Penerapan Metode *Random Forest* untuk Analisis Risiko pada dataset *Peer to peer lending*

http://dx.doi.org/10.28932/jutisi.v6i3.2890

Erick Renata[#1], Mewati Ayub ✉[#2]

[#] *Program Studi Magister Ilmu Komputer, Universitas Kristen Maranatha*
Jl. Surya Sumantri No.65, Bandung
[1]erickrenata.maranatha@gmail.com
[2]mewati.ayub@it.maranatha.edu

Abstract — **Peer to peer lending (P2PL) is one of financial technology (fintech) that develops very fast in society. On the other side, the P2PL project has many risks. The risk of the P2PL project can be analyzed using classification. There are two conditions of a loan, namely a good loan and a bad loan. This study uses two methods to analyze a P2PL dataset, which are Random Forest method and Logistic Regression method. Data is taken from P2PL loan dataset provided by Data World, which contains 887.379 entries with 74 features. The result of experiments is a model that can be used to predict and classify a P2PL loan as a good or bad one.**

Keywords— ***Fintech*; Logistic Regression; Peer to Peer Lending; Random forest**

I. PENDAHULUAN

Financial Technologi atau *Fintech* merupakan revolusi perkembangan teknologi dibidang ekonomi atau keuangan [1]. Setiap perusahaan *fintech* memiliki fokus yang berbeda-beda seperti; Pembayaran (*Payments*), Peminjaman (*Lending*), Perencanaan keuangan (*Personal Fiance*), Investasi Ritel, Pembiayaan (*Crowdfunding*), Remitasi, dan Riset Keuangan [2]. Saat ini jenis *Peer to peer lending* atau selanjutnya kita sebut P2PL merupakan salah satu *fintech* yang banyak dilirik oleh masyarakat. P2PL merupakan Layanan pinjam meminjam uang berbasis teknologi informasi, dimana proses transaksi menggunakan aplikasi sehingga peminjam dan pemberi pinjaman tidak bertemu secara langsung.

Mudahnya proses peminjaman uang dengan tanpa harus memiliki rekening bank dan peminjaman tanpa anggunan menjadi daya tarik yang membuat P2PL menjadi cepat peningkatannya di kalangan masyarakat. Regulasi khusus untuk P2PL sudah ditetapkan oleh Otoritas Jasa Keuangan yang selanjutnya akan disebut OJK melalui Peraturan Otoritas Jasa Keuangan Nomor 77/ P.OJK.01/ 2016 tentang Layanan Pinjam Meminjam Uang Berbasis Teknologi Informasi.

Dalam pelaksanaannya, P2PL memiliki banyak risiko yang membuat suatu proyek transaksi dikatakan gagal, baik itu yang berasal dari pihak penyelenggara ataupun pihak pengguna. Risiko terbesar yang kerap terjadi adalah risiko terjadinya kredit macet bahkan gagal bayar oleh peminjam. Beberapa contoh faktor yang mempengaruhi keberhasilan proyek P2PL antara lain; tujuan pinjaman, usia peminjam, pekerjaan peminjam, serta pendapatan tiap bulan peminjam.

Analisis risiko P2PL dengan teknik *machine learning* dapat dilakukan untuk mengetahui risiko dalam pelaksanaan P2PL seperti yang dibahas dalam [1][4]. Hsueh [1] menyebutkan bahwa regulasi yang berlaku untuk P2PL masih kurang jelas sehingga menempatkan konsumen pada risiko. Hal tersebut didasari dari beberapa fakor risiko utama seperti: 1) manajemen di perusahaan P2PL yang tidak seketat management di bank; hal ini berpotensi meningkatkan risiko penipuan kredit, 2) Biaya operasi yang tinggi, dan industrinya yang sangat kompetitif membuat profitabilitas sulit; hal ini meningkatkan risiko bisnis. Akhirnya, Risiko pasar mengacu pada risiko yang berlebihan. Oleh karena itu, studi ini mengeksplorasi tiga risiko utama yang terkait dengan pinjaman P2PL yaitu: risiko kredit, risiko bisnis, dan risiko pasar. Dalam penelitian ini digunakan dataset *Lending Club* [3] untuk data pengujian P2PL. Prediksi untuk analisis dataset *Lending Club* dilakukan dengan menggunakan metode *Random forest* dan metode *Logistic Regression* sebagai pembanding.

Fokus yang menjadi permasalahan utama adalah metode manakah yang memiliki tingkat akurasi tertinggi dalam melakukan klasifikasi dataset pinjaman P2PL, dan bagaimana cara mendapatkan model yang dapat memprediksi pinjaman itu baik atau buruk berdasarkan dataset pinjaman P2PL yang digunakan.

Berdasarkan fokus permasalahan tersebut, tujuan dari penelitian yang dilakukan adalah melakukan studi komparatif terhadap metode yang dapat memberikan tingkat akurasi tertinggi dalam klasifikasi dataset pinjaman P2PL, serta mendapatkan model prediksi pinjaman (baik atau

buruk) berdasarkan dataset pinjaman P2PL yang digunakan. Manfaat penelitian adalah menemukan metode dengan tingkat akurasi tertinggi yang dapat digunakan dalam klasifikasi dataset pinjaman P2PL, serta memberikan model prediksi pinjaman (baik atau buruk) berdasarkan dataset pinjaman P2PL yang digunakan.

II. KAJIAN LITERATUR

A. *Peer to Peer Lending (P2PL)*

Financial Technology (*Fintech*) atau Teknologi Keuangan, merupakan model layanan keuangan baru yang dikembangkan melalui inovasi teknologi informasi [1]. Setiap perusahaan *fintech* memiliki fokus yang berbeda-beda seperti; Pembayaran (*Payments*), Peminjaman (*Lending*), Perencanaan keuangan (*personal finance*), Investasi Ritel, Pembiayaan (*crowdfunding*), Remitasi, dan Riset Keuangan [2].

B. *Machine Learning*

Machine learning dapat digunakan untuk membantu fase penilaian risiko, seperti identifikasi risiko, analisis risiko, dan evaluasi risiko. *Machine learning* dapat membuat suatu model yang mampu memberikan masukan untuk menggantikan teknik penilaian risiko yang selama ini dilakukan secara manual [4]. Penelitian ini berfokus dalam membuat model *machine learning* dalam analisis risiko khususnya risiko yang terdapat pada P2PL.

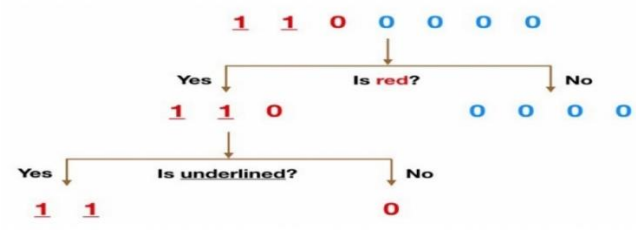
Machine Learning (ML) atau pembelajaran mesin merupakan salah satu cabang dari *Artificial Intelegence* (AI) yang mengembangkan aplikasi yang belajar dari data dan meningkatkan akurasinya dari waktu ke waktu tanpa perlu diprogram ulang. Dalam ML, algoritma dilatih untuk menemukan pola dan fitur dari sejumlah besar data untuk melakukan prediksi dan pengambilan keputusan berdasarkan data baru. Semakin banyak data, maka algoritma akan bekerja dengan lebih baik sehingga prediksi dan pengambilan keputusan akan lebih tepat [5].

Beberapa contoh teknik statistik yang banyak digunakan diantaranya: *Logistic Regression Analysis*, dan *Logistic Discriminant Analysis*. Sementara contoh teknik *Machine Learning* yang banyak digunakan diantaranya: *Support Vector Machnine* (SVM) dan *Random Forest*. Pada penelitian ini digunakan metode *Random Forest* sebagai metode utama dan metode *Logistic Regression* sebagai metode pembanding.

C. Metode *Random Forest*

Random Forest Algorithm merupakan salah satu algoritma *machine learning* yang digunakan sebagai klasifikasi. Klasifikasi merupakan bagian penting dari *machine learning* yang bertugas untuk mengelompokkan data bergantung pada kecenderungannya [7]. *Random Forest* pertama kali diinisias oleh Tin Kam Ho dan dikembangkan lebih lanjut oleh Leo Breiman dan Adele Cutler [8]. Kinerja *Random Forest* diadaptasi dari *decision tree*, dengan setiap *tree* dikembangkan dari sampel *bootstrap* berdasarkan data *training*.

Decision Tree merupakan bagian dari *Random Forest*, yaitu merupakan pohon logika untuk membedakan suatu data. Sebagai contoh, diberikan data yaitu dua angka 1 dan lima angka 0, dan setiap angka memiliki warnanya masing masing. Jika data ingin dikelompokkan, dan fitur yang diambil dari data tersebut adalah warna dan garis bawah, maka *tree* akan seperti Gambar 1.



Gambar 1 Cara kerja *Decission Tree*

Random Forest algorithm merupakan kumpulan dari *decision tree* yang beroperasi menjadi suatu gabungan fungsional. Setiap *decision tree* memiliki kesimpulan prediksi klasifikasi dan hasil prediksi akan digabungkan [6]. *Random Forest* a*lgorithm* memiliki hasil yang lebih baik dibandingkan dengan model individual lainnya dikarenakan *Random Forest* a*lgorithm* menggunakan *decision tree* yang tidak memiliki korelasi. Kesalahan prediksi dalam satu *decision tree* dapat ditutupi dengan kebenaran yang didapatkan daripada *decision tree* lainnya asalkan arah pembuatan *decision tree* benar.

Random Forest dipilih sebagai metode utama dalam memprediksi apakah suatu pinjaman itu baik atau tidak dikarenakan *Random Forest* berjalan efisien pada data yang jumlahnya banyak, dan dapat berjalan dengan baik dengan kelas yang populasinya tidak seimbang [7].

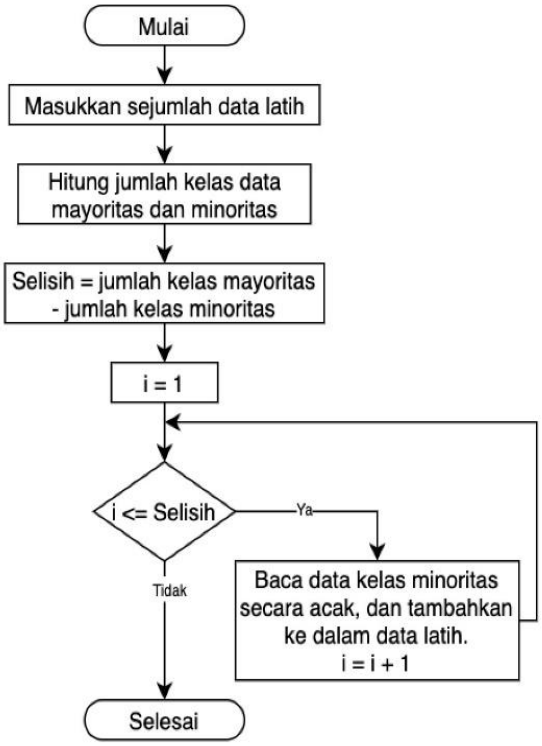
D. *Over-Under Sampling*

Klasifikasi pada data menggunakan metode *machine learning* memiliki kelemahan. Salah satunya adalah keadaan data yang tidak seimbang dimana model klasifikasi cenderung menihilkan peluang dari kelompok minoritas sehingga hasil prediksi akan cenderung kepada kategori mayoritas [8].

Dalam klasifikasi, data yang tidak seimbang terjadi jika banyaknya data yang tersedia untuk setiap kelas jauh berbeda. Jika terdapat dua kelas, data yang seimbang berarti pada setiap kelas tersedia sebanyak 50% data dari keseluruhan. Untuk sebagian besar metode *machine learning*, sedikit ketidakseimbangan data tidak menjadi masalah. Sebagai contoh, jika satu kelas mempunyai 60% data dan kelas lain 40%, maka hal tersebut tidak akan menyebabkan penurunan kinerja secara signifikan. Kondisi tidak seimbang yang tinggi muncul jika terdapat 90% data untuk satu kelas, dan hanya 10 % untuk kelas yang lain [19].

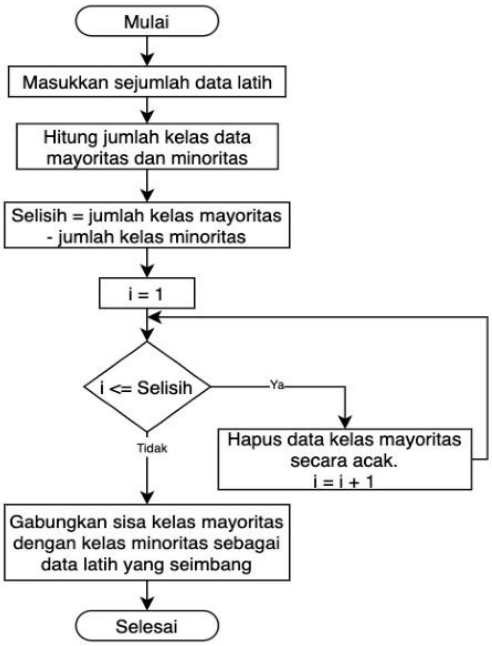
Jika bekerja menggunakan data yang tidak seimbang, maka hasil akurasi data tidak baik karena algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas [9]. Untuk mengatasi permasalahan tersebut, digunakan 2 macam teknik yaitu menggunakan SMOTE (*Synthetic Minority OverSampling Technique)* dan *Random Under Sampling Technique*.

SMOTE merupakan pendekatan dengan sistem kerja *over-sampling* pada kelas minoritasnya. Cara kerja SMOTE digambarkan menggunakan *flowchart* pada Gambar 2. Pertama dihitung selisih antara kelas mayoritas dengan kelas minoritas. Kemudian dilakukan perulangan sebanyak hasil penghitungan selisih sambil membaca data kelas minoritas [10].



Gambar 2 *Flowchart* SMOTE

Teknik *Under sampling* digambarkan dengan *flowchart* pada Gambar 3. Hampir sama dengan SMOTE, pertama dihitung selisih antara kelas mayoritas dengan kelas minoritas. Kemudian dilakukan perulangan sebanyak hasil penghitungan selisih kelas mayoritas dan minoritas. Selama perulangan, data kelas mayoritas dihapus secara acak, sehingga jumlah kelas mayoritas sama dengan jumlah kelas minoritas [10].



Gambar 3 *Flowchart Under Sampling*

E. Penelitian Terkait

Menurut Hsueh [1] revolusi perkembangan transaksi keuangan terjadi dengan sangat pesat karena dampak dari revolusi *Financial Technology* (*Fintech*). *Fintech* atau teknologi finansial adalah model layanan keuangan baru yang dikembangkan melalui inovasi teknologi informasi. Salah satu *Fintech* yang memiliki peningkatan yang pesat di kalangan masyarakat adalah *Peer to Peer Lending* (P2PL) atau kegiatan pinjam meminjam uang melalui media elektronik.

Dalam kegiatan P2PL sedikitnya terdapat 3 pihak terkait agar terlaksana, yaitu konsumen sebagai investor, konsumen sebagai penerima pinjaman, dan pihak penyelenggara. Namun dalam pelaksanaannya menurut Putri [11], P2PL dapat memungkinkan terjadinya wanprestasi atau kegagalan dalam proyek. Hal ini harus diperhatikan oleh konsumen mengenai regulasi yang dilakukan pihak penyelenggara selaku penanggung jawab kegiatan P2PL sebelum menyetujui dan turut serta dalam suatu proyek baik itu konsumen sebagai pemberi pinjaman/ investor ataupun konsumen sebagai penerima pinjaman.

Klasifikasi data dapat dilakukan guna memprediksi risiko yang terjadi agar dapat dijadikan pertimbangan sebelum turut serta dalam suatu proyek P2PL. Primajaya dan Sari [12] mengusulkan dilakukan klasifikasi data guna menemukan model prediksi suatu kasus menggunakan metode *Random Forest*. Metode ini dipilih karena dapat memberikan akurasi yang baik dalam klasifikasi dan menghasilkan kesalahan yang lebih rendah dibandingkan metode lain. Penelitiannya dilakukan untuk menemukan model prediksi hujan.

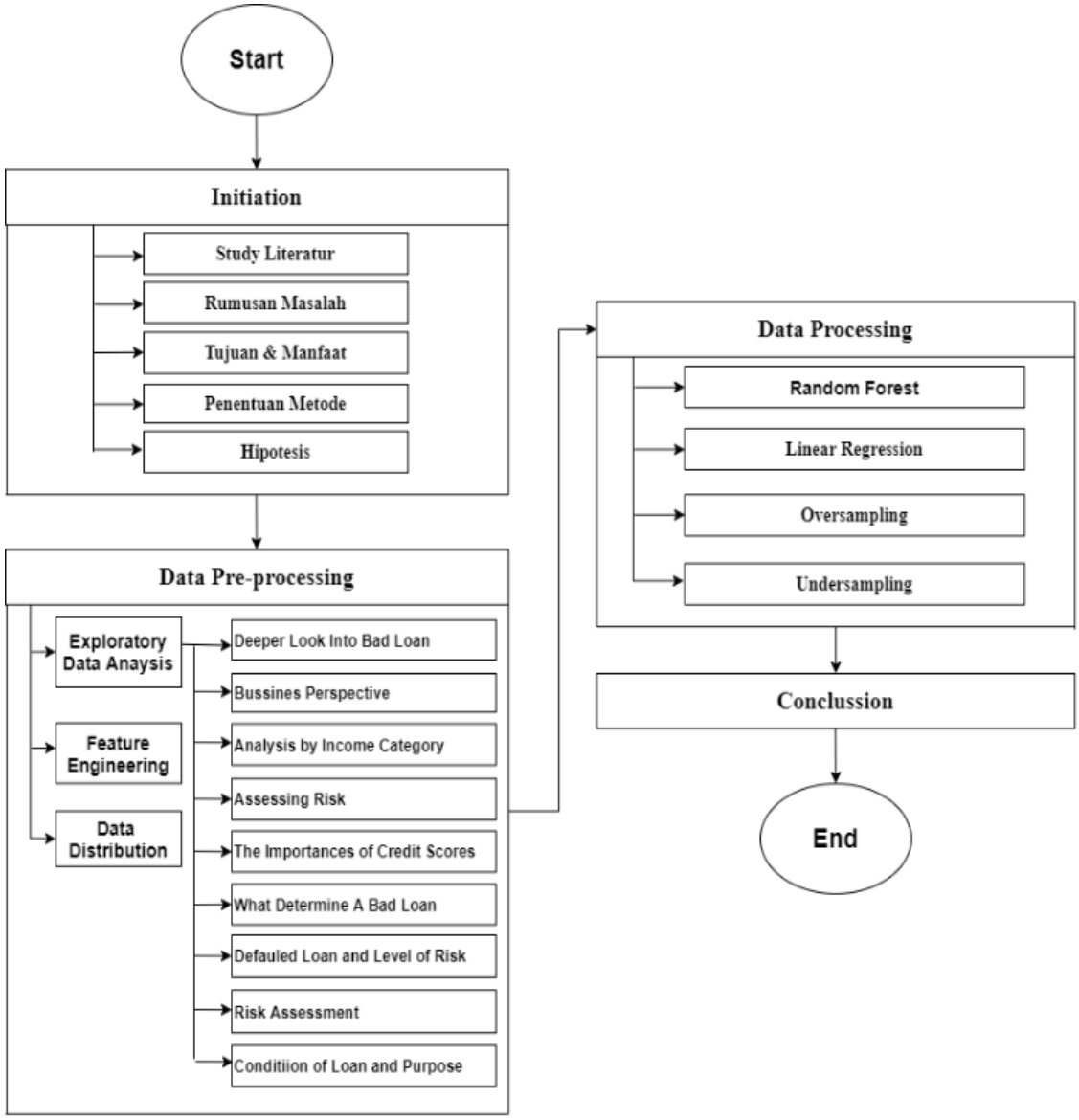
Namun menurut Sulistyowati dan Jajuli [9], klasifikasi pada data menggunakan metode *machine learning* memiliki

kelemahan. Salah satunya adalah keadaan dimana data kelas tidak seimbang. Sejalan dengan itu, Putri [13] mengemukakan bahwa bekerja menggunakan data yang tidak seimbang membuat hasil akurasi data tidak baik karena algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas. Untuk mengatasi permasalahan tersebut, digunakan dua macam teknik yaitu menggunakan SMOTE dan *Random Under Sampling Technique*.

III. METODOLOGI PENELITIAN

A. *Research design*

Dalam penelitian ini terdapat beberapa tahapan penelitian yang dilakukan diantaranya inisiasi, EDA (*Exploratory Data Analysis*), *feature engineering, data preprocessing*, *data distribution*, *training model*, evaluasi model, analisis, *oversampling/ undersampling*, kemudian yang terakhir adalah kesimpulan [14]. Untuk lebih jelasnya tahapan penelitian dapat dilihat pada Gambar 4.



Gambar 4 *Research Design*

B. *Proses Analisis Risiko P2PL*

Setelah pengumpulan data, dilakukan EDA untuk mendapatkan informasi-informasi baru yang terkandung dalam dataset pinjaman seperti definisi lebih lanjut tentang *bad loan* dan *good loan*, perspektif bisnis berdasarkan negara peminjam, analisis berdasarkan penghasilan peminjam, dan lain-lain. Dari informasi yang berhasil dikumpulkan dalam proses EDA, selanjutnya dilakukan *Feature Engineering*, yaitu proses pembobotan fitur yang menentukan fitur mana yang penting dan perlu diberikan

bobot lebih. Tahap berikutnya yaitu distribusi data, dataset dibagi menjadi dua yaitu *data training* dan *data test*, sebesar 80% untuk *data training* dan 20% untuk *data test*. Selanjutnya dilakukan analisis menggunakan metode *Random Forest* sebagai metode utama dan metode *Logistic Regression* sebagai metode pembanding. Menurut Putri [13], Metode *Random Forest* memiliki kelemahan salah satunya adalah keadaan dimana data kelas tidak seimbang yang dapat mempengaruhi hasil akurasi data. Oleh karena itu dilakukan juga teknik *oversampling* dan *undersampling* pada metode *Random Forest* dan metode *Logistic Regression.*

IV. DATA YANG DIPAKAI

Data yang digunakan pada penelitian ini adalah dataset yang didapat dari *Data World* [3] tentang pinjaman *Lending Club (LC)* dari tahun 2007-2015. Dataset tersebut terdiri dari 887.379 data pinjaman yang memiliki 25 kolom fitur. Beberapa sampel data dari dataset pinjaman dapat dilihat pada Tabel IA, IB, dan IC. Penjelasan tentang tiap kolom fitur pada Tabel IA, IB, dan IC dapat dilihat pada Tabel II.

TABEL I A
SAMPEL DATASET PINJAMAN P2PL

id	Member_id	loan_amnt	funded_amnt	funded_amnt_int	term	int_rate	installment	grade	sub_ grade
1070078	1305201	6500.0	6500.0	6500.0	60 months	14.65	153.45	C	C3
1069908	1305008	12000.0	12000.0	12000.0	36 months	12.69	402.54	B	B5
1064687	1298717	9000.0	9000.0	9000.0	36 months	13.49	305.38	C	C1
1069866	1304956	3000.0	3000.0	3000.0	36 months	9.91	96.68	B	B1
1069057	1303503	10000.0	10000.0	10000.0	36 months	10.65	325.74	B	B2

TABEL I B
SAMPEL DATASET PINJAMAN P2PL

emp_title	emp-length	home-oenership	annual_inc	verification_ status	issue_d	loan_status
Southwest Rural metro	5 years	OWN	72000.0	Not Verified	Dec-2011	Fully Paid
UCLA	10+ years	OWN	75000.0	Source Verified	Dec-2011	Fully Paid
Va. Dept of Conservation/Recreation	< 1 years	RENT	30000.0	Source Verified	Dec-2011	Charged off
Target	3 years	RENT	15000.0	Source Verified	Dec-2011	Fully Paid
SFMTA	3 years	RENT	10000.0	Source Verified	Dec-2011	Charged off

TABEL I C
SAMPEL DATASET PINJAMAN P2PL

pymnt _plan	url	desc	purpose	title	zip_ code	addr_ state	dtl
n	https://ww..	Borrowet added on 12/15/11>I had received a…	debt_ consolidation	High intrest Consolidation	853xx	AZ	16.12
n	https://ww..	NaN	debt_ consolidation	Consolidation	913xx	CA	10.78
n	https://ww..	Borrowet added on 12/15/11>Plan to pay off…	debt_ consolidation	Freedom	245xx	VA	10.08
n	https://ww..	Borrowet added on 12/15/11>I intend to pay…	credit_card	Citicard fund	606xx	IL	12.56
n	https://ww..	NaN	other	Other Loan	951xx	CA	7.06

TABEL III
DESKRIPSI DATASET

No	Fitur	Tipe Data	Keterangan
1	id	int64	Identitas unik dari tiap data pinjaman LC
2	member_id	int64	Identitas unik untuk anggota peminjam pada LC
3	loan_amnt	float64	Jumlah pinjaman yang diajukan oleh peminjam
4	funded_amnt	float64	Jumlah total yang diberikan untuk pinjaman tersebut
5	funded_amnt_inv	float64	Jumlah total yang dilakukan oleh investor untuk pinjaman tersebut
6	term	object	Jangka waktu pinjaman. Nilainya dalam bulan dan bisa 36 bulan atau 60 bulan
7	int_rate	float64	Suku bunga pinjaman
8	installment	float64	Pembayaran bulanan terhutang oleh peminjam
9	grade	object	Tingkat pinjaman yang ditetapkan LC
10	sub_grade	object	Sub tingkat pinjaman yang ditetapkan LC
11	emp_title	object	Judul pekerjaan yang diberikan oleh Peminjam saat mengajukan pinjaman.
12	emp_length	object	Lama kerja dalam tahun. Nilai yang mungkin antara 0 dan 10 di mana 0 berarti kurang dari satu tahun dan 10 berarti sepuluh tahun atau lebih.
13	home_ownership	object	Status kepemilikan rumah yang diberikan oleh peminjam pada saat pendaftaran. Nilai-nilainya adalah: RENT, OWN, MORTGAGE, LAIN-LAIN.
14	annual_inc	float64	Pendapatan tahunan yang dilaporkan sendiri yang diberikan oleh peminjam selama pendaftaran.
15	verification_status	object	Menunjukkan apakah pendapatan diverifikasi oleh LC, tidak diverifikasi, atau jika sumber pendapatan diverifikasi
16	issue_d	object	Bulan dimana pinjaman itu didanai
17	loan_status	object	Status pinjaman saat ini
18	pymnt_plan	object	Menunjukkan apakah rencana pembayaran telah disiapkan untuk pinjaman
19	url	object	Link detail pinjaman yang diajukan peminjam
20	desc	object	Deksripsi lengkap dari tiap pinjaman yang diajukan peminjam
21	purpose	object	Kategori yang disediakan oleh peminjam untuk permintaan pinjaman.
22	title	object	Judul pinjaman yang diberikan oleh peminjam
23	zip_code	object	3 angka pertama dari kode pos yang diberikan oleh peminjam dalam aplikasi pinjaman.
24	addr_state	object	Negara yang disediakan oleh peminjam dalam aplikasi pinjaman
25	dti	float64	Rasio yang dihitung menggunakan total pembayaran hutang bulanan peminjam pada total kewajiban hutang, tidak termasuk penggadaian dan pinjaman LC yang diminta, dibagi dengan pendapatan bulanan yang dilaporkan sendiri dari peminjam.

V. HASIL ANALISIS

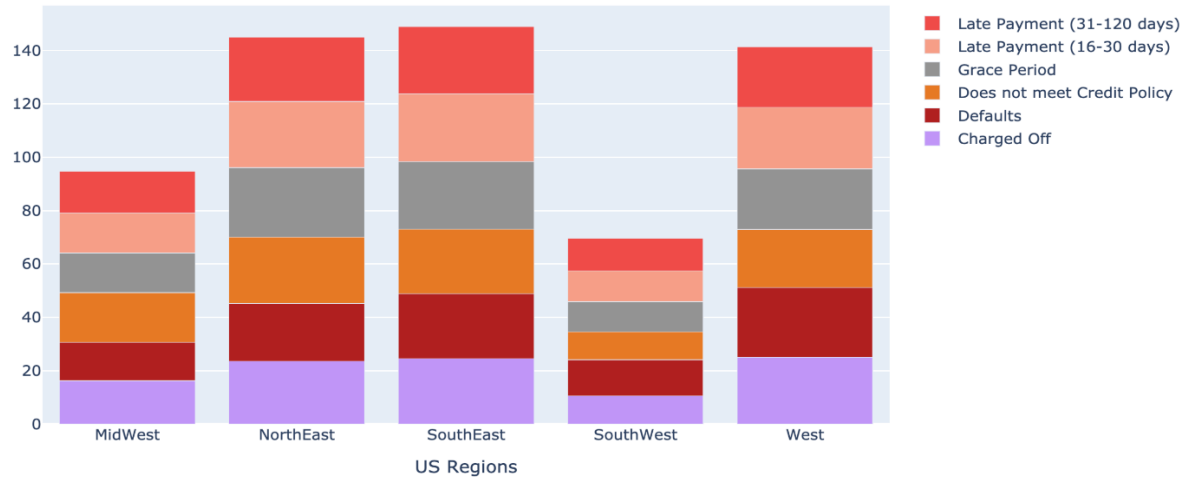
A. *Exploratory Data Analysis (EDA)*

EDA merupakan tahap awal dalam melakukan pengolahan data dimana pada tahap ini dilakukan eksplorasi mendalam terhadap dataset pinjaman untuk mendapatkan informasi baru apa saja yang dapat digunakan untuk tahap selanjutnya [14].

1) Deeper Look Into Bad Loan: Klasifikasi jumlah pinjaman yang termasuk kedalam kategori kredit macet untuk setiap wilayah berdasarkan status pinjamannya dilakukan untuk memudahkan analisis tingkat risiko di wilayah tertentu.

Dari Gambar 5 dapat diamati bahwa wilayah *West* dan *SouthEast* memiliki persentase lebih tinggi di sebagian besar status pinjaman "buruk" dan wilayah *NorthEast* memiliki persentase lebih tinggi dalam *Grace Period* dan Tidak memenuhi Kebijakan Kredit status pinjaman. Namun, keduanya tidak dianggap seburuk macet/*default.*

p-ISSN : 2443-2210
e-ISSN : 2443-2229

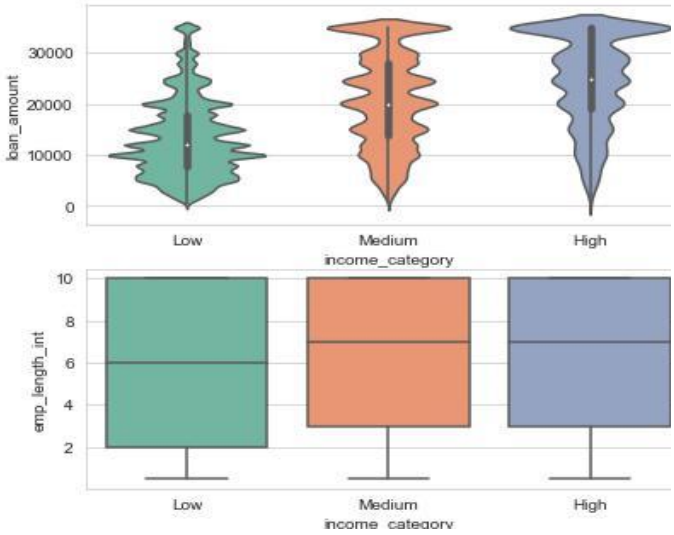


Gambar 5 Status *Bad Loan* Berdasarkan Region/ Negara Bagian

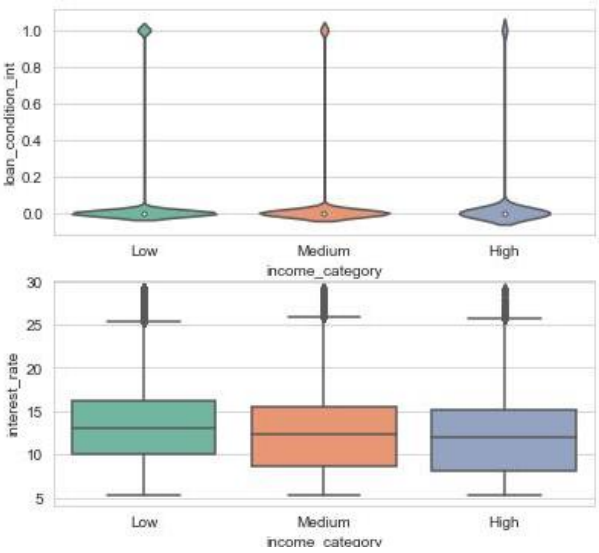
Dari Gambar 5 dapat disimpulkan bahwa wilayah *West* dan *SouthEast* memiliki status pinjaman yang paling tidak diinginkan, dengan persentase yang sedikit lebih tinggi dibandingkan dengan Wilayah *NorthEast.*

2) *Analysis by Income Category:* Klasifikasi dilakukan dengan cara dibagi menjadi 3 kriteria, yaitu: Berpenghasilan rendah, berpenghasilan menengah, dan berpanghasilan tinggi. Peminjam yang memiliki pendapatan tahunan lebih rendah atau sama dengan 100.000 USD termasuk ke dalam kategori berpenghasilan rendah. Peminjam yang memiliki pendapatan tahunan lebih dari 100.000 USD termasuk ke dalam kategori berpenghasilan rendah, sementara peminjam yang memiliki pendapatan tahunan lebih dari 200.000 USD termasuk ke dalam kategori berpenghasilan tinggi.

Plotting peminjam berdasarkan pinjaman dapat dilihat pada Gambar 6A dan Gambar 6B.



Gambar 6A Plotting peminjam berdasarkan peminjaman



Gambar 6B. *Plotting* peminjam berdasarkan peminjaman

Dari Gambar 6A dapat dilihat bahwa peminjam yang menjadi bagian dari kategori berpenghasilan tinggi mengambil jumlah pinjaman yang lebih tinggi daripada peminjam dari kategori berpenghasilan rendah dan berpenghasilan menengah. Sejalan dengan itu, orang-orang dengan pendapatan tahunan yang lebih tinggi cenderung membayar pinjaman dengan jumlah yang juga lebih tinggi. Meski demikian rata-rata suku bunga untuk peminjam dengan kategori pendapatan lebih rendah memiliki rata-rata suku bunga lebih tinggi sementara peminjam dengan pendapatan tahunan lebih tinggi memiliki suku bunga lebih rendah pada pinjaman mereka (lihat Gambar 6B).

Pinjaman oleh kelompok dengan kategori berpenghasilan rendah memiliki resiko yang sedikit lebih tinggi untuk menjadi kredit macet dibandingkan pinjaman

oleh kelompok dengan kategori lain, seperti pada Gambar 6A. Sementara peminjam dengan pendapatan tahunan tinggi dan sedang memiliki panjang pekerjaan yang lebih lama daripada orang-orang dengan pendapatan lebih rendah, seperti pada Gambar 6B.

3) What Determine A Bad Loan: Dalam dataset yang digunakan terdapat fitur *Loan Status*. Fitur ini mendefinisikan status dari setiap pinjaman yang dapat dilihat pada Tabel III.

TABEL IIII
STATUS PINJAMAN

No	Status Pinjaman	Keterangan
0	*Fully Paid*	Pinjaman telah lunas.
1	*Charged Off*	Pinjaman yang tidak ada kemungkinan untuk pembayaran lebih lanjut
2	*Does not meet the credit policy. Status:Fully Paid*	Saat pinjaman dilunasi, pengajuan pinjaman berikutnya tidak lagi memenuhi kebijakan kredit dan tidak akan disetujui di pasar.
3	*Does not meet the credit policy. Status:Charged Off*	Sementara pinjaman dibatalkan, aplikasi pinjaman berikutnya tidak lagi memenuhi kebijakan kredit dan tidak akan disetujui di pasar.
4	*In Grace Period*	Pinjaman tersebut telah lewat jatuh tempo tetapi masih dalam masa tenggang 15 hari.
5	*Late (31-120 days)*	Pinjaman belum dibayar dalam 31 sampai 120 hari (terlambat untuk pembayaran saat ini).
6	*Late (16-30 days)*	Pinjaman belum dibayar dalam 16 sampai 30 hari (terlambat untuk pembayaran saat ini).
7	*Default*	Pinjaman macet dan tidak ada pembayaran yang dilakukan selama lebih dari 121 hari.

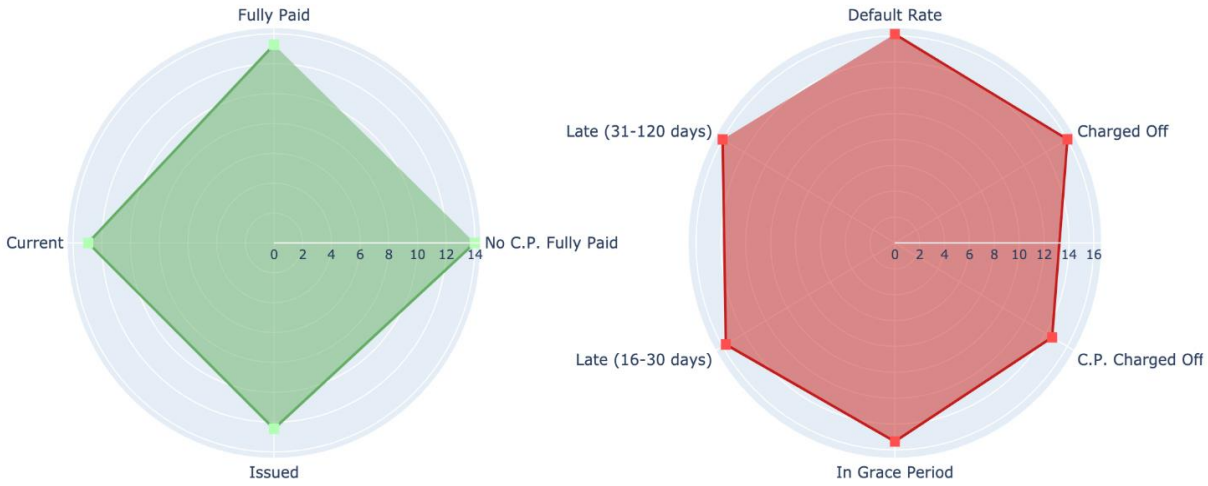
Pada bagian ini dilakukan analisis faktor-faktor utama yang menjadi penyebab suatu pinjaman dianggap sebagai Pinjaman buruk atau *bad loans.* Terdapat beberapa faktor yang dapat diperhatikan untuk melakukan analisis ririko suatu pinjaman agar tidak mengalami gagal bayar, diantara lain tingkat kredit yang rendah, persentase utang yang tinggi terhadap pendapatan, pendapatan tahunan rendah, suku bunga tinggi, kepemilikan rumah, dan tujuan peminjaman.

4) Risk Assessment: Pada bagian ini dilakukan pembandingan tingkat bunga rata-rata untuk status pinjaman milik masing-masing jenis pinjaman (pinjaman baik maupun pinjaman buruk) untuk melihat apakah ada perbedaan signifikan dalam rata-rata suku bunga untuk masing-masing kelompok.

Pada Gambar 7 dapat dilihat bahwa pinjaman buruk adalah status pinjaman dengan bunga berkisar 15-16% sementara Pinjaman baik merupakan status pinjaman dengan bunga antara 12 – 13%. Penilaian risiko lebih lanjut dibutuhkan karena tidak ada perbedaan yang signifikan dalam pembayaran bunga dari pinjaman baik maupun pinjaman buruk.

5) *Condition of Loan and Purpose:* Dalam analisis ini akan dilakukan analisis mengenai alasan peminjam mengajukan permohonan peminjaman untuk melihat apakan ada tujuan yang berkontribusi pada risiko pinjaman menjadi kredit macet/ pinjaman buruk atau tidak. Gambar 8 merupakan hasil analisis alasan pinjaman terhadap risiko pinjaman.

Pada Gambar 8, *Education*, *Renewable energy*, dan *Wedding* adalah alasan yang mengandung persen kredit macet tertinggi. Pada Gambar 9 terlihat perbedaan jumlah data yang cukup signifikan antara pinjaman baik dan pinjaman buruk. Hal ini akan berakibat pada pembayaran pinjaman di kemudian hari.



Gambar 7 Distribusi Status Pinjaman

p-ISSN : 2443-2210
e-ISSN : 2443-2229

Gambar 8 Alasan pinjaman terhadap status pinjaman

B. Feature Engineering

Feature engineering merupakan proses penentuan fitur dalam menentukan fitur mana yang penting dan perlu diberikan bobot lebih atau tidak [15]. Dalam penelitian ini, proses *feature engineering* dilakukan pada tahap *data preprocessing* dan juga tahap *data processing*. Berikut ini merupakan beberapa poin pembobotan yang dilakukan selama pengolahan data:

a. Pada tahap awal eksplorasi dilakukan pembuangan terhadap fitur yang tidak relevan *seperti id, member_id, emp_title, url, desc, zip_code, dan title.*

b. Melakukan penggantian nama fitur untuk proses visualisasi data seperti *loan_amnt* menjadi *loan_amount*, *funded_amnt* menjadi *funded amount*, *funded_amnt_inv* menjadi *investor_funds*, *int_rate* menjadi *interest_rate*, *annual_ic* menjadi *annual income.*

c. Melakukan transformasi nilai pada fitur *issue_d* ke dalam *year*. Karena nilai pada fitur *issue_d* terdiri dari bulan dan tahun, sedangkan untuk tahap visualisasi hanya dibutuhkan informasi tahun.

d. Melakukan pengkategorian status data pinjaman berdasarkan fitur *loan_status* dimana apabila status pinjaman termasuk ke dalam *Charged Off, Default, Does not meet the credit policy*. Status : *Charged Off, In Grace Period, Late (16-30 days),* dan *Late (31-120 days)* maka pinjaman tersebut dikategorikan sebagaian *bad loan*, sedangkan selain itu termasuk *good loan*. Kemudian hasil pengkategorian disimpan dalam fitur *loan_condition.*

e. Melakukan pengkategorian negara bagian berdasarkan region. Terdapat lima region yaitu *west, south_west, south_east, mid_west,* dan *north_east*.

f. Melakukan transformasi nilai pada fitur *employment_length* ke dalam bentuk tipe data integer berdasarkan lama peminjam bekerja dalam tahun.

g. Melakukan pengkategorian pendapatan bedasarkan fitur *annual_income* menjadi tiga kategori yaitu *high, medium* dan *low*. Dimana *high* memiliki syarat pendapatan lebih dari 200.000, *medium* memiliki pendapatan antara 100.000 sampai 200.000, dan *low* memiliki pendapatan di bawah 100.000.

h. Melakukan transformasi nilai pada fitur *loan_condition* menjadi tipe data integer dimana 0 mewakili untuk status *bad loan* dan 1 mewakili untuk status *good loan*.

i. Melakukan pengkategorian bunga berdasarkan fitur *interest_rate* menjadi dua kategori yaitu *high* dan *low*. Dimana *high* memiliki syarat bunga lebih dari 13,23 dan *low* memiliki syarat bunga lebih kecil atau sama dengan 13,23% kemudian hasil pengkategorian disimpan pada fitur *interest_payments*. 13,23% merupakan nilai rata-rata dari seluruh pinjaman.

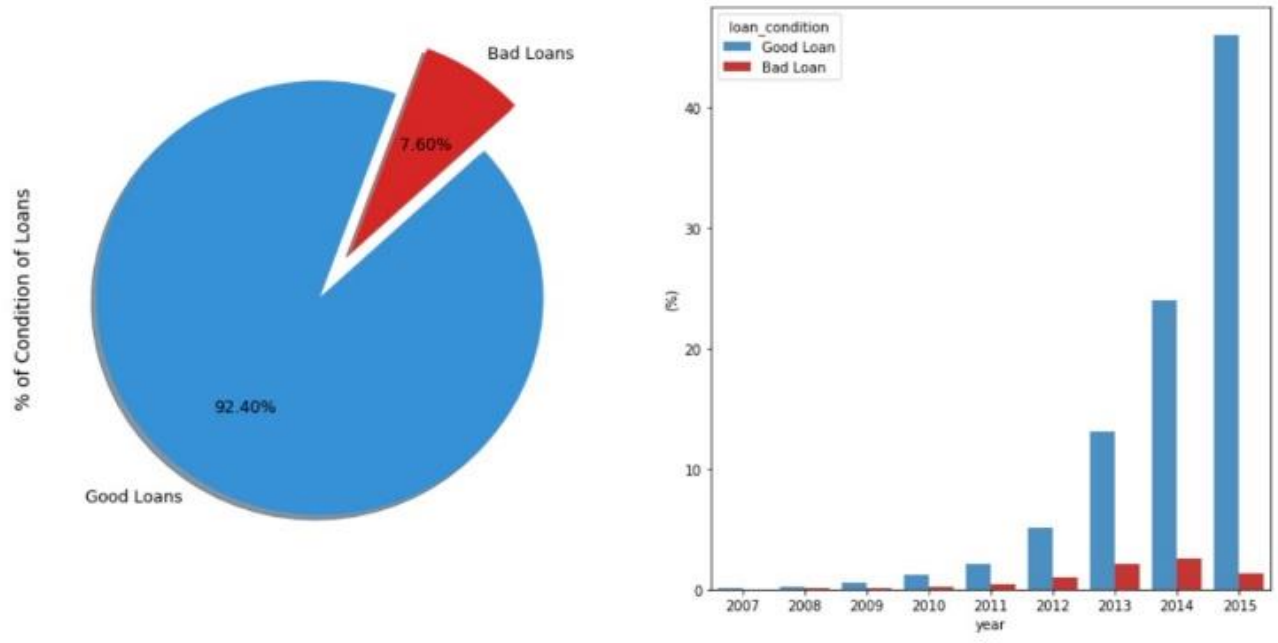
Melakukan pembuangan akhir terhadap fitur yang memiliki bobot rendah untuk menghasilkan dataset dengan fitur yang dibutuhkan untuk proses berikutnya. Gambar 10 menunjukkan fitur-fitur yang akan digunakan pada proses selanjutnya.

TABEL IV.A
FITUR YANG AKAN DIGUNAKAN

	loan_ amount	funded_ amount	investor_funds	term	Interest _rate	instalment	grade	Home_ ownership	Annual_ income
0	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	RENT	24000.0
1	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	RENT	30000.0
2	2400.0	2400.0	2400.0	36 months	15.96	84.83	C	RENT	12252.0
3	10000.0	10000.0	2400.0	36 months	13.49	339.31	C	RENT	49200.0
4	3000.0	3000.0	10000.0	60 months	12.69	67.79	B	RENT	80000.0

TABEL IV.B
FITUR YANG AKAN DIGUNAKAN

Verification_status	Dti	Emp_length_int	Income_category	Loan_condition_int	Interest_payments
Verified	27.65	10.0	low	0	Low
Souce verified	1.00	0.5	low	1	High
Not verified	8.75	10.0	low	0	High
Source verified	20.00	10.0	low	0	High
Source verified	17.94	1.0	low	0	Low



Gambar 9 Grafik perbandingan kondisi pinjaman

C. Data Distribution

Pada tahap ini dilakukan pembagian data menggunakan metode *Stratified Random Split* dikarenakan pembagian data awal tidak merata. Pada data awal, pinjaman lebih condong ke satu kondisi yaitu kondisi *good loan* dimana status pinjaman *good loan* memiliki presentase sebesar 92,40133% sedangkan status pinjaman *bad loan* memiliki presentase 7,59867%. Informasi mengenai informasi kondisi pinjaman dapat dilihat pada Gambar 9.

Pada tahap ini dilakukan pembagian data menggunakan metode *Stratified Random Split* dikarenakan pembagian data

awal tidak merata. Pada data awal, pinjaman lebih condong ke satu kondisi yaitu kondisi *good loan* dimana status pinjaman *good loan* memiliki presentase sebesar 92,40133% sedangkan status pinjaman *bad loan* memiliki presentase 7,59867%.

Metode *Stratified Random Split* dapat mengurangi bias pada data. Metode ini dapat mengurangi peluang data random yang dipilih hanya memiliki satu kondisi peminjaman saja yaitu data pinjaman *good loan*. Distribusi data itu sendiri dibagi menjadi *train set* dan *test set*. Data train set yang digunakan sebesar 80% dari keseluruhan data sedangkan untuk data *test set* digunakan 20% dari keseluruhan data. Untuk pembagian lebih jelasnya dapat dilihat pada Tabel V dimana 0 adalah *good loan* dan 1 adalah *bad loan*. Konversi tipe data dari kategorik menjadi numerik pada fitur status pinjaman dilakukan untuk mempermudah pengolahan data pada tahap *data processing*.

TABEL V
DISTRIBUSI DATA

Loan Condition	*Train Set Ratio*	*Test Set Ratio*
0	0.739211	0.184803
1	0.060789	0.015198

Setelah mendapatkan hasil distribusi data untuk *train set* dan *test set*, selanjutnya dilakukan *shuffle* data terhadap masing-masing *train set* dan *test set* untuk meningkatkan variasi data dan menghindari kemungkinan terjadinya *overfitting*. *Overfitting* adalah suatu kondisi dimana model data yang digunakan terlalu sesuai dengan model sekarang sehingga akan sulit untuk dapat membuat model data baru [16].

D. Data Processing dengan Random Forest

Pada bagian ini akan dilakukan pengolahan data dengan menggunakan metode *Random Forest* kemudian akan digunakan metode *Logistic Regression* sebagai metode pembandingnya. Tahap terakhir pada pengolahan data ini menggunakan metode *Undersampling* dan *Oversampling* terhadap metode *Random Forest* karena persebaran data antara status pinjaman *good loan* dan *bad loan* tidak merata.

TABEL VI
DESKRIPSI DATASET UNTUK DATA PROCESSING

No	Fitur	Tipe Data	Keterangan
1	loan_amount	float64	Jumlah pinjaman yang diajukan oleh peminjam
2	funded_amount	float64	Jumlah total yang diberikan untuk pinjaman tersebut
3	investor_funds	float64	Jumlah total yang dilakukan oleh investor untuk pinjaman tersebut
4	term	object	Jangka waktu pinjaman. Nilainya dalam bulan dan bisa 36 bulan atau 60 bulan
5	interest_rate	float64	Suku bunga pinjaman
6	installment	float64	Pembayaran bulanan terhutang oleh peminjam
7	grade	object	Tingkat pinjaman yang ditetapkan LC
8	home_ownership	object	Status kepemilikan rumah yang diberikan oleh peminjam pada saat pendaftaran. Nilai-nilainya adalah: RENT, OWN, MORTGAGE, LAIN-LAIN.
9	annual_income	float64	Pendapatan tahunan yang dilaporkan sendiri yang diberikan oleh peminjam selama pendaftaran.
10	verification_status	object	Menunjukkan apakah pendapatan diverifikasi oleh LC, tidak diverifikasi, atau jika sumber pendapatan diverifikasi
11	dti	float64	Rasio yang dihitung menggunakan total pembayaran hutang bulanan peminjam pada total kewajiban hutang, tidak termasuk penggadaian dan pinjaman LC yang diminta, dibagi dengan pendapatan bulanan yang dilaporkan sendiri dari peminjam.
12	emp_length_int	float64	Lama kerja peminjam dalam bentuk numerikal
13	income_category	object	Hasil kategori pendapatan peminjam dimana tedapat nilai HIGH, MEDIUM dan LOW
14	loan_condition_int	integer	Status pinjaman dalam bentuk numerikal dimana 0 adalah bad loan, sedangkan 1 adalah good loan
15	interest_payments	object	Hasil kategori suku bunga pinjaman dimana terdapat nilai HIGH dan LOW

Dataset yang digunakan untuk tahap pengolahan data dapat dilihat pada Tabel VI yang merupakan hasil dataset dari tahap *data preprocessing*. Tahap pertama pada proses analisis yaitu dengan membuat variabel x dan y untuk *train set* dan *test set*. Dimana variabel x merupakan variabel yang memiliki seluruh fitur selain '*loan_condition_int*', sedangkan variabel y merupakan variabel yang hanya memiliki fitur '*loan_condition_int'*.

Metode *Random Forest* digunakan sebagai metode utama dalam pengolahan data pinjaman P2PL dalam penelitian ini. Nilai akurasi yang didapatkan setelah

diproses dengan metode *Random Forest* dapat dilihat pada Tabel VII.

TABEL IVII
HASIL METODE RANDOM FOREST

Estimator	***Accuracy Score***
200	0,924017
150	0,924017
100	0,924006
50	0,924000
20	0,923719
10	0,922913
5	0,912066
1	0,845325

E. Analisis Dengan Metode Logistic Regression

Logistic Regression adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel respon dengan sekumpulan variabel prediktor dimana variabel respon bersifat biner [17]. Dalam penelitian ini variabel respon yang dimaksud adalah status pinjaman berupa 0 dan 1, dimana 0 adalah *bad loan* dan 1 adalah *good loan*. Sedangkan untuk variabel prediktornya terdiri dari besarnya pinjaman, jangka waktu pinjaman, suku bunga pinjaman, pendapatan per tahun, lama kerja peminjam, rasio hutang terhadap pendapatan, pembayaran bulanan, dan kepemilikan rumah.

Metode *Logistic Regression* digunakan sebagai metode pembanding *Random Forest* dalam pengolahan data pinjaman P2PL dalam penelitian ini. Nilai akurasi yang didapatkan setelah diproses dengan metode *Logistic Regression* adalah :

Accuracy Score = 0,923995

F. Oversampling

Proses *oversampling* atau disebut juga Teknik SMOTE dilakukan untuk menyeimbangkan kelas data minoritas dan mayoritas. Teknik ini dilakukan dengan cara mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas [9]. Fitur '*loan_condition_int*' pada Tabel VI menyebabkan penyebaran data tidak seimbang. Karena status pinjaman terlalu condong ke satu kondisi yaitu *good loan* dengan presentasi lebih dari 92%. Nilai akurasi yang didapatkan setelah menggunakan metode *Random Forest* dengan Teknik SMOTE dapat dilihat pada Tabel VIII.

TABEL VII
HASIL METODE RANDOM FOREST & OVER SAMPLING

Estimator	***Accuracy Score***
200	0,878490
150	0,877921
100	0,876625
50	0,876873
20	0,870371
10	0,863891
5	0,814408
1	0,719229

G. Under sampling

Sama seperti *over sampling*, proses *under sampling* juga dilakukan untuk menyeimbangkan kelas data minoritas dan mayoritas. Namun, cara kerja yang berbanding terbalik dengan *over sampling*, *under sampling* bekerja dengan mengurangi kelas mayoritas sehingga didapatkan sampel yang sama di setiap kelas. Hasil yang didapatkan setelah menggunakan metode *Random Forest* dengan Teknik *Under sampling* dapat dilihat pada Tabel IX.

TABEL IX
HASIL METODE RANDOM FOREST & UNDER SAMPLING

Estimator	***Accuracy Score***
200	0,668721
150	0,669803
100	0,669859
50	0,676784
20	0,682075
10	0,691614
5	0,614792
1	0,576432

H. Rekapitulasi Hasil Penelitian

Pada bagian ini akan dilakukan analisis terakhir terhadap hasil akhir *accuracy test* yang didapat dari hasil analisis metode dengan *Random Forest*, *Logistic Regression*, dan modifikasi metode *Random Forest* dengan Teknik *oversampling* dan *undersampling*. Berdasarkan hasil analisis pada bagian D sampai dengan G didapatkan data rekapitulasi yang dapat dilihat pada Tabel X.

TABEL X
HASIL REKAPITULASI DATA

Method	***Accuracy Score***
Random Forest	0.924012
Logistic Regression	0.923995
Random Forest + *Oversampling*	0.877904
Random Forest + Undersampling	0.666918

Jika diperhatikan pada Tabel X, hasil akurasi skor dari penggunaan metode *Random Forest* dengan teknik *over sampling* dan *under sampling* tidak meningkatkan nilai akurasi dibandingkan dengan yang tanpa teknik *under sampling* dan *over sampling*. Penggunaan keseluruhan data tanpa adanya penyesuaian memiliki hasil akurasi yang lebih baik yaitu dengan nilai akurasi 0.924012. Penggunaan metode *Random Forest* memiliki nilai akurasi tertinggi bahkan dibandingkan dengan yang menggunakan metode *Logistic Regression* yang hanya mendapat nilai akurasi sebesar 0.923995. Dengan itu disimpulkan bahwa penggunaan metode *Random Forest* terhadap dataset pinjaman sangat cocok implementasinya dalam menghasilkan model baru yang dapat digunakan untuk memprediksi suatu pinjaman apakah termasuk kategori baik atau tidak dalam konteks pinjaman P2PL.

VI. KESIMPULAN

A. Simpulan

Berdasarkan percobaan yang dilakukan, dapat disimpulkan bahwa analisis menggunakan metode *Random Forest* cocok untuk klasifikasi dengan dataset besar dan menghasilkan nilai akurasi yang sedikit lebih baik dari pada metode *Logistic Regression*.

Klasifikasi dengan penerapan SMOTE pada metode *Random Forest* tidak menghasilkan tingkat akurasi yang lebih baik daripada percobaan yang tidak menerapkan SMOTE pada metode *Random Forest.* Hal ini dikarenakan terdapat peluang dimana teknik ini memberikan bobot lebih pada kelas kecil, sehingga membuat model menjadi bias dan model tersebut akan memprediksi kelas kecil dengan akurasi lebih tinggi tetapi akurasi keseluruhan akan menurun.

Klasifikasi dengan penerapan *Random Under Sampling* pada metode *Random Forest* menghasilkan tingkat akurasi yang tidak lebih baik daripada percobaan yang tidak menerapkan *Random Under Sampling* pada metode *Random Forest*. Berbanding terbalik dengan teknik *over sampling,* teknik ini mengurangi atau menghapus data dengan bobot yang tinggi sehingga model yang digunakan dalam proses pengolahan memiliki bobot yang rendah dan mengakibatkan hasil nilai akurasi menurun secara keseluruhan.

Walaupun perbedaan akurasi tidak cukup signifikan, namun metode *Random Forest* memiliki sedikit kelebihan dalam mengolah data yang banyak dibandingkan dengan metode *Logistic Regression* [18]. Namun, selisih pada tingkat akurasinya sangat kecil sehingga dapat dikatakan kedua metode sama sama baik.

B. Saran

Saran untuk pengembangan penelitian selanjutnya memperhatikan dua hal berikut. Sumber dataset yang digunakan dapat menggunakan sumber dataset P2PL di Indonesia. Karena jika menggunakan dataset P2PL di Indonesia, hasil analisis data dapat digunakan sesuai kebutuhan. Hal yang mendasari tidak digunakannya dataset P2PL di Indonesia pada penelitian ini adalah karena adanya keterbatasan sumber dataset P2PL di Indonesia. Model yang dibuat dapat diimplementasikan pada aplikasi *production state* sehingga dapat di analisis performanya terhadap kasus nyata [19].

DAFTAR PUSTAKA

[1] S. C. Hsueh, "Effective Matching for P2P Lending by Mining Strong Association Rules," in *Proceedings of the 3rd International Conference on Industrial & Business Engineering*, 2017, pp.30-33.

[2] A. S. Marsudi and Y. Widjaja, "Industri 4.0 dan Dampaknya Terhadap Financial Technologi Serta Kesiapan Tenaga Kerja di Indonesia," *IKRAITH EKONOMIKA*, vol. 2, pp. 1-10, 2019.

[3] W. Kan. (2020) Lending Club Loan Data. Data World, 2007-2015. [Online]. Tersedia: https://data.world/data-society/lending-club-loan-data.

[4] J. Hegde and B. Rokseth, "Applications of machine learning methods for engineering risk assessment," *Safety Science,* vol.122, pp. 1-40, 2020.

[5] IBM Cloud Education. (2020) Machine Learning. [Online]. Tersedia:https://www.ibm.com/cloud/learn/machine-learning#toc-what-is-ma-qhM6PX35

[6] L. Zhu, D. Qiu, C. Ergu, C. Ying and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 142, pp. 503-513, 2019.

[7] J. R. Koza, F. H. Bennett, D. Andre and M. A. Kea, *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming,* Artificial Intelligence in Design, pp. 151-170, 1996.

[8] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Political Analysis*, vol. 9, no. 2, pp. 137-163, 2001.

[9] N. Sulistyowati and M. Jajuli, "Integrasi Naive Bayes dengan Teknik Sampling SMOTE untuk Menangani Data Tidak Seimbang," *Jurnal Nuansa Informatika*, vol. 14, no. 1, pp. 34-37, 2020.

[10] A. Saifudin and R. S. Wahono, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *Journal of Software Engineering*, vol. 1, no. 2, pp. 76-85, 2015.

[11] C. R. Putri, "Tanggung Gugat Penyelenggara Peer To Peer Lending Jika Penerima Pinjaman Melakukan Wanprestasi," *Jurist-Diction,* vol. I, no. 2, pp. 460-475, 2018.

[12] A. Primajaya and N. B. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM),* vol. 1, no. 1, pp. 27-31, 2018.

[13] S. A. Putri, "Integrasi SMOTE dan Information Gain pada Naive Bayes untuk Prediksi Cacat Software," *Journal of Software Engineering*, vol. 1, no. 2, pp. 86-91, 2015.

[14] J. W. Turkey, *Exploratory Data Analysis*, in The Future of Data Analysis, Thaeta.edu, 1977, pp. 5-23.

[15] A. Zheng and A. Casari, *Feature Engineering for Machine Learning,* Gravenstein Highway North: O'Reilly Media, Inc., 2018.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research,* vol. 15, pp. 1929-1958, 2014.

[17] F. S. Pamungkas, B. D. Prasetya and I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python," *PRISMA,* vol. 3, pp. 689-694, 2020.

[18] K. Kirasich, T. Smith and B. Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," *SMU Data Science Review*, vol. 1, no. 3, 2018.

[19] Ganesh. (2019) What is an imbalanced dataset. [Online]. Available: https://www.kaggle.com/getting-started/100018.