PROKASDADIK, Volume 1, Desember (2023) E-ISSN: 3063-5845

Analisis dan Deteksi Risiko *Fraud* Pada Data Program Indonesia Pintar (PIP) Menggunakan Algoritma *Machine Learning* (Studi Kasus Penyaluran Dana PIP di Kab. Cianjur)

Rizki Izandi Gumay[1], Sajarwo Anggai[2]
Teknik Informatika, Program Pascasarjana, Universitas Pamulang
e-mail: rigi.nazki@gmail.com, dosen02832@unpam.ac.id

Abstrak—Program Indonesia Pintar (PIP) merupakan program strategis nasional yang diharapkan mampu menjamin peserta didik dapat melanjutkan pendidikan sampai dengan menyelesaikan pendidikan menengah, dan menarik siswa putus sekolah atau tidak melanjutkan pendidikan agar kembali mendapatkan layanan pendidikan. Program ini akan berjalan dengan lancar sesuai dengan tujuannya jika yang diberikan bantuan tepat jumlah yaitu peserta didik yang menerima dana PIP sesuai dengan nilai bantuan yang disalurkan. Penelitian ini bertujuan untuk mengetahui kinerja algoritma klasifikasi yang digunakan yaitu *Support Vector Machine*(SVM), *Random Forest Classifier*(RFC) dan *Naïve Bayes* (NB). selanjutnya melakukan analisa perbandingan kinerja, terhadap data yang telah dilakukan normalisasi data dan penanganan *outlier* pada ketiga algoritma tersebut. Hasil penelitian menunjukkan bahwa model algoritma *RFC* memiliki kinerja paling baik dengan nilai akurasi sebesar 0.948 dan logloss 0.272 dibandingkan dengan SVM dan NB.

Kata Kunci—Machine Learning*; *Data Mining*; *Fraud Detection*; *Naïve Bayes; Random Forest; Support Vector Machine.

I. PENDAHULUAN

Program Indonesia Pintar (PIP) merupakan program strategis nasional yang diharapkan mampu menjamin peserta didik dapat melanjutkan pendidikan sampai dengan menyelesaikan pendidikan menengah, dan menarik siswa putus sekolah atau tidak melanjutkan pendidikan agar kembali mendapatkan layanan pendidikan. PIP bukan hanya bagi peserta didik di satuan pendidikan formal, namun juga non-formal yang berlaku bagi peserta didik di Sanggar Kegiatan Belajar (SKB), Pusat Kegiatan Belajar Masyarakat (PKBM), Lembaga Kursus dan Pelatihan (LKP) dan Balai Latihan Kerja (BLK), atau satuan pendidikan nonformal lainnya, sesuai dengan kriteria yang telah ditetapkan.

Berdasarkan informasi yang bersumber dari Pusat Layanan Pembiayaan Pendidikan (Puslapdik) Kementerian Pendidikan Kebudayaan Riset dan Teknologi (Kemendikbudristek) Jumlah penerima program bantuan siswa miskin/rentan miskin yang sudah disalurkan pada Tahun 2022 adalah sebesar 17.953.268 peserta didik dengan nilai realisasi yang tersalurkan senilai Rp9.628.223.300.000,00 dengan rincian jenjang SD/sederajat sebanyak 10.360.614 (57.71%) peserta didik, dengan nilai tersalurkan senilai Rp4.212.276.300.000,00 (43.75%), SMP/sederajat sebanyak 4.369.968 (24.3%) peserta didik, dengan nilai tersalurkan senilai Rp2.711.107.500.000,00 (28.2%), SMA/sederajat sebanyak 1.393.519 (7.8%) peserta didik, dengan nilai tersalurkan senilai Rp1.175.672.500.000,00 (12.2%) dan SMK sebanyak 1.829.167 (10.2%) peserta didik, dengan nilai tersalurkan senilai Rp1.529.1667.000.000,00 (15.9%). Program ini akan berjalan dengan lancar sesuai dengan tujuannya jika yang diberikan bantuan tepat jumlah yaitu peserta didik yang menerima dana PIP sesuai dengan nilai bantuan yang disalurkan.

Dalam pelaksanaan penyaluran dana bantuan sosial PIP, berdasarkan hasil audit Inspektorat Jenderal Kemendikbudristek ditemukan permasalahan penyaluran dana PIP yang diterima oleh peserta didik penerima PIP tidak diterima sesuai jumlah (potongan) yang telah disalurkan bahkan tidak menerima dana PIP sama sekali, tidak sedikit permasalahan tersebut yang telah diproses oleh pihak Aparat Penegak Hukum (Kepolisian atau Kejaksaan). Hal ini jelas bahwa program ini belum sepenuhnya tepat jumlah.

Untuk dapat memastikan salah satu indikator ketercapaian program tersebut diperlukan tindakan dari Inspektorat Jenderal Kemendikburistek untuk segera melakukan deteksi *fraud* (*fraud detection*) untuk mengurangi dampak risiko khususnya risiko reputasi, risiko operasional (karena ada kerugian untuk mengganti dana PIP Peserta Didik) dan risiko hukum (atas tuntutan dari orang tua/wali peserta didik karena terdapat kelemahan sistem di internal Kementerian).

Maka untuk mengatasi permasalahan diatas diperlukan sebuah cara yang efektif untuk mencegah *fraud* yang terjadi pada penyaluran dana bantuan sosial PIP. Saat ini ada bidang ilmu yang relatif baru yang bernama *data mining* yang bisa dimanfaatkan untuk memberikan solusi atas permasalahan diatas. *Data mining* adalah suatu teknik analisa yang relatif cepat dan mudah untuk menemukan pengetahuan, pola dan/atau relasi antar data, secara otomatis [1]. *Data mining* juga merupakan suatu teknik menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data (*database*) yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui [2], *data mining* berkontribusi untuk mengidentifikasi *fraud* dan segera

bertindak untuk menurunkan kemungkinan terjadinya *fraud* dengan menemukan pola dan mengidentifikasi data anomali [3].

Fraud detection dapat dikategorikan sebagai permasalahan klasifikasi biner karena *output* dari metode ini ada 3 yaitu risiko tinggi, risiko sedang dan risiko rendah. Pada keilmuan *data mining* terdapat beberapa metode yang umum digunakan untuk menyelesaikan masalah klasifikasi biner, antara lain *Naïve Bayes (NB), Random Forrest Classifier (RFC)* dan *Support Vector Machine (SVM)*. Pada penelitian ini akan disajikan tahapan implementasi *machine learning* dalam mendeteksi risiko *fraud* pada program bantuan sosial PIP mulai dari tahap ekplorasi data, pra proses, implementasi dan pengujian pada beberapa algoritma yang sudah dipilih.

Berdasarkan penjelasan diatas, penelitian ini bertujuan untuk melakukan analisis data dan implementasi algortima *machine learning* untuk mendeteksi risiko *fraud* pada program bantuan sosial PIP studi kasus penyaluran PIP di Kab. Cianjur.

II. METODE PENELITIAN

Tahapan langkah penelitian yang dilakukan tentang deteksi risiko *fraud* studi kasus penyaluran dana bantuan sosial PIP di Kab. Cianjur menggunakan *data mining* yaitu :

A. Tahapan Pengumpulan Informasi

Pada tahapan ini keluaran yang diharapkan adalah diperolehnya suatu data penyaluran bantuan sosial dana PIP di Kab. Cianjur Provinsi Jawa Barat dan konsep penerapan algoritma *machine learning* yang dapat mendeteksi risiko *fraud*.

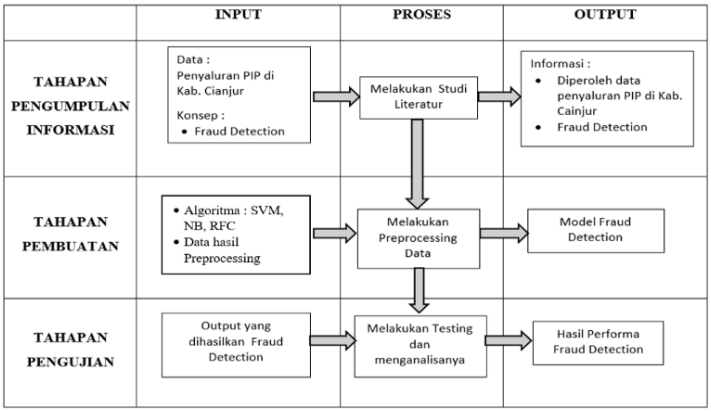
B. Tahapan Pembuatan

Pada tahapan ini keluaran yang diharapkan adalah telah terbentuk model risiko *fraud detection* dengan menggunakan data yang telah di olah dan algoritma yang telah dipilih sebelumnya.

C. Tahapan Pengujian

Pada tahapan ini keluaran yang diharapkan adalah dapat diketahui hasil kinerja algoritma yang dipilih dalam mendeteksi *fraud*.

Gambar 1 menggambarkan tahapan langkah penelitian yang dilakukan dalam studi tentang deteksi *fraud*.



Gambar 1.
Tahapan langkah penelitian

Dalam melakukan penelitian ini menggunakan aplikasi Orange datamining, aplikasi ini merupakan aplikasi *open source* yang terbukti mampu membantu peneliti menganalisa *data mining*[4], dengan penerapan algoritma tertentu melalui konsep visual programing. Dengan aplikasi orange, dapat diketahui akurasi dari mulai *preprocessing* hingga hasil akhir dengan metode tertentu dengan cepat dibandingkan perhitungan manual.

A. Deskripsi Data

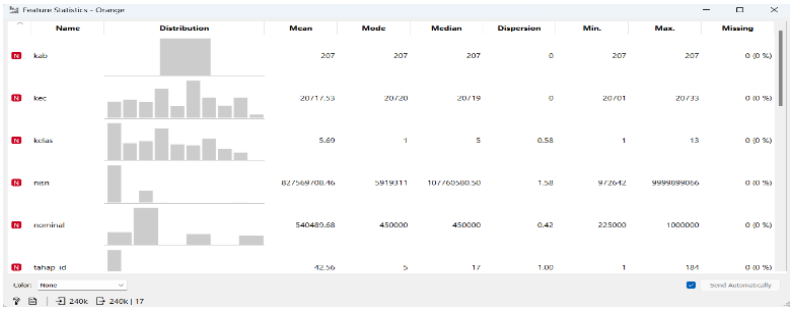
Dataset yang digunakan di dapat dari aplikasi SIPINTAR (http://pip.kemdikbud.go.id) dan dari Dapodik (https://dapo.kemdikbud.go.id/) yang terdiri dari 5 (lima) dataset untuk data latih dan data uji. Kelima dataset tersebut adalah data penyaluran dana PIP untuk jenjang SD/sederajat, jenjang SMP/sederajat, jenjang SMA/sederajat, jenjang SMK/sederajat dan data siswa per sekolah di Kab. Cianjur. Berdasarkan dataset tersebut diperoleh informasi sesuai pada tabel 1 sebagai berikut :

Tabel 1.
Informasi dataset yang diperoleh

Jenjang	Jumlah Siswa Penerima PIP	Jumlah Total Siswa	Jumlah Satuan Pendidikan
SD	143.053	253.480	1.017
SMP	54.191	98.903	402
SMA	13.228	33.542	94
SMK	25.261	59.929	186
PKBM	4.041	45.802	161
SLB	326	658	8
Total	240.084	492.392	1.869

Dari data penyaluran bantuan sosial PIP di Kab. Cianjur diketahui disalurkan dana PIP ke 240.084 peserta didik/baris data

(siswa) dari 1.869 satuan pendidikan (sekolah). Berdasarkan data tersebut dilakukan *Exploratory Data Analysis* (EDA), EDA merupakan proses menganalisis dan menampilkan data bertujuan mendapatkan pemahaman yang lebih baik tentang wawasan dari data [5]. EDA ini dapat digunakan untuk mengetahui pola data serta bentuk sebarannya. Untuk mengidentifikasi pola data dan sebarannya ada beberapa metode yang dapat digunakan diantaranya diagram batang dan histogram. diperoleh informasi statistik pada Gambar 2, sebagai berikut:

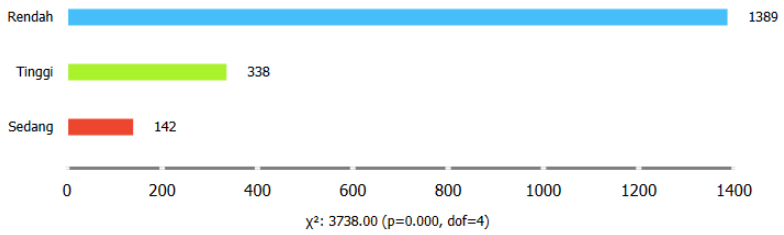


Gambar 2.
Informasi statistik dataset

B. Prapemrosesan Data

Setelah data awal didapatkan dan dieksplorasi untuk memahami pola data dan sebaran termasuk juga dengan melihat kejadian lampau sekolah yang melakukan tindakan *fraud*, maka dilakukan praproses data dengan tahapan sebagai berikut yaitu:

1. *Merging* data untuk menggabungkan *dataset* yang didapatkan.
2. Proses pembersihan data yang memiliki nilai kosong, juga dilakukan perubahan nilai dengan mengubah data menjadi bentuk *binary* yaitu '1,0'.
3. Proses *feature engineering* dengan membuat fitur baru dari fitur yang sudah ada dengan tujuan menambah informasi sekolah yang memiliki potensi *fraud* yang didapatkan dari dataset, berdasarkan hasil *feature engineering* diperoleh informasi persentase tingkat risiko *fraud* dengan kategori rendah sebanyak 1.389 sekolah atau 74.32%, kategori sedang sebanyak 142 sekolah atau 7.60% dan kategori tinggi sebanyak 338 sekolah atau 18.08% dari total keseluruhan sekolah sebanyak 1.869 sekolah dengan diagram pada Gambar 3 sebagai berikut:



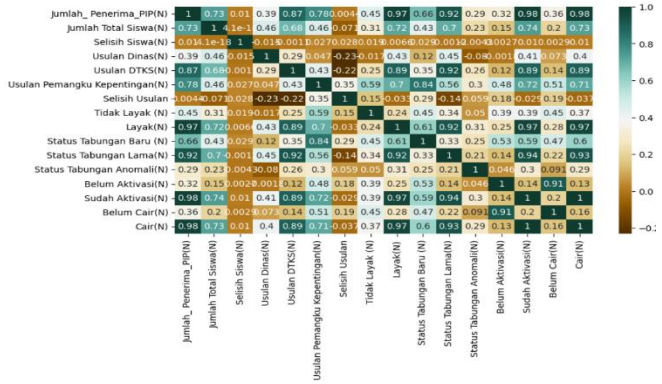
Gambar 3.
Distribusi Risiko

4. Proses *feature selection*, untuk melihat seberapa berpengaruhnya suatu fitur terhadap target dan mengetahui fitur mana saja yang memiliki informasi penting dalam dataset. Berikut pada Gambar 4 merupakan informasi fitur yang memiliki pengaruh terhadap target.

		#	Info. gain	Gain ratio	Gini	ANOVA
1	Selisih Usulan		0.444	0.489	0.173	1306.893
2	Usulan Pemangku Kepentingan(N)		0.200	0.100	0.076	212.753
3	Status Tabungan Baru (N)		0.094	0.047	0.041	133.442
4	Selisih Siswa(N)		0.073	0.037	0.012	0.625
5	Jumlah_ Penerima_PIP(N)		0.037	0.019	0.013	22.104

Gambar 4.
Ranking Fitur 5 teratas

5. Analisis *feature* untuk didapatkan informasi pengaruh dan korelasi antar fitur sehingga dapat diketahui seberapa tinggi pengaruh terhadap target dengan menggunakan pendekatan visualisasi heat maps supaya mudah untuk dipahami, seperti pada Gambar 5.



Gambar 5.
Heat Maps korelasi antar *feature.*

Pada gambar 5, menjelaskan bahwa warna menunjukan tingkat korelasi antar fitur, semakin kuning maka semakin tidak berpengaruh, sementara semakin hijau maka semakin berpengaruh.

6. Proses normalisasi data (*scaling*) merupakan proses penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu[6], dikarenakan data yang diperoleh cukup variatif satuanya sehingga memiliki potensi model yang dihasilkan menjadi kurang akurat. Teknik yang digunakan yaitu *rescaling min-max normalization*, metode normalisasi ini melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses[6], dengan persamaan (1) sebagai berikut:

$$Normalized(x) = \frac{minRange + (x - minValue)(maxRange - minRange)}{MaxValue - MinValue} \quad (1)$$

Proses ini sangat penting dilakukan untuk meningkatkan nilai *accuracy* dari *precision, recall, f1-score* dan *Logloss* pada table *predictive* dan *actual confusion matrix*.

7. Proses *encoding* data yang bertujuan untuk mengubah data berbentuk kategori menjadi data bilangan.
8. Mendeteksi *outlier* merupakan suatu pengamatan yang jarak titik pengamatannya relatif jauh dari pusat data dan titik pengamatannya menyimpang dari pola data[7]. *Outlier* dapat terjadi karena kesalahan manusia, kesalahan instrumen, perilaku curang, perubahan perilaku sistem atau kesalahan sistem, dan penyimpangan alami di dalam populasi.

C. Algoritma Machine Learning

Pada tahapan ini dilakukan proses klasifikasi terhadap data yang telah dilakukan prapemrosesan data menggunakan metode *Support Vector Machine* (SVM), *Naïve Bayes* (NB) dan *Random Forest Classifier* (RFC). Metode klasifikasi adalah teknik *data mining* yang mengelompokkan data berdasarkan lampiran dari data sampel. Metode klasifikasi juga merupakan proses mengklasifikasikan dokumen ke dalam satu atau lebih kategori tertentu atau dokumen dari kategori yang sama[8].

1) Support Vector Machine (SVM)

SVM merupakan algoritma yang menggunakan pemetaan nonlinear untuk mengubah data latih ke skala atau ukuran yang lebih tinggi [9]. SVM adalah metode yang ampuh untuk membangun *classifier*. Ini bertujuan untuk membuat batas keputusan antara dua kelas sehingga memungkinkan prediksi label dari satu atau beberapa fitur *vector* [10]. SVM memetakan data *input* nonlinear ke beberapa ruang dimensi yang lebih tinggi, dimana data dapat dipisahkan secara linier, sehingga memberikan kinerja klasifikasi atau regresi yang besar [11]. Konsep sederhana SVM, yakni usaha mencari *hyperplane* terbaik yang berfungsi sebagai batas dari dua buah kelas berdasarkan *support vectors* yang merupakan *vector* data berjarak paling mendekati *hyperplane* dan batas yang menyatakan *hyperplane* pemisah [12]. untuk mencari *hyperplane* yang didefinisikan sebagai berikut:

$$w.x_i + b = 0 \quad (2)$$

$$d = \frac{|(wi.xi) + b|}{\sqrt{||w||^2}} \quad (3)$$

$$y_i[(w_i \cdot x_i) + b] - 1 \geq 0 \quad (4)$$

$$L = \frac{1}{2}||w_i||^2 - \sum_{i=1}^{1} \alpha_i \, (y_i(x_i.w + b) - 1) \quad (5)$$

Ada beberapa metode SVM Multi Kelas yaitu salah satunya metode SVM Muti Kelas *One-Against-One* [1]. Pada metode *One-Again-One*, dengan cara membangun sejumlah model SVM biner yang nantinya akan dibandingkan satu kelas dengan kelas lainnya. Untuk mengklasifikasikan data ke k-kelas, maka harus membangun sejumlah $\frac{k(k-1)}{2}$ model SVM

biner.

2) Naïve Bayes (NB)

Naïve Bayes (NB) merupakan sebuah metode dengan teknik prediksi probabilitas dengan berdasarkan pada penerapan *teorema bayes* dimana antara suatu fitur dengan fitur lain dalam suatu data itu tidak saling keterkaitan, teknik metode ini merupakan salah satu bentuk sederhana untuk klasifikasi dengan persamaan dapat dilihat pada persamaan (6). [13]

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (6)$$

Di mana X adalah bukti, lalu H adalah hipotesa, sementara P(H|X) adalah probabilitas bahwa hipotesis H benar untuk bukti X atau dengan kata lain P(H|X) merupakan probabilitas posterior H dengan syarat X, selanjutnya penjelasan dari P(X|H) adalah probabilitas bahwa bukti X untuk hipotesis H atau probabilitas posterior X dengan syarat H, P(H) adalah probabilitas prior hipotesis H dan P(X) adalah probabilitas prior Bukti X.

3) Random Forest Classifier (RFC)

Metode RFC menerapkan metode *boostrap aggregating* (bagging) dan *random feature selection* yang merupakan perkembangan dari metode *Classification and Regression Trees*(CART) [14]. Klasifikasi *random forest* dilakukan melalui kombinasi pohon dengan dilakukannya percobaan pada sampel yang telah disediakan. Dalam penentuan kelas klasifikasi pada RFC yaitu berdasarkan voting dari beberapa pohon keputusan yang telah terbentuk dimana pemenang dari pohon tersebut ditentukan dengan vote yang paling banyak (*majority vote*). RFC membutuhkan dua parameter, yaitu jumlah *tree* yang akan dipakai dan jumlah variabel independent yang digunakan untuk proses pencabangan agar mencapai nilai yang optimal. Untuk mendapatkan nilai $mtry$ yaitu disarankan menggunakan persamaan sebagai berikut:

$$mtry_1 = \frac{1}{2}|\sqrt{p}| \quad (7)$$

$$mtry_2 = |\sqrt{p}| \quad (8)$$

$$mtry_3 = 2 \text{ x } |\sqrt{p}| \quad (9)$$

Dimana:
$mtry$: banyaknya variabel independent untuk setiap split
p: banyaknya variabel independent

D. Pengujian

Tahapan terakhir setelah model klasifikasi terbentuk, akan dilakukan pengujian dengan melakukan evaluasi keakuratan pengklasifikasian supaya memperoleh ketepatan klasifikasi yang akurat dan baik. Ukuran yang digunakan untuk mengukur kinerja klasifikasi yaitu *confusion matrix*. Analisis dilakukan dengan menghitung presisi, recall dan akurasi dari hasil klasifikasi menggunakan *confusion matrix*. *Confusion matrix* terdiri dari elemen yang diklasifikasikan secara benar dan tidak benar dari setiap kelas. Salah satu manfaat dari *confusion matrix* adalah memberikan kemudahan untuk melihat terjadinya kesalahan sistem dalam meletakkan kelas klasifikasi [15], seperti ditampilkan dalam tabel 2.

Tabel 2.
Confusion Matrix

		Predicted Class		
		Positive	***Negative***	***Total***
Actual Class	*Positive*	*True Positve (TP)*	*False Negative (FN)*	*TP+FN*
	Negative	*False Positive (FP)*	*True Negative (TN)*	*FP+TN*
	Total	*TP+FP*	*FN+TN*	*TP+FN+FP+TN*

Ukuran dalam mengevaluasi kinerja model berdasarkan *confusion matrix* ada berbagai macam diantaranya akurasi, sensitivitas, presisi, *f-measure*, dan nilai AUC. Rumus-rumus yang digunakan yaitu sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

$$Sensitivity = Recall = TP_{rate} = \frac{TP}{TP+FN} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$F-measure = 2\,X\,\frac{Precision\ x\ Recall}{Precision+\ Recall} \quad (13)$$

$$AUC = \frac{1+TP_{rate}-FP_{rate}}{2} \quad (14)$$

Parameter hasil akurasi, presisi, *recall* dapat dikatakan menghasilkan klasifikasi yang bagus atau tidak dengan menggunakan pedoman parameter hasil klasifikasi yang ditampilkan di tabel 3 [15].

Tabel 3.
Parameter Hasil Klasifikasi

Rentang	Hasil Klasifikasi
90-100%	*Excellent Classification*
80-90%	*Good Classification*
70-80%	*Fair Classification*
60-70%	*Poor Classification*
50-60%	*Failure*

Selain itu juga dilakukan Data uji coba yang digunakan dalam penelitian ini yaitu data asli setelah normalisasi untuk setiap model klasifikasi SVM, RFC dan NB. Untuk setiap percobaan dilakukan uji coba dengan data latih 80% dengan 20% dengan uji coba pada kondisi pengujian sebagai berikut:

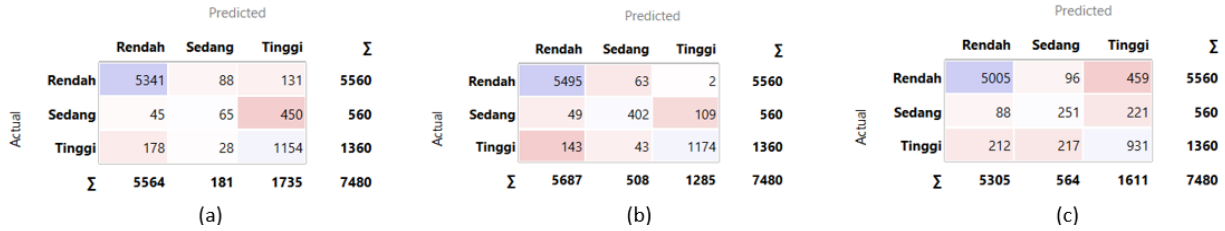
1. Data di uji dengan dataset sebelum dilakukan normalisasi data (*scaling*) dibandingkan dengan setelah dilakukan normalisasi data (*scaling*).
2. Data di uji dengan dilakukan penanganan *outlier* dengan membandingkan metode *covarian estimator, one class SVM, local outlier factor dan isolation forest* untuk mencari nilai *Accuracy dan Logloss* terbaik.

III. HASIL DAN PEMBAHASAN

Hasil dari berbagai observasi data yang telah dilakukan menggunakan ketiga model klasifikasi algoritma tersebut akan dibandingkan nilai kinerja satu sama lain dari model tersebut menggunakan *Confusion Matrix* untuk melihat model yang lebih baik digunakan pada data penyaluran dana bantuan sosial PIP untuk mendeteksi risiko *fraud.*

A. Evaluasi Hasil Prediksi

Pada tahapan ini, prediksi yang dihasilkan oleh model klasifikasi SVM, RFC dan NB menggunakan data asli setelah dilakukan normalisasi dan dianalisis menggunakan *Confusion Matrix* serta di hitung nilai *AUC, Accuracy,nilai F-1, Precision, Recall,* dan *Logloss*. Analisis dilakukan untuk melihat seberapa baik SVM, RFC dan NB dalam mengklasifikasi dalam mendeteksi risiko *fraud* pada penyaluran dana bantuan sosial PIP diperoleh hasil *Confusion Matrix* sebagaimana Gambar 6.



Gambar 6.
Confusion matrix (a) SVM, (b) RFC, (c) NB data latih 80% data uji 20%.

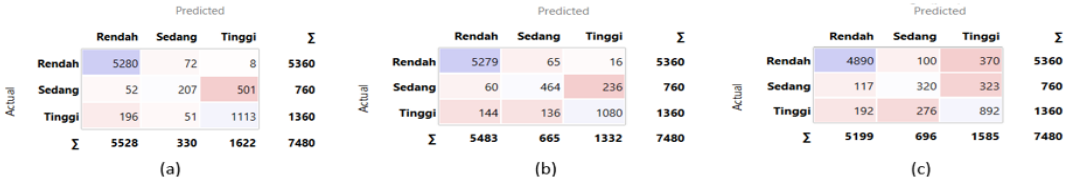
Tabel 4.
Hasil Uji coba perbandingan klasifikasi pada komposisi data 80% data latih 20% data uji

Model	AUC	CA	F-1 Score	Precision	Recall	Logloss
SVM	0.953	0.877	0.863	0.861	0.877	0.338
RFC	0.981	0.948	0.947	0.947	0.948	0.272
Naïve Bayes	0.904	0.827	0.832	0.840	0.827	0.547

Berdasarkan Tabel 4 dapat diketahui bahwa kinerja dari model RFC lebih baik dari model SVM dan NB. Akurasi klasifikasi sudah mencapai hasil yang sempurna, namun masih dapat ditingkatkan kembali mengingat masih ada kemungkinan ditemukan nilai eror. Hal tersebut dipengaruhi oleh adanya nilai *outlier* yang belum dilakukan penanganan terhadapnya dan banyaknya data uji dan data latih yang digunakan dalam proses simulasi yang dilakukan.

B. Hasil Uji Coba Dataset Sebelum Dilakukan Normalisasi Data

Pada bagian ini dilakukan uji coba klasifikasi kelas *fraud* pada data menggunakan perbandingan model klasifikasi SVM, RFC dan NB dengan perbandingan data latih 80% data dan data uji 20% pada data sebelum dilakukan normalisasi data (*scaling)*. Dari *confusion matrix* hasil klasifikasi pada Gambar 7 dihitung nilai ACU, *Balanced Accuracy*, *Recall, Precision*, *F-1 Score* dan *Logloss* pada hasil klasifikasi SVM, RFC dan NB.



Gambar 7.
Confusion matrix (a) SVM, (b) RFC, (c) NB data latih 80% data uji 20% sebelum dilakukan normalisasi data.

Tabel 5.
Hasil uji coba perbandingan sebelum dan sesudah normalisasi dengan data 80% data latih, 20% data uji

Model	AUC		CA		F-1 Score		Precision		Recall		Logloss	
	Sebelum	*Scaling*	Sebelum	*Scaling*	Sebelum	*Scaling*	Sebelum	*Scaling*	Sebelum	*Scaling*	Sebelum	*Scaling*
SVM	0.965	0.953	0.882	0.877	0.869	0.863	0.873	0.861	0.882	0.877	0.324	0.338
RFC	0.965	0.981	0.912	0.948	0.910	0.947	0.912	0.947	0.912	0.948	0.451	0.272
NB	0.906	0.904	0.816	0.827	0.819	0.832	0.816	0.840	0.816	0.827	0.549	0.547

Pada Tabel 5 yang menunjukkan hasil perbandingan metrik kinerja antara sebelum dilakukan *scaling* dan setelah dilakukan *scaling*. Tabel 5 menunjukkan hasil yang tidak jauh berbeda dengan Tabel 4 dimana didapatkan hasil model klasifikasi RFC lebih baik, karena setelah dilakukan *scaling* mengalami peningkatan pengukuran dibandingkan dengan model klasifikasi SVM dan NB.

Informasi yang didapatkan dari tabel 5 tersebut diketahui sebagian besar hasil uji coba dari klasifikasi menggunakan model klasifikasi SVM memiliki penurunan nilai kinerja setelah dilakukan *scaling* data dibandingkan sebelum dilakukan *scaling*, sementara klasifikasi menggunakan model klasifikasi NB mengalami penurunan pengukuran pada matrik AUC, selebihnya mengalami peningkatan setelah dilakukan *scaling* meskipun tidak signifikan kenaikannya.

C. Hasil Uji Coba Penanganan Outlier

Pada bagian ini dilakukan uji coba klasifikasi kelas *fraud* pada data menggunakan perbandingan model klasifikasi SVM, RFC dan NB setelah dilakukan penanganan *outlier* dengan menggunakan metode *Covarian Estimator, One Class SVM, Local Outlier Factor* dan *Isolation Forest* perbandingan data latih 80% data dan data uji 20% untuk mencari nilai nilai *Accuracy dan Logloss* terbaik.

Tabel 6.
Hasil uji coba penanganan outlier pada komposisi data 80% data latih 20% data uji

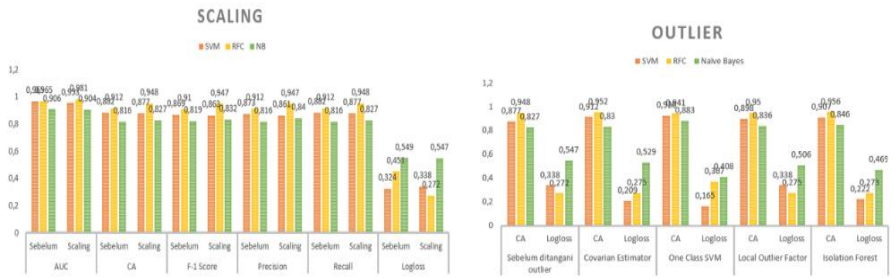
Model	Sebelum ditangani outlier		*Covarian Estimator*		*One Class SVM*		*Local Outlier Factor*		*Isolation Forest*	
	CA	Logloss	CA	Logloss	CA	Logloss	CA	Logloss	CA	Logloss
SVM	0.877	0.338	0.912	0.209	0.926	0.165	0.898	0.338	0.907	0.222
RFC	0.948	0.272	0.952	0.275	0.941	0.367	0.950	0.275	0.956	0.273
Naïve Bayes	0.827	0.547	0.830	0.529	0.883	0.408	0.836	0.506	0.846	0.469

Pada Tabel 6 yang menunjukkan hasil perbandingan metrik kinerja *Accuracy dan Logloss* pada model klasifikasi SVM, RFC dan NB dengan penanganan *outlier* menggunakan metode *covarian estimator, one class SVM, local outlier factor* dan *isolation forest.* Secara kinerja *Accuracy* model RFC masih lebih baik dalam memprediksi data yang benar dari seluruh prediksi dibandingkan SVM dan NB, namun dalam mengukur *Logloss* untuk memprediksi probablitas dari data target tidak mengalami peningkatan justru semakin menurun ketika diterapkan *outlier* metode *one class SVM*.

Sementara selain itu diperoleh informasi dari tabel 6, diketahui *Accuracy* model SVM dan NB dalam memprediksi data yang benar dari seluruh prediksi, mengalami peningkatan dan pengukuran *Logloss* semakin membaik dalam memprediksi probablitas dari data target terutama ketika diterapkan *outlier* metode *one class SVM*.

D. Pembahasan

Pada bagian ini akan ditunjukkan pembahasan berdasarkan hasil klasifikasi yang sudah dilakukan pada ujicoba sebelumnya, didapatkan hasil perbandingan dari model SVM, RFC dan NB pada klasifikasi tiga kelas dalam deteksi risiko *fraud* data penyaluran dana bantuan sosial PIP. Dari hasil uji coba tersebut diketahui bahwa nilai metrik kinerja pada pengujian model dengan perbandingan sebelum dan setelah *scaling* serta diterapkannya penanganan *outlier*, dari hasil tersebut dapat disimpulkan bahwa kemampuan dalam *generate model* untuk setiap metode cukup baik. Oleh karena itu didapatkan nilai rata-ratanya untuk melihat perbandingan pada setiap uji coba.

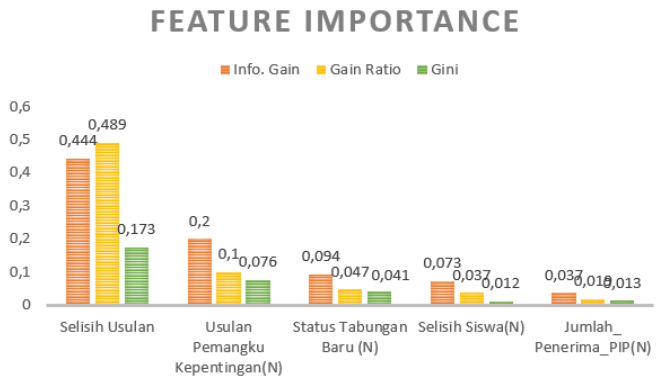


Gambar 8.
Perbandingan nilai matrik kinerja hasil percobaan

Pada Gambar 8 perbandingan nilai matrik kinerja dari data latih dan data uji. Terlihat bahwa semua uji coba klasifikasi model RFC memiliki tingkat pengukuran kinerja yang lebih baik dibandingkan dengan SVM dan NB, begitu juga setelah dilakukan penanganan *outlier*. Kinerja RFC terutama dengan metode *Isolation Forest* meningkat dari *Accuracy* 0.948 menjadi 0.956.

Namun pada penelitian ini penentuan model yang terbaik dengan memfokuskan pada nilai *Accuracy* dan *Logloss*. *Accuracy* dapat memperlihatkan secara umum bagaimana kinerja model pada data, sedangkan *Logloss* pada penelitian ini merepresentasikan seberapa dekat probabilitas prediksi dengan nilai benar terhadap target dalam hal ini memprediksi tingkat risiko *fraud*.

Sehingga dengan melihat pada Gambar 9, model RFC pada data penyaluran dana Bantuan Sosial PIP memiliki nilai *Accuracy* dan *Logloss* terbaik, dengan kata lain RFC adalah model yang dapat dengan sangat baik memprediksi tingkat risiko *fraud* dengan kategori risiko tinggi, sedang dan rendah.



Gambar 9.
5 (lima) *feature* penting

Untuk melihat pola yang dilakukan satuan pendidikan/sekolah yang melakukan tindak *fraud* maka dilakukan proses *feature importance* untuk melihat fitur apa yang berperan penting atas klasifikasi kecurangan tersebut. Pada Gambar 9 ditunjukkan 5 (lima) teratas *feature importance* pada dataset yaitu :

1. Fitur yang pertama adalah fitur yang menjelaskan selisih jumlah penerima PIP antara usulan pemangku kepentingan dengan usulan DTKS dan Dinas, semakin besar nilainya maka semakin ada kemungkinan terjadinya kecurangan.
2. Fitur yang kedua adalah fitur selisih jumlah siswa penerima PIP dengan jumlah total siswa di suatu sekolah, semakin besar nilainya maka semakin ada kemungkinan terjadinya kecurangan.
3. Dari hasil observasi tersebut dapat dilakukan pertimbangan kebijakan yang dikeluarkan Kemendikbudristek maupun stakeholder pendidikan di kabupaten/kota dan provinsi dalam melakukan kebijakan untuk mendeteksi adanya risiko *fraud*.

IV. KESIMPULAN

Model klasifikasi SVM, RFC dan NB dapat digunakan untuk klasifikasi deteksi risiko *fraud* pada data penyaluran dana bantuan sosial PIP dengan melakukan praproses pembersihan data, *feature engineering*, melakukan proses *encoding one-hot* pada beberapa fitur kategorikal, melakukan normalisasi data dengan teknik *rescaling min-max normalization*, lalu mengklasifikasi data menjadi tiga kelas berbeda yaitu kelas risiko tinggi, sedang dan rendah dengan menggunakan model klasifikasi SVM, RFC dan NB. Hasil perbandingan klasifikasi SVM, RFC dan NB didapatkan bahwa metode RFC lebih baik dibandingkan dengan SVM dan NB dilihat dari nilai *Accuracy* dan *Logloss* yang memiliki nilai terbaik didapatkan dengan menggunakan model RFC dengan nilai 0.948 dan 0.272 pada data dibandingkan SVM dengan nilai 0.877 dan 0.388 sementara model NB dengan nilai 0.827 dan 0.527.

V. SARAN

Beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya yaitu

1. Dapat dilakukan penyeimbangan data dikarenakan kelas target memiliki jumlah yang tidak seimbang/ketimpangan antara kelas rendah, sedang dan tinggi dengan menggunakan metode *undersampling* atau *oversampling*.
2. Dapat menggunakan model algoritma klasifikasi atau *clustering machine learning* lainnya untuk mendapatkan nilai akurasi yang lebih baik.
3. Dapat diterapkan dalam dataset penyaluran dana Bantuan PIP secara nasional, sehingga dapat mempresentasikan akurasi model *machine learning*.

Lalu saran bagi penyalur bantuan PIP dalam hal ini Kemendikbudristek untuk melakukan pendampingan dan pengawasan terhadap penyaluran dana PIP bagi sekolah yang mendapatkan jumlah penerima terbanyak dari usulan pemangku kepentingan.

DAFTAR PUSTAKA

[1] S. Suyanto, *Machine Learning tingkat Dasar Dan Lanjut*. Informatika Bandung, 2018.

[2] I Kadek Juni Arta, Gede Indrawan, Gede Rasben Dantes, "*Data Mining Rekomendasi Calon Mahasiswa Berprestasi Di STIMIK Denpasar Menggunakan Metode Technique For Others Reference By Similarity To Ideal Solution*," Jurnal Ilmu Komputer Indonesia (JIKI) Vol: 4, No. 1, Februari 2019.

[3] Matin N. Ashtiani And Bijan Raahemi, "*Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review*," *IEEE Access* Digital Object Identifier 10.1109/ACCESS.2021.3096799 Volume 10, 2022.

[4] Hozairi, Anwari, Syariful Alim, "*Orange Data Mining Implementation For Student Graduation Classification Using K-Nearest Neighbor, Decision Tree And Naive Bayes Models*," Jurnal Ilmiah Nero Vol. 6 No. 2, 2021.

[5] Muhammad Radhi, Amalia, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan Sinurat, Evta Indra, "*Analisis Big Data Dengan Metode Exploratory Data Analysis (EDA) Dan Metode Visualisasi Menggunakan Jupyter Notebook*," Jusikom Prima, Vol. 4 No. 2 Februari 2021 E-ISSN : 2580-2879.

[6] D. A. Nasution, H. H. Khotimah, dan N. Chamidah,"*Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma K-NN*," CESS (Journal Comput. Eng. Syst. Sci., vol. 4, no. 1, hal. 78– 82, 2019.

[7] Puji Puspa Sari, Erna Tri Herdiani, Nurtiti Sunusi," *Outlier Detection Using Minimum Vector Variance Algorithm with Depth Function and Mahalanobis Distance*," Jurnal Matematika, Statistika & Komputasi, E-ISSN 2614-8811 Vol. 17, No. 3, 418 - 427, May, 2021DOI: 10.20956/j.v17i3.12629.

[8] H. Derajad Wijaya and S. Dwiasnati, "*Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat*," *Jurnal Informatika*, vol. 7, no. 1, 2020, [Online]. *Available*: http://ejournal.bsi.ac.id/ejurnal/index.php/ji.

[9] S. A. Taher, K. A. Akhter, and K. M. A. Hasan, "*N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine*," in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1–5. doi: 10.1109/ICBSLP.2018.8554716.

[10] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "*Applications of support vector machine (SVM) learning in cancer genomics*," Cancer Genomics and Proteomics, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[11] S. Ghosh, A. Dasgupta, and A. Swetapadma, "*A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification*," in 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.

[12] L. Yahaya, N. D. Oye, and E. J. Garba, "*A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques*," Am. J. Artif. Intell., vol. 4, no.1, pp. 20–29, 2020, doi: 10.11648/j.ajai.20200401.12.

[13] Mohamad Efendi Lasulika, "*Komparasi Naïve Bayes, Support Vector Machine Dan K- Nearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran TV Kabel*" ILKOM Jurnal Ilmiah Volume 11 Nomor 1 April 2019 *e-ISSN 2548-7779.*

[14] Breiman, L. (2001). Random Forests. (R. E. Schapire, Ed.) Machine Learning, 45, 5-32. Retrieved April 22, 2021, from https://link.springer.com/article/10.1023/A:1010933404324.

[15] Pertiwi, D. P., Wiranto, & Anggrainingsih, R. (2019). Evaluation Of Campaign Categories On Kitabisa.Com By Naive Bayes Classifier Method. ITSMART: Jurnal Ilmiah Teknologi dan Informasi, 8(1), 26-33.

[16] Gorunescu, Data Mining: Concepts, Models and Techniques, Springer, 2011.