Optimalisasi Deteksi Penipuan Kartu Kredit Menggunakan Metode Klasifikasi Data Mining untuk Identifikasi Transaksi Ilegal

# Indah Werdiningsih[1], Endah Purwanti[2], Gede Rangga Wira Aditya[3], Auliya Rakhman Hidayat[4], R. Sulthan Rafi Athallah[5], Virda Adisty Sahar[6], Tio Satrio Wibisono[7], Darren Febriand Nura Somba[8]

Sistem Informasi, Fakultas Sains dan Teknologi Universitas Airlangga

# Email : indah-w@fst.unair.ac.id[1], endahpurwanti@fst.unair.ac.id[2], gede.rangga.wira-2020@fst.unair.ac.id[3], auliya.rakhman.hidayat-2020@fst.unair.ac.id[4] , raden.sulthan.rafi-2020@fst.unair.ac.id[5] ,

virda.adisty.sahar-2020@fst.unair.ac.id[6], tio.satrio.wibisono-2020@fst.unair.ac.id[7] , darren.febriand.nura-2020@fst.unair.ac.id[8]

***Abstract*— *The use of credit cards is increasing in today's digital era. This increase has resulted in many cases of fraud which have had a negative impact on credit card owners. To overcome this, many financial institutions have developed credit card fraud detection systems that can identify suspicious transactions. This study uses a classification method, namely random forest and decision tree to identify illegal transactions using a credit card, which then compares the results and attempts to create a model that can be useful for detecting fraud using a credit card that is more accurate and effective. The result of this study is that the accuracy provided by the Decision Tree Classifier is 0.98, while the accuracy provided by the Random Forest Classification is also 0.975. The conclusion obtained that the hypothesis stating Random Forest is a better classification algorithm than the Decision Tree can not be justified and is conditional, proven by different case studies that produce different accuracy.***

***Keywords—Credit Card, Classification, Decision Tree, Random Forest***

***Abstrak*— Penggunaan kartu kredit semakin meningkat dalam era digital saat ini. Peningkatan tersebut berdampak pada banyaknya kasus penipuan yang berdampak buruk bagi pemilik kartu kredit. Untuk mengatasinya, banyak lembaga keuangan telah mengembangkan sistem deteksi penipuan menggunakan kartu kredit yang dapat mengidentifikasi transaksi yang mencurigakan. Penelitian ini menggunakan metode klasifikasi yaitu *random forest* dan *decision tree* dalam mengidentifikasi transaksi ilegal menggunakan kartu kredit yang kemudian dibandingkan hasilnya dan berusaha untuk membuat model untuk mendeteksi penipuan menggunakan kartu kredit yang lebih akurat dan efektif. Hasil dari penelitian ini adalah akurasi yang diberikan oleh *Decision Tree Classifier* sebesar 0.98, sementara akurasi yang diberikan oleh *Random Forest Classification* juga sebesar 0.975. Kesimpulan yang didapat bahwa hipotesis yang menyatakan *Random Forest* merupakan algoritma klasifikasi yang lebih baik dari *Decision Tree* tidak dapat dibenarkan dan bersifat kondisional, dibuktikan dengan perbedaan studi kasus menghasilkan akurasi yang berbeda juga.**

***Kata Kunci—*Kartu Kredit*,* Klasifikasi*, Decision Tree, Random Forest***

1. PENDAHULUAN

Dalam era digital saat ini, penggunaan kartu kredit semakin meluas dan telah menjadi salah satu metode pembayaran yang paling populer di seluruh dunia. Kredit biasanya digunakan untuk merujuk pada transaksi keuangan elektronik yang dilakukan tanpa penggunaan kas fisik [1]. Namun, peningkatan penggunaan kartu kredit juga diikuti dengan peningkatan kasus penipuan yang menggunakan kartu kredit. Penipuan ini dapat berdampak buruk bagi pemilik kartu kredit dan lembaga keuangan yang terkait sehingga perlu adanya upaya untuk mengidentifikasi dan mencegah transaksi ilegal yang menggunakan kartu kredit [2].

Untuk mengatasi masalah ini, banyak lembaga keuangan telah mengembangkan sistem deteksi penipuan menggunakan kartu kredit yang dapat membantu mengidentifikasi transaksi yang mencurigakan atau ilegal. Namun, sebuah transaksi tidak dapat secara murni diklasifikasikan sebagai penipuan atau secara asli pada sistem, mereka hanya mencari kemiripan dan kemungkinan transaksi menjadi penipuan berdasarkan studi ekstensif perilaku pelanggan, kebiasaan belanja mereka dan juga menganalisis penipuan yang dilakukan sebelumnya, yang kemudian akan diamati pola mereka [3]. Selain itu, masalah yang sering muncul adalah kurangnya akurasi dalam mengklasifikasikan transaksi sebagai legal atau ilegal sehingga memungkinkan terjadinya kesalahan dalam menolak transaksi yang sebenarnya legal atau menerima transaksi yang seharusnya ditolak.

Penyelesaian masalah terkait klasifikasi/prediksi menggunakan *machine learning* saat ini menjadi acuan yang dapat diandalkan karena dirasa dapat memberikan akurasi terkait permasalahan [4]. *Machine learning* adalah subjek yang mempelajari cara menggunakan komputer untuk mensimulasikan kegiatan pembelajaran manusia dan untuk mempelajari metode peningkatan diri komputer untuk memperoleh pengetahuan dan keterampilan baru, mengidentifikasi pengetahuan yang ada, dan terus meningkatkan kinerja dan pencapaian [5]. Melalui penerapan yang telah diimplementasikan dalam berbagai aplikasi, *machine learning* atau pembelajaran mesin telah mencakup

hampir semua aspek dan wilayah ilmiah, yang dimana telah membawa dampak besar pada ilmu pengetahuan dan masyarakat [6]. Ada hubungan yang erat antara istilah *machine learning* dan *data mining*, jadi seringkali dalam literatur ilmiah, metode *machine learning* adalah disebut metode *data mining* [7].

Ada beberapa metode pengklasifikasian yang digunakan dalam *machine learning* untuk melakukan prediksi, diantaranya adalah *Decision Tree* dan *Random Forest*. *Decision tree learning* adalah algoritma induktif tipikal berdasarkan contoh, yang berfokus pada aturan klasifikasi yang ditampilkan sebagai pohon keputusan yang disimpulkan dari sekelompok gangguan dan contoh tidak teratur [8]. Di antara metode-metode *data mining* lainnya, klasifikasi *Decision Tree* memiliki berbagai keunggulan diantaranya sederhana untuk dipahami, mudah untuk diterapkan, membutuhkan sedikit pengetahuan, mampu menangani data numerik dan kategorikal, tangguh, dan dapat menangani *dataset* yang besar [9]. Sementara itu, klasifikasi *Random Forest* adalah pengklasifikasi ansambel yang menghasilkan banyak pohon keputusan, menggunakan subset sampel dan variabel pelatihan yang dipilih secara acak [10]. Klasifikasi *Random Forest* memiliki beberapa keunggulan. Pertama, algoritma ini dapat meningkatkan akurasi ketika terdapat data yang hilang dan juga dapat mengatasi pencilan (*outliers*). Selain itu, klasifikasi *Random Forest* efisien dalam penyimpanan data. Selain itu, algoritma ini juga melibatkan proses seleksi fitur, yang memungkinkan untuk memilih fitur-fitur terbaik dan meningkatkan performa model klasifikasi. Dengan adanya fitur seleksi ini, klasifikasi *Random Forest* dapat bekerja secara efektif pada data yang besar dengan parameter yang kompleks. Selain itu, *Random Forest* juga mampu bekerja secara paralel melalui metode *multiple random forest*. Namun, terkadang klasifikasi *Random Forest* dapat menghasilkan nilai yang tidak diharapkan dan tidak memprediksi rentang nilai respons pada data latih.[11].

Dalam konteks ini, penelitian ini bertujuan untuk menentukan metode klasifikasi yang dapat membantu dalam mengidentifikasi transaksi ilegal yang menggunakan kartu kredit secara akurat. Metode klasifikasi yang akan

dikembangkan akan didasarkan pada algoritma *Machine Learning*. Penelitian ini akan dilakukan dengan menggunakan *dataset credit card fraud* bersumber pada Kaggle yang mencakup informasi seperti jarak dari rumah tempat transaksi, jarak dari transaksi terakhir, rasio transaksi rata-rata harga beli terhadap harga pembelian, *repeat retailer*, transaksi menggunakan *chip* (kartu kredit), transaksi menggunakan nomor PIN, *online order*, dan *fraud* transaksi.

Pada penelitian sebelumnya [12], yaitu perbandingan *decision tree* dan *random forest* dalam menentukan faktor risiko yang berhubungan dengan diabetes tipe 2. Penelitian tersebut berkaitan dengan identifikasi faktor risiko yang berhubungan dengan diabetes melitus tipe 2 (T2DM) dengan menggunakan pendekatan *data mining* menggunakan *decision tree* dan *random forest* dengan menggunakan studi *Mashhad Stroke and Heart Atherosclerotic Disorders* (MASHAD). *Decision tree* dan *random forest* digunakan untuk memprediksi T2DM ketika beberapa karakteristik diamati pada 9528 subjek yang didapatkan dari database MASHAD. Didapatkan tingkat prevalensi T2DM sebesar 14%. Akurasi terbesar oleh *random forest* sebesar 71,1% dan *decision tree* sebesar 64,9%.

Penelitian ini menerapkan algoritma *decision tree* dan *random forest* untuk mengidentifikasi transaksi ilegal menggunakan kartu kredit. Hasil akurasi dari kedua algoritma tersebut akan dibandingkan akurasinya dan berusaha untuk membuat model yang dapat berguna untuk mendeteksi penipuan menggunakan kartu kredit yang lebih akurat dan efektif. Hasil penelitian diharapkan dapat memberikan kontribusi pada perkembangan teknologi *Machine Learning* dalam bidang keamanan finansial menggunakan algoritma klasifikasi terbaik yang didapatkan dari penelitian.

1. METODOLOGI PENELITIAN

Metode pengumpulan data pada penelitian ini adalah data transaksi menggunakan kartu kredit yang diambil dari situs penyedia *dataset* resmi, Kaggle. Dengan kolom transaksi 0 (legal) 1 (ilegal) dengan menggunakan *tools* Python versi 3.11.

1. *Dataset*

TABEL 1. *Dataset* yang Digunakan

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **distance\_fro m home** | **distance\_fro m\_last\_trans action** | **ratio\_to\_med ian\_purchase****\_price** | **repeat\_retail er** | **used\_chip** | **used\_pin\_nu mber** | **online\_order** | **fraud** |
| 0 | 57.877857 | 0.311140 | 1.945940 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 1 | 10.829943 | 0.175592 | 1.294219 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 5.091049 | 0.805153 | 0.427715 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| … | … | … | … | … | … | … | … | … |
| 999997 | 2.914857 | 1.472687 | 0.218075 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| 999998 | 5.258729 | 0.242023 | 0.475822 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 999999 | 52.108125 | 0.318110 | 0.386920 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 |

*Dataset* pada tabel 1 digunakan sebagai bahan penelitian klasifikasi untuk mengetahui terjadinya transaksi ilegal menggunakan kartu kredit diambil dari *website*

[https://www.kaggle.com/datasets/dhanushnarayana](https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud) [nr/credit-card-fraud](https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud).

1. *Attribute Selection*

Salah satu teknik dalam data mining yang sering digunakan pada tahap *preprocessing* adalah seleksi atribut/*feature*, yang bertujuan untuk mengurangi kompleksitas atribut. Proses seleksi atribut membantu dalam memilih atribut yang

memiliki dampak signifikan (atribut optimal) dan mengabaikan atribut yang tidak berpengaruh, sehingga mempercepat proses pemodelan.

Kemudian pada Gambar 1, dijalankan *feature selection* untuk menentukan variabel-variabel yang berpengaruh pada hasil atau *output* data yang digunakan yaitu *fraud*. Dengan begitu, didapatkan bahwa variabel *ratio\_to\_median\_purchase\_price* memiliki *feature importance* yang paling besar pada kedua algoritma yang digunakan.

TABEL 2. Deskripsi Atribut

|  |  |
| --- | --- |
| **Attribute** | **Deskripsi** |
| *distance\_from\_home* | Jarak dari rumah tempat terjadinya transaksi. |
| *distance\_from\_last\_transaction* | Jarak dari transaksi terakhir terjadi. |
| *ratio\_to\_median\_purchase\_price* | Rasio harga pembelian transaksi terhadap harga pembelian median. |
| *repeat\_retailer* | Apakah transaksi terjadi dari pengecer yang sama (1 = Ya, 0 = Tidak). |
| *used\_chip* | Apakah transaksi melalui chip (kartu kredit) (1 = Ya, 0 = Tidak). |
| *used\_pin\_number* | Apakah transaksi dilakukan dengan menggunakan nomor PIN (1 = Ya, 0 = Tidak). |
| *online\_order* | Apakah transaksi order online (1 = Ya, 0= Tidak). |
| *fraud* | Apakah transaksi tersebut penipuan (1 = Ya, 0 = Tidak). |



Gambar 1. Perbandingan *Feature Importance Random Forest* dan *Decision Tree*

1. *Data Pre Processing*

*Data Preprocessing* memegang peran utama dalam *data mining.* Dalam langkah pertama melakukan *data mining* adalah melakukan *training* pada *data* supaya dalam penemuan ilmu tidak akan sulit memahaminya dengan tidak adanya data yang tidak diinginkan, tidak relevan atau berantakan dan data yang tidak dapat diandalkan [13]. *Preprocessing* dianggap sebagai langkah yang diperlukan untuk mendapatkan model klasifikasi yang memiliki kinerja yang tinggi dalam memprediksi [14]. Dengan begitu, tujuan dari *data pre-processing* adalah untuk meminimalkan efek dari data mentah yang tidak berkualitas pada hasil analisis atau pemodelan. Proses *pre-processing* pada penelitian ini meliputi *data cleaning* terhadap data yang memiliki *missing value* dan *outlier* kemudian dilakukan normalisasi menggunakan metode Z-Score.

Pada tahap *data cleaning*, *dataset* yang terdiri dari satu juta record ini tidak terdapat *missing*

*values*, yang dapat dilihat pada Gambar 2. Kemudian pada Tabel 3 dapat dilihat bahwa terdapat temuan *outlier* pada data ini.

Gambar 2. Hasil *Missing Value*

Setelah dilakukan *cleaning data,* data yang tersisa kemudian dilakukan normalisasi dengan menggunakan metode Z-Score. Hasil setelah dilakukan normalisasi dapat dilihat pada Tabel 4..

TABEL 3. Hasil Temuan Outlier

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **distance\_from home** | **distance\_from****\_last\_transacti on** | **ratio\_to\_medi an\_purchase\_ price** | **repeat\_retalle r** | **used\_chip** | **used\_pin\_numbe r** | **online\_order** | **fraud** |
| 12 | 765,282559 | 0,371562 | 0,551245 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 13 | 2,131956 | 56,372401 | 6,358667 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| 24 | 3.803057 | 67,241081 | 1,87295 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| … |  |  |  |  |  |  |  |  |
| 999949 | 15.724799 | 1.875906 | 11.009366 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 |
| 999961 | 337.291331 | 4.606990 | 0.181799 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| 999973 | 10.148074 | 4.465290 | 12.022734 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 |

TABEL 4. Hasil Setelah Normalisasi Z-Score

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **distance\_fro m home** | **distance\_from\_ last\_transactio n** | **ratio\_to\_med ian\_purchase****\_price** | **repeat\_retaller** | **used\_chip** | **used\_pin\_number** | **online\_order** | **fraud** |
| 0 | 0,477882 | -0,182849 | 0,043491 | 0,366584 | 1,361576 | -0,334458 | -1.364425 | -0,309474 |
| 1 | -0,241607 | -0,188094 | -0,189300 | 0,366584 | -0,734443 | -0,334458 | -1.364425 | -0,309474 |
| 2 | -0,329369 | -0,163733 | -0,498812 | 0,366584 | -0,734443 | -0,334458 | 0.732909 | -0,309474 |
| … |  |  |  |  |  |  |  |  |
| 999997 | -0,362650 | -0,137903 | -0,573694 | 0,366584 | 1,361576 | -0,334458 | 0.732909 | -0,309474 |
| 999998 | -0,342098 | -0,185523 | -0,481628 | 0,366584 | -0,734443 | -0,334458 | 0.732909 | -0,309474 |
| 999999 | 0,481403 | -0,182849 | 0,513384 | 0,366584 | 1,361576 | -0,334458 | -1.364425 | -0,309474 |

1. *Split Data*

*Split data* adalah proses pembagian *dataset* menjadi dua bagian, yaitu *training set* dan *testing set*, dengan menggunakan proporsi 7:3 yang dilakukan secara acak. Tujuan dari *split data* adalah untuk melatih model dengan *data training* dan menguji model tersebut dengan *data test* yang belum pernah dilihat sebelumnya untuk mengevaluasi performa model. Pembagian data ini sangat penting dalam pengembangan model *machine learning* karena dapat membantu menghindari *overfitting* dan memastikan bahwa

model yang dibangun dapat menggeneralisasi data dengan baik [15].

1. *Decision Tree Classifier*

*Decision Tree* merupakan sebuah algoritma klasifikasi yang dapat dijelaskan sebagai pemisahan rekursif dari ruang sampel. *Decision Tree* terdiri dari simpul-simpul yang membentuk struktur pohon, di mana pohon tersebut memiliki simpul awal yang disebut akar dan terarah. Pohon keputusan didasarkan pada teknik data mining yang secara rekursif mempartisi sekumpulan data menggunakan metode *depth-first greedy* atau

pendekatan *breadth-first* [16]. Semua *node* akan terhubung dengan garis. Dalam pohon keputusan, setiap simpul internal akan membagi ruang sampel

* 1. *Heatmap*
1. HASIL DAN PEMBAHASAN

menjadi subruang yang lebih kecil, biasanya dua atau lebih subruang, berdasarkan suatu fungsi diskrit yang terkait dengan nilai atribut.

Klasifikasi pohon keputusan memecahkan masalah kompleks dengan memisahkannya menjadi yang sederhana dan menyelesaikannya dengan membangun pohon keputusan berdasarkan pengetahuan yang diperoleh melalui teknik penambangan data. Dasar model pohon keputusan adalah membangun pohon keputusan dengan presisi tinggi dan skala kecil [17].

1. *Random Forest Classifier*

Algoritma *Random Forest* (RF) merupakan pengembangan dari metode *Classification and Regression Tree* (CART) dengan menerapkan teknik *bootstrap aggregating* (*bagging*) dan seleksi fitur acak [10]. Metode ini mudah digunakan dan diakui keakuratannya dalam mengatasi sampel kecil dan fitur dengan dimensi tinggi. Selain itu, metode ini dapat dengan mudah diparalelkan, sehingga cocok untuk sistem yang kompleks dalam kehidupan nyata [18]. Kelebihan yang lain adalah pada algoritma RF tidak terdapat *pruning* atau pemangkasan variabel seperti pada algoritma *decision tree.*

Metode *Random Forest* (RF) mengombinasikan beberapa pohon (*tree*), berbeda dengan pohon tunggal yang hanya terdiri dari satu pohon, untuk melakukan klasifikasi dan prediksi kelas. Pada RF, pembentukan pohon dilakukan dengan melatih sampel data. Pengambilan sampel dilakukan dengan penggantian (*sampling with replacement*). Pemilihan variabel untuk melakukan pemisahan (*split*) dilakukan secara acak. Setelah semua pohon terbentuk, klasifikasi dilakukan. Keputusan klasifikasi akhir diambil dengan cara menghitung rata-rata (menggunakan *mean* aritmatika) probabilitas penugasan kelas yang dihitung oleh semua pohon yang dihasilkan. Setiap pohon memberikan suara untuk keanggotaan kelas. Kelas keanggotaan dengan suara terbanyak akan menjadi yang akhirnya dipilih [10].

1. *Evaluation*

Tahap ini dilakukan dengan melihat tingkat performa dari pola yang dihasilkan oleh model yang digunakan. Evaluasi perbandingan algoritma menggunakan parameter confusion matrix dengan mengacu pada akurasi, presisi, dan recall. *Confusion matrix* tetap menjadi metode yang efektif hingga saat ini untuk mengukur dan mengevaluasi kinerja model klasifikasi. [19].

*Heatmap* merupakan suatu representasi grafis dari data dengan nilai individu yang terkandung dalam suatu matriks [20]. Visualisasi *heatmap* digunakan untuk melihat korelasi masing-masing variabel dalam kontribusinya dengan *output* data. Pada Gambar 3, menunjukkan grafik *heatmap* yang berisi nilai korelasi antar variabel yang terdapat pada data mengenai kontribusinya dalam menentukan *output* data. Setelah menjalankan fitur *heatmap*, didapatkan kesimpulan bahwa variabel *ratio\_to\_median\_purchase\_price*, variabel yang mewakili rasio harga pembelian transaksi terhadap harga pembelian median, memiliki korelasi paling besar dengan variabel *fraud,* variabel yang mewakili apakah transaksi tersebut penipuan, dengan nilai sebesar 0.46.

1. *Decision Tree*

Gambar 4 yang menunjukkan hasil *Decision Tree*, didapatkan hasil mengenai klasifikasi dari *Decision Tree* yang dijalankan. Pada awalnya, dilakukan klasifikasi apakah *ratio\_to\_median\_purchase\_price* kurang dari sama dengan 0.777 yang diambil menggunakan 1000000 sampel merupakan *true* atau *false.* Pada klasifikasi *true,* diklasifikasi apakah *distance\_from\_home* kurang dari sama dengan 1.122 menggunakan 896842 sampel. Jika true, maka kelas *output-*nya adalah *fraud* dengan sampel berjumlah 852163. Jika false maka dilakukan klasifikasi apakah *online\_order* kurang dari sama dengan -0.316. Jika *true* maka kelas *output*-nya adalah *fraud* dengan sampel sejumlah 15628, dan jika *false* maka kelas *output*-nya adalah *not fraud* dengan sampel sejumlah 29053.

Pada klasifikasi *false* yang merupakan hasil dari klasifikasi penentuan apakah *ratio\_to\_median\_purchase\_price* kurang dari sama dengan 0.777, dilakukan klasifikasi penentuan apakan *online\_order* kurang dari sama dengan

-0.316 menggunakan sampel sejumlah 103158. Jika *true* maka dilakukan klasifikasi penentuan apakah *distance\_from\_home* kurang dari sama dengan

1.123 menggunakan 36190 sampel yang jika *true* maka akan menghasilkan kelas *output fraud* dengan sampel sejumlah 34415 dan jika false maka akan menghasilkan kelas *output not fraud* sejumlah1775 sampel. Dan jika penentuan apakah *online\_order* kurang dari sama dengan -0.316 adalah *false* maka dilakukan klasifikasi penentuan apakah *used\_pin\_number* lebih kecil dari sama dengan

1.328. Jika *true* maka akan menghasilkan kelas *output fraud* sejumlah 6765 sampel, sedangkan jika *false* maka akan menghasilkan kelas *output not fraud* sejumlah 60203 sampel.

Gambar 3. Grafik *Heatmap*

Gambar 4. *Tree* yang dihasilkan *Decision Tree*

Tabel 5 menjelaskan hasil setelah dijalankannya *confusion matrix.* Hasil yang didapatkan berupa nilai *True Positive* (TP) sebesar 270403, nilai *False Positive* (FP) sebesar 2518, nilai *True Negative* (TN) sebesar 3537, dan nilai *False Negative* (FN) sebesar 23542. Kemudian pada Tabel 6, didapatkan tingkat akurasi dari dijalankannya *Decision Tree*. Tingkat akurasi yang didapatkan sebesar 0.99 atau 99%.

TABEL 5. *Confusion Matrix Decision Tree*

|  |  |  |
| --- | --- | --- |
|  | *Positive* | *Negative* |
| True | 270403 | *3537* |
| False | *2518* | *23542* |

TABEL 6. Akurasi Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.99 | 0.99 | 0.99 | 273940 |
| 3 | 0.87 | 0.90 | 0.89 | 26060 |
| accuracy |  |  | 0.98 | 300000 |
| macro avg | 0.93 | 0.95 | 0.94 | 300000 |
| weighted avg | 0.98 | 0.98 | 0.98 | 300000 |

1. *Random Forest*

Gambar 5 didapatkan hasil mengenai *random forest* yang telah dijalankan. Pada klasifikasi pertama, yaitu apakah *ratio\_to\_median\_purchase\_price* kurang dari sama dengan 0.777. Jika *true*, maka akan dilakukan klasifikasi lanjutan apakah *online\_order* kurang dari sama dengan minus 0.316 dengan sampel 396948. Dengan hasil *true*, dilakukan klasifikasi tambahan *distance\_from\_last\_transaction* apakah kurang dari sama dengan 1.751 dengan sampel 138535, yang dimana jika *true* maka output akan menunjukkan *fraud* dengan sampel 136514, begitu pula jika *false* yang menunjukkan kelas *fraud* dengan sampel 2021.

Pada klasifikasi variabel *online\_order* kurang dari sama dengan minus 0.316 *false*, akan dilanjutkan klasifikasi tambahan menggunakan variabel *used\_pin\_number* kurang dari 1.328 dengan sampel 258413. Jika *true*, maka output akan menghasilkan kelas *fraud* dengan sampel 232194 dan gini 0.081, begitu pula dengan *false* dengan sampel 26219 dan gini 0.0.

Pada klasifikasi variabel *ratio\_to\_median\_purchase\_price false,* dilakukan klasifikasi lebih lanjut menggunakan variabel *used\_pin\_number* kurang dari 1.328. Jika *true*, maka akan dilanjutkan klasifikasi lebih lanjut menggunakan variabel *distance\_from\_home* kurang dari 1.121 dengan sampel 40976 yang dimana *true* dan *false* menghasilkan *output* yang sama yaitu *not fraud*, dengan *true* bersampel 38924 dengang gini

0.435 dan *false* bersampel 2052 dengan gini 0.0.

Pada hasil *false used\_pin\_number* kurang dari 1.328, dilakukan klasifikasi lebih lanjut menggunakan variabel

*distance\_from\_last\_transaction* kurang dari 1.761. Jika *true,* maka akan menghasilkan *output* kelas *fraud* dengan sampel 4589 dan gini 0.038. Jika *false,* maka akan menghasilkan *output not fraud* dengan sampel 77 dan gini 0.484.

Gambar 5. *Tree* yang dihasilkan *Random Forest*

Pada Tabel 7, didapatkan hasil setelah dijalankannya *confusion matrix.* Hasil yang didapatkan berupa nilai *True Positive* (TP) sebesar 273940, nilai *False Positive* (FP) sebesar 7459, nilai *True Negative* (TN) sebesar 0, dan nilai *False Negative* (FN) sebesar 18601. Kemudian pada Tabel 8, didapatkan tingkat akurasi dari dijalankannya *Random Forest*. Tingkat akurasi yang didapatkan sebesar 0.975 atau 97,5% namun

dibulatkan menjadi 0.98 atau 98%..

TABEL 7. *Confusion Matrix Random Forest*

|  |  |  |
| --- | --- | --- |
|  | *Positive* | *Negative* |
| True | 273940 | *0* |
| False | *7459* | *18601* |

TABEL 8. Akurasi Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.97 | 1.00 | 0.99 | 273940 |
| 3 | 1.00 | 0.71 | 0.83 | 26060 |
| accuracy |  |  | 0.98 | 300000 |
| macro avg | 0.99 | 0.86 | 0.91 | 300000 |
| weighted avg | 0.98 | 0.98 | 0.97 | 300000 |

III. KESIMPULAN

Kartu kredit semakin populer dalam era digital, tetapi penipuan menggunakan kartu kredit juga semakin meningkat. Penelitian ini bertujuan untuk menentukan metode klasifikasi yang dapat membantu dalam mengidentifikasi transaksi ilegal yang menggunakan kartu kredit secara akurat. Dataset yang digunakan dalam penelitian ini adalah data transaksi menggunakan kartu kredit yang diambil dari situs penyedia *dataset* resmi, Kaggle sebanyak satu juta *records*.

Di dalam *dataset* ini terdapat variabel *fraud* yang mewakili apakah transaksi tersebut penipuan atau tidak. Oleh karena itu variabel ini akan digunakan sebagai target. Setelah dilakukan visualisasi dengan *heatmap* didapatkan hasil bahwa variabel *ratio\_to\_median\_purchase\_price* memiliki korelasi paling besar dengan variabel *fraud,* dengan nilai sebesar 0,46. Dengan demikian, variabel *ratio\_to\_median\_purchase\_price* akan dijadikan sebagai variabel *root node* yaitu variabel yang menjadi keputusan besar pertama yang akan menentukan apakah *record* tersebut termasuk penipuan atau tidak di akhir.

Dari klasifikasi yang telah dilakukan menggunakan *tools* Python didapatkan kesimpulan Algoritma Klasifikasi *Decision Tree* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan Algoritma Klasifikasi *Random Forest* yaitu 98%. Kemudian, algoritma klasifikasi *Random Forest* memiliki tingkat akurasi yang sedikit lebih rendah dengan algoritma klasifikasi *Decision Tree* dengan tingkat akurasi 97,5%.

Dibandingkan dengan penelitian sebelumnya [12], yang mendapatkan hasil akurasi terbesar oleh *random forest* sebesar 71,1% dan *decision tree* sebesar 64,9%, pada penelitian ini justru didapatkan hasil yang berlawanan yaitu hasil akurasi terbesar diperoleh *decision tree* sebesar 98% dan *random forest* sebesar 97,5%. Dengan kasus ini, dibuktikan bahwa hipotesis yang menyatakan bahwa *Random Forest* merupakan algoritma klasifikasi yang lebih baik daripada *Decision Tree* tidak dapat dibenarkan dan bersifat kondisional. Hal ini dibuktikan dengan berbedanya studi kasus yang diambil mengakibatkan hasil akurasi yang berbeda juga.

Berdasarkan kesimpulan dari penelitian yang telah

dilakukan, ada beberapa rekomendasi untuk meningkatkan kualitas penelitian di masa yang akan datang. Hasil penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan ilmu komputer, terutama dalam penggunaan Python untuk pembelajaran dan edukasi. Selain itu, hasil penelitian ini juga dapat digunakan sebagai acuan dan literatur bagi penelitian yang ingin melakukan penelitian serupa dengan cakupan responden yang lebih luas dan metode serta rancangan penelitian yang berbeda. Dalam penelitian selanjutnya, diharapkan bisa ditambahkan kelompok responden yang berbeda untuk menghasilkan variasi yang lebih tinggi dan mendapatkan informasi yang lebih bermanfaat.

REFERENCES

* 1. K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit Card Fraud Detection Using AdaBoost and Majority Voting,” *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.
	2. A. De Sá, A. Pereira, and G. Pappa, “A Customized Classification Algorithm for Credit-Card Fraud Detection,” *Eng Appl Artif Intell*, vol. 72, May 2018, doi: 10.1016/j.engappai.2018.03.011.
	3. Y. Jain, N. Tiwari, S. Dubey, and S. Jain, “A comparative analysis of various credit card fraud detection techniques,” *International Journal of Recent Technology and Engineering*, vol. 7, pp. 402–407, May 2019.
	4. J. K. Afriyie *et al.*, “A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions,” *Decision Analytics Journal*, vol. 6, p. 100163, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100163.
	5. H. Wang, C. Ma, and L. Zhou, “A Brief Review of Machine Learning and Its Application,” in *2009 International Conference on Information Engineering and Computer Science*, 2009, pp. 1–4. doi: 10.1109/ICIECS.2009.5362936.
	6. J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, “A survey of machine learning for big data processing,” *EURASIP J Adv Signal Process*, vol. 2016, no. 1, p. 67, 2016, doi: 10.1186/s13634-016-0355-x.
	7. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” *Comput Struct Biotechnol J*, vol. 15, pp. 104–116, 2017, doi: https://doi.org/10.1016/j.csbj.2016.12.005.
	8. Q. Dai, C. Zhang, and H. Wu, “Research of Decision Tree Classification Algorithm in Data Mining,” *International journal of database theory and application*, vol. 9, pp. 1–8, 2016.
	9. J. Han, M. Kamber, and J. Pei, “Data mining concepts and techniques, third edition.” Morgan Kaufmann Publishers, Waltham, Mass., 2012. [Online]. Available:

<http://www.amazon.de/Data-Mining-Concepts-Techniques-Managem> ent/dp/0123814790/ref=tmm\_hrd\_title\_0?ie=UTF8&qid=136603903 3&sr=1-1

* 1. M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016, doi: https://doi.org/10.1016/j.isprsjprs.2016.01.011.
	2. P. Kashyap, *Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making*, 1st ed. USA: Apress, 2018.
	3. H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, “A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes,” *J Res Health Sci*, vol. 18, no. 2, p. 412, 2018.
	4. M. Durairaj and N. Ramasamy, “A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,” *International Journal of Control Theory and Applications*, vol. 9, no. 27, pp. 255–260, 2016.
	5. E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, “The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms’ Performance,” *Data (Basel)*, vol. 6, no. 2, 2021, doi: 10.3390/data6020011.
	6. S. Cho *et al.*, “A Hybrid Machine Learning Approach for Predictive Maintenance in Smart Factories of the Future,” in *Advances in Production Management Systems. Smart Manufacturing for Industry 4.0*, I. Moon, G. M. Lee, J. Park, D. Kiritsis, and G. von Cieminski, Eds., Cham: Springer International Publishing, 2018, pp. 311–317.
	7. J. R. Gaikwad, A. B. Deshmane, H. V Somavanshi, S. V Patil, and R.
		1. Badgujar, “Credit card fraud detection using decision tree induction algorithm,” *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 4, no. 6, pp. 2278–3075, 2014.
	8. D. L. Talekar and K. P. Adhiya, “Credit Card Fraud Detection System: A Survey,” *International journal of modern engineering research (IJMER)*, vol. 4, no. 9, 2014.
	9. G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.
	10. A. Nurmasani and Y. Pristyanto, “Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class,” *Pseudocode*, vol. 8, no. 1, pp. 21–26, Mar. 2021, doi: 10.33369/pseudocode.8.1.21-26.
	11. C.-S. Yu *et al.*, “Clustering Heatmap for Visualizing and Exploring Complex and High-dimensional Data Related to Chronic Kidney Disease,” *J Clin Med*, vol. 9, no. 2, 2020, doi: 10.3390/jcm9020403.