# Fraud Classification on Bank Accounts using Ensemble Learning Approach
# [Klasifikasi Penipuan pada Rekening Bank menggunakan Pendekatan Ensemble Learning]

Alfiahh Maghfiroh[1)], Yulian Findawati[*2)]

[1)]Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2)]Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: yulianfindawati@umsida.ac.id

***Abstract***. *Accounts are a collection of numbers commonly used for all transactions in the banking world. In bank accounts and the process of opening them there are attempts at criminal acts. In an effort to prevent criminal acts of fraud can be solved using data mining techniques, namely classification. Therefore, this research is made to classify bank accounts for fraud prevention. The detection results are classified into two classes, namely, prospective customers indicated fraud and prospective customers not indicated fraud. The classification method used in this research is extreme gradiant boosting (XGBoost) and random forest with a percentage ratio of 90% train data and 10% test data and tuning parameters processed by randomized search cross validation. This study obtained a train score of 99.50% and a test score of 99.59% for extreme gradiant boosting (xgboost) while random forest obtained a train score of 99.46% and a test score of 99.59%. These results show that the classification results of extreme gradiant boosting (XGBoost) are better than random forest.*
***Keywords -*** *Bank Account; Extreme Gradiant Boosting; Fraud; Classification; Random Forest*

***Abstrak***. *Rekening merupakan kumpulan nomor yang biasa digunakan untuk semua transaksi di dunia perbankan.Dalam rekening bank dan proses pembukaanya terdapat upaya tindak kriminal. Dalam upaya pencegahan tindak kriminal penipuan dapat dipecahkan menggunakan teknik data mining yaitu klasifikasi. Oleh karena itu penelitian ini dibuat untuk melakukan klasifikasi pada rekening bank untuk pencegahan penipuan. Hasil deteksi di klasifikasikan ke dalam dua kelas yaitu, calon nasabah terindikasi melakukan penipuan dan calon nasabah tidak terindikasi penipuan. Metode klasifikasi yang digunakan dalam penelitian ini yaitu extreme gradiant boosting (XGBoost) dan random forest dengan perbandingan prosentasi 90% data train dan 10% data test serta parameter tuning diproses dengan randomized search cross validation. Penelitian ini mendapatkan hasil skor train 99,50% dan skor test 99,59% untuk extreme gradiant boosting (xgboost) sedangkan random forest mendapatkan hasil skor train 99,46% dan skor test 99,59%. Dengan hasil tersebut menunjukkan bahwa hasil klasifikasi extreme gradiant boosting (XGBoost) lebih baik dibandingkan random forest.*
***Kata Kunci -*** *Rekening Bank; Extreme Gradiant Boosting; Penipuan; Klasifikasi;Random Forest*

## I. PENDAHULUAN

Rekening merupakan kumpulan nomor yang biasa digunakan untuk semua transaksi di dunia perbankan, mulai dari menabung, Tarik tunai, hingga melakukan pengecekan saldo rekening baik secara langsung maupun melalui online menggunakan m-banking. Pada setiap nasabah yang hendak membuka rekening tabungan di bank diberikan kombinasi angka berbeda – beda. Tujuannya yakni sebagai rekam jejak transaksi seseorang sehingga memungkinkan bank untuk lebih mudah mengumpulkan data nasabah maupun pelacakan mutasi saldo untuk kepentingan tertentu[1].Perkembangan teknologi ini berlangsung menyeluruh hampir di semua bidang. Perkembangan teknologi informasi, telekomunikasi, dan internet menyebabkan lahirnya aplikasi bisnis internet, salah satu aplikasi bisnis yang memanfaatkan perkembangan teknologi internet adalah perbankan. Seiring berkembangnya dunia perbankan, otomatis kejahatan dunia perbankan pun ikut berkembang[2][3].

Tindak pidana bisa disebut merupakan bentuk kesadaran dalam memberikan ciri tertentu pada peristiwa hukum pidana yang dilakukan oleh seseorang maupun dilakukan secara beregu/berkelompok. Disisi lain, penipuan adalah proses perbuatan atau cara penipuan melalui tindakam atau kata - kata yang tidak jujur (berbohong, pemalsuan, dll) dengan maksud untuk menipu, mengecoh, atau mendapatkan keuntungan[4].

Ada prinsip kehati-hatian dalam perbankan, atau prinsip bahwa bank harus berhati-hati dalam menjalankan tugas dan usahanya untuk melindungi dana masyarakat yang dipercayakan kepadanya [5]. Perbankan selalu berhati – hati pada setiap kegiatan usahanya untuk mencegah terjadinya berbagai macam resiko bank, termasuk dalam pembuatan rekening bank.Dalam rekening bank dan proses pembukaanya terdapat upaya tindak kriminal dengan berbagai macam yang dilakukan seseorang maupun kelompok. Dengan berbagai bentuk dan macam tersebut bank sudah semestinya memiliki upaya pencegahan dan prediksi atas yang dilakukan dalam kegiatan tindak criminal melibatkan rekening

tersebut. Upaya pencegahan yang dilakukan oleh bank dapat melihat ciri atas data yang diinputkan oleh nasabah. Ketika bank dapat memprediksi calon nasabah yang akan melakukan penipuan maka bank dapat mengurangi angka criminal yang terjadi bank dan dapat tetap menjaga uang nasabah, sehingga kepercayaan nasabah terhadap bank dapat meningkat[6].

Machine Learning menyediakan teknologi untuk menganalisis volume data yang besar atau mendeteksi data untuk mengubah data mentah menjadi informasi berharga. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk aplikasi seperti analisis pasar, deteksi penipuan dan analisis data pelanggan [7]. XGboost dan Random Forest adalah metode klasifikasi yang berlaku. Kedua metode tersebut memiliki banyak keunggulan dibandingkan metode klasifikasi lainnya, karena lebih robust terhadap outlier, waktu komputasi yang lebih sedikit, dan hasil yang akurat.

Random Forest (RF) adalah metode klasifikasi dan regresi berupa kumpulan pohon keputusan. Dalam model Random Forest, setiap pohon adalah Classification and Regression Trees (CART) yang menggunakan Decrease Gini Impurity untuk memilih diskriminan prediktor dari subset yang dipilih secara acak dari semua variabel prediktor yang tersedia. Juga, setiap pohon tidak menggunakan semua data asli, melainkan menggunakan data model bootstrap dengan pengembalian. Keputusan kelas dibuat berdasarkan suara terbanyak dari semua pohon yang terbentuk [8].

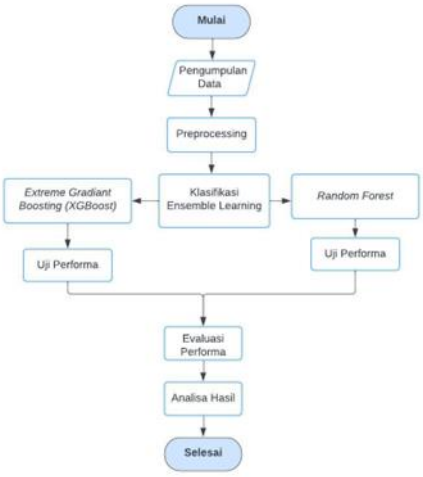
XGboost adalah salah satu metode boosting yaitu kumpulan decission tree yang pembangungan pohon berikutnya akan bergantung pada pohon sebelumnya. Pohon pertama di XGboost memiliki klasifikasi yang lemah dengan probabilitas awal yang ditentukan oleh peneliti dan kemudian memperbarui bobot setiap pohon yang dibuat untuk membuat kumpulan pohon klasifikasi yang kuat[9].

Adapaun penelitian yang menggunakan metode XGBoost oleh Ngakan Nyoman Pandika Pinata,dkk Tahun 2020 menunjukkan model XGBoost memiliki performa yang sangat baik dalam prediksi kecelakaan lalu lintas di Bali[10]. Metode XGBoost juga diterapkan `pada penelitian Muhamad Syukron,dkk pada Tahun 2020 untuk klasifikasi tingkat penyakit Hepatitis C dan mendapatkan hasil XGBoost memiliki recall kelas sirosis yang lebih baik dibanding Random Forest[9]. Pada tahun 2022 Mukhlis Febriady,dkk yang mendeteksi Penipuan pada Kartu Kredit menggunakan metode resampling SMOTE dan Random Forest menghasilkan kinerja terbaik[7]. Penelitian yang dilakukan pada tahun 2019 oleh Faried Zamachsari,dkk dengan penerapan deep learning dalam deteksi penipuan transaksi keuangan secara elektronik mendapatkan hasil terbaik tanpa proses SMOTE dengan menggunakan Deep Learning[11]. Adapun penelitian oleh Arief Kurniawan dan Yulianingsih tahun 2021 dalam pendugaan deteksi penepiuan kartu kredit dengann implementasi metode Random Forest menunjukkan bahwa model yang digunakan memiliki akurasi yang baik[6].

Berdasarkan uraian paragraf pada latar belakang diatas dan penjelasan beberapa penelitian terdahulu, Permasalahan tindak penipuan dan upaya pencegahannya dapat dipecahkan menggunakan Teknik data mining yaitu klasifikasi. Tujuan dari klasifikasi yaitu memprediksi label kelas dari suatu objek berdasarkan atribut yang ada[12], maka dengan begitu pada penlitian ini peniliti melakukan penelitian dengan studi kasus pada rekening bank menggunakan metode klasifikasi eXtreme Gradient Boosting (XGBoost) dan Random Forest dengan target yaitu Calon nasabah Terindikasi melakukan penipuan atau tidak terindikasi melakukan penipuan. Hasil klasifikasi yang didapat tersebut dapat membantu pihak bank untuk lebih berhati – hati dalam melayani dan dapat meningkatkan keamanan agar terhindar dari penipuan oleh nasabah terkait rekening bank.

## II. METODE

Metodologi penelitian adalah langkah-langkah yang peneliti terapkan dalam penelitian. Gambar dari tahapan penelitian yang diterapkan berjalan sesuai dengan gambar 1 di bawah ini beserta urainya.



**Gambar 1**. Tahapan Penelitian

**A. Pengumpulan Data**

Data adalah sebuah sumber daya, yang mana setiap jenis data yang berbeda akan akan memerlukan Teknik yang berbeda pula untuk membuatnya menjadi aplikasi yang memecahkan masalah yang diajukan oleh suatu algoritma[13]. Data yang digunakan untuk mendukung penelitian ini bersumber dari Kaggle ditulis oleh Sergio Jesus, dkk. Kumpulan data Penipuan Rekening Bank telah diterbitkan di NeurIPS 2022 yang kemudian di unggah pada Situs dan diakses pada 10 Desember 2022 (https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022) dengan judul "Bank Account Fraud Dataset Suite (NeurIPS 2022). Data terdiri dari 1 juta record. Namun pada penelitian ini diambil 22029 record.

**B. Preprocessing**

Data mentah tidak dapat digunakan langsung oleh sistem. Oleh karena itu, beberapa preprocessing harus dilakukan untuk sedikit memodifikasi data guna meningkatkan kualitas data yang digunakan. Pada penelitian ini dilakukan preprocessing untuk membersihkan data sebelum dilakukan modeling.

a. Encoding Categorical Data

Tahapan selanjutnya yaitu encoding categorical data atau mengubah data yang bersifat kategorik menjadi numerik. Hal ini dikarenakan mesin hanya dapat membaca data berupa angka saja[14].

b. Cek Correlation

Tahap selanjutnya yaitu, peneliti akan melakukan pengecekan korelasi data. Koefisien korelasi data yang lemah akan dihapus sehingga menyisakan korelasi data cukup, kuat dan sempurna.

**C. Normalisasi Data**

Langkah selanjutnya adalah normalisasi data. Normalisasi adalah proses penskalaan nilai atribut data sehingga data tersebut dapat berada dalam rentang tertentu. Beberapa teknik normalisasi data digunakan untuk normalisasi data, salah satunya adalah MinMaxScaler. MinMaxScaler adalah metode normalisasi yang melakukan transformasi linier pada data asli. Sehingga dapat menciptakan benchmark yang berimbang antar data [15].

**D. Processing**

Setelah melalui seluruh tahapan prepocessing, tahapan selanjutnya yaitu processing dimana pada tahapan ini dilakukan penerapan metode klasifikasi. Untuk mendapatkan parameter yang paling optimal, peneliti menggunakan teknik Randomized Search Cross Validation dimana teknik ini dapat melakukan hyperparameter tuning lebih cepat dibandingkan Grid Search Cross Validation[9]. Pada penelitian ini digunakan pembagian data train 90% dan data test 10% . Setelah mendapatkan parameter terbaik, selanjutnya dilakukan fitting model dengan algoritma machine learning diantaranya yaitu eXtreme Gradiant Boosting (XGBoost) dan Random Forest dengan prosentasi pembagian data train dan data test yang sudah ditentukan.

a.Klasifikasi dengan eXtreme Gradiant Boosting (XGBoost)

Extreme Gradient Boosting atau lebih umum XGBoost adalah implementasi lanjutan dari algoritme gradien boosting yang menggunakan pohon keputusan sebagai dasar klasifikasi. Diperkenalkan pada tahun 2014, telah banyak digunakan untuk menyelesaikan berbagai masalah klasifikasi atau regresi karena kecepatan, efisiensi, dan skalabilitasnya[10]. Gradient boosting merupakan algoritma yang dapat menemukan solusi optimal untuk berbagai masalah[16]. Ide dasar dari algoritma ini adalah menyesuaikan parameter pembelajaran secara iteratif untuk mengurangi fungsi biaya..

b. Klasifikasi dengan Random Forest

Random Forest adalah klasifikasi yang terdiri dari kumpulan decision tree, dimana prediksi random forest diperoleh dengan keputusan mayoritas prediksi individu pohon[17]. Random Forest bergantung pada nilai vektor acak yang memiliki distribusi yang sama di semua pohon, adalah pengklasifikasi yang terdiri dari pengklasifikasi pohon {h(x, $\theta$ k ), k = 1, . . .} di mana $\theta$k adalah vektor acak yang terdistribusi secara independen dan setiap pohon unit memilih kelas yang paling populer dari input x[7].

**E. Tahap Evaluasi**

Evaluasi data mining ini digunakan untuk mengukur akurasi atau jumlah kesalahan pada model yang kita buat. Dengan cara ini kita tahu seberapa optimal model kita untuk memecahkan masalah. Selain itu, kami dapat menguji keandalan data yang tersedia untuk kami di sini. Apakah bahan data cocok dan dapat diandalkan untuk mengekstraksi informasi. Evaluasi data mining ini dapat digunakan untuk membandingkan atau membedakan algoritma yang digunakan. Salah satu teknik yang dapat digunakan untuk mengukur kinerja model, khususnya dalam kasus klasifikasi (supervised learning) pada machine learning, adalah denganconfusion matrix[18].

## III. Hasil dan Pembahasan

### A. *Preprocessing*

Pada tahap ini data akan dilakukan *Preprocessing* guna membersihkan data sebelum dilakukan modelling. Tahap yang dimaksud adalah Encoding Categorical data dan Cek Korelasi.

**Encoding Categorical Data**

Teknik yang dilakukan dalam Encoding Categorical Data pada penelitian ini adalah *One Hot Encorder*. *One hot encorder* akan mengubah data dalam bentuk string ke dalam bentuk biner/numerik sehingga data akan terbagi menjadi kolom sebanyak jumlah jenis dalam string yang akan diubah. Data yang diubah yakni *employment status, payment type, housing status, source, device.* Dimana pada data Target terdapat 0 untuk tidak terindikasi penipuan dan 1 untuk terindikasi penipuan.

**Tabel 1.** Data *Employment Status*

| Target | Employment status |
|---|---|
| 1 | CA |
| 0 | CB |
| 1 | CC |

Tabel 1 menunjukkan data employment status atau status pekerjaan calon nasabah sebelum dilakukan encoding categorical data menggunakan one hot encoder.

**Tabel 2.** Data *Employment Status One Hot Encoder*

| Target | CA | CB | CC | CD | CE | CF | CG |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

Tabel 2 menunjukkan data employment status setelah dilakukan encodinga categorical data

**Tabel 3.** Data *Payment Type*

| Target | Payment Type |
|---|---|
| 1 | AC |
| 0 | AB |
| 1 | AB |

Tabel 3 menunjukkan data Payment type atau tipe pembayaran dari calon nasabah sebelum dilakukan encoding categorical data menggunakan one hot encoder

**Tabel 4.** *Payment Type One Hot Encoder*

| Target | AA | AB | AC | AD | AE |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 |

Tabel 4 menunjukan data Payment type setelah dilakukan encoding categorigal data

**Tabel 5.** Data *Housing Status*

| Target | Housing status |
|---|---|
| 1 | BA |
| 0 | BB |
| 1 | BC |

Tabel 5 menunjukkan data housing status atau status perumahan milik calon nasabah sebelum dilakukan encoding catergorigal data menggunakan one hot encoder

**Tabel 6.** *Housing Status One Hot Encorder*

| Target | BA | BB | BC | BD | BE | BF | BG |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

Tabel 6 menunjukkan data housing status atau status perumahan calon nasabah setelah dilakukan encoding categorical data

**Tabel 7.** Data *Sourc**e***

| Target | Source |
|---|---|
| 1 | Internet |
| 0 | Internet |
| 1 | Internet |

Tabel 7 menunjukkan data source atau sumber data caon nasabah sebelum dilakukan encoding categorical data menggunakan one hot encoder

**Tabel 8.** Data *Source One Hot Encorder*

| Target | Internet | Teleapp |
|---|---|---|
| 1 | 1 | 0 |
| 0 | 1 | 0 |
| 1 | 1 | 0 |

Tabel 8 menunjukkan data source atau sumber calon nasabah setelah dilakukan encoding categorical data

**Tabel 9.** Data *Device*

| Target | Device |
|---|---|
| 1 | Windows |
| 0 | Linux |
| 1 | Windows |

Tabel 9 menunjukkan data device atau data perangkat calon nasabah sebelum dilakukan encoding categorical data menggunakan one hot encorder

**Tabel 10.** Data *Device One Hot Encorder*

| Target | Linux | Macintosh | Other | Windows | X11 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 |

Tabel 10 menunjukkan data device atau data perangkat calon nasabah setelah dilakukan encoding categorical data

**Cek Korelasi**

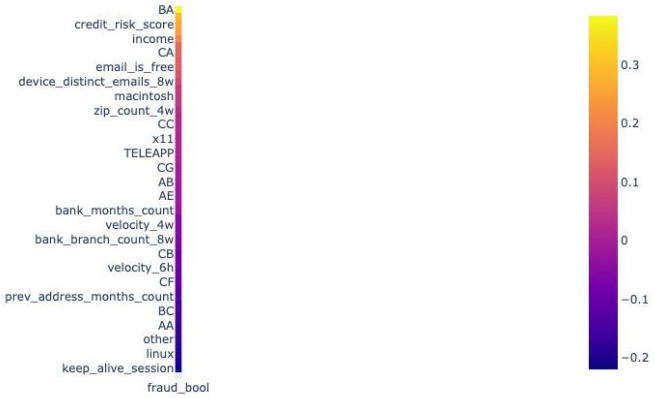
Korelasi adalah teknik statistik yang digunakan untuk menguji ada/tida knya hubungan antara dua variabel atau lebih dan arah hubungannya. Besarnya hubungan antara dua variabel dinyatakan dengan angka yang disebut koefisien korelasi. Adapun tabel koefisien korelasi data adalah sebagai berikut :

**Tabel 11.** Koefisien Korelasi

| **Koefisien** | **Keterangan** |
|---|---|
| 0 | Tidak ada korelasi antara dua variable |
| >0 – 0,25 | Korelasi sangat lemah |
| >0,25 - 0,5 | Korelasi cukup |

| | |
|---|---|
| >0,5 – 0,75 | Korelasi kuat |
| 1 | Korelasi hubungan sempurna positif |
| -1 | Korelasi hubungan sempurna negatif |

Hasil atau output yang diperoleh berdasarkan kode sumber dan data penilitian sebagai berikut :




**Gambar 1.** Hasil Korelasi Data

Berdasarkan gambar 3 dapat terlihat bahwa nilai koefisien korelasi tertinggi didapat oleh kode BA yang merupakan bagian dari data housing status atau status perumahan dengan koefisien 0,3 dimana dalam tabel 11 koefisien tersebut menunjukkan korelasi cukup, yang berarti kode BA dari data housing status memiliki keterkaitan atau hubungan paling tinggi dengan antar data atau variable lain yang diuji. Sedangkan nilai koefisen terendah berasal dari data keep alive session dan linux bagian dari data device calon nasabah dengan koefisien -0,2, dimana koefisien tersebut menunjukkan korelasi sempurna negative atau dianggap tidak memiliki korelasi antar variable atau data yang diuji.

**B. Normalisasi**

Langkah selanjutnya adalah normalisasi data. Karena penelitian ini menggunakan data dalam jumlah besar, diperlukan normalisasi data untuk menjamin konsistensi dan kualitas data. Normalisasi biasanya diperlukan ketika skala atau rentang atribut dataset berbeda. Pada penelitian ini digunakan metode penskalaan minmax untuk normalisasi. Metode normalisasi minmax mengubah kumpulan data menjadi skala 0-1 [19]. Rumus berikut dapat digunakan dalam perhitungan MinMaxScaler:

$$x_{sc} = \frac{X - Xmin}{Xmax - Xmin} \qquad (1)$$

Keterangan :
Xsc = Nilai normalisasi.
X = Nilai value dalam dataset.
Xmin = Nilai minimal dalam suatu kolom.
Xmax = Nilai maximal dalam suatu kolom.

**Tabel 12.** *After* Normalisai *Minmax Scaler*

| Target | Credit Risk Score | Income | Email is free |
|---|---|---|---|
| 0 | 0.65 | 0.125 | 0 |
| 1 | 0.66 | 1 | 1 |
| 0 | 0.60 | 0.5 | 0 |
| 0 | 0.80 | 1 | 0 |

**C. Klasifikasi Metode *eXtreme Gradiant Boosting (XGBoost)***

Langkah pertama dari tahap klasifikasi dengan XGBoost adalah membuka data yang telah dilakukan ekstraksi fitur ke dalam jupyter notebook menggunakan library pandas. Ketika data sudah berhasil diload, maka dilakukan pembagian data antara data X dan data y dimana data X merupakan kolom fitur dan data y merupakan kolom target.

Setelah melakukan pembagian data X dan y, kemudian dilakukan pembagian data train dan data test pada data X menggunakan modul scikit-learn yaitu train_test_split dengan prosentase 90% untuk data train dan 10% untuk data test.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state= 42)
```

**Gambar 2.** Pembagian data train 90% dan data test 10%

Untuk mendapatkan parameter yang paling optimal pada kasus ini, peneliti menggunakan Teknik Randomized Search Cross Validation. Teknik ini dapat mencari parameter optimal dari algoritma yang digunakan untuk kasus yang sedang dianalisa. Randomized Search Cross Validation adalah bagian dari modul scikit-learn yang bertujuan secara otomatis dan sistematis melakukan validasi beberapa model dan setiap hyperparameter. Random Search tidak mencoba semua hyperparameter yang tersedia, tetapi melakukan pencarian acak sesuai dengan ruang hyperparameter yang telah ditentukan dan berhenti setelah loop sesuai dengan iterasi yang diinginkan[20] Berikut merupakan parameter tuning yang akan digunakan.

**Tabel 13.** Hyperparameter Tuning XGBoost

| Hyperparameter | Value |
|---|---|
| max_depth | Integer(low=1, high=10) |
| learning_rate | Real(low=-2, high=0, prior='log-uniform') |
| n_estimators | Integer(low=100, high=200) |
| subsample | Real(low=0.3, high=0.8, prior='uniform') |
| gamma | Integer(low=1, high=10) |
| colsample_bytree | Real(low=0.1, high=1, prior='uniform') |
| reg_alpha | Real(low=-3, high=1, prior='log-uniform) |
| reg_lambda | Real(low=-3, high=1, prior='log-uniform') |

Proses tuning dan modelling menggunakan extreme gradiant boosting (XGBoost) dimulai dengan membuat *pipline* yang berisi *imputer, scaler(minmax scaler)* untuk normalisasi, dan *column transformer. Pipline* untuk mengolah data numerik, *ColumnTransformer* untuk mengolah data kategorik. Lalu masuk akan dilakukan proses tuning parameter untuk mencari model terbaik menggunakan *randomizedsearchcv.* Pada penelitian ini, peneliti menggunakan *randomized search cross validation* dengan *cross validation* adalah 5. Sehingga dataset dibagi menjadi 5 bagian data sama banyak. Jika bagian 1 menjadi data test maka bagian 2 hingga 5 menjadi data train. Sedangkan jika bagian 2 menjadi data test maka bagian 1 dan 3 hingga 5 menjadi data train. Begitu seterusnya hingga bagian 5 menjadi data test.Setelah proses tuning dan modelling selesai akan menghasilkan nilai akurasi model untuk data train, best score, dan data testing

**D. Klasifikasi Metode Random Forest**

Tahapan pada klasifikasi menggunakan metode random forest hampir sama dengan tahapan pada klasifikasi menggunakan extreme gradient boosting (xgboost) yaitu memisahkan data fitur dan data target. Dalam artian membagi data X dan data Y. Data fitur dibagi menjadi data train dan data test menggunakan scitit-learn yaitu train_test_split. Besaran pembagian data yaitu 90% untuk data train dan 10% untuk data test.Untuk mendapatkan parameter yang paling optimal pada metode random forest juga menggunakan Teknik RandomizedSearchCV. Berikut parameter tuning yang digunakan.

**Tabel 14.** Hyperparameter Tuning Random Forest

| Hyperparameter | Value |
|---|---|
| n_estimators | Integer(low=100, high=200) |
| max_depth | Integer(low=20, high=80) |
| max_features | Real(low=0.1, high=1, prior='uniform') |
| min_samples_leaf | Integer(low=1, high=20) |

Proses tuning dan modelling yang kedua menggunakan *random forest.* Proses dimulai dengan membuat *pipline,* yang berisi proses *imputer, scaler(minmax scaler),* dan *column transformer. Imputer* digunakan untuk mengisi nilai yang hilang dengan nilai rata – rata. *Scaler* digunakan untuk mengubah skala data agar memiliki nilai yang sama. *Column*

*transformer* digunakan untuk mengubah bentuk data. Setelah itu, *pipline* tersebut dimasukkan ke dalam model klasifikasi *random forest*. Model ini kemudian dilatih dengan menggunakan *randomizedsearchcv* dengan parameter yang telah ditentukan. Setelah proses tersebut selesai, model kemudian diuji dengan data train dan data test. Hasil pengujian ini akan menunjukkan akurasi dari model klasifikasi *random forest* yang telah dibuat.

Setelah dilakukan proses tuning dan modelling menggunakan klasfikasi *extreme gradiant boosting* dan *random forest*, berikut hasil dari data train dan data test yang telah diuji.

**Tabel 15.** Hasil Best Score Modelling

| Algoritma | *Score Train* | Score Test |
|---|---|---|
| XGBoost | 99,50% | 99,59% |
| Random Forest | 99,46% | 99,59% |

Berdasarkan tabel diatas dapat diketahui bahwa terdapat skor test 99,50% untuk algoritma extreme gradiant boosting dan dengan skor test 99,59%. Sedangkan algoritma random forest dengan skor train 99,46% dan skor test 99,59%. Kedua metode mendapatkan hasil yang fit dan tidak mengalami overfiting.

**E. Evaluasi**

Confusion matrix atau sering juga disebut error matrix. Pada dasarnya, confusion matrix memberikan informasi komparatif dari hasil klasifikasi yang dibuat oleh sistem (model) dengan hasil klasifikasi sebenarnya. Confusion matrix berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada sekumpulan data uji yang diketahui nilai sebenarnya[18]. Langkah ini digunakan untuk mengukur kinerja model pembelajaran mesin yang dibuat. Untuk mendapatkan hasil terbaik, data latih dan data uji nantinya akan diuji.

Konsep yang merepresentasikan hasil proses klasifikasi persentase pada confusion matrix adalah true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). confusion matrix dapat digunakan untuk menghitung berbagai kinerja Metrik untuk mengukur performa model yang dibangun. Ukuran kinerja meliputi presisi, akurasi, dan recall.

a. Accuracy menunjukkan seberapa akurat model dapat mengklasifikasikan dengan benar. Oleh karena itu, accuracy adalah rasio prediksi yang benar (positif dan negatif) terhadap total data. Dengan kata lain, accuracy adalah derajat kedekatan nilai prediksi dengan nilai sebenarnya. Nilai akurasi diperoleh dengan Persamaan

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

b. Precision menunjukkan tingkat akurasi antara data yang diminta dan hasil prediksi yang diberikan oleh model. data yang diminta dan hasil prediksi yang diberikan oleh model. Jadi, akurasi adalah rasio prediksi positif yang benar terhadap hasil prediksi positif total. Dari semua kelas positif yang diprediksi dengan benar, berapa banyak data yang benar-benar positif. Nilai precision dapat diperoleh dengan persamaan.

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

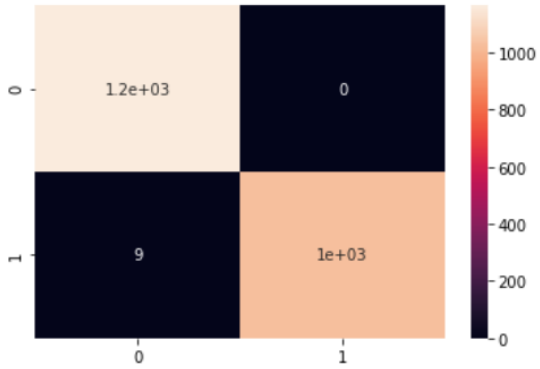
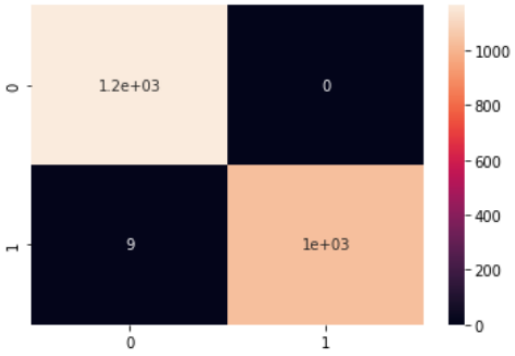
c. Recall menunjukkan keberhasilan model dalam mengambil informasi. Jadi, recall adalah rasio prediksi benar-positif terhadap semua data positif-benar. Nilai recall dapat diperoleh dengan persamaan.

$$Recall = \frac{\mathrm{TP}}{TP+FN} \qquad (4)$$

confusion matrix dapat memberi tahu seberapa baik kita melakukan model. Secara khusus confusion matrix juga memberikan informasi tentang TP, FP, TN, dan FN. Ini sangat berguna, karena hasil klasifikasi seringkali tidak dapat dinyatakan dengan baik oleh satu angka saja.

**Tabel 16.** Hasil *Confusion Matrix*

| Algoritma | Confusion Matrix |
|---|---|

| | |
|---|---|
| Extreme Gradiant Boosting |   |
| Random Forest |   |

**Tabel 17.** Hasil Classification Report

| **Algoritma** | ***Classification Report*** | | | |
|---|---|---|---|---|
| XGBoost | | precision | recall | f1-score | support |
| | 0.0 | 0.99 | 1.00 | 1.00 | 1164 |
| | 1.0 | 1.00 | 0.99 | 1.00 | 1039 |
| | accuracy | | | 1.00 | 2203 |
| | macro avg | 1.00 | 1.00 | 1.00 | 2203 |
| | weighted avg | 1.00 | 1.00 | 1.00 | 2203 |
| Random Forest | | precision | recall | f1-score | support |
| | 0.0 | 0.99 | 1.00 | 1.00 | 1164 |
| | 1.0 | 1.00 | 0.99 | 1.00 | 1039 |
| | accuracy | | | 1.00 | 2203 |
| | macro avg | 1.00 | 1.00 | 1.00 | 2203 |
| | weighted avg | 1.00 | 1.00 | 1.00 | 2203 |

Berdasarkan tabel 16 dan 17 diatas didapatkan hasil dari pengukuran dua metode *machine learning* dalam *confusion matrix* dan *Classification Report* skor akurasi, presisi,recall, dan F1-score algoritma xgboost dan random forest menghasilkan nilai skor yang sama baiknya, baik dari akurasi, presisi, recal, maupun F1-score yaitu 100% untuk akurasi, 100% untuk presisi, 99% untuk recall, dan 100% untuk F1-score, begitu juga dengan hasil klasifikasi random forest dengan akurasi 100%, presisi 100%, recall 99%, dan F1-score 100%

## VII. SIMPULAN

Hasil penelitian ini menunjukkan bahwa klasifikasi penipuan pada rekening bank dengan klasifikasi dua kelas yaitu, nasabah terindikasi tindak penipuan dan tidak terindikasi penipuan menggunakan algoritma extreme gradiant

boosting(XGBoost) dan random forest terhadap data bank pada kaggle dapat diproses. Dengan Randomized SearchCV peneliti dapat menemukan parameter terbaik untuk algoritma yang digunakan. Hasil didapatkan berdasarkan pengolahan data dari preprocessing hingga evaluasi dengan menunjukkan adanya performa yang baik. Hal tersebut dapat dilihat dari banyaknya jumlah data yang diproses yakni 22029 record menghasilkan skor train 99,50% dan skor test 99,59% untuk extreme gradiant boosting (xgboost) sedangkan random forest mendapatkan hasil skor train 99,46% dan skor test 99,59%. Dengan hasil akurasi pada penelitian ini yang baik tersebut, pada penelitian selanjutnya diharapkan dapat ditemukan improvisasi baik dalam tahap preprocessing dan processing yang sanggup mengatasi kelemahan pada penelitian ini serta dapat menerapkan metode ensemble learning yang lain untuk menjadi tolak ukur penerapan metode yang terbaik pada klasifikasi penipuan rekening bank.

## UCAPAN TERIMA KASIH

Puji dan syukur kami panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat dan rahmat-Nya, kami dapat menyelesaikan penelitian ini. Peneliti juga mengucapkan terima kasih kepada prodi Informatika Universitas Muhammadiyah Sidoarjo yang telah memberikan dukungan maupun bimbingan selama penelitian ini dilakukan.

## REFERENSI

[1] R. Chaisari, “Tindak Pidana Penipuan Rekening Bank (Suatu Penelitian Di Wilayah Hukum Polresta Banda Aceh),” Universitar syiah kuala, 2017.

[2] J. Handoko, K. Kevin, P. Paulus, and Z. Salsabila, “Sistem Deteksi Nomor Telepon dan Rekening Bank Terindikasi Penipuan Berbasis Aplikasi Android dan Web,” *J. SIFO Mikroskil*, vol. 23, no. 2, pp. 183–196, 2022, doi: 10.55601/jsm.v23i2.913.

[3] A. Prakosa, “Analisis Pengaruh Persepsi Teknologi Dan Persepsi Risiko Terhadap Kepercayaan Pengguna M-Banking,” *J. Manaj.*, vol. 9, no. 2, 2019, doi: 10.26460/jm.v9i2.1030.

[4] L. Septriyana, “Pengambilalihan Kredit Oleh Karyawan Alih Daya (Outsourcing) Pt Bank Mandiri Yang Berakibat Pada Tindak Pidana Penipuan,” *Indones. Priv. Law Rev.*, vol. 1, no. 2, pp. 99–106, 2020, doi: 10.25041/iplr.v1i2.2056.

[5] Suparyanto dan Rosad (2015, “ANALISIS PENERAPAN PRINSIP KEHATI-HATIAN BANK PADA LAYANAN E-BANKING DI PT BANK BNI SYARIAH CABANG MATARAM,” *Suparyanto dan Rosad (2015*, vol. 5, no. 3, pp. 248–253, 2020.

[6] A. Kurniawan and Y. Yulianingsih, “Pendugaan Fraud Detection pada kartu kredit dengan Machine Learning,” *Kilat*, vol. 10, no. 2, pp. 320–325, 2021, doi: 10.33322/kilat.v10i2.1482.

[7] M. Febriady, Samsuryadi, and D. P. Rini, “Klasifikasi Transaksi Penipuan Pada Kartu Kredit Menggunakan Metode Resampling Dan Pembelajaran Mesin,” *J. Media Inform. Budidarma*, vol. 6, pp. 1010–1016, 2022, doi: 10.30865/mib.v6i2.3765.

[8] T. S. Lestari and D. A. N. Sirodj, “Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest,” *J. Ris. Stat.*, vol. 1, no. 2, pp. 160–167, 2022, doi: 10.29313/jrs.v1i2.525.

[9] M. Syukron, R. Santoso, and T. Widiharih, “PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IMBALANCE CLASS DATA,” vol. 9, pp. 227–236, 2020.

[10] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, “Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: 10.24843/jim.2020.v08.i03.p04.

[11] F. Zamachsari and N. Puspitasari, “Penerapan Deep Learning dalam Deteksi Penipuan Transaksi Keuangan Secara Elektronik,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 203–212, 2021, doi: 10.29207/resti.v5i2.2952.

[12] G. A. Shafila, “IMPLEMENTASI METODE EXTREME GRADIENT BOOSTING ( XGBOOST ) UNTUK KLASIFIKASI PADA DATA BIOINFORMATIKA ( Studi Kasus : Penyakit Ebola , GSE 122692 ),” *Dspace.Uii.Ac.Id*, 2020, [Online]. Available: https://dspace.uii.ac.id/bitstream/handle/123456789/29276/16611022 Gregy Addis Shafila.pdf?sequence=1&isAllowed=y.

[13] J. Paul Mueller and L. Massaron, *Algorithm dummies*. Canada: john willey&sons,inc., 2017.

[14] V. YUGESH, “A Complete Guide to Categorical Data Encoding,” *DEVELOPERS CORNER*, 2021. https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/.

[15] O. V. Putra, T. Harmini, and A. Saroji, “Outlier Detection On Graduation Data Of Darussalam Gontor University Using One-Class Support Vector Machine,” *Procedia Eng. Life Sci.*, vol. 2, no. 2, pp. 89–92, 2021, doi: 10.21070/pels.v2i0.1139.

[16] E. Agustin, A. Eviyanti, and N. L. Azizah, "Deteksi Penyakit Epilepsi Melalui Sinyal EEG Menggunakan Metode DWT dan Extreme Gradient Boosting," vol. 7, pp. 117–127, 2023, doi: 10.30865/mib.v7i1.5412.
[17] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*. New york, 2014.
[18] R. E. Prasetyo, "[Belajar DM] Evaluasi Model Data Mining," 2022. https://rudyekoprasetya.wordpress.com/2021/04/13/belajar-dm-evaluasi-model-data-mining/.
[19] H. Mukhtar *et al.*, "Jurnal Computer Science and Information Technology ( CoSciTech ) Peramalan Kedatangan Wisatawan ke Suatu Negara Menggunakan Metode Support Vector," vol. 3, no. 3, pp. 274–282, 2022.
[20] A. Arimuko, A. S. W. Wibawa, and A. Firmansyah, "Analisis Perbandingan Penentuan Hiposentrum Menggunakan Metode Grid Search, Geiger, dan Random Search: Studi Kasus pada Letusan Gunung Sinabung 2017," *Diffraction*, vol. 1, no. 2, pp. 22–28, 2019, doi: 10.37058/diffraction.v1i2.1290.

***Conflict of Interest Statement:***
*The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.*