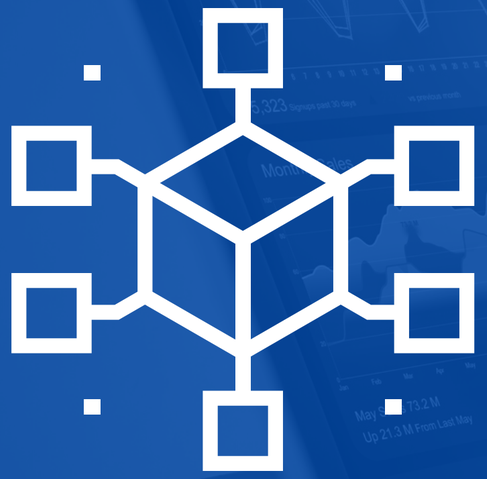
Final Task - Data Scientist ID/X
PARTNER - VIX RAKAMIN

Rakamin Academy

id/x partners



Credit Risk Analysis and Prediction with Machine Learning Approaches for Loan Decisions

KOMANG RYANDHI SUANDITA

Background Project

Penilaian risiko kredit adalah tantangan mendasar dalam industri keuangan. Pemberi pinjaman memiliki tugas penting untuk mengevaluasi kelayakan kredit peminjam untuk membuat keputusan pemberian pinjaman yang terinformasi dan mengelola paparan risikonya.

Prediksi risiko kredit yang tidak akurat dapat mengakibatkan kerugian keuangan besar, memburuknya portofolio pinjaman, dan ketidakstabilan ekonomi. Oleh karena itu, mengembangkan model prediksi risiko kredit yang akurat dan dapat diandalkan sangat penting untuk menyederhanakan operasi, mengoptimalkan sumber daya, dan memastikan pengambilan keputusan yang baik di lembaga pemberi pinjaman.

Problem Understanding

01 Goals

02 Understanding with Data

03 Analytic Approach

04 Modelling

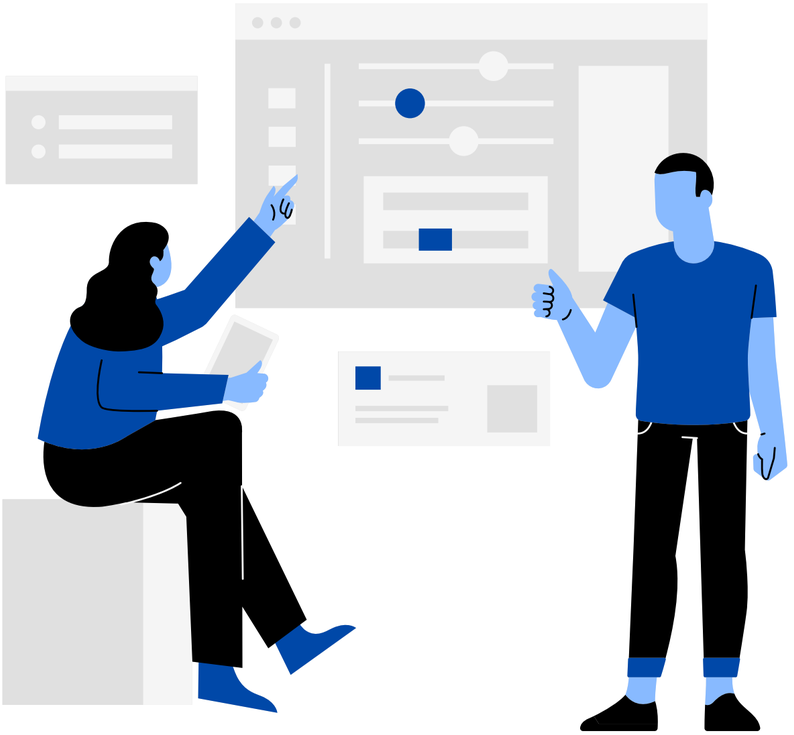
Goals

1. Memperlancar arus masuk dan keluar keuangan perusahaan
2. Membuat model prediktif yang dapat menentukan kemampuan suatu pengguna dalam membayar kredit

Understanding with Data

Data yang akan digunakan adalah data pinjaman perusahaan dari tahun 2007 sampai tahun 2014 dengan nama dataset ‘loan_data_2007_2014.csv’

	Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	...	total_bal_il	il_util	open_rv_12m	open_rv_24m
0	0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	...	NaN	NaN	NaN	NaN
1	1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	...	NaN	NaN	NaN	NaN
2	2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	...	NaN	NaN	NaN	NaN
3	3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	...	NaN	NaN	NaN	NaN
4	4	1075358	1311748	3000	3000	3000.0	60 months	12.69	67.79	B	...	NaN	NaN	NaN	NaN



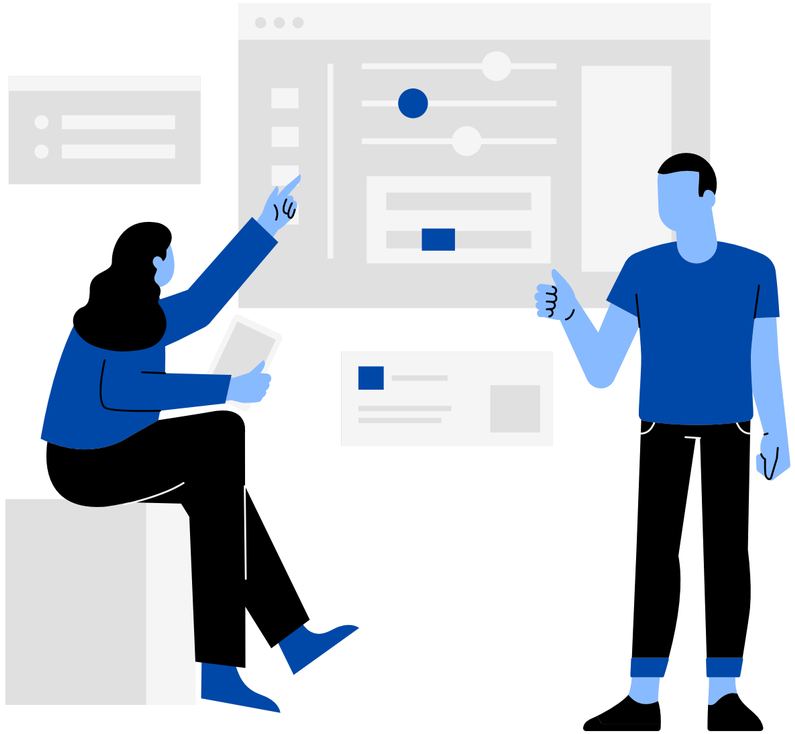
Analytic Approach

Untuk mengatasi tantangan yang dihadapi, kami memutuskan untuk mengadopsi pendekatan analisis yang lebih baik, yaitu dengan membangun model machine learning. Hal ini dipilih karena kebutuhan kami tidak hanya sebatas pada pemahaman inferensial atau deskriptif, melainkan kami juga perlu mengembangkan prediksi yang akurat. Dengan menerapkan model machine learning, kami berharap dapat mengeksplorasi pola-pola kompleks dalam data untuk memberikan solusi yang lebih mendalam dan dapat diandalkan.

Analytic Approach

Modelling

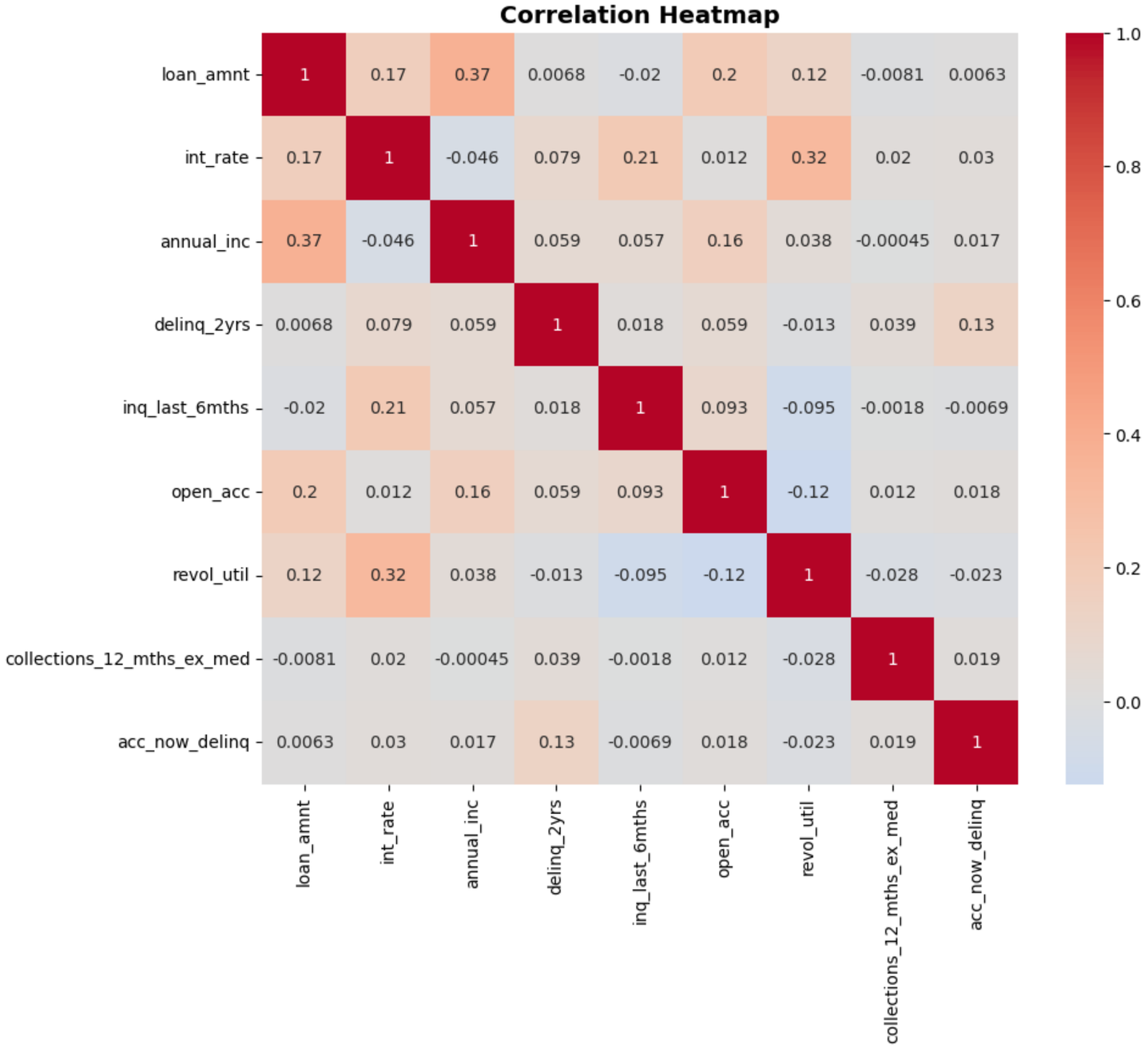
Untuk pemodelan, kami akan mengimplementasikan 5 model machine learning regresi yang termasuk dalam kategori unsupervised, yaitu Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, dan K-Nearest Neighbors. Penggunaan kombinasi model ini dirancang untuk memungkinkan kami mengeksplorasi berbagai pendekatan dan memilih model yang paling sesuai dengan karakteristik data yang kami miliki.



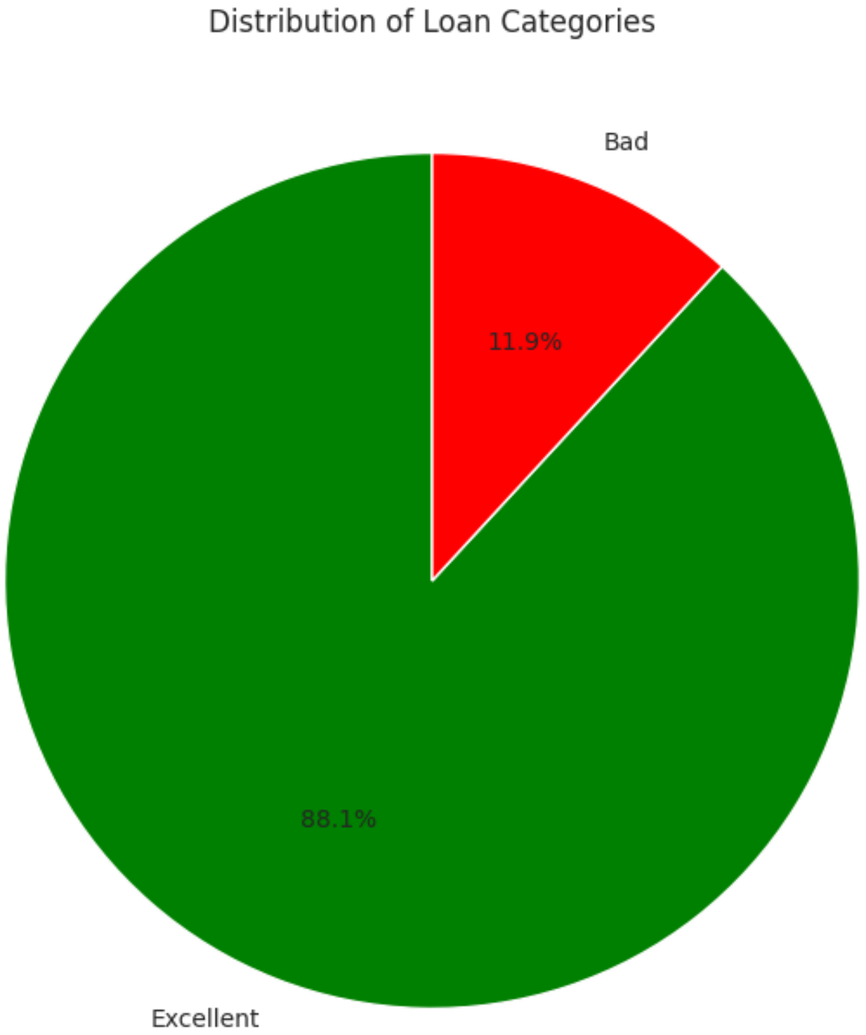
Quick Overview

- Our dataset consist of 466285 rows and 75 features
- There are 9776224 missing values
- There are 0 Duplicated Value

Data Correlation

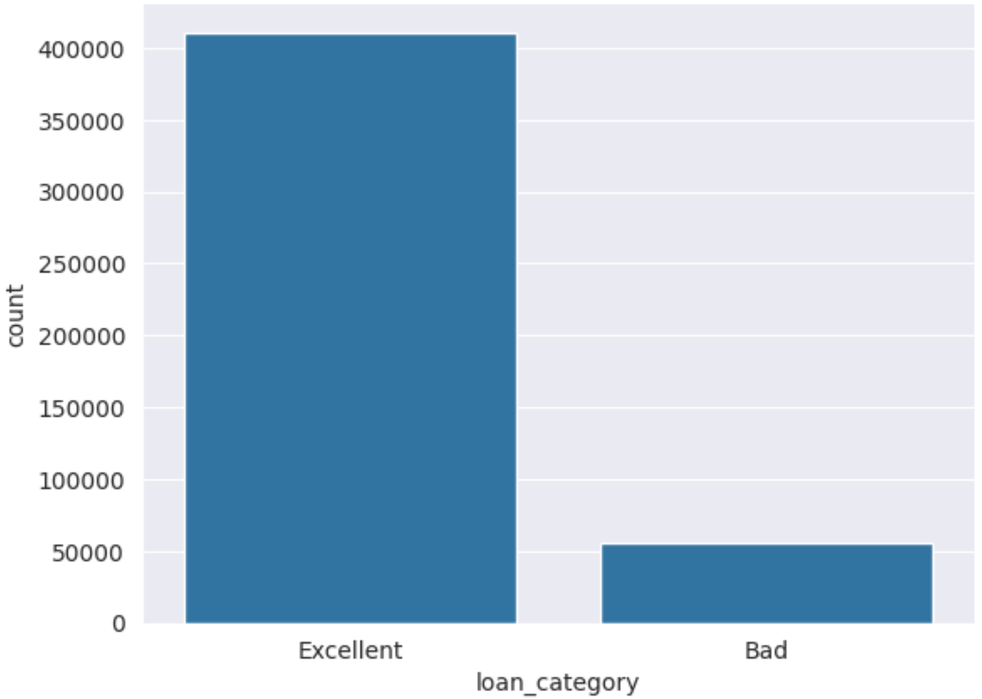


Data Cleaning

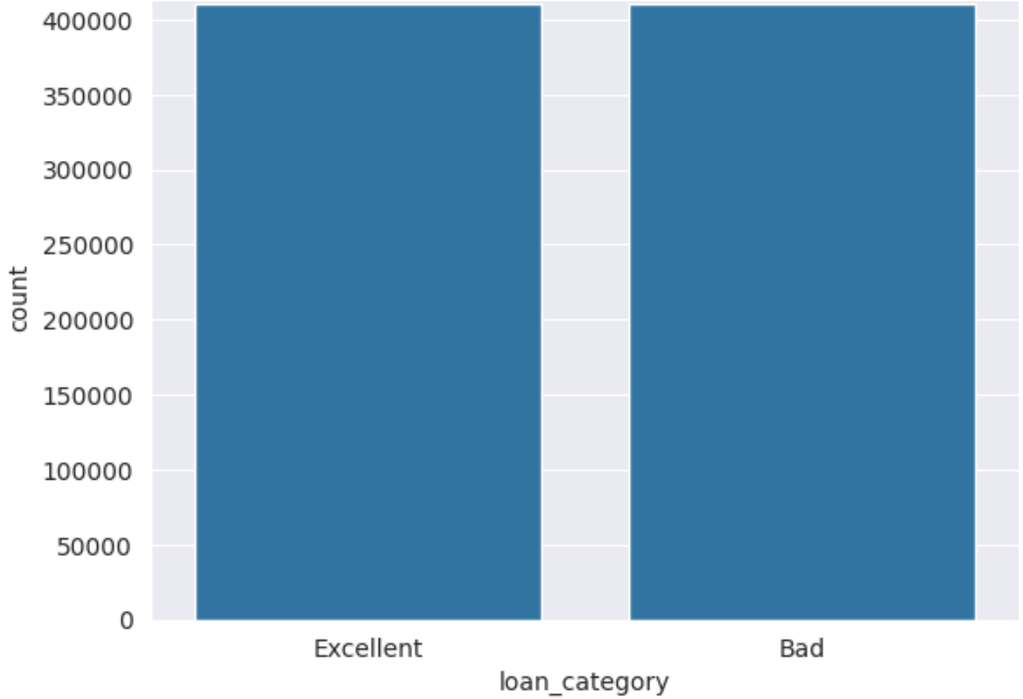


There is 88,1% Loans with status ‘Excellent’ and 11,9% bad loans category.

Oversampling for loan_category



Name: loan_category, dtype: int64



Machine Learning Modelling

Dalam langkah pemodelan, dataset telah dibagi menjadi set pelatihan dan pengujian dengan proporsi 80:20. Fitur-fitur dinormalisasi menggunakan Standard Scaler. Lima model machine learning regresi yang berbeda (Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, K-Nearest Neighbors) diuji dan dievaluasi pada set pengujian.

By using big data, you can make good social media content that suits your brand. You will know:

Define features and target variable

Train and test split

80% Training, 20% Testing

Reference from baeldung.com

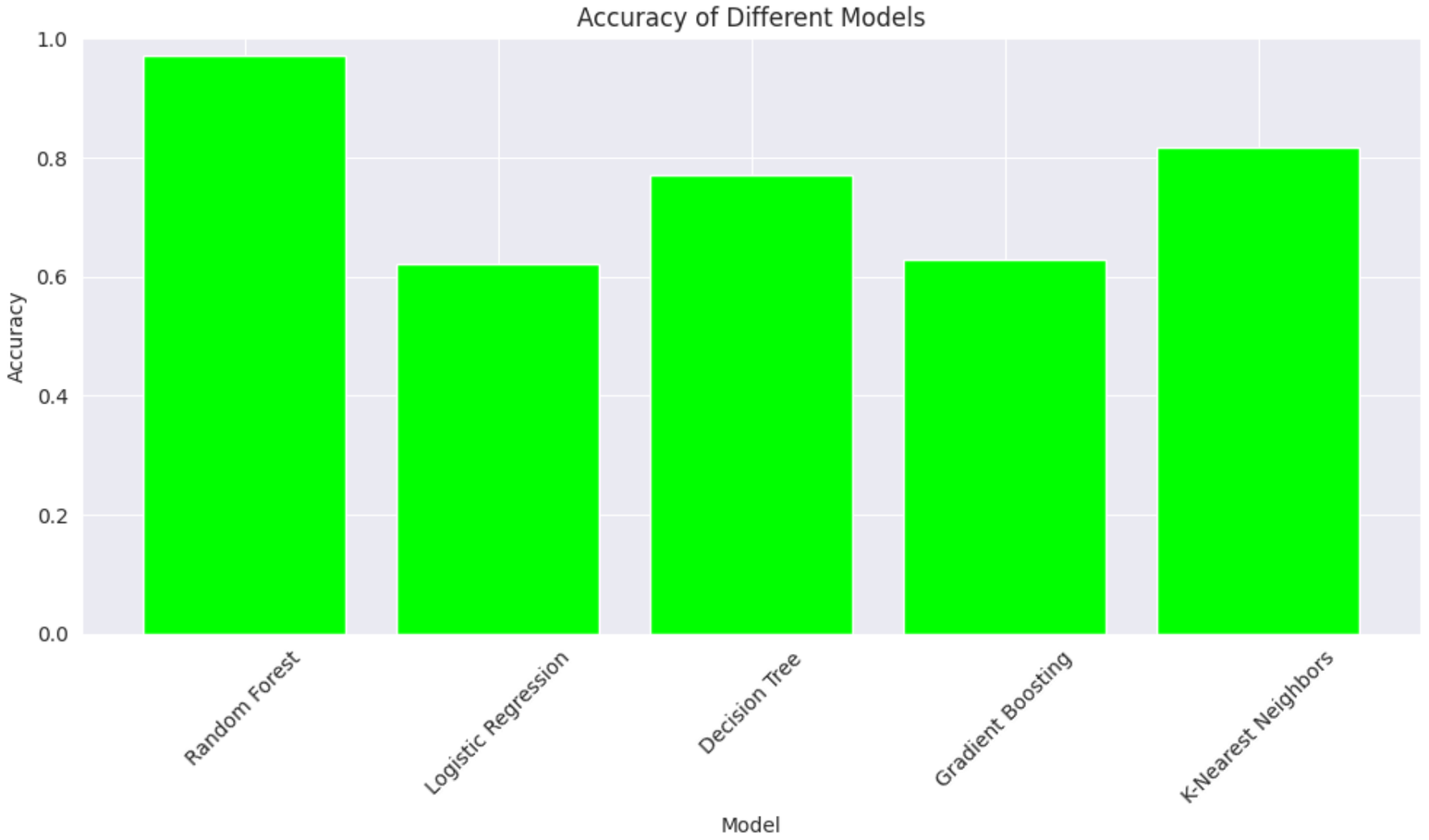
Normalization

Initialize dictionary

Evaluate each model

Random Forest, Logistic Regression, Decision Tree, Gradient Boosting, and K-Nearest Neighbors.

Machine Learning Result



	Model	Accuracy (%)
0	Random Forest	97
1	Logistic Regression	62
2	Decision Tree	77
3	Gradient Boosting	63
4	K-Nearest Neighbors	82

The best model is Random Forest with accuracy 97%.

Thanks

Github Link