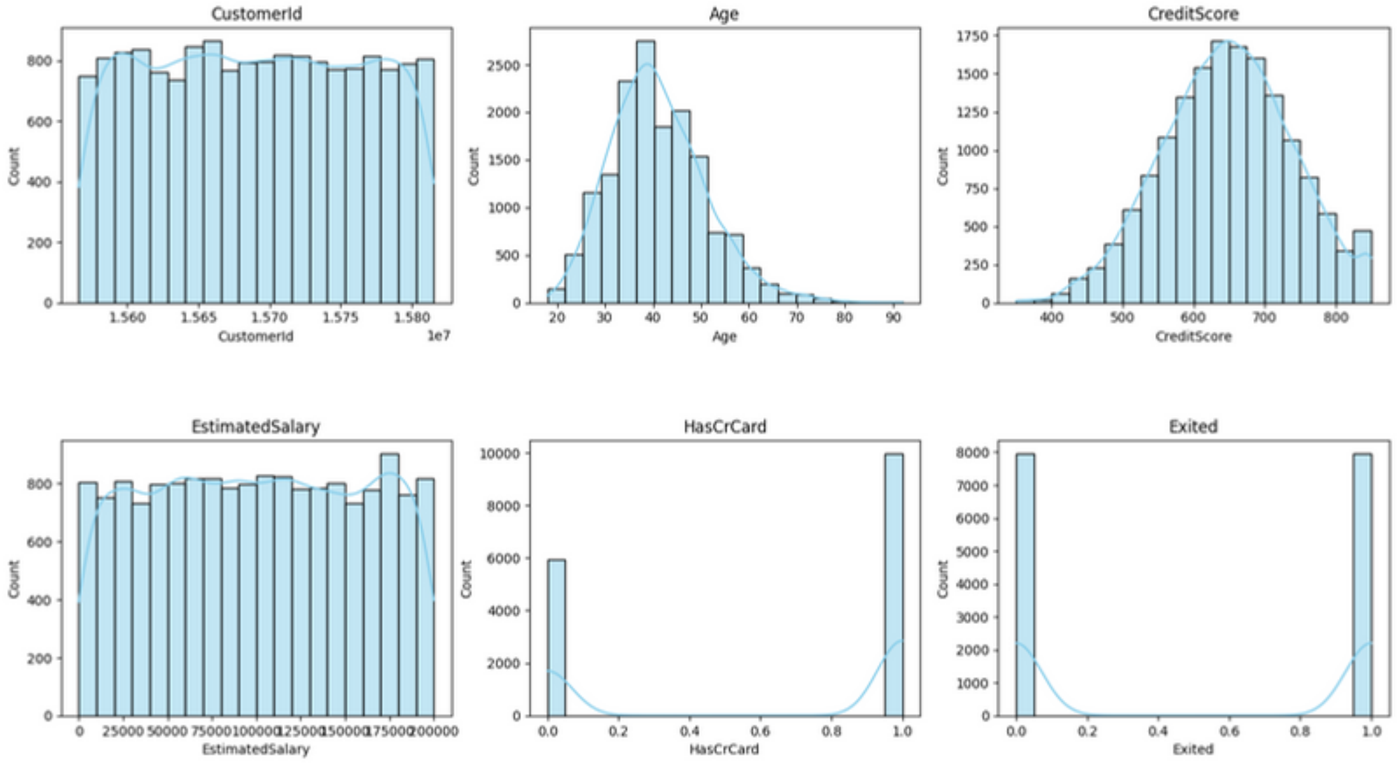
Handling Imbalance Dataset



Distribusi Variabel Dependen Terhadap Beberapa Variabel Independen

Mempersiapkan Data

Handling Imbalance Dataset

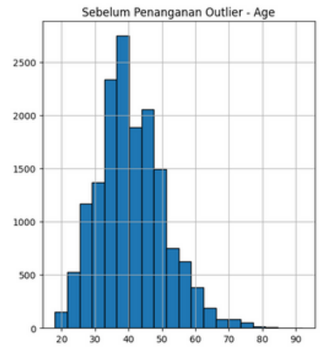


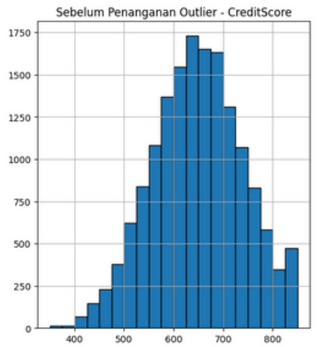
Distribusi Variabel Numerik

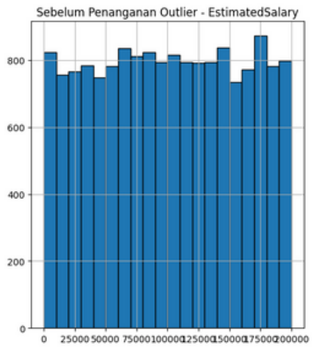
SMOTE berhasil menyeimbangkan jumlah sampel antara kelas Exited (1) dan kelas Not Exited (0), mengurangi ketidakseimbangan kelas yang dapat memengaruhi kinerja model klasifikasi.

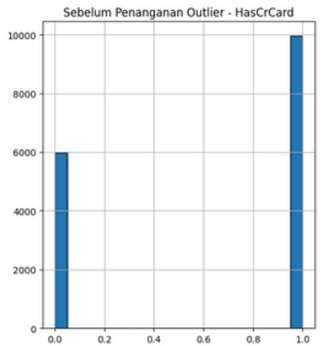
Mempersiapkan Data

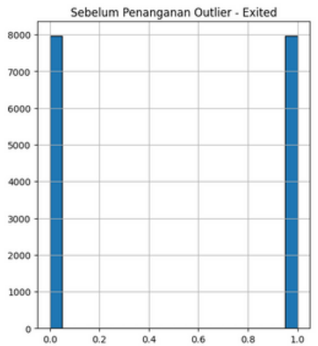
Handling Outliers

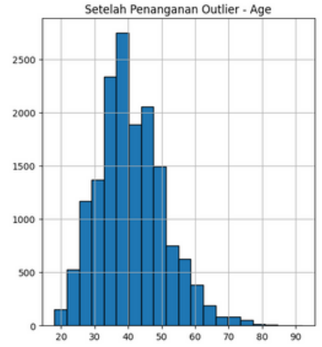


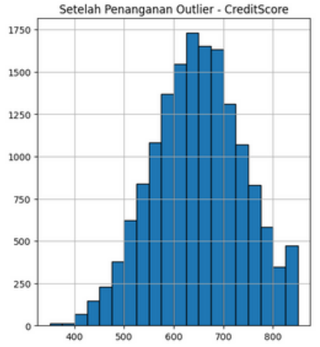


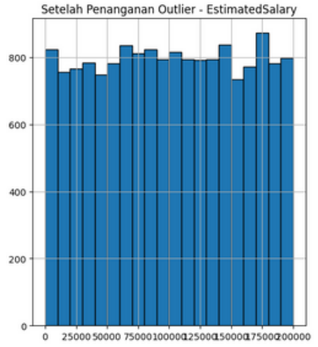


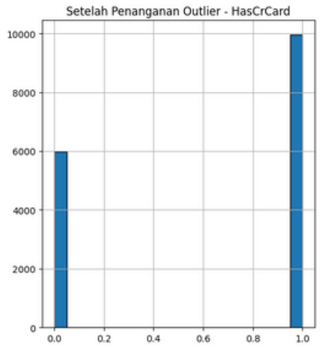


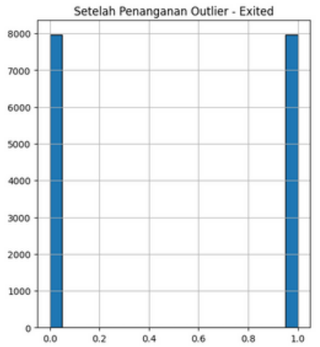




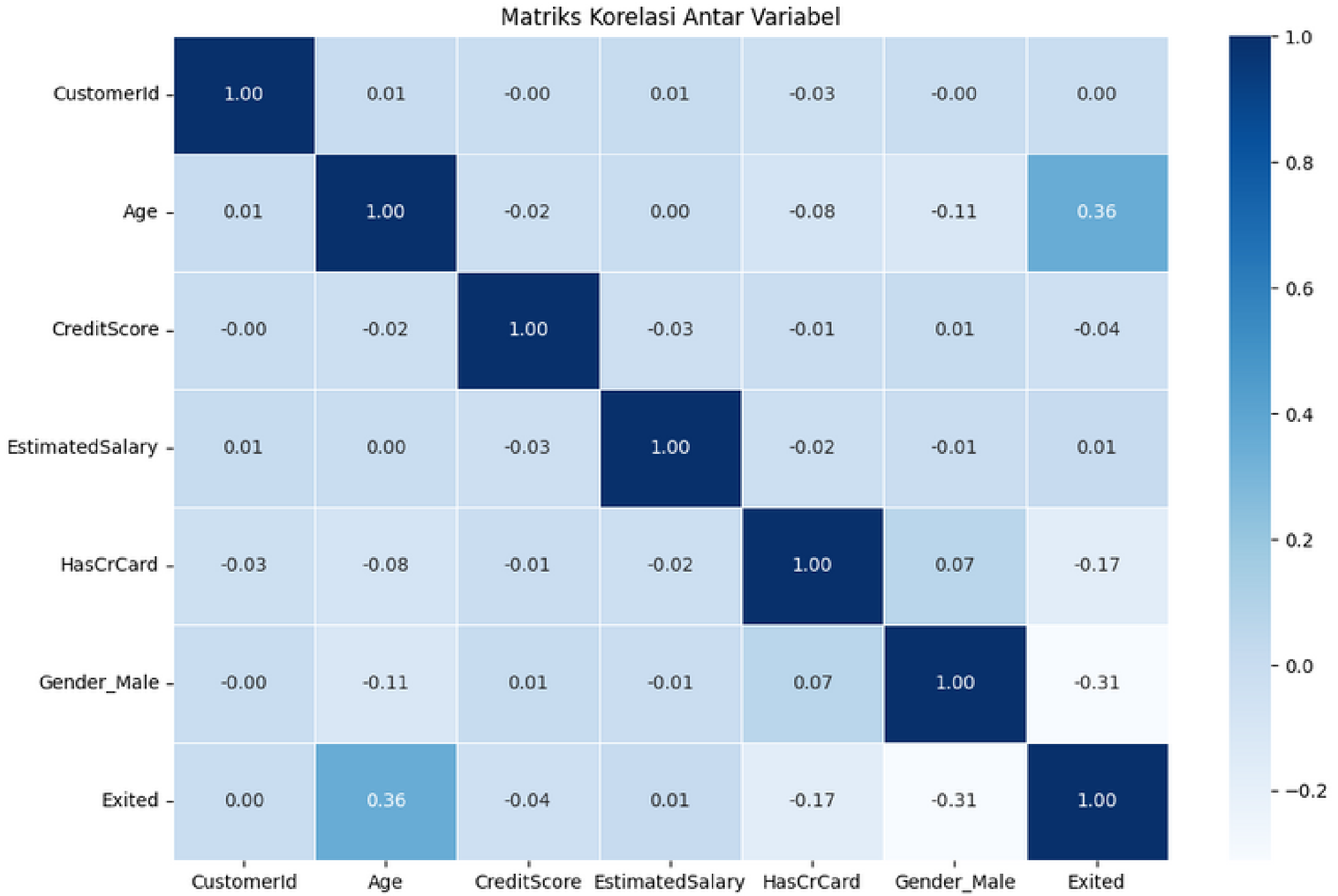


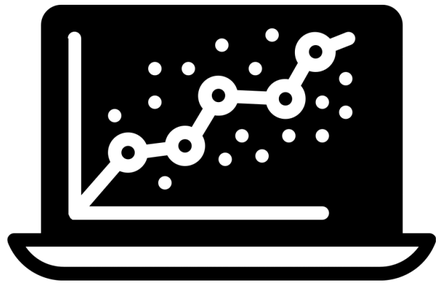




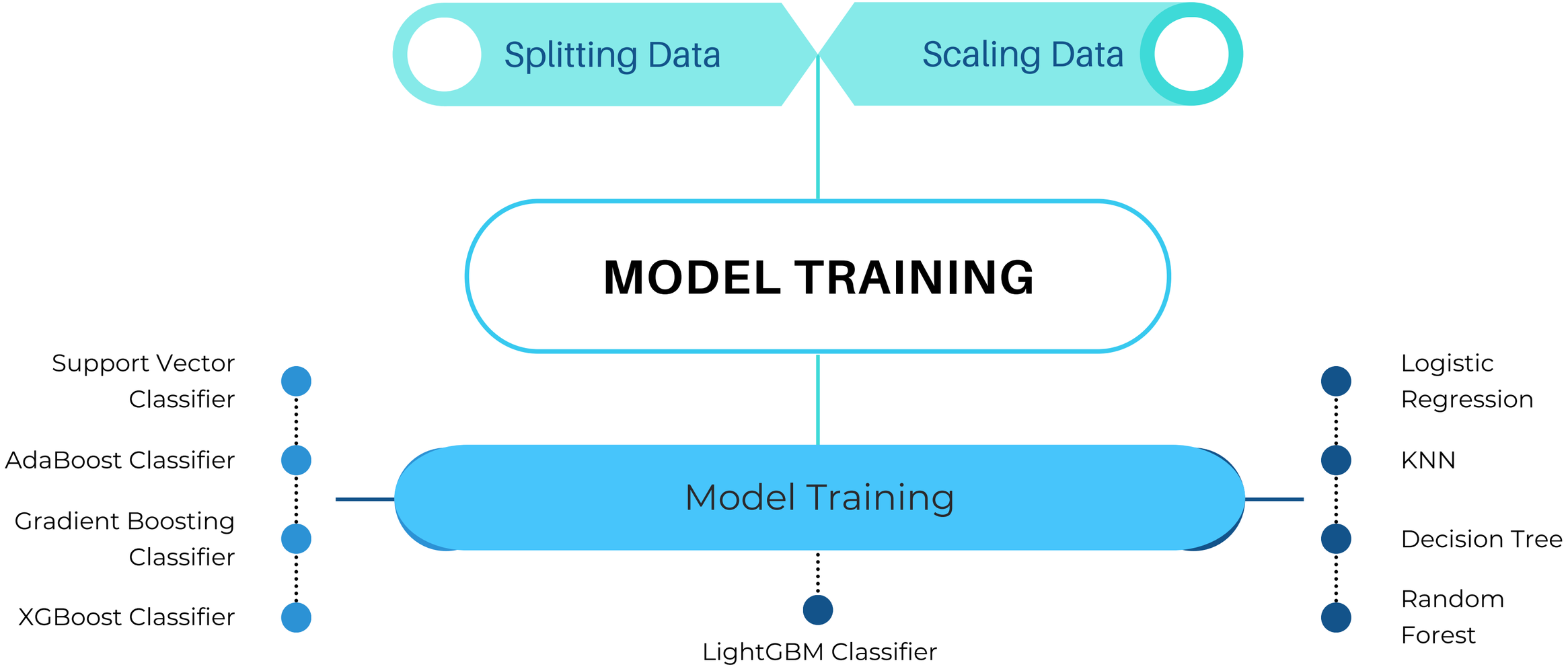


Matriks Korelasi antar Variabel





Model Training


Splitting Data
Scaling Data
MODEL TRAINING
Model Training
Support Vector Classifier
AdaBoost Classifier
Gradient Boosting Classifier
XGBoost Classifier
LightGBM Classifier
Logistic Regression
KNN
Decision Tree
Random Forest

Model Training

Splitting & Scaling Data

```
def model_prepare(df_model):
    y = df_model[dependent_variable_name]
    X = df_model.loc[:, df_model.columns != dependent_variable_name]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
    sc = StandardScaler()
    X_train = sc.fit_transform(X_train)
    X_test = sc.transform (X_test)
    return X_train, X_test, y_train, y_test
```

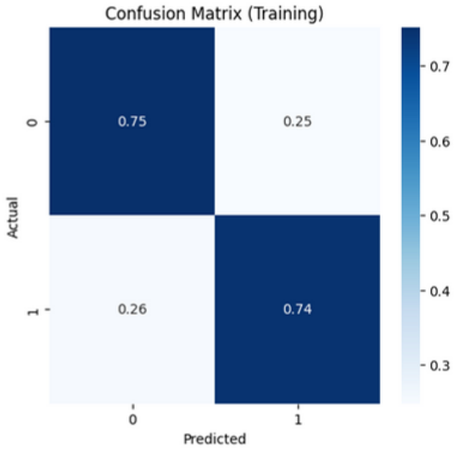
Model Training

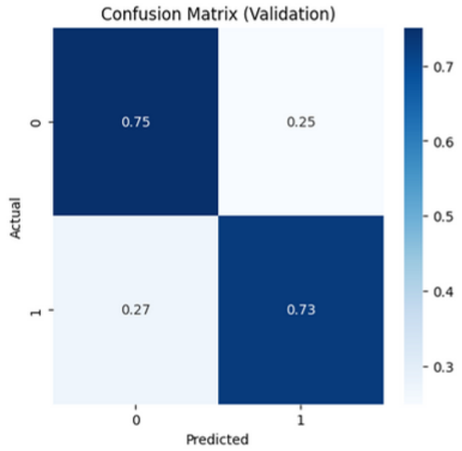
No.	Model	Parameter	Before	After
9	LightGBM Classifier	accuracy	0.743566	0.749843
		precision	0.724664	0.731334
		recall	0.764327	0.769478
		f1-score	0.743967	0.749922

Model Training

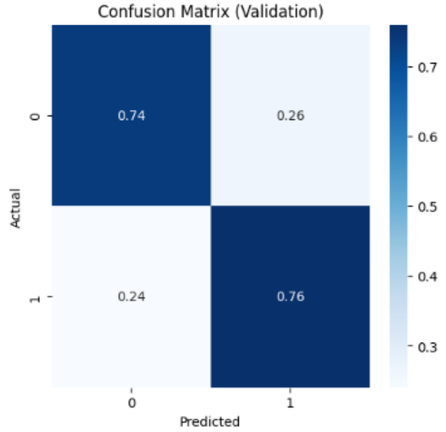
Model Training

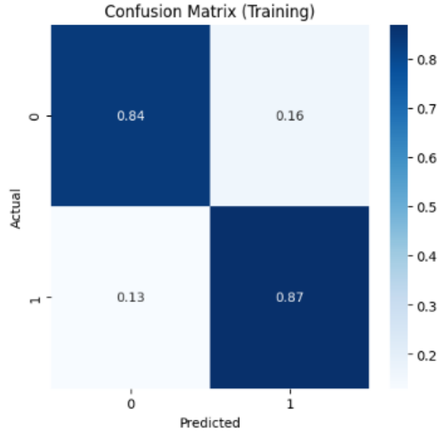
Adaboost



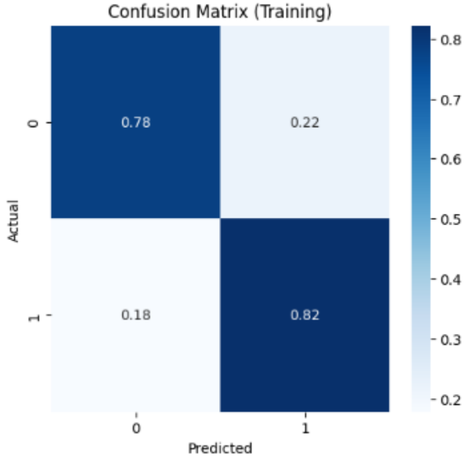


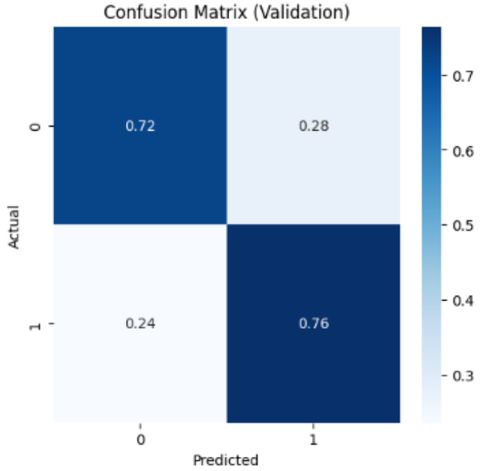
XGboost





Lightboost





Model Deployment

Final Project - Algowizard Team

Model Deployment

Overview

.streamlit	1/6/2024 2:13 AM	File folder	
Archive	1/6/2024 7:35 PM	File folder	
Datasets	12/30/2023 4:49 PM	File folder	
model	1/5/2024 9:08 PM	File folder	
algowizard.jpg	1/6/2024 4:58 PM	JPG File	87 KB
app.py	1/6/2024 5:59 PM	Python Source File	13 KB
Cover.jpg	12/30/2023 5:35 PM	JPG File	5,684 KB
final_model.pkl	1/6/2024 6:05 PM	PKL File	499 KB
Final_Project_AlgoWizard.ipynb	1/6/2024 6:05 PM	Jupyter Source File	2,283 KB
output1.png	1/6/2024 5:31 PM	PNG File	108 KB
output2.png	1/6/2024 5:31 PM	PNG File	61 KB
output3.png	1/6/2024 5:31 PM	PNG File	119 KB
requirements.txt	1/6/2024 3:35 PM	Text Document	1 KB
training_result.csv	1/6/2024 1:24 AM	Microsoft Excel Com...	1 KB

app.py > main

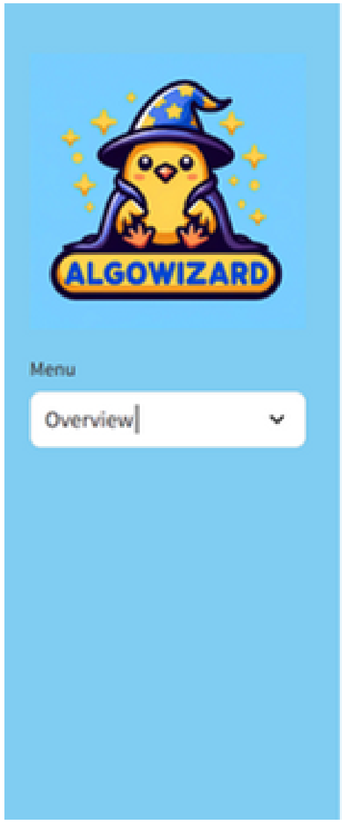
```python
import streamlit as st
import streamlit.components.v1 as stc
import pickle
import pandas as pd
import numpy as np

with open('final_model.pkl','rb') as file:
    Final_Model = pickle.load(file)

def main():
    # stc.html(html_temp)
    # st.title("Customer Churn Prediction App")
    st.markdown("""
            <p style="font-size: 44px; color: #023047;font-weight: bold">Customer Churn Prediction App</p>
            """, unsafe_allow_html=True)
    st.markdown("Aplikasi ini dibuat oleh tim Algowizard untuk Final Project Data Science Bootcamp Digital Skola")

    with st.sidebar:
        st.image("algowizard.jpg")

        menu = ["Overview","Machine Learning"]
        choice = st.sidebar.selectbox("Menu", menu)


    if choice == "Overview":
        st.header("Overview")
        st.markdown("Aplikasi prediksi churn memanfaatkan pembelajaran mesin dan kecerdasan buatan untuk menganalisis data pelanggan dan m
```

Streamlit

GitHub

Model Deployment

Overview



Customer Churn Prediction App

Aplikasi ini dibuat oleh tim Algowizard untuk Final Project Data Science Bootcamp Digital Skola

Overview

Aplikasi prediksi churn memanfaatkan pembelajaran mesin dan kecerdasan buatan untuk menganalisis data pelanggan dan mengidentifikasi mereka yang berisiko pergi. Hal ini memungkinkan bisnis untuk secara proaktif melibatkan pelanggan ini dengan intervensi yang ditargetkan dan strategi retensi, meminimalkan churn dan meningkatkan nilai umur pelanggan.

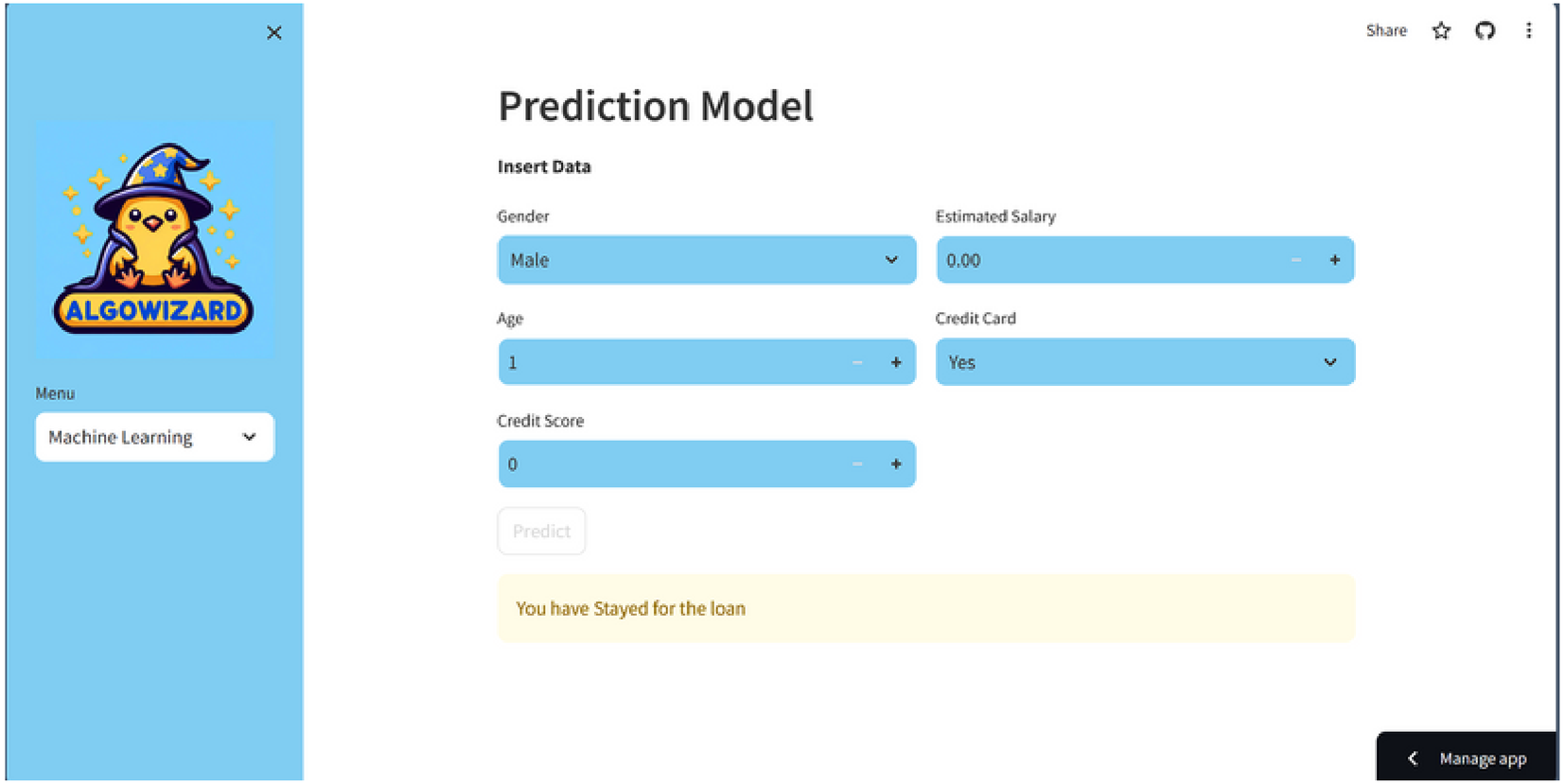
Sekilas tentang Dataset yang digunakan

	CustomerId	Gender	Age	CreditScore	EstimatedSalary	HasCrCard	Exited
0	15634602	Female	42	619	101,348.8800	1	1
1	15647311	Female	41	608	112,542.5800	0	0
2	15619304	Female	42	502	113,931.5700	1	1
3	15701354	Female	39	699	93,826.6300	0	0

https://bit.ly/finalprojectalgowizard

Model Deployment

Machine Learning - Prediction Model



https://bit.ly/finalprojectalgowizard

Kesimpulan

1. Variable Age memiliki korelasi terbesar dengan churn pada customer.
2. LightGBM Classifier memiliki performa terbaik dibandingkan model lainnya untuk melakukan churn prediction terhadap dataset yang dimiliki.
3. Model Deployment dapat digunakan oleh user untuk memprediksi apakah customer akan exit atau tidak berdasarkan Gender, Salary, Age, Credit Card, dan Credit Score

https://bit.ly/notebookfinalprojectalgowizard
https://bit.ly/finalprojectalgowizard

Pembagian Tugas

Nama	Pembagian Tugas	Persentase
Nur Aula	EDA & Data Pre-Processing	14.28%
Ni Putu Juliyant Ananda Rika Pangastuti	EDA & Data Pre-Processing	14.28%
Fazarianti Fariha Hilman	EDA & Data Pre-Processing	14.28%
Hani Musyaffa Hadi	EDA & Data Pre-Processing	14.28%
Julius Timothy	Model Training & Evaluation	14.28%
Eko Budiono	Model Training & Evaluation	14.28%
Kemas Muhammad Rizki Fadhila	Model Deployment	14.28%

Thank you!

Algowizard - Team 2

- Nur Aula
- Ni Putu Juliyant Ananda Rika Pangastuti
- Fazarianti Fariha Hilman
- Julius Timothy
- Kemas Muhammad Rizki Fadhila
- Eko Budiono
- Hani Musyaffa Hadi

Final Project - Algowizard Team