Analisis Risiko Pemberian Pinjaman Pada Lending Club Menggunakan Machine Learning Dengan Metode Logistic Regression Dan Random Forest

[1] Faishal Syams Afif1, [2] Dede Brahma Arianto
[1]Universitas Padjadjaran, [2]Universitas Islam Indonesia
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat

[1] faishal20001@mail.unpad.ac.id, [2]dede.brahma2@gmail.com

Abstrak

Penelitian ini bertujuan untuk menganalisis risiko pada data pemberian pinjaman Lending Club, sebuah platform peer-to-peer (P2P) lending terkemuka di dunia. Dataset yang digunakan mencakup informasi lengkap tentang pinjaman dari tahun 2007 hingga 2018. Metode klasifikasi risiko yang digunakan adalah Logistic Regression dan Random Forest, yang berfungsi untuk memprediksi risiko pinjaman berdasarkan variabel independen.

Sebelum dilakukan pemodelan, dilakukan preprocessing data yang meliputi pembersihan data seperti penghapusan fitur yang tidak relevan, data duplikat, penanganan nilai yang hilang (missing value), dan outliers handling. Hasil pengujian menunjukkan akurasi sebesar 80.37% untuk Logistic Regression dan 80.23% untuk Random Forest. Meskipun demikian, kedua model memiliki keseimbangan yang rendah dalam nilai precision, recall, dan f1-score pada segi rata-rata makro (macro avg). Hal ini disebabkan oleh ketidakseimbangan kelas dalam interpretasi performa model dan evaluasi yang digunakan.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengelolaan risiko pada Lending Club dan platform P2P lending lainnya. Dengan pengembangan model yang lebih baik dan penanganan ketidakseimbangan kelas yang tepat, metode machine learning dapat membantu mengidentifikasi pinjaman yang berisiko tinggi di masa depan. Ini akan membantu pengelolaan risiko secara efektif dan meningkatkan keberlanjutan industri P2P lending.

***Kata Kunci**: Lending Club, Peer-to-peer lending, Analisis Risiko Pinjaman, Machine learning*

Abstract

This research aims to analyze the risks associated with lending data from Lending Club, a leading peer-to-peer (P2P) lending platform in the world. The dataset used covers comprehensive information about loans from 2007 to 2018. The risk classification methods employed are Logistic Regression and Random Forest, which are used to predict loan risks based on independent variables.

Prior to modeling, data preprocessing was conducted, including data cleaning by removing irrelevant features, duplicated data, handling missing values, and outliers. The testing results showed an accuracy of 80.37% for Logistic Regression and 80.23% for Random Forest. However, both models exhibited low balance in precision, recall, and f1-score values in terms of macro average. This was due to the class imbalance in interpreting model performance and the evaluation metrics used.

This research is expected to contribute to the management of risks in Lending Club and other P2P lending platforms. With improved model development and appropriate handling of class imbalances, machine learning methods can help identify high-risk loans in the future. This will assist in effective risk management and enhance the sustainability of the P2P lending industry.

***Key Words**: Lending Club, Peer-to-peer lending, Loan Risk Analysis, Machine learning*

PENDAHULUAN

Menurut Deloitte, Fintech adalah istilah yang menggambarkan perusahaan baru yang menggunakan teknologi untuk mengubah dan meningkatkan kegiatan keuangan. Deloitte. (2017). Salah satu bidang fintech yang semakin populer adalah P2P (Peer-to-Peer) lending. P2P lending adalah platform yang menghubungkan peminjam dan pemberi pinjaman secara langsung melalui platform online, mengeliminasi peran bank tradisional sebagai perantara. P2P lending memberikan akses ke pinjaman cepat dan mudah bagi individu atau usaha kecil yang sulit memperoleh pinjaman dari lembaga keuangan konvensional. (Cambridge Centre for Alternative Finance. 2018). (The Global Alternative Finance Industry Report 2017.)

LendingClub adalah salah satu platform peer-to-peer (P2P) lending terbesar di dunia yang berbasis di Amerika Serikat. Platform ini menyediakan layanan pemberian pinjaman secara online dengan menghubungkan peminjam dan investor secara langsung. LendingClub memungkinkan individu dan usaha kecil untuk meminjam dana tanpa melalui lembaga keuangan tradisional. Namun, seperti halnya bisnis pinjaman pada umumnya, terdapat Risiko pinjaman yang harus diperhatikan. Risiko pinjaman adalah kemungkinan terjadinya kerugian finansial bagi pemberi pinjaman akibat adanya kegagalan peminjam untuk melunasi pinjaman sesuai dengan perjanjian (Jurnal Keuangan dan Perbankan, 2018). Oleh karena itu, analisis risiko yang cermat diperlukan dalam memberikan pinjaman, tidak hanya dalam platform Lending Club, tetapi juga pada platform P2P lending lainnya.

Dengan menggunakan Machine Learning dapat membantu kita dalam menganalisis risiko pinjaman menjadi lebih mudah. Machine learning adalah suatu pendekatan dalam bidang kecerdasan buatan (artificial intelligence) yang mengizinkan sistem komputer untuk belajar dan mengambil keputusan atau melakukan prediksi berdasarkan data tanpa diprogram secara eksplisit (Mitchell, T. M. 1997). Pendekatan ini didasarkan pada pengembangan algoritma yang mampu mengidentifikasi pola dan mempelajari hubungan dalam data, sehingga sistem dapat beradaptasi dan meningkatkan kinerjanya seiring waktu.

Penelitian yang menggunakan machine learning memiliki beberapa keunggulan yang membuatnya menjadi pilihan yang menarik dalam berbagai bidang, termasuk dalam analisis risiko pemberian pinjaman pada Lending Club. Berikut adalah alasan mengapa penelitian harus menggunakan machine learning:

1. Kemampuan untuk memproses dan menganalisis data kompleks : Pada dataset historis pemberian pinjaman yang dipakai terdapat banyak faktor yang dapat mempengaruhi risiko, seperti riwayat kredit, profil pengguna, dan masih banyak

lagi. Dengan menggunakan Machine learning dapat dengan mudah memproses dan menganalisis data dalam skala besar, mengidentifikasi pola yang rumit, dan menemukan hubungan yang mungkin sulit untuk ditemukan oleh metode tradisional.

2. Prediksi yang lebih akurat : Dengan menggunakan machine learning, model dapat melakukan prediksi berdasarkan pola yang ditemukan dalam data. Dengan menganalisis data historis dan mengidentifikasi pola yang berkaitan dengan risiko pemberian pinjaman, model dapat memberikan prediksi yang lebih akurat tentang kemungkinan pembayaran pinjaman oleh peminjam.

3. Efisiensi dan skalabilitas : Dengan menggunakan machine learning, analisis risiko pemberian pinjaman dapat dilakukan secara efisien dan dengan skalabilitas yang tinggi. Model yang telah dilatih dapat dengan cepat memberikan prediksi dan evaluasi risiko dalam waktu yang relatif singkat.

Di dalam Machine Learning terdapat berbagai macam model-model klasifikasi dan regresi yang digunakan untuk mengklasifikasikan kualitas pinjaman menjadi kategori baik atau buruk. Namun, pada setiap model, seperti Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, dan k-NN, memiliki kelebihan dan kelemahan masing-masing (Lu et al., 2020).

Machine Learning telah menjadi topik penelitian yang populer dalam beberapa tahun terakhir. Banyak penelitian yang dilakukan oleh para peneliti di berbagai bidang untuk mengembangkan dan menerapkan metode-metode yang berbeda-beda.

Pada penelitian sebelumnya oleh Zhang et al. (2017). Melakukan penelitian untuk analisis risiko pinjaman dan platform P2P lending menggunakan Random Forest dan SVM. Dari hasil penelitian menunjukkan bahwa pendekatan machine learning dapat memberikan prediksi risiko kredit yang lebih akurat.

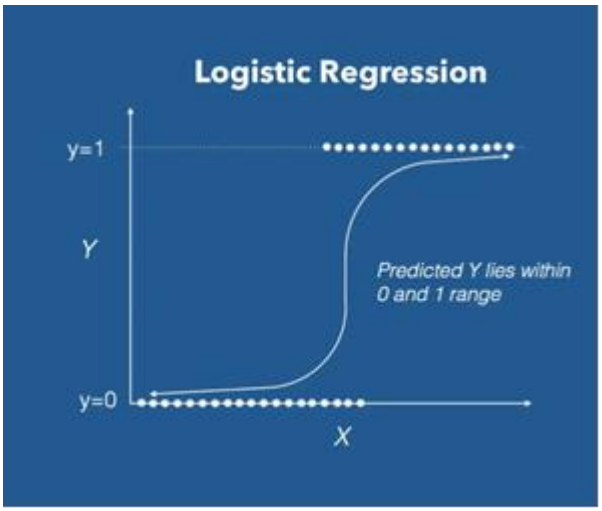
Penelitian lainnya dilakukan oleh Chen et al. (2018). Penelitian ini membandingkan berbagai algoritma machine learning seperti Naive Bayes, Decision Tree, dan Neural Network dalam mengklasifikasikan risiko pinjaman pada platform P2P lending. Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan hasil yang lebih baik dalam mengidentifikasi pinjaman berisiko.

Selain itu, penelitian oleh Li et al. (2019) berhasil mengembangkan model prediksi default pinjaman menggunakan algoritma Logistic Regression, Random Forest, dan Gradient Boosting. Penelitian tersebut menunjukkan bahwa model machine learning dapat memberikan prediksi default yang lebih akurat dibandingkan metode tradisional.

Pada penelitian kali ini, fokus utama akan tertuju pada analisis risiko pemberian pinjaman pada platform Lending Club menggunakan teknik machine learning. Data historis

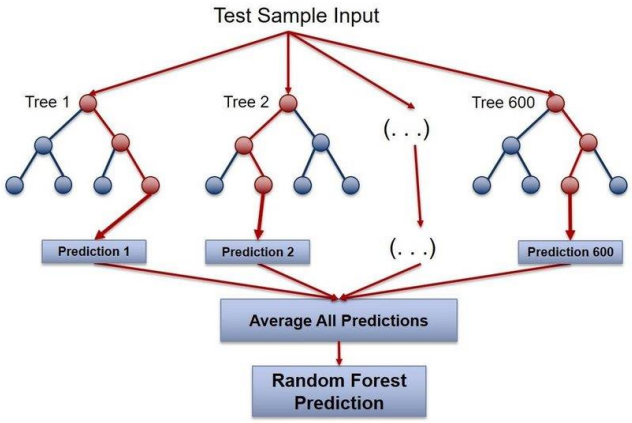
mengenai pemberian pinjaman akan diproses menggunakan teknik machine learning seperti Logistic Regression dan Random Forest.

Logistic Regression digunakan untuk memodelkan hubungan antara variabel dependen kategorik dengan satu atau lebih variabel independen untuk dapat memprediksi probabilitas atau kemungkinan terjadinya suatu peristiwa atau kejadian berdasarkan variabel independen. Metode ini sering digunakan dalam analisis risiko, prediksi kredit, dan prediksi kejadian medis, di antara banyak aplikasi lainnya.



Gambar 2. Arsitektur Logistic Regression
[sumber: machinelearningplus.com]

Random forest adalah metode yang menggabungkan beberapa pohon keputusan (decision trees) untuk membuat prediksi yang lebih akurat. Setiap pohon dalam random forest diberi bobot secara acak pada subset data pelatihan. Metode ini sering digunakan dalam klasifikasi dan regresi, dan memiliki kemampuan untuk mengatasi overfitting dan mengatasi masalah variabel yang tidak relevan.



Gambar 2. Arsitektur Random Forest
[sumber: researchgate.net]

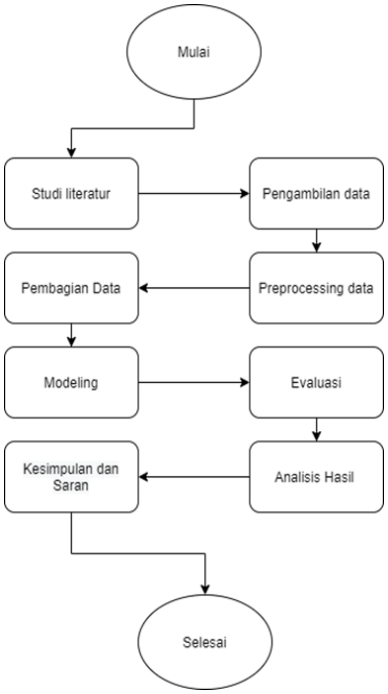
Tujuan dari penelitian ini adalah melakukan analisis risiko pemberian pinjaman pada Lending Club menggunakan pendekatan machine learning dengan metode Logistic Regression dan Random Forest. Secara khusus, tujuan penelitian ini meliputi:

1. Membangun model prediktif menggunakan metode Logistic Regression dan Random Forest untuk memprediksi risiko pinjaman. Model ini akan digunakan untuk mengklasifikasikan pinjaman menjadi kategori "good loan" atau "bad loan" berdasarkan variabel-variabel yang diidentifikasi.

2. FMengukur dan membandingkan performa model Logistic Regression dan Random Forest dalam melakukan prediksi risiko pinjaman. Performa model akan dievaluasi menggunakan metrik evaluasi yang tepat, seperti akurasi, precision, recall, dan F1-score.

METODE PENELITIAN

Metodologi penelitian ini dirancang untuk memberikan kerangka kerja yang sistematis dan terstruktur dalam melakukan analisis risiko pemberian pinjaman pada Lending Club menggunakan pendekatan Machine Learning dengan metode Logistic Regression dan Random Forest. Metodologi ini akan memandu proses pengumpulan, analisis, dan interpretasi data yang diperlukan untuk mencapai tujuan penelitian seperti ditunjukkan pada Gambar 3.



Gambar 3. Flowchart Metode Penelitian

Tahap awal, dalam metodologi penelitian adalah melakukan studi literatur yang mendalam tentang topik yang akan diteliti, yaitu analisis risiko pemberian pinjaman pada Lending Club menggunakan metode machine learning. Studi literatur ini bertujuan untuk memahami konsep-konsep dasar, teori-teori, dan penelitian terdahulu yang relevan dengan topik tersebut. Sumber literatur yang digunakan dapat berupa jurnal ilmiah, buku referensi, artikel, dan publikasi terkait lainnya.

Tahap kedua, dalam metodologi penelitian ini adalah pengumpulan data pinjaman dari Lending Club. Data yang digunakan mencakup informasi lengkap tentang pinjaman, termasuk karakteristik peminjam, riwayat kredit, tujuan pinjaman, dan performa pembayaran. Data dapat diperoleh dari sumber resmi Lending Club atau sumber data lain yang terpercaya.

Tahap ketiga, Untuk menghindari bias dan mempengaruhi evaluasi performa model, dilakukan *preprocessing* data yaitu data *cleaning* seperti menghapus *Irrelevant feature*, *Duplicated*, *missing value* dan *outliers handling* yang kemudian diproses lebih lanjut.

Tahap keempat, akan dibagi menjadi dua subset, yaitu data pelatihan (training data) dan data pengujian (testing data). Data pelatihan akan digunakan untuk melatih model dan mengoptimalkan parameter, sedangkan data pengujian akan digunakan untuk menguji performa model yang telah dilatih.

Pada tahap kelima, dilakukan pemodelan menggunakan metode Logistic Regression dan Random Forest. Model Logistic Regression akan digunakan untuk memodelkan hubungan antara variabel independen dengan variabel dependen kategorik, yaitu klasifikasi risiko pinjaman. Model Random Forest akan digunakan untuk membangun kumpulan pohon keputusan yang masing-masing akan memberikan prediksi risiko pinjaman.

Setelah pemodelan, performa model akan dievaluasi menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Metrik-metrik ini akan memberikan informasi tentang sejauh mana model dapat memprediksi risiko pemberian pinjaman dengan akurat.

Kemudian Hasil evaluasi performa model akan dianalisis dan diinterpretasikan untuk memahami kekuatan dan kelemahan masing-masing model. Perbandingan antara Logistic Regression dan Random Forest akan dilakukan untuk mengetahui perbedaan dalam prediksi risiko pemberian pinjaman.

Pada tahap terakhir ini, kesimpulan penelitian akan dirumuskan berdasarkan hasil analisis dan interpretasi. Saran juga dapat diberikan untuk peneliti selanjutnya dan juga bagi pengelolaan risiko pemberian pinjaman di Lending Club dan platform P2P lending lainnya.

HASIL DAN PEMBAHASAN

1. Data *Understanding*

Dalam dunia nyata, seringkali ditemukan data yang kotor dan tidak lengkap. Langkah awal yang dilakukan dalam masalah prediksi terkait penilaian status pinjaman adalah membersihkan data yang akan diolah hingga diperoleh informasi yang paling berguna sebisa mungkin. Pada penelitian ini menggunakan dataset "Lending Club Dataset" yang diperoleh dari Kaggle dimana Dataset terdiri dari 24 kolom dan 396029 baris. Detail variabel dataset ditunjukkan pada Tabel 1.

Tabel 1**.** Deskripsi Dataset

Kolom	**Keterangan**
loan_amnt	Jumlah uang awal yang diminta oleh peminjam untuk pinjaman.
term	Durasi pinjaman dalam bulan, baik 36 bulan atau 60 bulan.
int_rate	Tingkat bunga yang ditetapkan untuk pinjaman.
installment	Jumlah pembayaran bulanan yang harus dibayarkan oleh peminjam jika pinjaman disetujui.
grade	Grade yang ditetapkan untuk pinjaman oleh LC (LendingClub), menunjukkan kelayakan kredit peminjam.
sub_grade	Subgrade yang ditetapkan untuk pinjaman oleh LC, memberikan detail lebih lanjut dalam grade pinjaman.

emp_title	Jabatan pekerjaan peminjam saat mengajukan pinjaman.
emp_length	Lama kerja dalam tahun, berkisar dari 0 (kurang dari satu tahun) hingga 10 (sepuluh tahun atau lebih).
home_ownership	Status kepemilikan rumah yang disediakan oleh peminjam, seperti RENT (sewa), OWN (milik sendiri), MORTGAGE (hipotek), atau OTHER (lainnya).
annual_inc	Pendapatan tahunan yang dilaporkan sendiri oleh peminjam selama aplikasi pinjaman.
verification_status	Menunjukkan apakah pendapatan peminjam telah diverifikasi oleh LC, tidak diverifikasi, atau jika sumber pendapatan telah diverifikasi.
issue_d	Bulan dimana pinjaman didanai.
loan_status	Status terkini dari pinjaman. dengan 2 nilai yaitu Fully Paid dan Charged Off
purpose	Kategori yang disediakan oleh peminjam untuk permohonan pinjaman.
title	Judul pinjaman yang disediakan oleh peminjam.

dti	Rasio utang-terhadap-pendapatan, dihitung dengan membagi total pembayaran utang bulanan peminjam (tidak termasuk hipotek dan pinjaman yang diminta) dengan pendapatan bulanan yang dilaporkan sendiri.
earliest_cr_line	Bulan di mana garis kredit terlama yang dilaporkan oleh peminjam dibuka.
open_acc	Jumlah garis kredit terbuka dalam file kredit peminjam.
pub_rec	Jumlah catatan publik yang merugikan yang terkait dengan peminjam.
revol_bal	Total saldo kredit yang berputar (revolving) dari peminjam.
revol_util	Tingkat penggunaan garis kredit berputar, yang mewakili jumlah kredit yang digunakan oleh peminjam relatif terhadap semua kredit berputar yang tersedia.
total_acc	Jumlah total garis kredit saat ini dalam file kredit peminjam.
initial_list_status	Status penayangan awal dari pinjaman, dengan nilai mungkin "W" atau "F".
application_type	Menunjukkan apakah pinjaman adalah aplikasi individu atau aplikasi bersama dengan dua peminjam bersama.

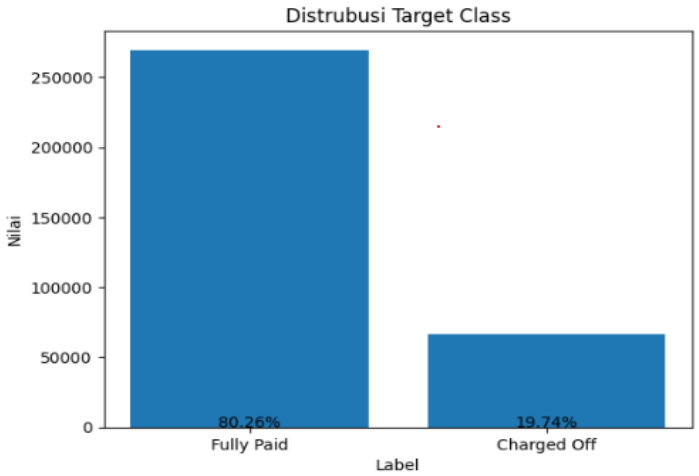
mort_acc	Jumlah akun hipotek yang dimiliki oleh peminjam.
pub_rec_bankruptcies	Jumlah kebangkrutan catatan publik yang terkait dengan peminjam.
address	Alamat tempat tinggal peminjam

Dalam penelitian ini, tujuan utamanya adalah untuk memprediksi apakah suatu pinjaman termasuk dalam kategori bad loan atau good loan. Kriteria yang digunakan untuk menentukan apakah suatu pinjaman dianggap sebagai bad loan adalah jika nilai kolom "loan_status" pada data memiliki nilai "Charged off". "Charged off" mengindikasikan bahwa pemberi pinjaman, seperti bank atau lembaga keuangan, menganggap pinjaman tersebut sebagai kerugian yang tidak dapat dipulihkan. Hal ini terjadi ketika peminjam gagal membayar pinjaman secara teratur dan tidak ada upaya yang berhasil dilakukan untuk mendapatkan pembayaran tertunggak.

Sementara itu, kriteria untuk good loan adalah jika nilai kolom "loan_status" pada data memiliki nilai "Fully Paid". "Fully paid" mengindikasikan bahwa peminjam telah melunasi seluruh jumlah utang atau pinjaman yang dimiliki. Ini berarti bahwa peminjam telah membayar semua pokok pinjaman beserta bunga dan biaya lainnya yang terkait.

Dengan menggunakan metode Logistic Regression dan Random forest, penelitian ini akan mencoba membangun model machine learning yang dapat memprediksi dengan akurat apakah suatu pinjaman akan menjadi bad loan atau good loan berdasarkan data historis yang tersedia.

Dalam dataset yang digunakan, terdapat distribusi target pada kolom "loan_status". Berdasarkan gambar distribusi, terlihat bahwa sebagian besar baris dalam dataset memiliki nilai "Fully Paid" sebesar 80,26%. Sedangkan nilai "Charged Off" hanya terdapat pada 19,74% dari total 396,030 baris dalam dataset.



Gambar 4. Distribusi Target Class

Dari Gambar 4. Distribusi Target Class tersebut dapat dilihat bahwa jumlah pinjaman yang telah dilunasi sepenuhnya ("Fully Paid") lebih dominan dibandingkan dengan jumlah pinjaman yang dianggap sebagai kerugian yang tidak dapat dipulihkan ("Charged Off"). Distribusi ini dapat mempengaruhi proses pembuatan model dan evaluasi akurasi dalam memprediksi risiko pinjaman.

2. Data *Preprocessing*

Pada tahap preprocessing data, ada 2 hal yang dilakukan yaitu Data Cleaning dan Data *Manipulation* (encoding).

2.1 Data *Cleaning*

Hal yang pertama dilakukan adalah melakukan Data *Cleaning* untuk memastikan keakuratan, konsistensi, kegunaan, dan kualitas data dalam data set. Yang dilakukan pada proses Data *Cleaning* adalah :

1. Menghapus irrelevant feature.

2. Menghapus duplicated data

3. Menghapus missing values

4. Menghapus outliers

Pada tahap pertama penghapusan irrelevant feature, penghapusan feature - feature berikut 'address', 'emp_title', 'title', 'purpose', 'sub_grade', 'issue_d' dan 'earliest_cr_line'. Feature tersebut dihapus karena tidak relevan atau memiliki terlalu banyak feature yang unik sehingga tidak berguna dalam model prediksi, oleh karena itu dihapus.

Kemudian tahap kedua menghapus duplicated data, data yang terduplikat harus dihapus dikarenakan ketika terdapat duplikat data, hal itu dapat menyebabkan hasil menjadi bias atau distorsi sehingga mengurangi keakuratan analisis prediksi nantinya.

Tabel 2. Persentase Missing Value Dataset

Kolom	Persentase missing value
emp_length	4.621115
revol_util	0.069692
mort_acc	9.543469
pub_rec_bankruptcies	0.135091

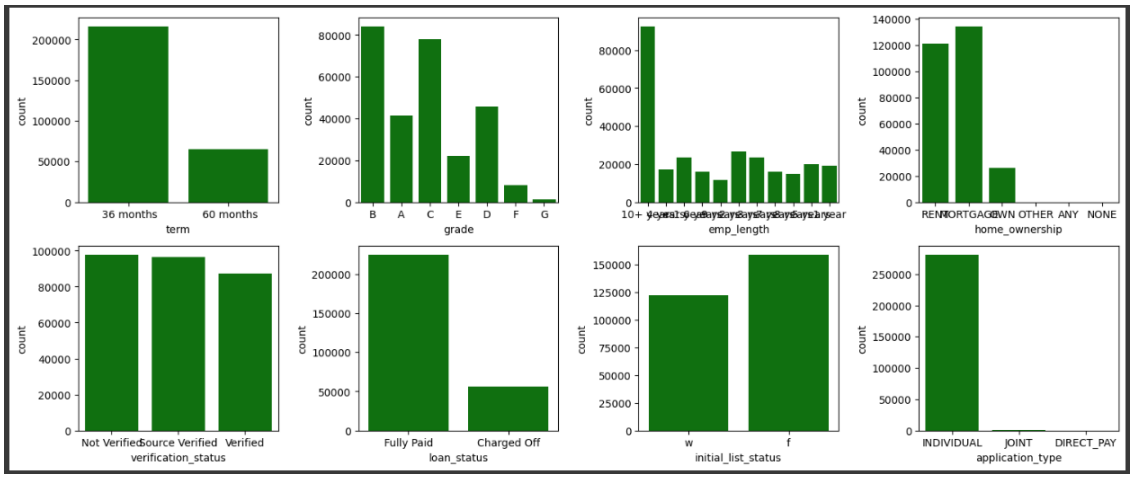
Jika ada *missing valu*e melebihi 50%, maka fitur atau kolom akan dihapus. Lalu dilanjutkan dengan menghapus baris data yang memiliki *missing value* nilai fitur. pada penelitian kali ini terdapat missing value pada dataset di beberapa kolom dengan tingkat missing value yang kecil yaitu kurang dari 10% seperti yang ada pada tabel 2. maka langsung dilakukan penghapusan bagi baris yang memiliki missing values.

Proses Outliers handling dilakukan karena dapat menyebabkan ketidakseimbangan atau ketidakberaturan dalam pola atau hubungan data. Dalam upaya menjaga kebersihan dan konsistensi data, penghapusan outliers dapat membantu dalam membangun dataset yang lebih terstruktur dan representatif.

Sebelum dilakukan nya data cleaning, dataset “lending club” memiliki 24 kolom dan 396029 baris. Kemudian setelah dilakukan data cleaning dataset berkurang menjadi 20 kolom dan 324570 baris.

2.2 Data *Manipulation* (*Encoding*)

Setelah dilakukan pembersihan data, selanjutnya dilakukan Data Manipulation (encoding), Donald D. Knuth, seorang ilmuwan komputer terkenal, mendefinisikan encoding sebagai proses pengubahan data ke dalam bentuk yang dapat diolah oleh komputer.



Gambar 5. Distribusi Tipe Data Kategori

Pada tipe data kategorikal seperti pada Gambar 5. diperlukan penanganan berupa encoding data, data kategorikal menjadi data numerik(Duan, 2019). Kolom ‘terms’, ‘grade’, ‘emp_length’, “home_ownership”, “verification_status” “loan_status”, “initial_list_status” dan “application_type” akan dikonversi dengan encoding dimana nilai pada kolom tersebut diubah menjadi numerik.

Contoh pada penelitian ini untuk kolom target yaitu loan_status, yang terdapat dua kategori yaitu 'Fully Paid' dan 'Charged Off', maka dapat dilakukan encoding dengan memberikan nilai 0 untuk 'Fully Paid' dan nilai 1 untuk 'Charged Off'. Hal yang serupa dilakukan untuk kolom-kolom lainnya, di mana setiap kategori akan diberikan nilai numerik yang sesuai.

Penerapan encoding ini memungkinkan algoritma machine learning untuk memproses dan memahami data kategorikal. Namun, penting untuk memilih metode encoding yang sesuai tergantung pada jenis data dan konteks penelitian.

3. Perancangan Model

3.1 Pembagian Dataset

Dalam rangka mengevaluasi kinerja algoritma, digunakan teknik train-test split menggunakan library sklearn. Dataset dibagi secara acak menjadi dua subset data. Pertama, dataset digunakan sebagai data training (fit model) sebesar 70%. Kedua, dataset digunakan sebagai data uji (test) sebesar 30%.

Source code yang digunakan :

```
from sklearn.model_selection import train_test_split
x = df.drop(["loan_status"],axis=1)
y = df["loan_status"]
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=42)
```

Selain itu, digunakan juga MinMaxScaler dari sklearn.preprocessing untuk melakukan penskalaan fitur dalam dataset.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
```

3.2 Pembentukan Model Logistic Regression

Model *Logistic Regression* dari sklearn.linear_model digunakan untuk membangun model regresi logistik.

Source code yang digunakan :

```
from sklearn.linear_model import LogisticRegression
model=LogisticRegression()
model.fit(x_train,y_train)
pred_lr = model.predict(x_test)
```

3.3 Pembentukan Model *Random Forest*

Model *Random Forest Classifier* dari sklearn.ensemble digunakan untuk membangun model random forest.

Source code yang digunakan :

```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(max_depth=6)

model.fit(x_train,y_train)

pred_lr = model.predict(x_test)
```

3.4 Pembentukan Model Evaluasi

Setelah model-model tersebut dibuat, langkah selanjutnya adalah melakukan evaluasi kinerja model. Terdapat berbagai metrik yang dapat digunakan untuk mengukur kinerja model, seperti AUC, RMSE, MSE, ROC, dan beberapa metrik tambahan seperti F-*score*, *recall*, dan *precision*.

Namun, dalam penelitian ini, penilaian model akan berfokus pada matrik *Accuracy* sebagai metrik utama, yang mengukur sejauh mana model dapat memprediksi secara benar terhadap data yang diberikan. Selain itu, beberapa metrik tambahan seperti F-*score*, *recall*, dan *precision* juga akan digunakan untuk memberikan gambaran yang lebih lengkap tentang kinerja model dalam mengklasifikasikan *bad loan* dan *good loan*.

Penggunaan modul sklearn memudahkan perhitungan metrik-metrik tersebut, sehingga peneliti dapat secara efisien mengevaluasi dan membandingkan kinerja model *Logistic Regression* dan *Random Forest* berdasarkan metrik-metrik yang relevan.

3.4.1 Model Evaluasi Logistic Regression :

Untuk model Regresi Logistik, nilai prediksi diperoleh menggunakan fungsi predict(). Untuk mengevaluasi kinerja model, dihitung skor akurasi dan laporan klasifikasi (classification report).

Source code yang digunakan :

```
from sklearn.metrics import accuracy_score


accuracy_lr= accuracy_score(pred_lr,y_test)*100
```

```
print('Accuracy of Regularized Logistic Regression is: {:.2f}'.format(accuracy_lr))
from sklearn.metrics import classification_report
print(classification_report(y_test, pred_lr))
```

3.4.2 Model Evaluasi Random Forest :

Untuk model Random Forest, juga dihitung skor akurasi dan laporan klasifikasi.

Source code yang digunakan :

```
from sklearn.metrics import accuracy_score
accuracy= accuracy_score(pred_lr,y_test)*100
print('Accuracy of RandomForestClassifier: {:.2f}'.format(accuracy))
from sklearn.metrics import classification_report
print(classification_report(y_test, pred_lr))
```

Dalam penelitian ini, untuk membangun model regresi logistik dan random forest, digunakan modul sklearn, yang merupakan salah satu pustaka populer dalam bahasa pemrograman Python untuk machine learning dan data science.

4. Hasil Pengujian

Hasil pengujian dinyatakan dalam Tabel 3 dan 4, menunjukkan bahwa dengan menggunakan model Logistic Regression dan Random Forest, dihasilkan accuracy terbaik pada model Logistic Regression yaitu 80,37%, sedangkan Random Forest yaitu 80.23%.

Tabel 3. Hasil Model Logistic Regression

	precision	recall	f1-score	support
(“Good Loan”) 0	0.81	0.98	0.89	78019
(“Bad loan”) 1	0.54	0.09	0.15	19352

accuracy			**80.37**	97371
macro avg	0.67	0.54	0.52	97371
weighted avg	0.76	0.80	0.74	97371

Tabel 4. Hasil Model Random Forest

	precision	recall	f1-score	support
(“Good Loan”) 0	0.80	1.00	0.89	78019
(“Bad loan”) 1	0.68	0.01	0.02	19352
accuracy			**80.23**	97371
macro avg	0.73	0.54	0.46	97371
weighted avg	0.78	0.80	0.72	97371

Dari hasil pengujian evaluasi model Logistic Regression dan Random Forest yang diberikan pada tabel 3 dan 4, terlihat beberapa metrik evaluasi klasifikasi seperti presisi (precision), sensitivitas (recall), skor F1 (F1-score), dan jumlah sampel (support) untuk kedua kelas (0 dan 1). Berikut adalah interpretasi dari masing-masing metrik:

Precision (presisi) mengukur sejauh mana hasil positif yang diprediksi benar.

- Pada model Logistic Regression untuk kelas 0, memiliki nilai precision sebesar 0,81, yang berarti 81% dari prediksi positif untuk kelas 0 adalah benar. Sedangkan untuk kelas 1, memiliki nilai precision sebesar 0,54, yang berarti 54% dari prediksi positif untuk kelas 1 adalah benar.
- Pada model Random Forest : Untuk kelas 0, memiliki nilai precision sebesar 0,80, yang berarti 80% dari prediksi positif untuk kelas 0 adalah benar. Sedangkan untuk kelas 1, memiliki nilai precision sebesar 0,67, yang berarti 67% dari prediksi positif untuk kelas 1 adalah benar.

Recall (sensitivitas) mengukur sejauh mana model dapat mengidentifikasi dengan benar semua sampel positif yang ada.

- Pada model Logistic Regression : untuk kelas 0, memiliki nilai recall sebesar 0,98, yang berarti model mampu mengidentifikasi 98% dari semua sampel positif kelas 0. Namun, untuk kelas 1, recall hanya sebesar 0,09, yang berarti model hanya mampu mengidentifikasi 9% dari semua sampel positif kelas 1.
- Pada model Random Forest : untuk kelas 0, memiliki nilai recall sebesar 1,00, yang berarti model mampu mengidentifikasi 100% dari semua sampel positif kelas 0. Namun, untuk kelas 1, recall hanya sebesar 0,01, yang berarti model hanya mampu mengidentifikasi 1% dari semua sampel positif kelas 1.

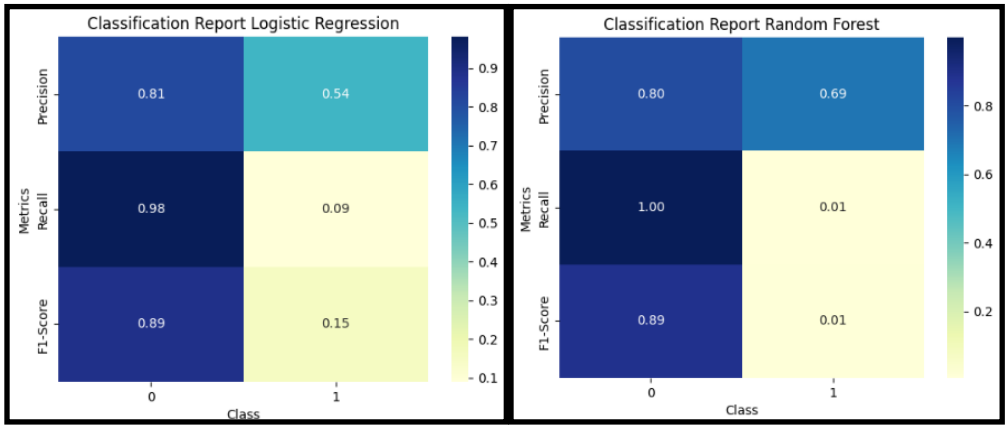
F1-score adalah metrik yang menggabungkan precision dan recall menjadi satu angka yang merepresentasikan keseimbangan antara kedua matriks tersebut.

- Pada model Logistic Regression : F1-score kelas 0 adalah 0,89, sedangkan untuk kelas 1 hanya 0,15. F1-score yang lebih tinggi menunjukkan kinerja yang lebih baik dalam memprediksi kelas yang relevan.
- Pada model Random Forest : F1-score kelas 0 adalah 0,89, sedangkan untuk kelas 1 hanya 0,02. F1-score yang lebih tinggi menunjukkan kinerja yang lebih baik dalam memprediksi kelas yang relevan.

Support mengindikasikan jumlah sampel yang termasuk dalam setiap kelas.

- Pada model Logistic Regression dan Random forest memiliki nilai support yang sama yaitu terdapat 78.019 sampel, sedangkan untuk kelas 1 hanya terdapat 19.352 sampel.

Dalam evaluasi keseluruhan (macro avg dan weighted avg), rata-rata precision dan recall. (kelas 1) memiliki nilai yang lebih rendah dibandingkan dengan kelas mayoritas (kelas 0), karena kinerja yang buruk pada kelas minoritas (kelas 1). Hal ini menunjukkan adanya ketidakseimbangan kelas dalam dataset.



Gambar 6. Hasil Visual Logistic Regression dan Random Forest.

Pada evaluasi akurasi (accuracy), terdapat perbedaan accuracy antara Logistic Regression (80,37%) dan Random Forest (80,23%) pada penelitian tersebut sebenarnya sangat kecil. Meskipun secara numerik Logistic Regression memiliki tingkat akurasi yang sedikit lebih tinggi, perbedaan tersebut tidak signifikan.

Dalam konteks ini, tidak dapat diambil kesimpulan definitif tentang model yang lebih baik hanya berdasarkan perbedaan akurasi yang kecil. Sebagai gantinya, perlu mempertimbangkan faktor lain seperti precision, recall, dan F1-score untuk kedua model serta konteks dan tujuan dari analisis klasifikasi yang dilakukan.

Dari perbedaan hasil akurasi yang kecil ini, dapat disimpulkan bahwa kinerja kedua model dalam memprediksi risiko pinjaman memiliki performa yang serupa. Namun, masih diperlukan evaluasi lebih lanjut dan pemilihan metrik yang tepat untuk mengukur kinerja dan kecocokan model dengan konteks dan tujuan penelitian secara lebih komprehensif.

KESIMPULAN

Dalam evaluasi model Logistic Regression dan Random Forest, terlihat adanya ketidakseimbangan kelas dalam dataset, di mana kelas mayoritas (kelas 0) memiliki kinerja yang lebih baik daripada kelas minoritas (kelas 1). Hal ini tercermin dari nilai presisi, sensitivitas, dan skor F1 yang lebih rendah untuk kelas 1 dibandingkan dengan kelas 0.

Pada evaluasi akurasi (accuracy), terdapat perbedaan accuracy antara Logistic Regression (80,37%) dan Random Forest (80,23%) pada penelitian tersebut sebenarnya sangat kecil. Meskipun secara numerik Logistic Regression memiliki tingkat akurasi yang sedikit lebih tinggi, perbedaan tersebut tidak signifikan secara praktis.

Meskipun kedua model memiliki akurasi yang serupa, matriks evaluasi yang lain seperti precision, recall, dan F1-score pada segi makro rata-rata (macro avg) menunjukkan keseimbangan yang rendah. Hal ini menunjukkan adanya kesulitan dalam memprediksi dengan baik kedua kelas, baik kelas mayoritas maupun kelas minoritas.

Dari perbedaan hasil akurasi yang kecil ini, dapat disimpulkan bahwa kinerja kedua model dalam memprediksi risiko pinjaman memiliki performa yang serupa. Namun, masih diperlukan evaluasi lebih lanjut dan pemilihan metrik yang tepat untuk mengukur kinerja dan kecocokan model dengan konteks secara lebih komprehensif.

Juga perlu diingat bahwa evaluasi performa model tidak hanya bergantung pada metrik evaluasi yang digunakan, tetapi juga harus mempertimbangkan konteks dan tujuan analisis klasifikasi, serta memperhatikan ketidak seimbangan kelas dalam dataset.

SARAN

Ketika melakukan evaluasi performa model, perlu mempertimbangkan matrik evaluasi yang relevan dan sesuai dengan konteks dan tujuan analisis klasifikasi. Dalam kasus ini,

penggunaan metrik seperti presisi, sensitivitas, dan skor F1 memberikan informasi yang berguna untuk memahami performa model dalam memprediksi kelas yang relevan.

Selain itu, penting juga untuk memperhatikan ketidakseimbangan kelas dalam interpretasi hasil evaluasi. Dalam kasus ini, kelas mayoritas memiliki jumlah sampel yang jauh lebih besar daripada kelas minoritas, sehingga evaluasi performa model lebih dipengaruhi oleh kelas mayoritas dalam metrik rata-rata tertimbang (weighted avg).

DAFTAR PUSTAKA

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. Risks, 6(2), 1–20. https://doi.org/10.3390/risks6020038

Duan, J. (2019). Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. Journal of the Franklin Institute, 356(8), 4716–4731. https://doi.org/10.1016/j.jfranklin.2019.01.046 Economics, A., & Xxvi, V. (2019). A Deep Neural Network (DNN) based classification model in application to loan default prediction. Theoretical and Applied Economics, XXVI(4), 75–84

Amelia, R., Widyawati, S., & Sulistyorini, V. (2018). Analisis Risiko Kredit pada Pemberian Pinjaman dengan Pendekatan Credit Scoring. Jurnal Keuangan dan Perbankan, 22(2), 201-212.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Zhang, L., Zhu, Z., & Yao, Q. (2017). Credit Risk Evaluation of P2P Lending in China: A Machine Learning Approach. International Journal of Services Operations and Informatics, 8(2), 123-144.

Chen, Y., Wang, G., Li, S., & Zhang, B. (2018). Loan Classification Using Machine Learning Techniques for Peer-to-Peer Lending Platforms. Mathematical Problems in Engineering, 2018.

Li, J., Zhang, X., Li, G., & Zhang, S. (2019). Predicting Loan Default in P2P Lending with Machine Learning Techniques. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4854-4863). IEEE.

Prabhakaran, S. (2017, September 13). Logistic Regression – A Complete Tutorial With Examples in R. Retrieved from https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/

Deloitte. (2017). Fintech by the numbers. [Online] Available: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-fsi-gx-2017-fintech-by-numbers.pdf