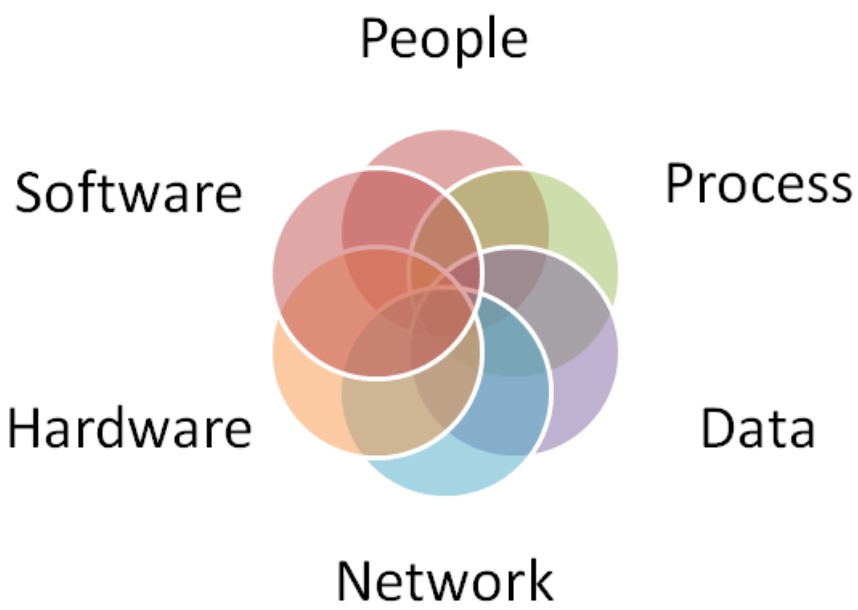
OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

ISSN 1979-3979

jurnal sisfo

Inspirasi Profesional Sistem Informasi

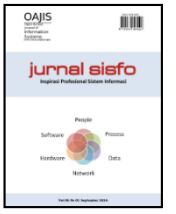


Vol. 10 No. 01 Desember 2021

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 10 No. 01 (2021) i–ii



Pimpinan Redaksi

Reny Nadlifatin

Dewan Redaksi

Faizal Mahananto

Ahmad Muklason

Eko Wahyu Tyas Damaningrat

Amalia Utamima

Penyunting

Rully Agus Hendrawan

Radityo Prasetyanto W.

Sekretariat

Jurusan Sistem Informasi – Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember (ITS) – Surabaya
Telp. 031-5999944 Fax. 031-5964965
Email: editor@jurnalsisfo.org
Website: http://jurnalsisfo.org

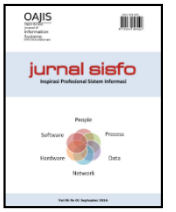
Jurnal SISFO juga dipublikasikan di *Open Access Journal of Information Systems* (OAJIS)

Website: http://is.its.ac.id/pubs/oajis/index.php

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 10 No. 01 (2021) i–ii



Mitra Bestari

Riyanto Jayadi, S. Kom, M.IM., Ph.D. (Universitas Bina Nusantara)

Nur Aini Rakhmawati, S.Kom., M.Sc.Eng., Ph.D. (Institut Teknologi Sepuluh Nopember)

Bobby Ardiansyahmiraja, S.M., M.MT. (Universitas Surabaya)

Radityo Prasetianto Wibowo, S.Kom., M.Kom. (Institut Teknologi Sepuluh Nopember)

Satria Fadil Persada, S.Kom., M.BA., Ph.D. (Mapua University)

Retno Aulia Vinarti, S.Kom., M.Kom., Ph.D. (Institut Teknologi Sepuluh Nopember)

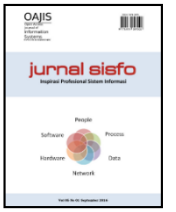
Raras Tyasnurita, S.Kom., M.BA., Ph.D (Institut Teknologi Sepuluh Nopember)

Ilma Mufidah, S.T., M.T., Ph.D (Universitas Telkom)

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 10 No. 01 (2021) iii



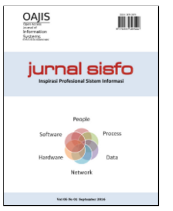
Daftar Isi

Halaman ini sengaja dikosongkan

OAJIS
Open Access
Journal of
Information
Systems
www.is.its.ac.id/pubs/oajis/

SISFO
Inspirasi Profesional Sistem Informasi

Jurnal SISFO Vol. 10 No. 01 (2021) 11–24



Perbandingan Klasifikasi Kredibilitas Pengguna Kartu Kredit Menggunakan *Decision Tree* dan *Neural Network*

Sisca Threecya Agatha, Raras Tyasnurita*

Departemen Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuh Nopember Surabaya

Abstract

Classification is the division of objects into different classes. Various methods can be used to do the classification, including the Decision Tree and Neural Network. Decision Tree adopts the concept of a tree in which there are nodes, internal nodes, and leaves. Random Forest is a collection of Decision Trees or development of Decision Trees. While the Neural Network tries to imitate the workings of the human nervous system. In a dataset containing credit card clients, data processing is performed, starting from selecting the dataset to analysis. At the end of the process, the classification results between the Decision Tree and Neural Network methods are compared, so the best method for completing research on the dataset appears. The classification carried out makes a machine learning model that functions as a means of identification and prediction. Based on testing, it was found that the classification of the credit card dataset can be resolved using the Random Forest method because the classification results have pretty good accuracy of 81% and the lowest false positive value.

Keywords: Machine Learning, Decision Tree, Neural Network, Classification, Credit Card

Abstrak

Klasifikasi merupakan pembagian objek ke dalam kelas yang berbeda-beda. Berbagai metode dapat digunakan untuk melakukan klasifikasi, diantaranya *Decision Tree* dan *Neural Network*. *Decision Tree* mengadopsi konsep dari sebuah pohon yang di dalamnya terdapat *node, internal node,* dan *leaf*. *Random Forest* adalah kumpulan *Decision Tree* atau pengembangan dari *Decision Tree*. Sedangkan *Neural Network* mencoba untuk meniru cara kerja sistem syaraf manusia. Pada dataset yang berisi tentang kartu kredit pelanggan, dilakukan pengolahan data yang diawali dari pemilihan dataset hingga analisis. Pada akhir proses, dibandingkan hasil klasifikasi antara metode *Decision Tree* dan *Neural Network*, sehingga muncul metode terbaik untuk menyelesaikan penelitian tentang dataset. Klasifikasi yang dilakukan membuat model pembelajaran mesin yang berfungsi sebagai alat identifikasi dan prediksi. Berdasarkan pengujian, diperoleh bahwa klasifikasi pada dataset kartu kredit bisa disolusikan dengan menggunakan metode *Random Forest* karena hasil klasifikasi memiliki akurasi cukup baik yaitu sebesar 81% dan nilai *false positive* paling rendah.

Kata kunci: Pembelajaran Mesin, Decision Tree, Neural Network, Klasifikasi, Kartu Kredit

Histori Artikel : Disubmit 06 Maret 2020 ; Direvisi 06 Juli 2020; Diterima 26 November 2021; Tersedia Online 26 Desember 2021

1. Pendahuluan

Masyarakat di negara dunia memiliki kecenderungan untuk berperilaku konsumtif. Hal ini juga dipengaruhi oleh beberapa faktor seperti perkembangan dunia teknologi yang memunculkan berbagai variasi produk keuangan inovatif yang ditawarkan oleh perbankan dengan semakin mudahnya transaksi, perkembangan dunia industri yang menyebabkan masyarakat memiliki dorongan untuk selalu mengedepankan gaya hidup sehingga seseorang bisa mengeluarkan uang banyak dalam sekejap hanya untuk berbelanja [1].

Kartu kredit adalah alat pembayaran sebagai pengganti uang, yang dapat digunakan untuk membayar barang atau jasa yang dibeli di tempat yang menerima pembayaran dengan kartu kredit. Badan Pusat Statistik (BPS) menyatakan

*Corresponding author
Email address: raras@is.its.ac.id (Raras Tyasnurita)
https://doi.org/10.24089/j.sisfo.2020.09.002 (DOI)

pada kuartil tiga 2018 tercatat 11,94% kenaikan dari nilai transaksi kartu kredit, debit dan uang elektronik [2]. Di Indonesia, jumlah kartu kredit yang beredar sampai akhir bulan Agustus 2018 mencapai 17,28 juta unit [3]. Untuk dapat memiliki kartu kredit pengguna harus mengajukan permohonan kepemilikan kartu kredit ke bank penerbit tertentu. Permohonan tersebut tentunya disertai dengan syarat-syarat seperti identitas diri, besar pendapatan dan lain- lain. Namun, kartu kredit juga menuntut bagi pengguna aktif untuk mengerti dan memahami penggunaan kartu kredit sesuai dengan aturan yang ditetapkan oleh bank penerbit kartu kredit. Beberapa aturan mengenai pemegang kartu kredit yaitu membayar biaya tagihan tepat waktu, membayar konsekuensi atas keterlambatan pembayaran, dan lainnya [4]. Banyak dari pengguna kartu kredit yang mengabaikan aturan ini, sehingga penundaan pembayaran kartu kredit dapat berakibat pada posisi, status pemegang kartu kredit, dan persetujuan untuk pinjaman maupun layanan kredit selanjutnya. Hal ini disebut *default* kartu kredit.

Sebagai fasilitas kredit tanpa jaminan, kartu kredit memiliki risiko besar di balik tingginya pengembalian bank. Jumlah kartu sirkulasi kartu kredit yang terus meningkat telah menyebabkan peningkatan jumlah default kartu kredit, dan jumlah besar tagihan dan data informasi pembayaran juga telah membawa kesulitan tertentu kepada pengendali risiko [5]. Oleh karena itu, diperlukan cara untuk mengolah data yang terkait dengan pengguna kartu kredit, dan mendapatkan informasi dari pengolahan tersebut yang berguna untuk mengendalikan risiko, mengurangi tingkat *default*, dan mengendalikan pertumbuhan tingkat non-performa telah menjadi salah satu perhatian utama bank. Pengolahan data seperti ini dapat dikerjakan dengan menggunakan pembelajaran mesin.

Dataset "Default of Credit Card Clients" berisi tentang data pembayaran standar, data kredit, dan riwayat pembayaran kartu kredit di Taiwan dari April 2005 hingga September 2005 [6]. Semua informasi mengenai riwayat pembayaran kartu kredit dari pengguna kartu kredit akan selalu tercatat dari bulan ke bulan. Pemilihan *dataset* ini bertujuan untuk melakukan klasifikasi dengan menggunakan metode *Decision Tree* dan *Neural Network. Machine learning* dapat menjadi sebuah 'alat' apabila bekerja dengan data yang relatif besar bahkan tidak terbatas [7]. Hal ini membuat *machine learning* digemari dalam ranah ilmu pengetahuan karena penyusunan model matematis dapat dilakukan dengan otomatis. Tujuan digunakan *machine learning* adalah untuk melakukan klasifikasi. Klasifikasi tersebut akan membuat model pembelajaran mesin atau biasa disebut *machine learning* yang berfungsi sebagai alat identifikasi dan prediksi tentang kredibilitas pengguna kartu kredit. Model ini dapat diketahui pengguna yang sering melakukan penundaan pembayaran kartu kredit dan dapat ditentukan pula apakah pengguna tersebut layak untuk mendapatkan pinjaman selanjutnya.

2. Tinjauan Pustaka

Tinjauan pustaka berisi tentang teori-teori yang mendukung penelitian.

2.1. Dataset "Default of Credit Card Clients"

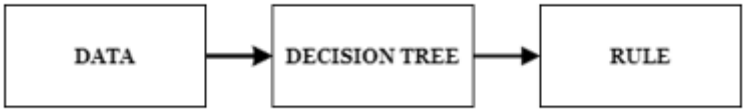
Dataset adalah sekumpulan data yang terstruktur dalam sebuah hubungan yang menyeluruh sebagai sebuah entitas [7]. Data yang digunakan pada penelitian adalah berasal dari *UCI Machine Learning Repository. Dataset "Default of Credit Card Clients"* berisi tentang data pembayaran standar, data kredit, dan riwayat pembayaran kartu kredit di Taiwan dari April 2005 hingga September 2005 [6]. Semua informasi mengenai riwayat pembayaran kartu kredit dari pengguna kartu kredit akan selalu tercatat dari bulan ke bulan. Variabel yang terdapat dalam *dataset* umumnya tentang data diri calon debitur, batas angka kredit, status riwayat pembayaran, dan jumlah tagihan.

2.2. Klasifikasi

Klasifikasi adalah aktivitas menilai objek data dan memasukkan objek tersebut ke dalam kelas dari beberapa kelas yang tersedia. Terdapat dua *task* utama yang dilakukan oleh klasifikasi, yaitu (1) pembangunan model yang dijadikan *prototype* untuk disimpan sebagai memori dan (2) menggunakan model tersebut untuk melakukan prediksi suatu objek data lain dengan tujuan objek tersebut diketahui termasuk ke dalam kelas mana di dalam model [8].

2.3. Decision Tree

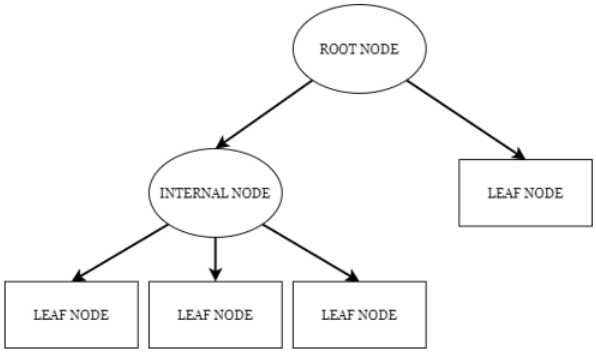
Decision tree mengubah data atau fakta menjadi pohon – pohon keputusan yang merepresentasikan sebuah aturan. Konsep dari *Decision Tree* terlihat seperti pada 1. Data yang dinyatakan dalam bentuk tabel diubah menjadi model pohon dengan menentukan atribut yang terpilih mulai dari akar, cabang, hingga menuju keputusan [10]. Setelah model pohon terbentuk selanjutnya yaitu mengubahnya menjadi rule. Simpul akar dan cabang akan menjadi premis dari rule sedangkan simpul daun akan menjadi bagian konklusi. Selain itu, metode ini juga berguna untuk eksplorasi data dan menemukan hubungan antara jumlah calon variable input dengan sebuah variabel target [11].



Gambar 1. Konsep Decision Tree

Decision tree terdiri dari himpunan *If...Then*, dimana setiap *path* atau jalur dalam *tree* dihubungkan dengan suatu aturan yang terdiri dari kumpulan node dan kesimpulan dari aturan yang terdiri dari kelas yang terhubung dengan *leaf* dari *path*. Terdapat tiga jenis node pada *Decision tree* [9] yang ditunjukkan pada Gambar 2, antara lain :

a. *Root*, merupakan *node* yang terletak paling atas, tidak ada input untuk node ini dan mempunyai output lebih dari satu atau tidak ada sama sekali
b. *Internal node*, atau biasa disebut *node* percabangan. Terdapat satu input dan output minimal dua.
c. *Leaf*, merupakan terminal *node* atau *node* terakhir, terdapat satu input namun tidak ada output.



Gambar 2. Node pada *Decision Tree*

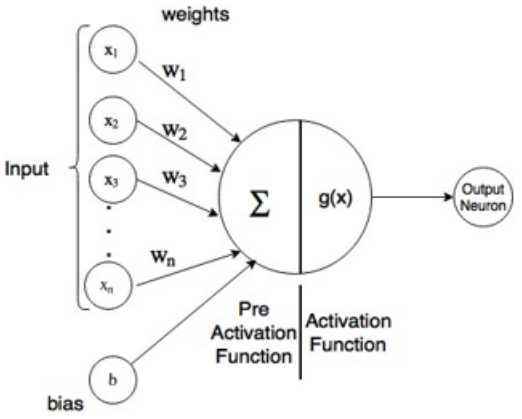
Random Forest adalah *a collection of Decision Tree* atau pengembangan dari *Decision Tree. Random forest* merupakan skema untuk membangun prediktor *ensamble* dengan seperangkat pohon keputusan *(decision tree)* yang tumbuh pada data yang dipilih secara acak [10]. *Random Forest* juga termasuk algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi *random forest* dilakukan melalui kombinasi pohon dengan dilakukannya percobaan pada sampel data yang telah disediakan. Penggunaan pohon yang semakin banyak akan mempengaruhi akurasi menjadi lebih baik dan optimal. Untuk masalah klasifikasi, pohon yang dibangun adalah pohon klasifikasi dan hasil prediksi *random forest* adalah berdasarkan *majority vote* (suara terbanyak) [11].

2.4. Artificial Neural Network

Artificial Neural Network merupakan representasi dari tiruan jaringan syaraf otak manusia. Otak manusia berisi jutaan neuron atau sel syaraf yang bertugas memproses informasi. Satu inti sel yang terdapat pada sel syaraf akan melakukan pemrosesan informasi dengan saling berinteraksi dengan sel syaraf lain sehingga mendukung kemampuan kerja otak. ANN yang menerapakan konsep sel syaraf juga memiliki beberapa neuron. Neuron yang membawa informasi akan menyimpan informasi pada suatu nilai tertentu pada bobot tertentu. Bobot merupakan hubungan dalam jaringan syaraf di ANN. Struktur dari ANN terdapat dalam Gambar 3.

Secara umum, *Artificial Neural Network* terdiri dari tiga lapisan, yaitu [15] :

a. *Input layer*, terdiri dari neuron – neuron yang menerima input dari luar. Neuron akan mengirimkan input ke *hidden layer* untuk dilakukan pemrosesan.
b. *Hidden layer*, terdiri dari neuron yang menerima masukan dari input layer, lalu melakukan pengolahan di dalamnya dan menjadi output yang diteruskan ke lapisan berikutnya.
c. *Output layer*, terdiri dari neuron yang menerima output dari hidden layer dan memberikan respon kepada *user*.



Gambar 3. Struktur Artifical Neural Network

Neuron pada ANN memiliki cara kerja yang sama dengan neuron biologis. Informasi yang merupakan input dakan dikirim ke neuron dengan bobot tertentu. Input akan diolah di suatu fungsi dan bobot yang datang akan dijumlahkan lalu dibandingkan dengan nilai ambang tertentu. Apabila input melewati nilai ambang, maka neuron akan diaktifkan. Neuron yang aktif akan dikirimkan ke output.

3. Metodologi

Metodologi yang digunakan untuk penelitian ini antara lain [12]:

Gambar 4. Metodologi

3.1. Data Gathering

Data gathering atau pengumpulan data adalah proses pengumpulan data serta informasi mengenai objek yang akan dilakukan penelitian. Keluaran dari proses pengumpulan data adalah dipilihnya *dataset "Default of Credit Card Clients"* sebagai objek dalam penelitian. *Dataset* yang didapatkan berupa file berformat csv yang akan diolah pada proses selanjutnya.

3.2. Data Exploration

Eksplorasi data yaitu merupakan salah satu metode yang digunakan untuk mencari tahu secara mendalam mengenai data. Di dalam *data mining* ekplorasi data merupakan tahapan untuk memahami data sebelum di lakukan praproses. Pemahaman terhadap data akan membantu dalam penentuan teknik analisis data. Eksplorasi data akan menyajikan data yang nantinya akan bermanfaat menjadi informasi yang berguna serta bisa digunakan untuk mengetahui pola data, dari pola data tersebut bisa teridentifikasi bentuk datanya.

3.3. Data Pre-Processing

Tahapan yang dilakukan untuk memperbaiki kualitas data, dilakukan sebelum proses pemodelan data. Pada *pre-processing* dilakukan proses cleaning data yang mencakup aktivitas menghapus duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data. Proses mengolah data agar menghasilkan sebuah informasi yang dapat berguna pada tahap processing. Berfokus pada ekstraksi pola-pola data. Pola data diidentifikasi oleh sistem, kemudian diinterpretasikan sebagai pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan manusia, salah satu contohnya yaitu prediksi dan klasifikasi.

3.4. Data Modelling

Data modelling berarti membangun model yang akan digunakan untuk prediksi kelas *dataset 'default of credit card client'*. Model ini dibuat dengan tiga algoritma yaitu *decision tree* yang menggunakan *library* 'rpart', *random forest* yang menggunakan *library* 'randomForest' dan *neural network* yang menggunakan *library* 'nnet'. Pada tahapan ini, data akan disiapkan dan dibagi menjadi dua bagian yaitu ada data yang akan digunakan sebagai data pembelajaran *(training)* dan pengujian *(testing). Dataset* dibagi dengan rasio perbandingan 70:30.

3.5. Analyzing and Testing Model

Proses selanjutnya, yaitu melakukan *training* pada *training set* untuk menghasilkan akurasi terbaik. Setelah itu, mengambil model dengan akurasi terbaik dan dilakukan pengujian dengan *dataset* pengujian *(test set)*. Kemudian, ditampilkan *confusion matrix* pada prediksi yang dilakukan.

4. Hasil dan Pembahasan

Hasil yang dibahas pada bab ini mencakup setiap tahapan dalam metodologi yaitu *Data Gathering, Data Exploration, Data Pre-Processing, Data Modelling,* dan *Analyzing and Testing Model.*

4.1. Data Gathering

Artikel ini akan membahas mengenai data kartu kredit pelanggan dari bulan April hingga bulan September tahun 2005 di Taiwan. Data diperoleh dari situs *UCI Machine Learning Repository. Dataset* terdiri atas 1 variabel respon *(default)* dan 23 variabel penjelas (X1 – X23), serta 30.000 kasus data. Penjelasan dari setiap variabel dalam data ditunjukkan oleh Tabel 1.

4.2. Data Exploration

Dalam melakukan eksplorasi data, hal pertama yang dilakukan adalah menggunakan *missingness map* untuk melakukan pengecekan terhadap *dataset*, apakah memiliki nilai yang kosong atau tidak. Dari hasil *missingness map* menunjukkan bahwa tidak ada data yang memiliki nilai yang kosong. *Missing value* dapat terjadi karena beberapa alasan, yaitu informasi yang tidak terkumpul dan adanya atribut yang tidak dapat diterapkan untuk semua kasus. Permasalahan *missing value* relatif umum ditemui pada suatu penelitian dan dapat memiliki efek signifikan pada kesimpulan yang diambil [17]. Setelah dilakukan pemodelan *missingness map*, selanjutnya dilakukan *plotting* untuk berbagai grafik untuk mencari informasi berguna yang dapat dimanfaatkan pada tahap *pre-processing* maupun *processing*. Plotting dilakukan dengan menggunakan metode *geometry bar plot, geometry box plot* dan *jitter plot* untuk setiap atribut yang ada pada *dataset*. Fungsi dari *barplot* dalam R [13] yaitu menyajikan data dalam bentuk diagram batang dari seluruh variabel yang telah di definisikan sebelumnya pada tabel 1. *Geometry bar plot* Gambar 5 menampilkan *dataset* kolom *sex* dengan *range* data yang ditampilkan dari 0 hingga 30.000 terbagi tiap 1000 data

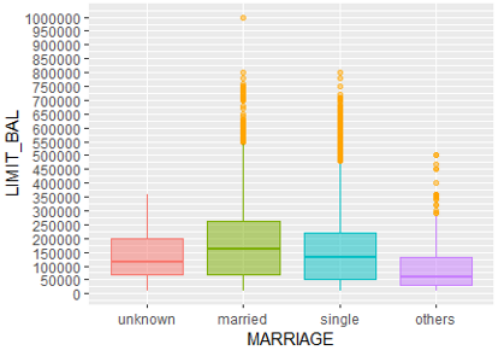
Gambar 5. Geometry Bar Plot

Boxplot atau *Box and Whisker Plot* menggambarkan bentuk distribusi data *(skewness)*, ukuran tendensi sentral dan ukuran penyebaran data penelitian. *Geometry box plot* pada Gambar 6 menampilkan dataset kolom *marriage* dan kolom limit *balanced*. *Box plot* menunjukkan penyebaran jumlah kredit yang diterima berdasarkan status perkawinan.

Tabel 1. Variabel Dataset Beserta Deskripsi

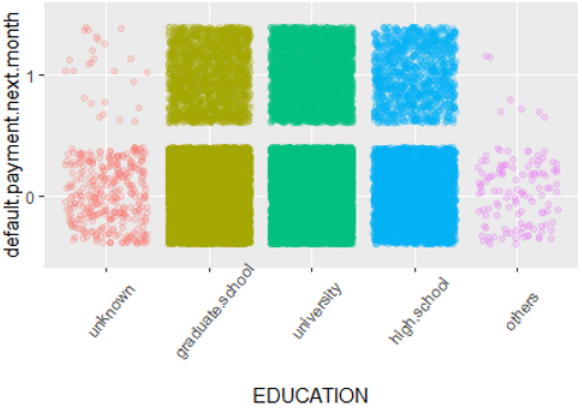
Variabel	Deskripsi
LIMIT_BAL	Jumlah kredit yang diberikan ($)
SEX	Jenis kelamin klien 1 = Male ; 2 = Female
EDUCATION	Pendidikan 1 = Graduate School ; 2 = University ; 3 = High School ; 4 = Others
MARRIAGE	Status Pernikahan 1 = Merried ; 2 = Single ; 3 = Others
AGE	Umur (tahun)
PAY_0	Riwayat pembayaran di bukan September 2005 -1 = pay duly ; 1 = payment delay for one month ; 2 = payment delay for two month ... 8 = payment delay for eight months dan seterusnya
PAY_2	Riwayat pembayaran di bulan Agustus 2005 (skala sama seperti di atas)
PAY_3	Riwayat pembayaran di bulan Juli 2005 (skala sama seperti di atas)
PAY_4	Riwayat pembayaran di bulan Juni 2005 (skala sama seperti di atas)
PAY_5	Riwayat pembayaran di bulan Mei 2005 (skala sama seperti di atas)
PAY_6	Riwayat pembayaran di bulan April 2005 (skala sama seperti di atas)
BILL_AMT1	Jumlah tagihan di bulan September 2005 ($)
BILL_AMT2	Jumlah tagihan di bulan Agustus 2005 ($)
BILL_AMT3	Jumlah tagihan di bulan Juli 2005 ($)
BILL_AMT4	Jumlah tagihan di bulan Juni 2005 ($)
BILL_AMT5	Jumlah tagihan di bulan Mei 2005 ($)
BILL_AMT6	Jumlah tagihan di bulan April 2005 ($)
PAY_AMT1	Jumlah pembayaran sebelumnya di bulan September 2005 ($)
PAY_AMT2	Jumlah pembayaran sebelumnya di bulan Agustus 2005 ($)
PAY_AMT3	Jumlah pembayaran sebelumnya di bulan Juli 2005 ($)
PAY_AMT4	Jumlah pembayaran sebelumnya di bulan Juni 2005 ($)
PAY_AMT5	Jumlah pembayaran sebelumnya di bulan Mei 2005 ($)
PAY_AMT6	Jumlah pembayaran sebelumnya di bulan April 2005 ($)
Default.payment.next.month	1 = yes ; 0 = no

Sumbu X merupakan status perkawinan sedangkan sumbu Y merupakan jumlah kredit yang diterima nasabah. Hasil dari grafik menunjukkan bahwa nasabah dengan status perkawinan *married* atau menikah memiliki persebaran jumlah kredit paling luas antara 75.000*hingga* 275.000



Gambar 6. Geometry Box Plot

Sedangkan *jitter plot* memvisualisasikan data dalam sebuah titik titik. Semakin rapat titik tersebut artinya semakin banyak data pada variabel yang dijadikan penelitian. *Jitter plot* Gambar 7 memvisualisasikan banyaknya nasabah yang default maupun tidak berdasarkan tingkat pendidikan. Dapat dilihat bahwa dari seluruh kategori yang ada pada tingkat pendidikan memiliki titik-titik yang ada lebih padat di variabel default.payment.next.month yang bernilai “No” atau 0. Sehingga didapatkan informasi bahwa lebih banyak klien yang memiliki kemampuan untuk membayar tagihannya bulan berikutnya.

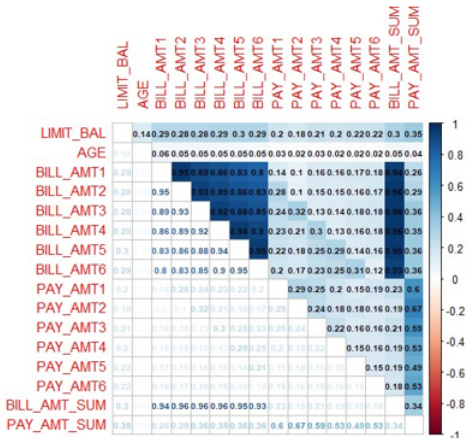


Gambar 7. Jitter Plot

Tujuan dari grafik *correlogram* adalah untuk menemukan korelasi atau kesamaan antar atribut yang tertata, untuk lebih memudahkan peneliti dalam melakukan penelitian dalam bentuk visual [14]. Dalam *dataset* ini, tidak ada data yang bersifat negatif, oleh karena itu tidak ada tabel yang berwarna merah. Grafik *correlogram* ditunjukkan pada Gambar 8.

4.3. Data Pre-Processing

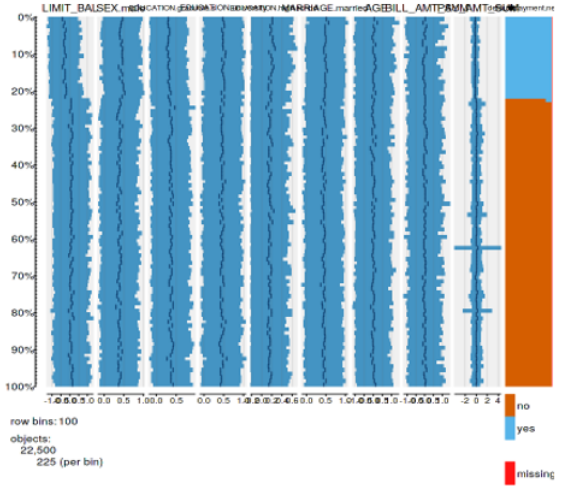
Pada tahap ini, dilakukan penyiapan data seperti normalisasi variabel, membuat DummyVars untuk variabel yang bersifat kategorikal, menghilangkan variabel pada DummyVars yang memiliki varians yang mendekati nol untuk pemodelan prediksi yang lebih baik. *Pre-processing* dilakukan dengan metode *center* dan *scale*. Normalisasi dilakukan pada atribut LIMIT_BAL, AGE, BILL_AMT1 sampai BILL_AMT6 dan PAY_AMT1 sampai PAY_AMT6



Gambar 8. Grafik Correlogram

4.4. Data Modelling

Membagi data menjadi *training set* dan *test set* untuk melakukan prediksi. Pembagian *training set* dan *test set* memiliki rasio 3:1, atau dengan kata lain *training set* terdiri dari 70% data pada dataset dan *test set* terdiri dari 30%. Dari pembagian data tersebut, dilakukan visualisasi grafik terhadap *training set* dengan menggunakan *table plot* seperti yang ditunjukkan pada Gambar 9 dengan mengurutkan kepada default.payment.next.month untuk melihat kelayakan dari *training set*.

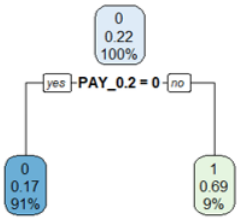


Gambar 9. Table Plot untuk Training Set

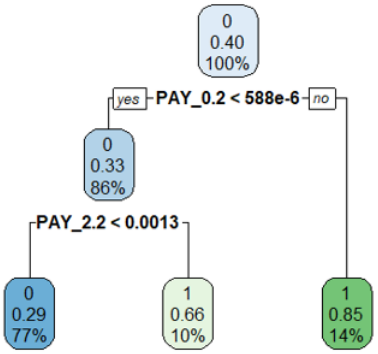
4.5. Analyzing and Testing Model

Pada tahap ini, dilakukan *training* pada *training set* yang telah dianalisa pada tahap sebelumnya dengan menggunakan ketiga algoritma, yaitu *decision tree, random forest,* dan *neural network*. Visualisasi menggunakan *decision tree* ditunjukkan pada Gambar 10 dan 11.

Model pertama berupa plot *decision tree* dengan hasil kelas klasifikasi "0" sebanyak 91% dan kelas "1" sisanya sebanyak 9%. *Branch yes* pada *decision tree* menggambarkan bahwa nasabah bisa membayar tagihan bulan berikutnya atau ditandai dengan kelas 0 yang artinya tidak *default*. *Branch no* sebaliknya. Hasil visualisasi model *decision tree* diketahui sebagai *unbalanced* data karena terdapat dominasi pada salah satu kelas, hal itu bisa disebabkan oleh *dataset* awal sebelum dibuat model memang memiliki satu kelas yang dominan, sehingga hasil klasifikasi memiliki kecenderungan pada satu kelas.



Gambar 10. Visualisasi Decision Tree Unbalanced Data



Gambar 11. Visualisasi Model Decision Tree Balanced Data

Pemodelan data kedua dibuat untuk menghindari adanya kecenderungan pada satu kelas dengan cara menyeimbangkan data atau *balanced data*. Setelah data seimbang, baru dilakukan pemodelan pada data train. Hasil dari pemodelan kedua menggunakan *decision tree* menunjukkan bahwa hasil klasifikasi nasabah yang bisa membayar tagihan bulan berikutnya sebanyak 77% sisanya sebanyak 23% berada pada kelas "1" atau kondisi *default*. Evaluasi dari model yang dibuat menggunakan *confusion matrix* yang dapat dilihat pada Tabel 2, 4, 6, dan 8. *Confusion matrix* memberikan informasi mengenai perbandingan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi data sebenarnya. *Confusion matrix* memberi tahu seberapa baik model yang telah dibuat. Pada permasalahan klasifikasi biner yang hanya menghasilkan dua output kelas seperti "ya" dan "tidak" untuk setiap input yang diberikan maka biasanya kelas utama dinotasikan sebagai data positif. Pada *dataset default of credit card client* ini kelas yang dinotasikan sebagai data positif adalah kelas dengan label "0".

Pada evaluasi model pertama diketahui bahwa:

- 6765 data positif diprediksi benar, menjelaskan bahwa nasabah yang bisa membayar tagihan berada pada *class* 0 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar tagihan dan berada pada *class* 0.
- 1448 data negatif diprediksi sebagai data positif, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar dan berada pada *class* 0.
- 543 data negatif diprediksi benar, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan bulan berikutnya berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut berada pada *class* 1 dan tidak bisa membayar tagihan bulan berikutnya.

Tabel 2. Confussion Matrix Model Decision Tree Unbalanced Data

		Actual	
	Predicted	Positive	Negative
	Positive	6765	1448
	Negative	244	543

Tabel 3. Statistik Model Decision Tree Unbalanced Data

Statistik	Decision Tree Unbalanced Data
Accuracy	0.812
Precision	0.68996
Recall	0.27273
No Information Rate	0.7788

Tabel 4. Confussion Matrix Model Decision Tree Balanced Data

		Actual	
	Predicted	Positive	Negative
	Positive	6378	1135
	Negative	631	856

- 244 data positif namun diprediksi sebagai data negatif, menjelaskan bahwa nasabah yang bisa membayar tagihan pada bulan berikutnya serta berada pada kelas 0 diprediksi tidak bisa membayar tagihan dan berada pada *class* 1.

Selain evaluasi tersebut, diketahui juga *confusion matrix* dari model pertama memberikan akurasi sebesar 81,2% dengan prediksi model sebelumnya terdapat 91% mampu membayar tagihannya bulan depan dan 9% tidak mampu membayar tagihan pada bulan selanjutnya. Selain akurasi, diketahui pula presisi sebesar 68,9% dan recal 27,3%. Pada statistik hasil Tabel 3 menunjukkan bahwa No. *Information Rate* atau kemungkinan model bisa melakukan prediksi dengan benar tanpa informasi tambahan sebesar 77,8% artinya jika persentasenya dibawah akurasi menunjukkan bahwa model tersebut cukup reliabel.

Evaluasi model kedua yaitu:

- 6378 data positif diprediksi benar, menjelaskan bahwa nasabah yang bisa membayar tagihan berada pada *class* 0 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar tagihan dan berada pada *class* 0.
- 1135 data negatif diprediksi sebagai data positif, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar dan berada pada *class* 0.
- 856 data negatif diprediksi benar, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan bulan berikutnya berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut berada pada *class* 1 dan tidak bisa membayar tagihan bulan berikutnya.
- 631 data positif namun diprediksi sebagai data negatif, menjelaskan bahwa nasabah yang bisa membayar tagihan pada bulan berikutnya serta berada pada kelas 0 diprediksi tidak bisa membayar tagihan dan berada pada *class* 1.

Pada model kedua dengan data yang sudah dilakukan *balancing*, memberikan nilai akurasi sebesar 80,38% dengan hasil klasifikasi sebelumnya pada model bahwa 77% klien mampu membayar tagihan di bulan selanjutnya sedangkan sisanya tidak mampu membayar. Selain akurasi, diketahui pula presisi sebesar 57,5% dan recall 42,9%. Pada statistik hasil Tabel 5 menunjukkan bahwa No. *Information Rate* persentasenya di bawah akurasi menunjukkan bahwa model tersebut cukup reliabel.

Evaluasi model metode Random Forest yaitu:

- 4363 data positif diprediksi benar, menjelaskan bahwa nasabah yang bisa membayar tagihan berada pada *class* 0 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar tagihan dan berada pada *class* 0.

Tabel 5. Statistik Model Decision Tree Balanced Data

Statistik	Decision Tree Balanced Data
Accuracy	0.8038
Precision	0.57566
Recall	0.49224
No Information Rate	0.7788

Tabel 6. Confussion Matrix Metode Random Forest

		Actual	
	Predicted	Positive	Negative
	Positive	4363	889
	Negative	241	507

Tabel 7. Statistik Metode Random Forest

Statistik	**Random Forest**
Accuracy	0.8117
Precision	0.6778
Recall	0.3632
No Information Rate	0.7773

- 889 data negatif diprediksi sebagai data positif, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar dan berada pada *class* 0.
- 507 data negatif diprediksi benar, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan bulan berikutnya berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut berada pada *class* 1 dan tidak bisa membayar tagihan bulan berikutnya.
- 241 data positif namun diprediksi sebagai data negatif, menjelaskan bahwa nasabah yang bisa membayar tagihan pada bulan berikutnya serta berada pada kelas 0 diprediksi tidak bisa membayar tagihan dan berada pada *class* 1.

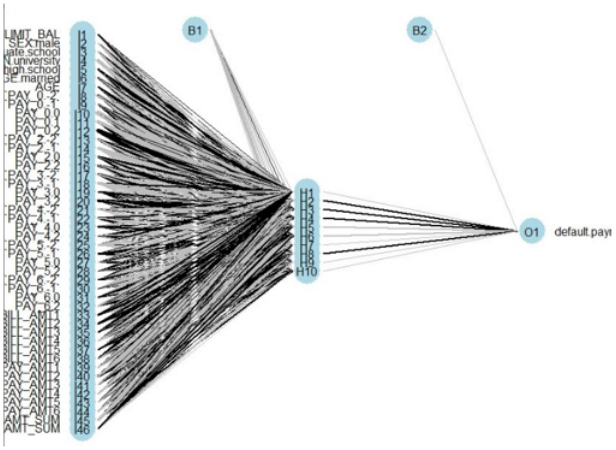
Teknik *Random Forest* adalah pengembangan dari model *Decision Tree* yang telah dibuat. Pada teknik *Random Forest*, didapatkan informasi statistik berupa akurasi sebesar 81.17% dengan presisi sebesar 67.78%, Recall sebesar 36.32%. Pada statistik hasil Tabel 7 menunjukkan bahwa No. *Information Rate* persentasenya di bawah akurasi menunjukkan bahwa model tersebut cukup reliabel. Sedangkan pada teknik *Neural Network*, dapat divisualisasikan seperti Gambar 12. Algoritma yang digunakan pada teknik *Neural Network* adalah *backpropagation*. Tujuannya untuk memperkecil tingkat error pada proses *training*. Algoritma *backpropagation* menggunakan sistem *multi layer* yaitu *input layer, hidden layer*, dan *output layer*. Pada penelitian ini jumlah *hidden layer* yang digunakan adalah *single hidden layer.* Adanya *hidden layer* dapat menurunkan tingkat error pada proses training. Hal tersebut karena *hidden layer* melakukan perhitungan dari layer input dan menyesuaikan bobot yang bisa diarahkan mendekati target output yang diinginkan. Penentuan arsitektur jaringan yang optimal dilakukan dengan mencari kombinasi pada layer yang digunakan, yaitu menentukan jumlah input yang akan diuji yang berpengaruh pada nilai output dan menentukan jumlah *hidden layer*. Proses pembelajaran pada *neural network* akan mengatur input yang digunakan untuk memetakan output. Pada *dataset* terdapat 23 variabel sebagai *input layer* dan dilanjutkan pemrosesan data pada *hidden layer* serta menghasilkan 1 output yang pada penelitian ini adalah *default of payment.*

Evaluasi model metode *Neural Network* yaitu:

- 6091 data positif diprediksi benar, menjelaskan bahwa nasabah yang bisa membayar tagihan berada pada *class* 0 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar tagihan dan berada pada *class* 0.
- 965 data negatif diprediksi sebagai data positif, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut bisa membayar dan berada pada *class* 0.

Tabel 8. Confussion Matrix Model Neural Network

		Actual	
	Predicted	Positive	Negative
	Positive	6091	965
	Negative	918	1026



Gambar 12. Visualisasi Metode Neural Network

Tabel 9. Statistik Metode Neural Network

Statistik	Neural Network
Accuracy	0.8117
Precision	0.6778
Recall	0.3632
No Information Rate	0.7773

- 1026 data negatif diprediksi benar, menjelaskan bahwa nasabah yang tidak bisa membayar tagihan bulan berikutnya berada pada *class* 1 dan dari model yang dibuat memprediksi nasabah tersebut berada pada *class* 1 dan tidak bisa membayar tagihan bulan berikutnya.
- 918 data positif namun diprediksi sebagai data negatif, menjelaskan bahwa nasabah yang bisa membayar tagihan pada bulan berikutnya serta berada pada kelas 0 diprediksi tidak bisa membayar tagihan dan berada pada *class* 1.

Selain itu, didapatkan informasi statistik yaitu akurasi sebesar 79.08% dengan presisi sebesar 52.78% dan Recall sebesar 51.53%. Pada statistik hasil Tabel 9 menunjukkan bahwa No. *Information Rate* persentasenya di bawah akurasi menunjukkan bahwa model tersebut cukup reliabel. Pada penelitian yang dilakukan oleh Sharma (2018) dengan menggunakan dataset sama ditemukan bahwa penelitian tersebut melakukan *sensitivity analysis* dan menghilangkan beberapa atribut variabel untuk mengukur performa algoritma [15]. Penelitian lain juga dilakukan oleh Dr. Maruf Pasha (2017), menggunakan *dataset* yang sama dengan algoritma yang berbeda. Algoritma yang digunakan pada penelitian tersebut ada FLDA, J48, *Logistic Regression, Naive Bayes,* MLP dan IBK seperti ditunjukkan pada Tabel 10 [16]. Hasil dari penelitian tersebut menunjukkan bahwa dengan menggunakan acuan akurasi, algoritma MLP *(Multilayer Perceptron)* dapat menghasilkan akurasi terbaik. Penelitian lainnya dilakukan oleh Torvekar (2019), dengan menggunakan dataset sama. Penelitian ini membandingkan 4 algoritma seperti ditunjukkan pada Tabel 11 [17]. Kedua penelitian sebelumnya menggunakan acuan akurasi untuk melakukan evaluasi prediksi klasifikasi. Jika dibandingkan, hasilnya tidak jauh berbeda dengan penelitian pada artikel ini yaitu akurasi yang didapatkan cukup optimal dengan rata-rata 80%.

Berdasarkan hasil evaluasi model, diketahui perbandingan *confusion matrix* seperti yang ditunjukkan pada Tabel 12. Setiap metode memiliki keunggulan masing-masing, selain melihat evaluasi matrix pada Tabel 12 dilakukan pula perbandingan akurasi, *recall*, dan presisi pada setiap metode klasifikasi. Perbandingan dari hasil ketiga metode ditunjukkan pada Tabel 13.

5. Kesimpulan

Setelah dilakukan pemrosesan data dan pengujian maka ditarik kesimpulan untuk melihat hasil terbaik dari perbandingan metode yang digunakan dalam menyelesaikan permasalahan dan saran yang dapat diberikan untuk penelitian selanjutnya.

Tabel 10. Hasil Penelitian Lain (1)

Metode	Correcat Classification (Accuracy %)	Incorrect Classification (Accuracy %)
FLDA	72.4	27.6
J48	80.3	19.7
Logistic Regression	81	19
Naive Bayes	69.4	30.6
MLP	81.7	18.3
IBK	72.9	27.1

Tabel 11. Hasil Penelitian Lain (2)

Metode	WEKA tool (Accuracy %)	KNIME tool (Accuracy %)
Naive Bayes	62.42	76.6
Logistic Regression	80.83	81.7
SVM	80.69	79.4
Random Forest	81.58	82.7

Tabel 12. Perbandingan Evaluasi Confussion Matrix

Metode	True Positive	False Positive	False Negative	True Negative
Decision Tree Unbalanced Data	6765	1448	244	543
Decision Tree Balanced Data	6378	1135	631	856
Random Forest	4363	889	241	507
Neural Network	6091	965	918	1026

Tabel 13. Perbandingan hasil Decision Tree, Random Forest dan Neural Network

Metode	Decision Tree Unbalanced Data	Decision Tree Balanced Data	Random Forest	Neural Network
Accuracy	0.812	0.8038	0.8117	0.7908
Precision	0.68996	0.57566	0.6778	0.5278
Recall	0.27273	0.42993	0.3632	0.5153
F1	0.39093	0.49224	0.4729	0.5215
No Detection Rate	0.7788	0.7788	0.7773	0.7773
Balanced Accuracy	0.61896	0.66995	0.6554	0.6922

5.1. Simpulan

Metode *Decision Tree* memiliki akurasi prediksi yang paling tinggi yaitu sebesar 81,2% diikuti oleh *Random Forest* sebesar 81,17%, dan *Neural Network* yakni 79,08%. Ketiga metode menunjukkan sedikit perbedaan akurasi. Meskipun begitu, teknik atau metode pembelajaran menggunakan *Neural Network* memiliki angka *Recall* terbesar jika dibandingkan dengan kedua metode pembelajaran lainnya, yaitu sebesar 51,53%. Data yang tidak seimbang menyebabkan nilai presisi dan recall menjadi rendah. Oleh karena itu, pemilihan metode pembelajaran terbaik dikembalikan kepada tujuan penelitian. Apabila ingin memilih metode yang bisa menggambarkan dan mengklasifikasi model dengan akurat bisa memilih *Decision Tree* dengan acuan akurasi yang tinggi. Apabila ingin memilih metode dengan prediksi klasifikasi benar terbanyak dapat menggunakan acuan *Recall* tertinggi yaitu metode *Neural Network.* Namun, jika melihat dari kasus klasifikasi kartu kredit, evaluasi *confusion matrix* berupa *false positive* bisa dipertimbangkan karena *false positive* merupakan hal cukup berbahaya dimana nasabah yang tidak bisa membayar tagihan diprediksi bisa membayar tagihan oleh model yang dibuat. Tentunya hal tersebut menjadi salah satu penyebab terjadinya kredit macet karena kesalahan prediksi dan dampak besarnya bisa mempengaruhi ekonomi suatu negara. Maka dari itu, metode *Random Forest* bisa dipilih karena memiliki hasil *confusion matrix* yang cukup baik dengan hasil *false positive* paling rendah.

5.2. Saran

Penelitian ini ditujukan untuk mencari metode atau teknik pembelajaran terbaik untuk memprediksi potensi pengguna yang mampu membayar tagihan kartu kredit di bulan berikutnya dengan menggunakan metode *Decision Tree, Random Forest* dan *Neural Network.* Oleh karena itu, banyak atribut yang kurang dimaksimalkan potensinya dalam dataset ini dikarenakan cakupan penelitian yang sangat spesifik. Perlu adanya penelitian lebih lanjut seperti melakukan *sensitivity analysis* atau parameter tuning untuk mengetahui lebih mendalam tentang *dataset* ini sehingga mendapatkan manfaat yang lebih besar yang berkaitan dengan kartu kredit dan penggunanya.

References

[1] F. Margaretha, S. M. Sari, Faktor penentu tingkat literasi keuangan para pengguna kartu kredit di indonesia, Journal of Accounting and Investment 16 (2) (2015) 132–144. doi:10.18196/jai.2015.0038.132-144.

[2] Perbankan yakin kartu kredit masih terus tumbuh sampai tahun depan. (Accessed: 05-Jul-2020).
URL `https://keuangan.kontan.co.id/news/perbankan-yakin-kartu-kredit-masih-terus-tumbuh-sampai-tahun-depan`

[3] Milenial, hindari 4 hal ini saat menggunakan kartu kredit halaman all - kompas.com. (Accessed: 05-Jul-2020).
URL `https://ekonomi.kompas.com/read/2018/11/05/070700226/milenial-hindari-4-hal-ini-saat-menggunakan-kartu-kredit?page=all`

[4] R. Sitorus, Perlindungan nasabah kartu kredit diinjau dari undang-undang no.8 tahun 1999 tentang perlindungan konsumen, Lex Privatum 3 (1) (2015) 232–239. doi:-.

[5] S. Yang, H. Zhang, Comparison of several data mining methods in credit card default prediction, Intelligent Information Management 10 (-) (2018) 115–122. doi:10.4236/iim.2018.105010.

[6] Uci machine learning repository: default of credit card clients data set (Accessed: 05-Jul-2020).
URL `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

[7] A. Renear, Definitions of dataset in the scientific and technical literature, American Society for Information Science and Technology 47 (1) (2010) 1–4. doi:10.1002/meet.14504701240.

[8] S. Defiyanti, Integrasi metode clustering dan klasifikasi untuk data numerik, in: Conference on Information Technology and Electrical Engineering, 2017, pp. 256 – 261.

[9] A. Saikhu, J. Lianto, U. Hanik, Fuzzy decision tree dengan algoritma c4 . 5 pada data diabetes indian pima, in: Konferensi Nasional Sistem dan Informatika, 2011, pp. 297 – 301.

[10] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[11] B. Baba, G. Sevil, Predicting ipo initial returns using random forest, Borsa Istanbul Revdoi:10.1016/j.bir.2019.08.001.

[12] D. T. Larose, C. D. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, 2nd Edition, John Wiley & Sons, Inc., 2014.

[13] P. GioAdhitya, R. Effendie, Belajar bahasa pemrograman r, Belajar Bhs. Pemrograman R 1 (1) (2017) 100—219. doi:10.31227/osf.io/ktmy2.

[14] P. GioAdhitya, R. Effendie, Corrgrams: Exploratory displays for correlation matrices, The American Statistician 5 (-) (2012) 316–324. doi:-.

[15] S. Sharma1, V. Mehra, Default payment analysis of credit card clients - (July) (2018) 316–324. doi:10.13140/RG.2.2.31307.28967.

[16] D. M. Pasha, M. Fatima, A. M. Dogar, F. Shahzad, Performance comparison of data mining algorithms for the predictive accuracy of credit card defaulters, IJCSNS International Journal of Computer Science and Network Security 17 (3) (2017) 178. doi:-.

[17] N. Torvekar, P. Game, Predictive analysis of credit score for credit card defaulters, International Journal of Recent Technology and Engineering (IJRTE) 7 (5) (2019) 283—286. doi:-.