e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 23 Agustus 2023

Implementasi Algoritma Random Forest Terhadap Prediksi *Good Loan/Bad Loan* Kredit Nasabah Bank Di Jakarta

Sultan Farel Syah Reza[1],Widya Cholil[2]
Program Studi S1 Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta 12450
sultanfarel@upnvj.ac.id[1], widyacholil@upnvj.ac.id[2]

Abstrak. Kredit merupakan pemberian (penjaminan) barang atau jasa oleh satu pihak dengan uang untuk memenuhi segala kebutuhan, keinginan, dan aspirasi masyarakat, berdasarkan persaingan masyarakat yang semakin kompetitif. Risiko kredit adalah risiko kerugian yang terkait dengan ketidakmampuan dan/atau keengganan peminjam untuk memenuhi kewajibannya untuk membayar kembali dana pinjaman secara penuh pada atau setelah tanggal jatuh tempo. Dalam pemberian kredit, bank harus mengidentifikasi, mengelola, dan memastikan risiko kredit pada seluruh produk dan harus melalui proses pengendalian manajemen risiko yang layak. Oleh karena itu, dibutuhkannya sebuah sistem dimana yang mampu memprediksi risiko kredit yang ditimbulkan oleh nasabah bank yang tidak mampu membayar pinjaman kredit agar bank tidak merugi. Menggunakan data yang didapatkan dari ID/X untuk membuat sebuah model *machine learning* menggunakan algoritma *Random Forest.* Keluaran yang dihasilkan dari model yang telah dibuat adalah pengklasifikasian nasabah bank terbilang *good loan / bad loan.* Model klasifikasi yang diperoleh akan dievaluasi menggunakan nilai *accuracy, precision, recall,* dan *F1-Score.* Hasil evaluasi terbaik didapatkan oleh Model rasio perbandingan 70% data latih dan 30% data uji dengan nilai akurasi sebesar 84,32%, nilai *precision* sebesar 96,79%, nilai *recall* sebesar 86,44% dan F1-*score* sebesar 91.3%.

Kata Kunci: Kredit, Manajemen Risiko, *Machine Learning,* Prediksi, *Random Forest*

1 Pendahuluan

Bank merupakan perantara yang bertanggung jawab untuk menerima simpanan dari nasabah dan meminjamkannya kepada nasabah (badan) lain yang membutuhkan uang. Kegiatan utama dari bank yang fundamental ialah menyediakan dana kepada masyarakat dalam bentuk kredit [1].

Kredit adalah pemberian (penjaminan) barang atau jasa oleh satu pihak dengan uang untuk memenuhi segala kebutuhan, keinginan, dan aspirasi masyarakat, berdasarkan persaingan masyarakat yang semakin kompetitif [2].

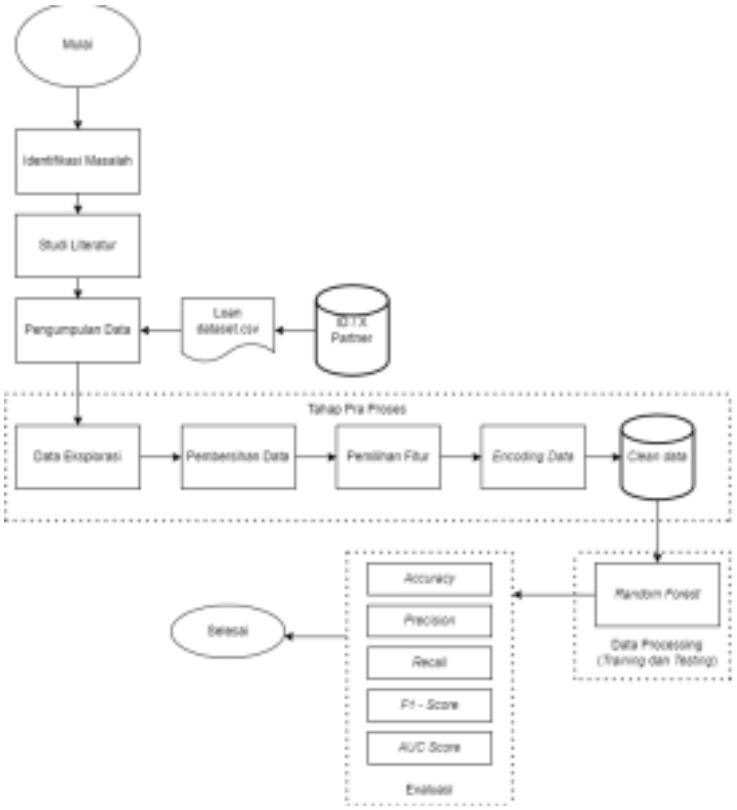
Risiko kredit adalah risiko kerugian yang terkait dengan ketidakmampuan dan/atau keengganan peminjam untuk memenuhi kewajibannya untuk membayar kembali dana pinjaman secara penuh pada atau setelah tanggal jatuh tempo [1].

Data mining adalah kegiatan analisis yang bertujuan untuk melihat kumpulan informasi yang berguna untuk menemukan hubungan yang tidak diharapkan dan yang dapat meringkas informasi dengan cara yang berbeda dari sebelumnya, sehingga mudah dipahami oleh pemilik informasi dan bermanfaat [3]. *Random Forest* adalah metode *ensemble* yang dikembangkan dari klasifikasi *decision tree.* Metode ini dapat menangani data biner, kategorikal, dan numerik [4]

Dalam penelitian terdahulu yang dilakukan oleh Budi Prasojo dan Emy Haryatmi menghasilkan akurasi sebesar 83% untuk prediksi kelayakan pemberian kredit pinjaman dengan metode *Random Forest* [5]. Penelitian lainnya dilakukan oleh Rochdi Wasono mendapatkan akurasi 98.16% untuk *Random Forest* dan 95.93% untuk *Naïve Bayes* untuk klasifikasi debitur berdasarkan kualitas kredit [6]. Penelitian lainnya yang dilakukan oleh Andreas Beny Kurniawan mendapatkan akurasi sebesar 92.67% menggunakan random forest [9]. Penelitian selanjutnya yang dilakukan oleh Joseph Sanjaya mendapatkan akurasi 92.29% untuk random forest dan 89.71% untuk Adaboost [10]. Oleh karena itu, peneliti menggunakan algoritma *Random Forest* karena memiliki akurasi

tertinggi dibandingkan dengan algoritma lainnya.

2 Metode Penelitian



Gambar 1. Tahapan Metode Penelitian

2.1 Identifikasi Masalah

Pada tahapan ini, penulis mendefinisikan permasalahan apa yang terjadi. Pada tahapan ini, penulis menemukan permasalahan yang terjadi pada peminjaman kredit nasabah kepada bank. Seringkali ditemukan nasabah tidak sanggup membanyar pinjaman kredit yang telah diajukan. Oleh karena itu, dibutuhkannya sistem untuk melakukan prediksi kesanggupan nasabah terhadap kredit yang diajukan.

2.2 Studi Literatur

Studi literatur digunakan dalam penelitian ini bertujuan untuk memperoleh informasi berdasarkan rumusan masalah yang telah didefinisikan. Selain itu tahapan ini akan membantu penulis dalam menambah pengetahuannya dan memahami permasalahan.

2.3 Pengumpulan Data

Proses pengumpulan data ini bertujuan untuk mengumpulkan data sebagai bahan pendukung pada penelitian. Data yang didapatkan berasal dari tempat magang penulis sebelumnya yaitu perusahaan *consulting* bernama ID/X Partners yang memiliki klien bank. Data yang digunakan berjumlah 4900 *records* dengan 10 atribut dan 1 label. Data yang diperoleh terdiri dari : *loan amount, funded amount, term, employee length, home ownership, annual income, verification status, purpose, last_payment_date, last_payment_amount* dan label *loan status.*

2.4 Praproses Data

Pada tahap *preprocessing data* dilakukan berbagai tahapan untuk menyiapkan data sebelum data siap dipakai untuk dilakukan pemodelan. Tahapan pra proses dimulai dengan pembersihan data dari *missing value,* data yang masih bersifat NaN, atribut yang tidak relevan. Selanjutnya masuk ke tahap seleksi fitur dengan melihat korelasi antar atribut menggunakan rumus Pearson. menghilangkan data outlier, melakukan encoding data pada fitur tertentu, standarisasi data, dan pembagian data menjadi data latih dan data uji.

2.5 EDA

Pada tahap EDA, peneliti melakukan eksplorasi lebih lanjut terhadap data yang sudah dilakukan tahap praproses. Pada tahap ini, data akan divisualisasikan menggunakan *chart* agar peneliti dapat mudah mendapatkan informasi dari data. Korelasi Data untuk mengetahui keterhubungan antar kolom fitur dengan kolom target, mengecek Distribusi Data, atau persebaran data dan Uji Normalitas Data untuk memastikan bahwa data yang dimiliki memiliki standar yang sama.

2.6 Training dan Testing

Pada tahap ini, peneliti akan melatih dan menguji model dengan data yang sudah dilakukan pra-proses menggunakan algoritma yang telah ditentukan yakni *Random Forest* untuk mendapatkan hasil prediksi yang kemudian akan digunkan untuk membentuk model *Ensemble Learning*.

2.7 Pemodelan

Tujuan dari penelitian ini ialah untuk memprediksi risiko kredit nasabah bank, sehingga metode yang digunakan adalah metode klasifikasi. *Ensemble Learning* adalah algoritma yang akan digunakan pada penelitian ini. Prediksi yang didapatkan dari hasil *training* dan *testing* akan digunakan untuk pembentukan model dan mendapatkan akurasi model prediksi. *Random Forest* adalah kumpulan *Decision Tree*, dimana fungsi pemisahan dilakukan dengan perspektif indeks Gini dan isu kedua jumlah level dibatasi oleh d yang mewakili parameter algoritma [7].

2.8 Evaluasi

Setelah tahap pemodelan dilakukan, dilakukan proses evaluasi. Evaluasi dilakukan untuk melihat seberapa akurat prediksi yang dilakukan oleh model. Metode evaluasi yang akan digunakan adalah *accuracy, precision, recall, F1 score*, dan memanfaatkan *confusion matrix* dan *classification report* supaya lebih mudah dianalisa. *Output* dari *Confusion Matrix* menghasilkan *Accuracy, Precision, Recall,* dan *F1-Score* [8].

3 Hasil dan Pembahasan

Data yang digunakan dalam penelitian ini adalah data yang didapatkan dari perusahaan consulting IDX/Partner yang memiliki klien salah satu bank di Jakarta. Dataset yang diberikan ialah dataset pinjaman kredit nasabah dari tahun 2014 sampai tahun 2021 dengan total data 4900 baris dengan total 11 kolom, yaitu *loan amount, term, employee length, annual income, loan status, funded amount, home ownership, verification status, purpose, last payment date, last payment amount.*

3.1 Praproses Data

Tahapan yang harus dilakukan terlebih dahulu sebelum masuk ke tahap pembuatan model prediksi *good loan / bad loan* nasabah kredit bank di Jakarta menggunakan algoritma *Random Forest.* Tahapan praproses bertujuan untuk mempersiapkan data yang masih mentah atau memiliki *noise* menjadi data yang siap digunakan untuk menjadi pembuatan model. Tahapan pra proses itu sendiri terdiri dari 4 tahapan diantaranya *Data Cleaning, Exploratory Data Analysis, Feature Selection,* dan *Encoding.*

3.1.1 *Data Cleaning*

Pada *data cleaning* dilakukan beberapa tahapan untuk mempersiapkan data untuk dipakai pada saat pemodelan. Langkah – langkah pada *data cleaning* yaitu *check missing value* dan *check duplicate data.*

Tabel 1. Data *Missing Value* sebelum *cleaning* dan sesudah

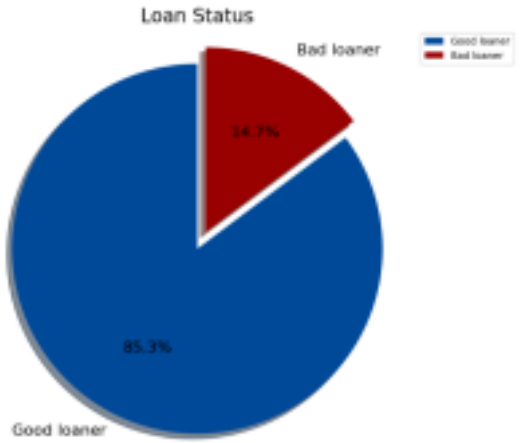
label	**Data sebelum *dicleaning***	**Data sesudah *dicleaning***
Loan_amnt	**1**	**0**
Funded_amnt	**1**	**0**
Term	**1**	**0**
Emp_length	**217**	**0**
Home_owner	**2**	**0**
Annual_inc	**8**	**0**
Verif_status	**2**	**0**
Loan_status	**1325**	**0**
Purpose	**99**	**0**
Lst_pymt_date	**1413**	**0**
Lst_pymt_amt	**4**	**0**

Tabel 2. Jumlah data yang duplikat

Label data yang duplikat	Jumlah data yang duplikat
0	**0**

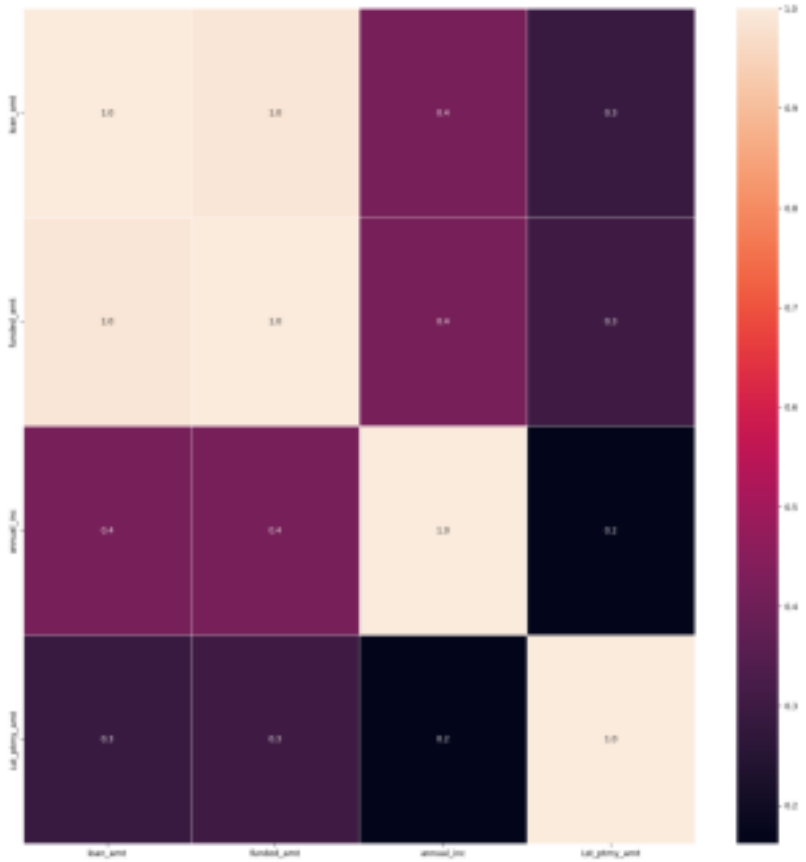
3.1.2 *Exploratory Data Analysis*

Exploratory Data Analysis yang berfungsi untuk memahami data secara lebih mendalam sebelum dilakukan analisis yang lebih lanjut menggunakan metode statistik dan visualisasi.



Gambar 2. Visualisasi *pie chart loan status* antara *good loan* dengan *bad loan*

3.1.3 *Feature Selection*



Gambar 3. Visualisasi *heatmap* korelasi *Pearson*

Dari visualisasi *heatmap* diatas, didapatkan fitur fitur yang memiliki korelasi *Pearson* yaitu : *loan amount, funded amout, annual income,* dan *last payment amount.*

3.1.4 *Encoding*

Tahapan praproses terakhir sebelum data masuk ke tahap pemodelan ialah *encoding.* Tahapan *encoding* bertujuan untuk merubah kolom kategorik menjadi angka agar dapat diolah atau dimengerti oleh komputer dan mampu dipelajari oleh model *Machine Learning.* Penulis mengerjakan tahapan *Encoding* menggunakan *One-Hot Encoding* yaitu dengan mengubah kolom kolom yang berjenis kategorikal menjadi numerik. Kolom – kolom kategorikal yang akan diubah menjadi numerik yaitu kolom *loan status, employee length, home ownership,*

verification status, purpose dan *term.*

Tabel 3. *Encoding Value Loan Status*

Loan Status	***Encoding Value***	***Value Counts***
Good Loan	1	2814
Bad Loan	0	481

Tabel 4. *Encoding Value Employee Length*

Employee Length	***Encoding Value***	***Value Counts***
1 *year*	0	195
10+ *years*	1	1063
2 *years*	2	276
3 *years*	3	236
4 *years*	4	246
5 *years*	5	260
6 *years*	6	217
7 *years*	7	171
8 *years*	8	151
9 *years*	9	142
<1 *year*	10	338

Tabel 5. *Encoding Value Home Ownership*

Home Ownership	***Encoding Value***	***Value Counts***
Mortgage	0	1557
Own	1	278
Rent	2	1460

Tabel 6. *Encoding Value Verification Status*

Verification Status	***Encoding Value***	***Value Counts***
Verified	0	3295

Tabel 7. *Encoding Value Purpose*

Purpose	***Encoding Value***	***Value Counts***
Car	0	33
Credit Card	1	756
Debt Consolidation	2	1917
Home Improvement	3	152
House	4	12
Major Purchase	5	53
Medical	6	38
Moving	7	19
Other	8	157
Renewable Energy	9	5
Small Business	10	109
Vacation	11	16
Wedding	12	28

Tabel 8. *Encoding Value Term*

Term	***Encoding Value***	***Value Counts***
36 months	0	2059
60 months	1	1236

3.2 Pemodelan

Pembuatan model menggunakan algoritma *Random Forest* untuk klasifikasi *good loan* / *bad loan* nasabah kredit bank di Jakarta. Penelitian ini menggunakan 3 rasio perbandingan untuk pembagian data *training* dan data *testing*. Pertama, dengan rasio 80% data latih dan 20% data uji. Kedua, dengan rasio 70% data latih dan 30% data uji. Ketiga, dengan rasio 60% data latih dan 40% data uji. Tujuan peneliti melakukan 3 percobaan *train test split* data rasio yang berbeda beda ialah untuk membandingkan pembagian data dengan rasio mana yang memiliki nilai akurasi paling terbaik.

Tabel 9. Rasio perbandingan data

Perbandingan	**Rasio**	**Total Data**	***Good Loan***	***Bad Loan***
		3295	2814	481
1	80% data latih	2636	2251	385
	20% data uji	659	563	96
2	70% data latih	2307	1970	337
	30% data uji	989	844	145
3	60% data latih	1977	1688	289
	40% data uji	1318	1126	192

3.3 Evaluasi

Melihat hasil evaluasi setiap model Machine Learning yang telah dibuat dengan tujuan untuk mengevaluasi performa model. Untuk mengevaluasi model, penulis menggunakan Confusion Matrix untuk menghitung nilai akurasi, precision, recall dan F1 Score menggunakan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Nilai TP atau data yang benar diprediksi sebagai positif. Contohnya dalam model yang penulis buat, nilai TP adalah jumlah nasabah kredit good loan yang terprediksi good loan. Nilai TN atau data yang benar diprediksi sebagai negatif. Contohnya adalah jumlah nasabah kredit bad loan yang terprediksi bad loan. Nilai FP atau data yang salah diprediksi sebagai positif. Contohnya, jumlah nasabah kredit yang seharusnya terbilang bad loan namun terprediksi menjadi good loan. Nilai FN atau data yang salah diprediksi sebagai negatif. Contohnya, jumlah nasabah kredit yang seharusnya terbilang good loan namun terprediksi menjadi bad loan.

Tabel 10. Perbandingan hasil evaluasi

Evaluasi	**Data Latih 80%**	**Data Latih 70%**	**Data Latih 60%**	**Data Uji 20%**	**Data Uji 30%**	**Data Uji 40%**
Akurasi	100%	100%	100%	84.21%	84.32%	83.99%
Precision	100%	100%	100%	97.15%	96.79%	97.15%
Recall	100%	100%	100%	86.11%	86.44%	85.91%
F1 Score	100%	100%	100%	91.2%	91.3%	91.17%

Dapat dilihat dari tabel diatas, nilai akurasi terbaik didapatkan dari data hasil dan data uji dengan rasio 70:30 dengan nilai akurasi data latih 100% dan data uji 84.32%. Nilai akurasi pada data latih mendapatkan 100%

dikarenakan jumlah *data training* yang jauh lebih banyak dibandingkan jumlah *data testing* sehingga hasil akurasi menjadi lebih optimal.

4 Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan data yang didapatkan, indikator yang dapat menimbulkan resiko peminjaman kredit nasabah bank di Jakarta yaitu : term, home ownership, employee length dan purpose. Berdasarkan analisis model Machine Learning pada prediksi good loan / bad loan kredit nasabah bank di Jakarta menggunakan algoritma Random Forest memiliki nilai akurasi 100% pada data latih dan 84.32% pada data uji, nilai precision sebesar 100% pada data latih dan 96.79% pada data uji, nilai recall 100% pada data latih dan 86.44% pada data uji, dan yang terakhir nilai F1 Score 100% pada data latih dan 91.3% pada data uji. Algoritma *Random Forest* yang merupakan bagian dari metode Ensemble Learning mampu memberikan model yang cukup baik untuk prediksi good loan / bad loan kredit nasabah bank di Jakarta menggunakan algoritma Random Forest dengan akurasi yang didapatkan sebesar 84.32% maka dapat terhitung Very Good Model [5]. Rasio perbandingan model yang digunakan adalah rasio perbandingan dengan 70% data latih dan 30% data uji , karena pada rasio tersebut model memberikan hasil evaluasi terbaik dari pada rasio perbandingan yang lainnya.

4.2 Saran

Memperbanyak jumlah data. Pada penelitian ini, data yang dipakai hanya 4900 baris data sebelum dilakukan tahap pra proses sehingga setelah dilakukannya pra proses maka data akan semakin berkurang. Oleh karena itu, dengan banyaknya jumlah data maka model akan memiliki pembelajaran terhadap banyak jenis data atau karakteristik terbaru dari nasabah kredit bank di Jakarta. Mencoba penerapan algoritma *Boosting* dari metode *Ensemble Learning* untuk meningkatkan dan membandingkan hasil evaluasi terhadap prediksi kredit nasabah bank di Jakarta. Melakukan *balancing data* menggunakan teknik SMOTE untuk mengecek apakah dengan melakukan *data balancing* akan mempengaruhi hasil evaluasi atau tidak.

Referensi

[1] Sari, I.M., Siregar, S. and Harahap, I. (2020) 'Manajemen Risiko Kredit Bagi Bank Umum', *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) 2020*, pp. 553–557. Available at: https://prosiding.seminar-id.com/index.php/sainteks/article/download/497/493.

[2] Siti Qomah, N. (2021) 'Klasifikasi Pengelolaan Kredit Menggunakan Metode Naïve Bayes', *Jurnal Data Science & Informatika (Jdsi)*, 2(1), pp. 35–40.

[3] Almira, A., Suendri, S., & Ikhwan, A. 2021. Implementasi Data Mining Menggunakan Algoritma Fp-Growth pada Analisis Pola Pencurian Daya Listrik. Jurnal Informatika Universitas Pamulang, 6(2), 442-448.

[4] Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel. Jurnal Teknik Informatika dan Sistem Informasi, 7(1).

[5] Prasojo, B. and Haryatmi, E. (2021) 'Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest', Jurnal Nasional Teknologi dan Sistem Informasi, 7(2), pp. 79–89. Available at: https://doi.org/10.25077/teknosi.v7i2.2021.79-89.

[6] Wasono, R. (2022). Perbandingan Metode Random Forest dan naive bayes untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit.

[7] Uddin, M. S., Chi, G., Al Janabi, M. A. M., & Habib, T. (2020). Leveraging random forest in micro‐enterprises credit risk modelling for accuracy and interpretability. International Journal of Finance & Economics. doi:10.1002/ijfe.2346

[8] Putra, J.W.G. (2020). Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4, hlm. 45–46

[9] Kurniawan. (2020). Pendekatan Random Forest untuk Memprediksi Nasabah yang berpotensi Membuka Tabungan Deposito

[10] Sanjaya, J. et al. (2020) 'Prediksi Kelalaian Pinjaman Bank Menggunakan Random Forest dan Adaptive Boosting', Jurnal Teknik Informatika dan Sistem Informasi, 6(1), pp. 50–60. Available at: https://doi.org/10.28932/jutisi.v6i1.2313.