SKRIPSI

IMPLEMENTASI DETEKSI FRAUD PENGGUNAAN JARINGAN INTERNET FIBER DI PERUSAHAAN PENYEDIA LAYANAN INTERNET

Disusun dan diajukan oleh:

ADY AHMADI SUWARDI
D121171517

PROGRAM STUDI SARJANA TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2023

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI DETEKSI FRAUD PENGGUNAAN JARINGAN INTERNET FIBER DI PERUSAHAAN PENYEDIA LAYANAN INTERNET

Disusun dan diajukan oleh

ADY AHMADI SUWARDI

D121171517

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 22 November 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,

Dr. Amil Ahmad Ilham, S.T., M.IT.
Nip. 197310101998021001

Pembimbing Pendamping,

A. Ais Prayogi, S.T., M.Eng.
Nip. 198305102014041001

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng.
Nip. 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;

Nama	:	Ady Ahmadi Suwardi
NIM	:	D121171517
Program Studi	:	Teknik Informatika
Jenjang	:	S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Implementasi Deteksi Fraud Penggunaan Jaringan Internet Fiber di Perusahaan Penyedia Layanan Internet

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, .29...November. 2023

Menyatakan

Ady Ahmadi Suwardi

ABSTRAK

ADY AHMADI SUWARDI. *Implementasi Deteksi Fraud Penggunaan Jaringan Internet Fiber di Perusahaan Penyedia Layanan Internet* (dibimbing oleh Amil Ahmad Ilham dan A.Ais Prayogi Alimuddin)

Sektor telekomunikasi dapat dikatakan sebagai salah satu sektor bisnis yang besar di Indonesia dengan perkembangan bisnis yang sangat pesat. Perkembangan yang sangat pesat ini mengakibatkan perusahaan-perusahaan di bidang Telekomunikasi mengalami kebocoran pendapatan yang relatif besar setiap tahunnya. *Fraud* merupakan salah satu penyebab kebocoran pendapatan yang relatif besar dan paling berbahaya bagi perusahaan Telekomunikasi. Selain menimbulkan kerugian finansial secara langsung dalam jumlah besar, *fraud* juga berdampak pada citra perusahaan terhadap *customer* maupun *investor*.

Penelitian ini bertujuan untuk merancang sistem pendeteksi kecurangan/*fraud* pada sektor telekomunikasi menggunakan pendekatan *machine learning*. *Supervised Machine Learning* adalah jenis algoritma yang mampu menghasilkan pola dan hipotesis umum dengan menggunakan dataset yang ada untuk memprediksi kejadian di masa depan yang sangat sesuai penggunaannya pada penelitian ini yaitu klasifikasi deteksi *fraud*.

Algoritma *Supervised Machine Learning* yang akan digunakan dalam penelitian ini adalah *Random Forest, XG Boost, Logistic Regression,* dan *Backpropagation Neural Network*. Dalam proses pembuatan modelnya, dataset yang diperoleh akan diproses dengan pembersihan data (cleaning), penyeimbangan data(balancing) menggunakan teknik Oversampling Adaptive Synthetic (ADASYN), dan transformasi data menggunakan penskalaan standard. Evaluasi model akan diukur berdasarkan *confusion matrix*, *recall*, accuracy, kurva ROC, dan *execution time*. Penelitian ini bertujuan untuk mendapatkan perbandingan algoritma yang paling optimal dari keempat algoritma yang digunakan untuk diterapkan sebagai sistem pendeteksi *fraud* penggunaan jaringan internet fiber di perusahaan penyedia layanan internet.

Didapatkan hasil menunjukkan masing-masing algoritma mendapatkan nilai *recall*, *accuracy,* dan rata-rata *execution time* untuk *Random Forest* adalah 91,92%, 99,39%*,* dan 2,3864582 detik, *XG Boost* adalah 93,33%, 98,12%*,* dan 17,1476647 detik, *Logistic Regression* adalah 90,17%, 96,14%*,* dan 0,4112701 detik dan *Backpropagation Neural Network* adalah 93,33%, 97,82%, dan 6 menit 34,52594 detik.

Kata Kunci: *Fraud*, *Supervised Machine Learning*, *ADASYN*, *Standard Scaling Logistic Regression*, *Random Forest*, *XGBoost*, *Backpropagation Neural Network.*

ABSTRACT

ADY AHMADI SUWARDI. *Implementation of Fraud Detection at Fiber Internet Networks in Internet Service Provider Companies* (supervised by Amil Ahmad Ilham and A.Ais Prayogi Alimuddin)

The telecommunications sector can be said to be one of the big business sectors in Indonesia with very rapid business development. This rapid development has resulted in telecommunications companies experiencing a relatively large leakage of revenue each year. Fraud is one of the causes of revenue leakage which is relatively large and the most dangerous for telecommunications companies. Apart from causing large direct financial losses, fraud also has an impact on the company's image to customers and investors.

This study aims to design a fraud detection system in the telecommunication sector using a machine learning approach. Supervised Machine Learning is a type of algorithm that is capable of generating general patterns and hypotheses using existing datasets to predict future events, which is very suitable for its use in this study, namely the classification of fraud detection. The Supervised Machine Learning algorithms that will be used in this research are Random Forest, XG Boost, Logistic Regression, and Backpropagation Neural Network.

In the process of creating the model, the dataset obtained will be processed by cleaning data, balancing data using the Oversampling Adaptive Synthetic (ADASYN) technique, and transforming data using standard scaling. Model evaluation will be measured based on confusion matrix, recall, accuracy, ROC curve, and execution time. This study aims to obtain a comparison of the most optimal algorithm of the four algorithms used to be applied as a fraud detection system using fiber internet networks in internet service provider companies.

The results shows that each algorithm has a recall, accuracy and execution time for Random Forest is 91,92% , 99,39% and 2,3864582 seconds, XG Boost is 93,33%, 98,12% and 17,1476647 seconds, Logistic Regression is 90,17%, 96,14% and 0,4112701 seconds and Backpropagation Neural Network is 93,33%, 97,82% and 6 minutes 34,52594 seconds.

Keyword: *Fraud*, *Supervised Machine Learning*, *ADASYN*, *Standard Scaling Logistic Regression*, *Random Forest*, *XGBoost*, *Backpropagation Neural Network.*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
n	jumlah kelas target
p_i	proporsi kelas i terhadap partisi D
v	jumlah partisi
D_j	total partisi ke j
D	total baris pada semua partisi
$\hat{y}_i^{(t)}$	F*inal tree model*
$\hat{y}_i^{(t-1)}$	Model pohon yang dihasilkan sebelumnya
$f_t(xi)$	Model baru yang dibangun
t	Jumlah total model dari base tree model
$\hat{y}_i$	Nilai actual
$l(\hat{y}_i^{(t)}, \hat{y}_i)$	*Lost function*
$\Omega(f_i)$	istilah regularisasi
T	Jumlah *leaf*
ω	Bobot *leaf*
λ dan γ	Koefisien, dengan nilai default ditetapkan untuk λ=1 dan γ=0

v_{0j}	Bobot layer input bias ke unit tersembunyi ke-j
x_i	unit input ke-i
v_{ij}	bobot unit input ke-i ke layer tersembunyi ke-j
z_j	nilai unit tersembunyi ke-j menggunakan fungsi aktivasi sigmoid
e	nilai konstanta = 2,718
$y_n e\ t_k$	nilai unit *output* ke-k
w_{0k}	bobot unit tersembunyi bias ke unit *output* ke-k
w_{jk}	bobot unit tersembunyi ke-j unit *output* ke-k
y_k	nilai unit *output* ke-k menggunakan fungsi aktivasi sigmoid
δ_k	nilai *error* unit *output*
z_j	nilai target *output*
α	*learning rate*
$\Delta wj\ k$	perubahan bobot unit tersembunyi ke-j ke unit output ke-k
d	derajat keseimbangan
	jumlah data yang memiliki kelas minoritas

mr	jumlah data yang memiliki kelas mayoritas
mx	banyaknya sintetis data yang akan dibuat
G	level keseimbangan
B	*density distribution*
rx	*ratio*
r	data sintetis baru
S	data minority class yang masuk ke dalam perulangan
xi	data dari data latih yang dipilih secara acak
xu	
s	
λ	angka acak dari 0 sampai 1

DAFTAR LAMPIRAN

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah SWT yang senantiasa memberikan rahmat serta hidayahnya serta diberi kelancaran dan kesehatan sehingga penulis dapat menyelesaikan karya tulisnya yang berjudul **"Implementasi Deteksi Fraud Penggunaan Jaringan Internet Fiber di Perusahaan Penyedia Layanan Internet"** guna memenuhi salah satu persyaratan dalam menyelesaikan jenjang Strata-1 di Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Penulis menyadari sepenuhnya bahwa skripsi ini masih jauh dari kesempurnaan karena menyadari segala keterbatasan yang ada. Dalam penulisan skripsi ini penulis menghadapi berbagai kendala dan masalah, namun karena usaha yang maksimal dan kemampuan yang Tuhan berikan kepada penulis serta bantuan dan dukungan dari berbagai pihak, maka penulisan skripsi ini dapat selesai. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

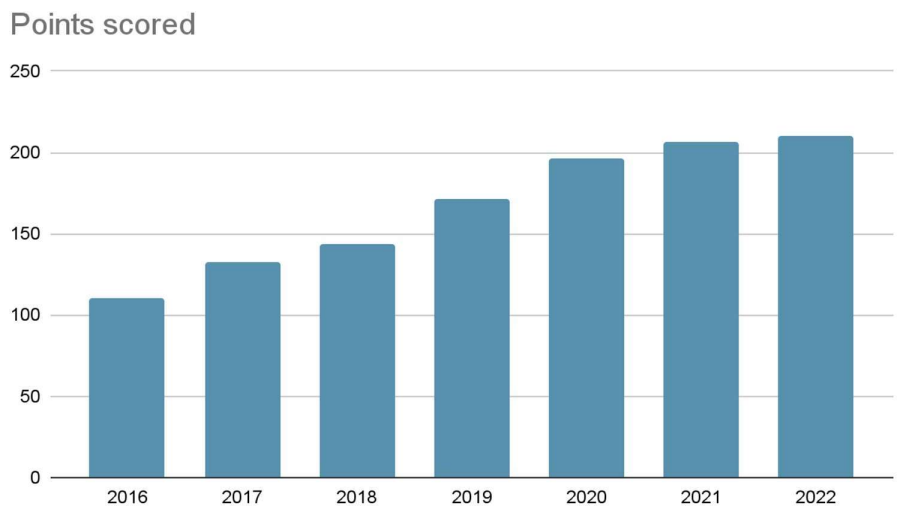
1. Allah SWT ,Tuhan pencipta alam semesta yang senantiasa memberikan rahmat serta hidayahnya kepada penulis.
2. Kedua orang tua penulis, Bapak Suwardi Abubakar dan Ibu Marhaeni yang selalu memberikan kasih sayang, nasehat, motivasi, dukungan, dan doa kepada penulis.
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing utama dan Bapak A.Ais Prayogi Alimuddin, S.T., M.Eng. selaku pembimbing pendamping yang senantiasa menyediakan waktu, tenaga, pikiran, dan perhatian yang luar biasa dalam mengarahkan penulis dalam penyusunan tugas akhir ini.
4. Bapak Robert, Bapak Zainuddin dan Ibu Yuanita serta segenap staf Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah membantu kelancaran penyelesaian tugas akhir penulis.
5. Segenap Dosen dan Staf Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah banyak membantu semasa perkuliahan hingga penyelesaian tugas akhir penulis.
6. Saudara seperjuangan penulis RECOGN17ER yang telah menemani dan mendukung perjalanan penulis sekaligus tempat berbagi keluh kesah selama menjadi mahasiswa teknik di Departemen Informatika Fakultas Teknik Universitas Hasanuddin.
7. Seluruh pihak yang tidak sempat disebutkan satu persatu yang telah banyak meluangkan tenaga, waktu, dan pikiran selama penyusunan tugas akhir ini.

BAB I
PENDAHULUAN

1.1. Latar Belakang

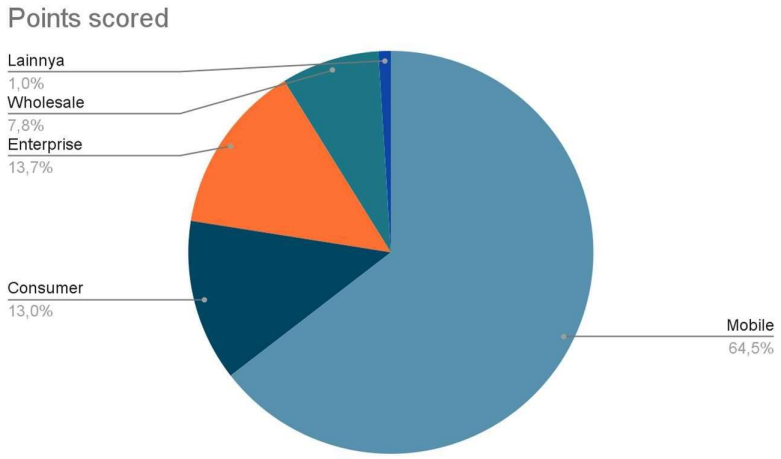
Internet dapat dikatakan sebagai kebutuhan hidup utama masyarakat dunia saat ini, tak terkecuali di Indonesia. Internet dapat mempengaruhi banyak aspek diantaranya aspek pendidikan, kesehatan, rekreasi, perdagangan dan banyak hal penting lainnya serta merubah kebiasaan lama dalam berperilaku di masyarakat (Jain, 2017). Terbukti dari data statistik yang dikeluarkan oleh Kementerian Komunikasi dan Informatika (Kominfo) (2022), menyebutkan bahwa Indonesia terdata sebagai salah satu negara dengan pengguna internet paling aktif di media sosial, data menunjukkan 95% dari konsumen digital di Indonesia memanfaatkan internet untuk mengakses jejaring sosial dengan rata-rata menghabiskan waktu 6 jam sehari.

Data dari Badan Pusat Statistika (BPS), Produk Domestik Bruto atas dasar harga konstan di kuartal kedua tahun 2021, sektor telekomunikasi memberikan kontribusi sebesar Rp. 172,39 Triliun atau tumbuh sebesar 6,87% dari kuartal yang sama pada tahun sebelumnya. Membuktikan bahwa sektor telekomunikasi adalah salah satu sektor yang dapat bertahan dari dampak penyebaran pandemi Covid-19. Sementara itu Pusat Domestik Bruto atas dasar harga berlaku di sektor telekomunikasi sebesar 185,21 triliun pada kuartal kedua tahun yang sama. Jumlah tersebut berkontribusi sebesar 4,4% terhadap Produk Domestik Bruto nasional yang mencapai Rp.4.175,43 triliun. Selain itu, berdasarkan hasil survey Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) hingga awal tahun 2022 seperti yang ditunjukkan pada gambar 1 pengguna internet mencapai 210 juta atau 77% dari total populasi yang berjumlah sekitar 273 juta penduduk Indonesia.



Gambar 1 Jumlah pengguna internet di Indonesia tahun 2016-2022 sumber Asosiasi Penyelenggara Jasa Internet Indonesia(APJII))

Pada sektor telekomunikasi, yang memimpin pasar adalah PT. Telekomunikasi Indonesia (Telkom Indonesia) dengan pangsa pasar yaitu 92,9% pada bagian jaringan internet fiber (Laporan Tahunan PT. Telkom tahun 2019) dan 44,7% pada bagian jaringan internet mobile (Survey APJII, 2022). Sebagai pemegang pangsa pasar terbesar di Indonesia, Telkom Indonesia sendiri merupakan perusahaan Badan Usaha Milik Negara (BUMN) yang bergerak di bidang industri telekomunikasi dengan pendapatan tahunan mencapai Rp.135 Triliun (Laporan Tahunan 2019). Dari angka tersebut, terlihat pada gambar 2, sebagian besar pendapatan disumbang oleh segmen *mobile* melalui anak perusahaannya yaitu Telkomsel yang mencapai 64,4% dari total pendapatan, kemudian disusul oleh segmen *consumer* melalui layanan Indihome sebesar 13%, dan sisanya melalui segmen *enterprise, wholesale,* dan *digital service* lainnya (Laporan Tahunan 2019).



Gambar 2 Kontribusi setiap segmen bisnis terhadap pendapatan PT Telekomunikasi Indonesia (dalam Persentasi)

Dengan posisi ini, sebagai pemimpin pasar pada sektor telekomunikasi yang berbasis pengguna dan potensi pasar yang sangat besar tentunya harus sangat memperhatikan keseimbagan pengeluaran dan kualitas yang diberikan kepada konsumen. Hal ini dikarenakan keseimbangan antara pendapatan dan kualitas yang diberikan merupakan tujuan utama dari setiap perusahaan yang melakukan kontak langsung dengan konsumen dari barang maupun jasa yang ditawarkan dengan tujuan pendataan pendapatan yang transparan, meningkatkan profit dan mengurangi pengeluaran, serta menyesuaikan pendapatan dan pengeluaran dengan perkembangan teknologi dan inovasi(Mattison, 2005). Kebocoran pendapatan di perusahaan telekomunikasi disebabkan oleh faktor-faktor seperti tunggakan, pelanggan tidak membayar (*unpaid customer*), pelanggaran/kecurangan (f*raud*), kesalahan pada tagihan *(billing error*), pergantian pelanggan (*customer churn*), dan lain-lain(N. Lu et al., 2014). Setelah ditinjau lebih jauh dari data internal perusahaan, Telkom Indonesia mengalami kebocoran pendapatan mencapai Rp.9,1 Triliun pada tahun 2020. Namun diantara beberapa faktor tersebut fraud merupakan faktor yang paling berdampak terhadap keberlangsungan perusahaan kedepannya karena merupakan aktivitas terencana yang digunakan untuk keuntungan pribadi

(Gajbhiye, 2013). Hal ini tentunya menyebabkan kemerosotan pendapatan perusahaan yang kemudian akan berdampak terhadap kualitas dan performa perusahaan kedepannya.

Mengingat pentingnya deteksi dini dalam sistem deteksi fraud, besarnya jumlah dataset yang akan dianalisis, dan didukung oleh fakta bahwa semua proses *fraud detection* di PT. Telkom Indonesia dilakukan secara manual & investigasi langsung, maka penelitian ini akan membahas tentang pemanfaatan *machine learning* dalam *fraud detection* di perusahaan penyedia layanan internet. Didukung oleh data yang menyebutkan bahwa pada tahun 2020 PT. Telkom Indonesia mengalami kebocoran pendapatan sebesar 9,1 Triliun Rupiah, dimana 3 Triliun diantaranya diakibatkan fraud oleh pelanggan PT. Telkom Indonesia. S*ubscription fraud* merupakan klasifikasi penyebab terbesar, dimana pelanggan menggunakan lebih dari yang seharusnya mereka gunakan (*overuse*) (Kabari, 2016).

Machine learning dapat menggunakan data yang dikumpulkan dari perilaku pengguna untuk menentukan pola perilaku pengguna dan dapat menyimpulkan secara mandiri apabila terjadi kejanggalan data yang terjadi menggunakan algoritma *machine learning* untuk mengidentifikasi titik abnormal yang terjadi pada data. *Fraud* sering kali terjadi pada kartu kredit dan *e-commerce* dikarenakan jumlah perputaran uang yang sangat banyak dan cepat, dan mungkin saja terjadi pada sektor lain dengan ciri-ciri yang serupa (Meng et al., 2020).

Penelitian ini menggunakan *machine learning* untuk mencari informasi lebih dalam melalui *raw data* yang ada menggunakan *dataset* yang sangat besar. Teknik *machine learning* sendiri dapat dibagi menjadi empat jenis berdasarkan pengaplikasiannya yaitu *supervised learning* yang merupakan model yang dibuat berdasarkan kelas atau label dari informasi yang sudah ada, *unsupervised learning* yang proses pembelajarannya dilakukan dengan menggunakan data yang belum diberi label dengan melakukan identifikasi pola dan informasi secara langsung dan menyeluruh, *Semi-supervised learning* yang merupakan penggabungan cara belajar *supervised* dan *unsupervised* machine learning dan *reinforcement learning* yang

menggali informasi melalui *trial & error* (Shing Lim et al., 2021). Dari ketiga jenis tersebut, penelitian pendeteksi *fraud* dalam bisnis telekomunikasi pada umumnya menggunakan *supervised learning* dengan label dan atribut yang telah ditentukan terlebih dahulu.

Algoritma yang akan digunakan dalam penelitian ini yaitu: *logistic regression, random forest, XGBoost,* dan *Backpropagation Neural Network. Logistic regression* dan *random forest* sering kali dibandingkan karena kedua algoritma ini menggunakan metode yang berbeda yaitu linear dan non-linear dalam pengaplikasiannya dalam pengklasifikasian kelas dengan hasil yang baik serta memiliki ketahanan terhadap *overfitting*(Trigila et al., 2015). Sedangkan *XGBoost* digunakan sebagai algoritma perbandingan yang dari beberapa penelitian yang menggunakan algoritma ini mendapatkan hasil yang lebih baik dari random forest yang sama sama berbasis pohon keputusan(Huang et al., 2020), *Backpropagation Neural Network(BPNN)* digunakan sebagai algoritma pembanding dengan 3 algoritma sebelumnya cenderung memiliki waktu eksekusi lebih lama, namun tergantung pada arsitekturnya bisa sangat kuat dan mampu menangkap pola kompleks dalam data, BPNN dapat menangkap hubungan non-linear yang lebih kompleks (S. Lu et al., 2019).

Berdasarkan latar belakang yang telah diuraikan sebelumnya, penulis akan melakukan penelitian untuk melihat perbandingan dari 4 algoritma yang digunakan untuk menentukan model paling optimal dalam sistem deteksi *fraud* pada sektor Telekomunikasi di PT. Telekom Indonesia yaitu *Random Forest, xgboost, logistic regression,* dan *backpropagation neural network*. Algoritma yang digunakan diputuskan berdasarkan penelitian-penelitian sebelumnya terkait dengan *fraud detection* yang juga menggunakan algoritma yang sama namun dengan dataset yang berbeda yaitu dataset dalam sektor *e-commerce* dan perbankan. Maka penulis menjadikan penelitian ini sebagai tugas akhir dengan judul **"Implementasi Deteksi Fraud Penggunaan Jaringan Internet Fiber di Perusahaan Penyedia Layanan Internet"**.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, masalah yang ingin diatasi yaitu *fraud* di perusahaan penyedia layanan internet. Hal ini penting karena dapat mengganggu keberlangsungan perusahaan dan tentunya menyebabkan banyak kerugian, Oleh karena itu Rumusan masalah yang akan dihadapi adalah:

1. Bagaimana cara mendeteksi *fraud* menggunakan sistem *fraud detection* untuk meminimalisir kebocoran pendapatan perusahaan di perusahaan penyedia layanan internet?
2. Apa algoritma paling efisien yang dapat digunakan untuk mendeteksi *fraud* sedini mungkin pada perusahaan penyedia layanan internet?

1.3. Tujuan

Berdasarkan rumusan masalah diatas, maka tujuan dari penelitian ini yaitu:

1. Membuat sistem deteksi *fraud* berdasarkan *dataset* yang ada demi meminimalisir kebocoran pendapatan perusahaan di perusahaan penyedia layanan internet.
2. Menemukan algoritma paling efisien yang dapat digunakan untuk mendeteksi *fraud* sedini mungkin pada perusahaan penyedia layanan internet.

1.4. Manfaat

Manfaat dari pengerjaan tugas akhir ini adalah memberikan kontribusi pada literatur ilmiah tentang deteksi *fraud* di bidang telekomunikasi khususnya penyedia layanan internet, dan mendukung pengembangan teknologi deteksi *fraud* yang dapat meningkatkan kinerja sistem di masa depan.

1.5. Ruang Lingkup

Ruang lingkup dari pengerjaan tugas akhir ini adalah :

a. Penelitian ini akan menggunakan 4 algoritma yang memiliki cara kerja yang berbeda-beda yaitu *Random Forest, xgboost, logistic regression* dan *backpropagation neural network*.
b. Penelitian ini akan menggunakan dataset pelanggan jaringan internet fiber di PT. Telkom Indonesia regional 7 Kawasan Indonesia Timur(KTI) dengan total 136.244 data yang digunakan.
c. Variabel-variabel penelitian terdiri dari :
    - Nomor Identitas Pelanggan
    - STO
    - Data Usage
    - Billing
    - Internet Speed
    - Umur Berlangganan
    - Metode Pembayaran
    - Flagging

BAB II
TINJAUAN PUSTAKA

2.1 Industri Telekomunikasi

Faktor dasar yang ingin dicapai setiap perusahaan di sektor telekomunikasi demi mewujudkan tujuan setiap perusahaan adalah keterbukaan informasi, keuangan perusahaan dibuat transparan dimaksudkan agar setiap pihak diluar perusahaan mengetahui gambaran yang lengkap mengenai prospek dan kinerja perusahaan(Lestari J et al., 2019). Dalam industri telekomunikasi sendiri, persaingan sangat ketat terjadi karena bisnis ini sangat bergantung pada kualitas dan kuantitas yang ditawarkan, dan juga didorong oleh perkembangan teknologi yang sangat pesat membuat perusahaan telekomunikasi harus terus beradaptasi dengan cepat terhadap perkembangan tersebut (Weiss, 2009).

Sektor telekomunikasi adalah salah satu bidang yang pertama kali mengadopsi teknologi data mining. Hal ini kemungkinan besar karena perusahaan telekomunikasi secara rutin menghasilkan dan menyimpan sejumlah besar data berkualitas tinggi, memiliki basis pelanggan yang sangat besar, dan beroperasi dalam lingkungan yang berubah dengan cepat dan sangat kompetitif (Vitor & De Sousa, 2014).

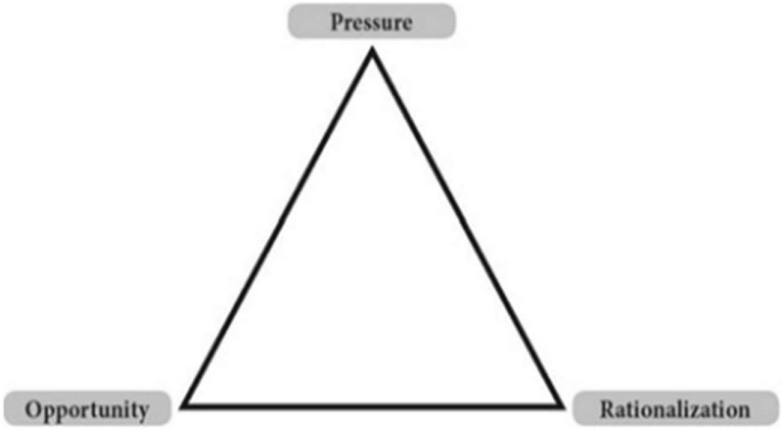
2.2 Fraud

2.2.1 Definisi Fraud

Kecurangan/*fraud* menurut *The Association of Certified Fraud Examiners* (ACFE) adalah segala sesuatu yang dilakukan seseorang/kelompok, baik itu merugikan bagi orang lain atau demi mendapat keuntungan pribadi dengan cara menutupi kebenaran, tipu daya, kelicikan atau mengelabui dengan cara tidak jujur yang lain (Wardhani, 2014). Sumber lain memaparkan bahwa *fraud* adalah istilah yang mengacu pada penipuan yang disengaja untuk mendapatkan sesuatu yang bernilai, biasanya berupa uang. *Fraud* umumnya dilakukan melalui penggunaan informasi palsu, representasi yang salah, atau perilaku tidak jujur yang

dimaksudkan untuk menyesatkan atau menipu. Karena *fraud* diklasifikasikan sebagai aktivitas ilegal dengan tujuan merugikan individu atau kelompok, ada banyak undang-undang yang mencakup istilah dan artinya. Salah satunya undang-undang di Slovakia mendefinisikan *fraud* sebagai tindakan yang disengaja dengan tujuan memperkaya individu atau orang lain dengan merugikan properti orang/institusi lain melalui misrepresentasi orang lain atau mengambil keuntungan dari kesalahan orang lain, sehingga menyebabkan kerugian bagi orang itu (Eldin Mohammed Abd El-Hamid Ahmed Abdou et al., 2019).

Fraud sendiri muncul dari segitiga *fraud* diantaranya yaitu tekanan, peluang dan rasionalitas yang lebih dikenal dengan *the cressey fraud triangle* (Awalluddin et al., 2022). Yang terlampir pada gambar 3 Teori ini pada umumnya diimplementasikan dalam segala bentuk *fraud* dan fokus kepada faktor yang mempengaruhi motivasi pelaku *fraud.*



Gambar 3 *The cressey fraud triangle*

Fraud umumnya diklasifikasikan menjadi 5 jenis yaitu superimposed *fraud, subscription fraud, technical fraud, internal fraud,* dan *social engineering.* namun pada kasus pada penelitian ini, *subscription fraud* merupakan jenis fraud yang paling sering ditemui pada sektor telekomunikasi (Kabari, 2016).

2.2.2 Fraud pada Industri Telekomunikasi

Salah satu faktor penyebab kerugian pendapatan di sektor telekomunikasi adalah bersumber dari *fraud* oleh pelanggan (Chua & Wareham, 2004). *Fraud* pada industri telekomunikasi global dan dampaknya terhadap bisnis cenderung menjadi langkah awal yang mendasari berbagai jenis dan metode penipuan lainnya untuk mendukung industri dengan pemahaman yang lebih baik tentang pola dibalik aktivitas penipuan(Martinez-Plumed et al., 2021). Untuk mencegah terjadinya *fraud*, penting bagi penyedia layanan dalam hal ini penyedia layanan internet untuk memahami aspek-aspek dasar *fraud* itu sendiri agar terhindar dari *fraud*, modernisasi dan bentuk teknologi baru seperti *blockchain* dan *artificial intelligent* semakin populer dan menjadi alat yang menjanjikan dalam meminimalisir terjadinya *fraud* (Kryvinska & Greguš, 2008). Karena dampak yang ditimbulkan oleh *fraud* ini lah, maka perancangan sistem *fraud prevention & detection* sangatpenting (Estévez et al., 2006).

Sebagai jenis *fraud* yang paling sering terjadi, jenis pelanggan dapat dikategorikan masuk ke dalam *subscription fraud* dengan karakteristik antara lain (Rosset et al., 1999) :

- *Usage* atau jumlah penggunaan yang tidak normal atau melewati batas yang seharusnya dalam satu periode;
- *Billing* yang terlihat normal, atau pelanggan tidak mendapatkan *overcharge* atau tagihan tambahan sama sekali
- Penggunaan pribadi yang intensif dan sangat tinggi
- Terjadi di daerah perkotaan

Ciri-ciri diatas umumnya terjadi di semua sektor bisnis telekomunikasi, dan implementasinya telah diaplikasikan di sektor jaringan internet mobile, namun tidak menutup kemungkinan sistem ini dapat diterapkan dalam bisnis telekomunikasi pada bagian jaringan internet fiber sebagaimana sistem ini memiliki cara kerja yang mirip dengan sistem *fault detection* (kesalahan atau anomali yang terjadi).

Dalam kasus penelitian yang dilakukan kali ini yaitu deteksi fraud pada jaringan internet fiber di perusahaan penyedia layanan internet dalam hal ini di PT. Telkom Indonesia, fraud yang terjadi seringkali disebabkan oleh beberapa faktor diantaranya yaitu pemanfaatan kekuasaan ataupun kemampuan yang dilakukan oleh oknum pegawai PT. Telkom, mengambil keuntungan pribadi oleh pelanggan terhadap bug pada sistem jaringan internet fiber, ataupun kesalahan pada sistem yang biasanya terjadi karena *human error*.

2.3 Machine Learning

2.3.1 Pengertian Machine Learning

Machine learning adalah disiplin ilmu yang mempelajari bagaimana menggunakan komputer untuk mensimulasikan kegiatan belajar manusia, dan mempelajari metode pengembangan diri komputer untuk memperoleh pengetahuan baru dan keterampilan baru, mengidentifikasi pengetahuan yang ada, dan terus meningkatkan kinerja dan akurasi yang didapatkan (Somvanshi & Chavan, 2016). Umumnya di dalam aplikasinya, *machine learning* sendiri memiliki keunggulan yaitu dapat melakukan analisis multivariat dengan cepat dan efisien walaupun menggunakan data yang sangat besar (Venkata Suryanarayana et al., 2018).

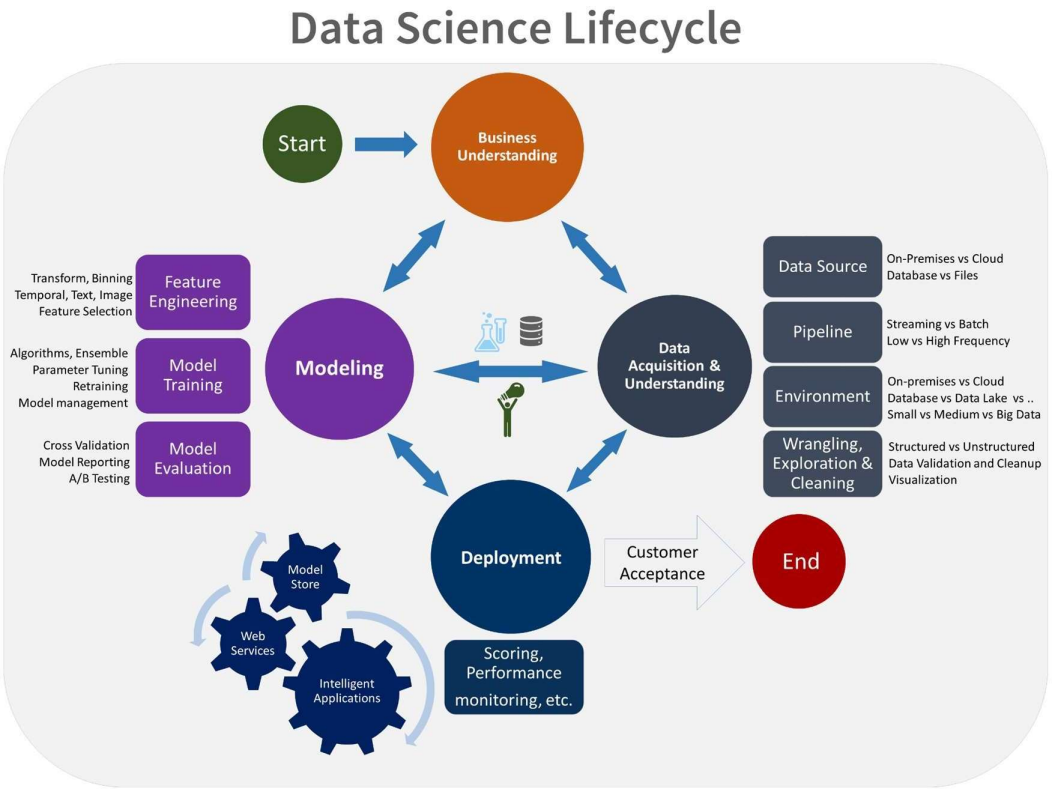
Secara umum, *machine learning* dapat dikategorikan menjadi tiga jenis diantaranya *supervised machine learning, unsupervised machine learning,* dan *reinforcement learning* (Khatri et al., 2020.). Kasus *fraud detection* di perusahaan telekomunikasi dapat dilakukan melalui *supervised machine learning*, dimana data historis dari keseluruhan pelanggan dan label *flagging* fraud pelanggan digunakan sebagai *data training* untuk membuat model yang kemudian akan diaplikasikan pada data baru (Kou et al., 2004).

2.3.2 Tahapan *Machine Learning*

Ada beberapa *framework* yang dapat digunakan dalam *machine learning*, salah satunya adalah kaidah dari *Team Data Science Process* (TDSP) Microsoft. Proces

ini merupakan *framework data science* yang paling lengkap karena merupakan simplifikasi dari 3 jenis *framework data science lainnya* seperti *Knowledge Data Discovery* (KDD), *Cross Industry Standard Process for Data Mining (CRISP-DM),* dan *Front Data Mining Process* (FDMS)(Martinez-Plumed et al., 2021). TDSP ini dilakukan dalam lima tahap seperti yang ditunjukkan pada Gambar 4 yaitu:

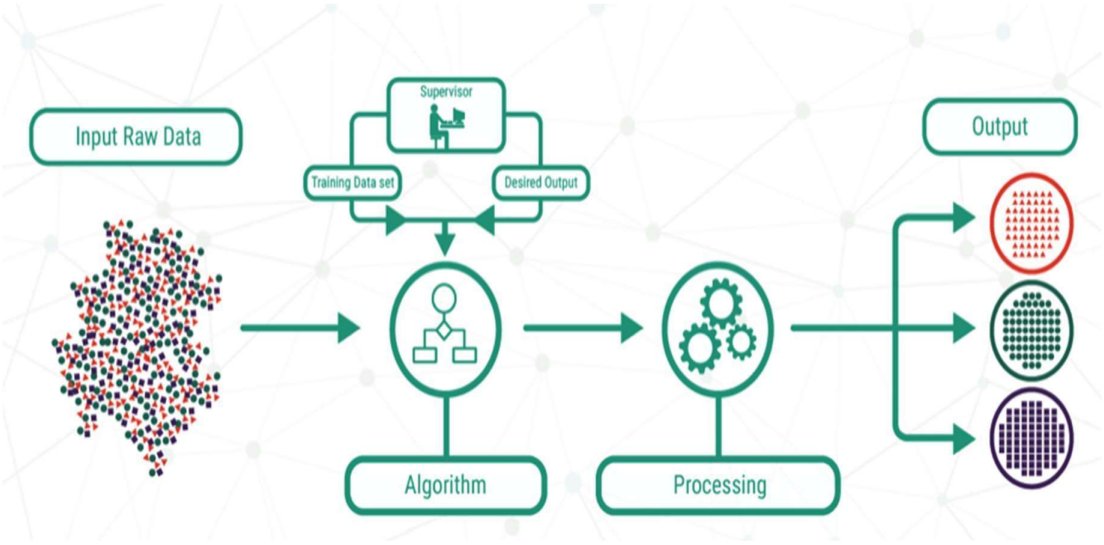
- *Business Understanding*
- *Data Ingest & Understanding*
- *Modelling*
- *Deployment*
- *Customer Acceptance*



Gambar 4 Tahapan framework team data science process

2.4 Supervised Machine Learning

Jenis *machine learning* ini juga biasa disebut sebagai pembelajaran prediktif karena memprediksi kelas objek yang tidak diketahui berdasarkan informasi mengenai kelas objek serupa yang sudah ada sebelumnya yang terangkum dalam *raw data*. Inspirasi utama di balik jenis *machine learning* ini adalah belajar dari informasi tentang tugas yang telah diberikan di masa lalu. Sebuah mesin membutuhkan data dasar tentang tugas yang akan ditugaskan padanya. Masukan atau pengalaman dasar ini diberikan dalam bentuk *data training*. Ini adalah informasi atau data masa lalu dari tugas tertentu. Metode *machine learning* ini sangat cocok, karena sering diaplikasikan pada kasus klasifikasi berbasis biner (0 atau 1). Dalam aplikasinya, ada beberapa basis algoritma yang dapat digunakan seperti yang berbasis regresi, *decision tree*, dan *neural network*. Cara kerja supervised dapat dilihat pada gambar 5.



Gambar 5 Cara kerja supervised machine learning

2.4.1 Algoritma *Logistic Regression*

Algoritma *logistic regression* adalah model analisis statistik yang menyatakan hubungan antara variabel respon dengan dua atau lebih kategori dan satu atau lebih variabel penjelas pada skala kategoris atau interval. Metode ini

menghasilkan probabilitas untuk setiap kategori respons yang bertindak sebagai panduan klasifikasi, menempatkan pengamatan dalam kategori respons tertentu berdasarkan nilai peluang terbesar yang dihasilkan. Regresi logistik telah diaplikasikan di beberapa bidang antara lain penelitian sosial, penelitian medis, prediksi kebangkrutan, segmentasi pasar, dan pola perilaku pelanggan (Noviandi, 2018). Berbeda dengan algoritma regresi linier yang pengaplikasiannya digunakan untuk memprediksi atau memperkirakan nilai, regresi logistik digunakan untuk tugas klasifikasi. Proses klasifikasi data ini dapat dilakukan dengan menghitung fitur/variabel pada data yang ada menjadi probabilitas yang divisualisasikan dalam ruang dua dimensi. Data tersebut kemudian dibagi menjadi dua kelompok berdasarkan garis sigmoid tersebut (Gunawan et al., 2020).

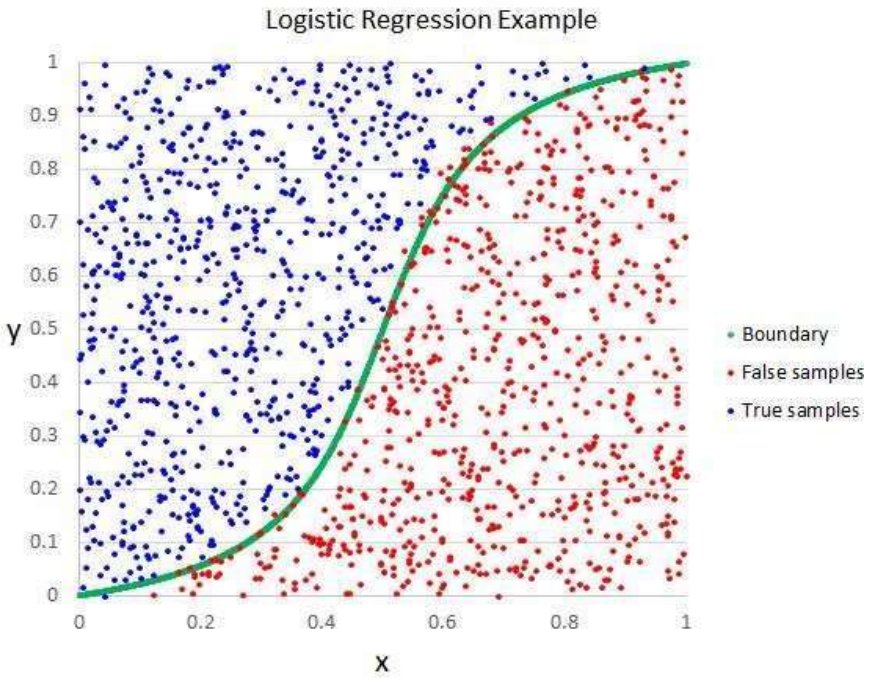
Seperti yang telah dijelaskan sebelumnya, *Logistic regression* adalah model statistik yang menggunakan fungsi logistik, atau fungsi logit, dalam matematika sebagai persamaan antara *x* dan *y*. Fungsi logit memetakan *y* sebagai fungsi sigmoid dari *x*. Sehingga apabila bisa disusun plot persamaan *logistic regression*:

$$f(x) = \frac{1}{1+e^{-x}} \quad (1)$$

Dimana: $f(x)$ = persamaan regression

e^{-x} = koefisien

Yang akan membentuk kurva persamaan seperti yang ditujukkan pada gambar 6 Seperti yang dapat dilihat, fungsi logit mengembalikan nilai antara 0 atau 1 untuk variabel dependen, terlepas dari nilai-nilai variabel independen. Ini adalah cara regresi logistik memperkirakan nilai variabel dependen. Metode regresi logistik juga memodelkan persamaan antara beberapa variabel independen dan satu variabel dependen.



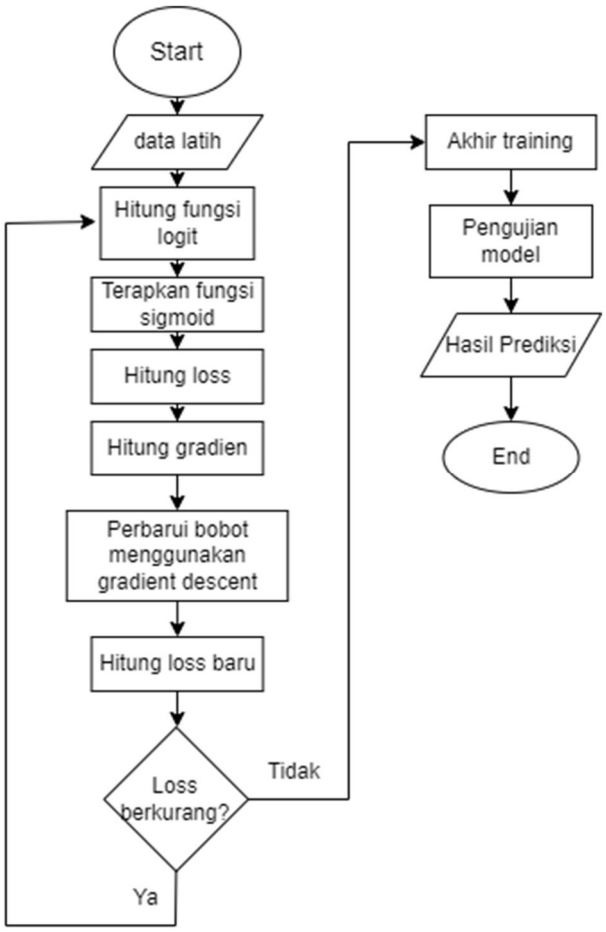
Gambar 6 Visualisasi logistic regression

Dalam kasus penelitian kali ini, beberapa variabel eksplanatori mempengaruhi nilai variabel dependen. Untuk memodelkan set data input tersebut, rumus regresi logistik mengasumsikan hubungan linier antara variabel independen yang berbeda. Jalan keluar dari permasalahan tersebut adalah memodifikasi fungsi sigmoid dan menghitung variabel *output* akhir sebagai

$$y = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_{\mathrm{n}} x_{\mathrm{n}} \qquad (2)$$

Simbol β mewakili koefisien regresi. Model logit dapat menghitung balik nilai koefisien ini ketika Anda memberikan set data eksperimental yang cukup besar dengan nilai yang diketahui dari variabel dependen dan independen.

Berikut adalah *flowchart* algoritma *logistic regression* pengaplikasiannya dalam *fraud detection*:



Gambar 7 Diagram proses *logistic regression*

Pada gambar 7 dijelaskan diagram proses logistic regression, pertama mengumpulkan dan mempersiapkan data latih yang berisi contoh-contoh yang memiliki fitur-fitur dan label kelas. Kemudian, menghitung fungsi logit yang merangkum hubungan antara fitur-fitur dan peluang kelas. Langkah berikutnya adalah menerapkan fungsi sigmoid, yang membawa hasil logit ke dalam skala probabilitas yang lebih intuitif.

Langkah selanjutnya, mengukur seberapa baik model apabila dibandingkan dengan data menggunakan fungsi kerugian (dalam penelitian ini menggunakan Cross-Entropy Loss). Perhitungan gradient dilakukan untuk mengetahui arah dan

seberapa jauh harus dilakukan pembaruan bobot model agar fungsi kerugian berkurang dengan menggunakan metode optimasi seperti Gradient Descent.

Setelah pembaruan bobot, menghitung ulang fungsi kerugian untuk melihat efek dari pembaruan tersebut. Jika loss berkurang, Anda kembali ke tahap perhitungan fungsi logit dan iterasi berlanjut. Namun, jika loss tidak berkurang cukup signifikan, diarahkan ke tahap pengujian model. Di sini model yang telah dilatih akan diuji pada data yang belum pernah dilihat selama pelatihan untuk menilai sejauh mana performanya.

Proses ini berulang-ulang hingga model mencapai kinerja yang memadai atau hingga kriteria penghentian tercapai. Selama proses ini, model logistic regression secara bertahap mempelajari pola dalam data latih dan mengoptimalkan bobotnya untuk meminimalkan loss, dengan tujuan akhir menghasilkan prediksi yang lebih baik pada data baru. Pada gambar 7 dijelaskan diagram proses logistic regression, pertama mengumpulkan dan mempersiapkan data latih yang berisi contoh-contoh yang memiliki fitur-fitur dan label kelas. Kemudian, menghitung fungsi logit yang merangkum hubungan antara fitur-fitur dan peluang kelas. Langkah berikutnya adalah menerapkan fungsi sigmoid, yang membawa hasil logit ke dalam skala probabilitas yang lebih intuitif.

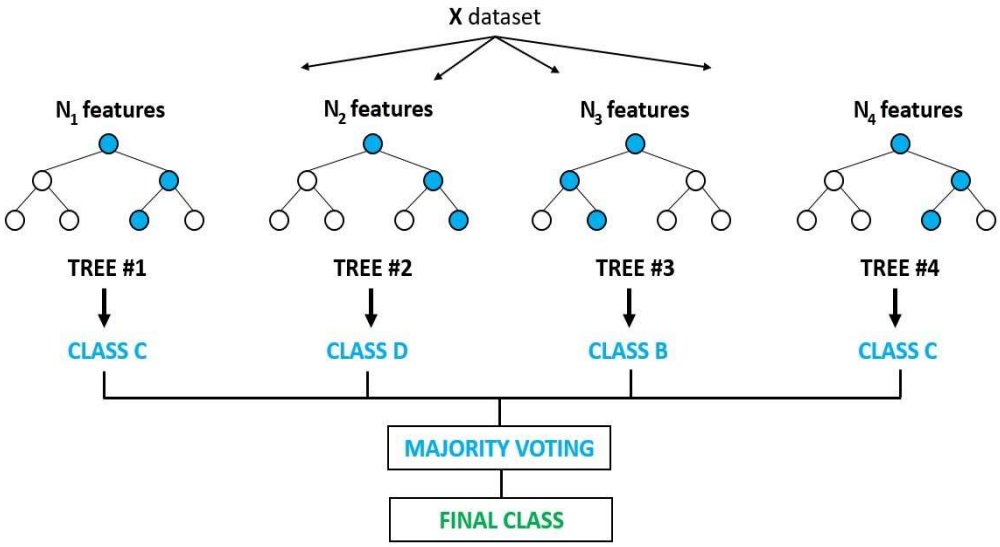
Langkah selanjutnya, mengukur seberapa baik model apabila dibandingkan dengan data menggunakan fungsi kerugian (dalam penelitian ini menggunakan Cross-Entropy Loss). Perhitungan gradient dilakukan untuk mengetahui arah dan seberapa jauh harus dilakukan pembaruan bobot model agar fungsi kerugian berkurang dengan menggunakan metode optimasi seperti Gradient Descent.
Setelah pembaruan bobot, menghitung ulang fungsi kerugian untuk melihat efek dari pembaruan tersebut. Jika loss berkurang, Anda kembali ke tahap perhitungan fungsi logit dan iterasi berlanjut. Namun, jika loss tidak berkurang cukup signifikan, diarahkan ke tahap pengujian model. Di sini model yang telah dilatih akan diuji pada data yang belum pernah dilihat selama pelatihan untuk menilai sejauh mana performanya.

Proses ini berulang-ulang hingga model mencapai kinerja yang memadai atau hingga kriteria penghentian tercapai. Selama proses ini, model logistic

regression secara bertahap mempelajari pola dalam data latih dan mengoptimalkan bobotnya untuk meminimalkan loss, dengan tujuan akhir menghasilkan prediksi yang lebih baik pada data baru.

2.4.2 Algoritma *Random Forest*

Algoritma *Random Forest* merupakan pengembangan dari algoritma *decision tree* sehingga dapat diklasifikasikan sebagai algoritma *ensemble learning* yang dianggap sebagai solusi elegan untuk banyak permasalahan *machine learning* terutama klasifikasi. Sebuah teknik yang meningkatkan kinerja prediktif dari model tunggal dengan melatih beberapa model dan menggabungkan prediksi yang dihasilkan, algoritma *Random Forest* menggunakan konsep yang selanjutnya disebut sebagai *bootstrap aggregating (bagging)* untuk melatih beberapa keputusan. Kemudian pohon dapatkan hasil dengan rata-rata atau voting oleh *single learner* (Breiman & Cutler, 2005). Visualisasi algoritma ini dapat dilihat melalui gambar 8.



Gambar 8 Visualisasi algoritma random forest

Cara kerja algoritma *Random Forest* ini adalah dengan melakukan *bootstrap aggregating (bagging)* dimana beberapa *decision tree* akan dibangun secara acak. Hasil akhir kemudian diputuskan dengan majority voting yang dihitung dari beberapa *decision tree* yang dibangun. Oleh karena itu, *Random Forest* diharapkan dapat membuat lebih sedikit kesalahan prediksi dari *decision tree* dan mengurangi variasi model karena konsep rata-rata yang diterapkan (Hastie et al., 2009).

Langkah-langkah dalam penerapan metode *Random Forest* antara lain:

1. Membuat data sampel dengan cara pengambilan acak dengan pengembalian dari dataset.
2. Gunakan sampel data untuk membangun pohon ke i (i=1, 2, 3, ..., k)
3. Ulangi langkah 1 dan 2 sebanyak k kali.

Perhitungan yang digunakan ketika membangun pohon keputusan dengan metode CART adalah informasi gain yang menggambarkan ukuran dalam pemilihan atribut yang digunakan setiap node sebuah pohon untuk klasifikasi. Misalkan N adalah node untuk memisahkan setiap kelas berdasarkan atribut dari suatu data yang dilambangkan D. Pemisahan (split) node dilakukan berdasarkan atribut yang memiliki informasi gain tertinggi. Rumus untuk mendapatkan informasi gain sebagai berikut:

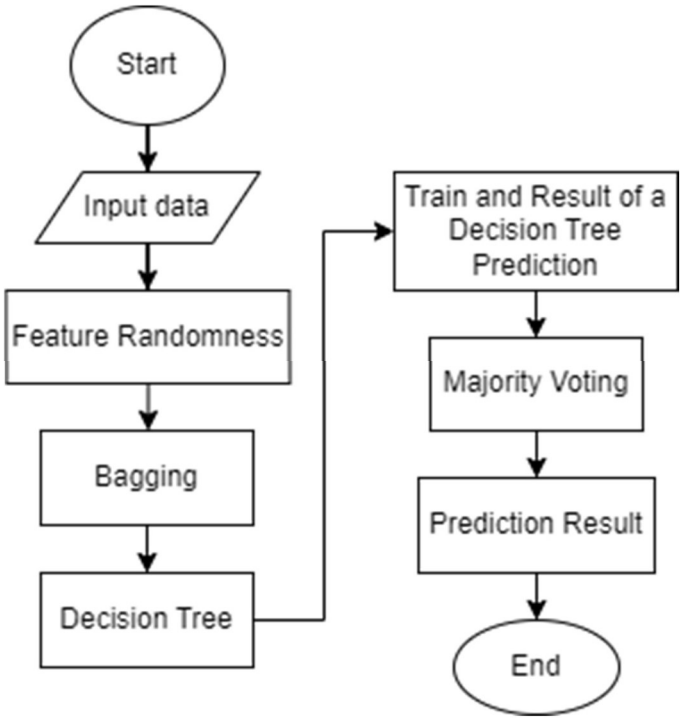
$$Gain\ (A) = Info\ (D) - InfoA(D) \tag{3}$$

Dengan nilai info (D) dapat diperoleh dengan menggunakan rumus 2.4 dan 2.5 untuk mendapat nilai $Info_A(D)$ di bawah ini:

$$Info(D) = -\sum_{i=1}^{n} p_i \, log2(p_i) \quad (4)$$

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \quad (5)$$

Nilai information gain pada atribut dengan nilai kontinu atau numerik harus menentukan nilai pembelah (split point) terbaik untuk pengelompokkan nilai. Split point terbaik diperoleh dari dengan cara mengurutkan data terlebih dahulu. Kemudian median atau nilai tengah setiap pasangan nilai yang saling berdekatan dianggap sebagai kemungkinan split point yang dapat digunakan. Apabila atribut A merupakan atribut dengan nilai kontinu maka seluruh nilai A diurutkan, kemudian menentukan nilai tengahnya sehingga kemungkinan jumlah partisi pada persamaan **2.5** adalah dua atau v = 2 (j=1 dan 2) (Suliztia, M, 2009).

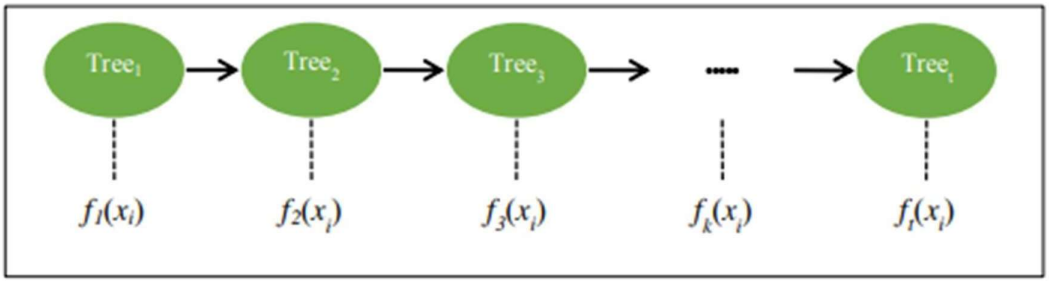


Gambar 9 Diagram Proses Random Forest

Pada gambar 9 dijelaskan teknik bagging yang melibatkan pemilihan atribut secara acak dijelaskan sebagai metode untuk mengembangkan model *Random Forest*. Dalam pendekatan ini, dataset dibagi menjadi beberapa bag atau subset, dan setiap bag dilatih dengan model *Decision tree* yang identik. Hal ini bertujuan untuk mencegah overfitting dalam model dan menciptakan beberapa model terlatih yang beragam, meskipun menggunakan jenis model yang sama. Setelah semua *Decision tree* terlatih pada masing-masing bag, teknik majority voting digunakan untuk menggabungkan hasil prediksi dari semua model. Ini memungkinkan pemilihan prediksi yang paling sering muncul dari semua model, menghasilkan nilai prediksi akhir (y_pred) yang lebih tepat. Dalam *Random Forest*, proses ini diaplikasikan pada setiap *Decision tree* dalam ensemble, menghasilkan model yang lebih stabil dan tangguh dalam membuat prediksi pada dataset baru. Melalui pendekatan ini, model *Random Forest* dapat mengatasi kekurangan model *decision tree* dan memberikan prediksi yang lebih tepat dan dapat diandalkan.

2.4.3 Algoritma eXtreme Gradient Boosting (XG Boost)

Extreme Gradient Boosting atau lebih dikenal sebagai XGBoost adalah algoritma yang berdasar pada *gradient boosting tree*, yang dapat memainkan peran penting dalam peningkatan gradien. XGBoost berlandaskan pada teori klasifikasi dan *regression tree* dan dapat menjadi solusi untuk masalah klasifikasi yang sangat efektif. Selain dengan tujuan optimasi, tujuan lain dari *XGBoost* dibagi dua bagian yang berbeda, yang mewakili penyimpangan model dan istilah reguler untuk mencegah overfitting(Qiu et al., 2021).



Gambar 10 Visualisasi algoritma XG Boost

Nilai prediksi pada langkah t diumpamakan dengan:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \quad (6)$$

$f_k(x_i)$ menggambarkan model pohon. Untuk diperoleh dari perhitungan berikut:

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_1) = \hat{y}_i^{(0)} + f_1(x_1)$$

$$\hat{y}_i^{(2)} = f_1(x_1) + f_2(x_2) = \hat{y}_i^{(1)} + f_2(x_2)$$

....

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (7)$$

Dimana:

$\hat{y}_i^{(t)}$ = F*inal tree model*

$\hat{y}_i^{(t-1)}$ = Model pohon yang dihasilkan sebelumnya

$f_t(xi)$ = Model baru yang dibangun

t = Jumlah total model dari base tree model

Untuk algoritma XGBoost, penentuan jumlah pohon dan depth merupakan hal penting. Permasalahan dalam menemukan algoritma yang optimum dapat diubah dengan pencarian klasifikasi baru yang dapat mengurangi loss function, dengan target fungsi kerugian ditunjukkan pada persamaan berikut:

$$Obj^{(t)} = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=i}^{t} \Omega(f_i) \quad (8)$$

Dimana:

$\hat{y}_i^{(t)}$ = Nilai prediksi

$\hat{y}_i$ = Nilai aktual

$l(\hat{y}_i^{(t)}, \hat{y}_i$ = *Lost function*

$\Omega(f_i)$ = istilah regularisasi

Karena model ensemble tree pada persamaan 8 merupakan fungsi sebagai parameter dan tidak dapat dioptimalkan menggunakan metode pengoptimalan tradisional pada ruang Euclidean. Sehingga digantikan dengan model dilatih dengan cara aditif dengan menggunakan $\hat{y}_i^{(t)}$ pada prediksi ke-i dan iterasi ke-t (Chen & Guestrin, 2016). Dalam meminimalkan loss function maka ditambahkan f_t sehingga didapatkan persamaan sebagai berikut:

$$Obj^{(t)} = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t-1)} + (f_t(x_i)) + \Omega(f_t) + constant \quad (9)$$

Selanjutnya target akhir dari loss function diubah menjadi persamaan 9 kemudian dilatih sesuai dengan target loss function berikut:

$$Obj^{(t)} = \sum_{i=1}^{t} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^{\,2}(x_i) \right] + \Omega(f_t) \quad (10)$$

Dimana g_i dan h_i merupakan urutan pertama dan kedua statistik gradient pada loss function.

Istilah regularisasi $\Omega(f_i)$ dapat dihitung menggunakan persamaan 10 yang digunakan untuk mengatasi kompleksitas model dan dapat meningkatkan kegunaan pada dataset yang lainnya.

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \quad (11)$$

Dimana:

T = Jumlah *leaf*

ω = Bobot *leaf*

λ dan γ = Koefisien, dengan nilai default ditetapkan untuk λ=1 dan γ=0

Deret Taylor second order dapat digunakan untuk pendekatan fungsi tujuan. Bobot optimal leaf j dan nilai optimal yang sesuai dapat diperoleh dengan persamaan berikut.

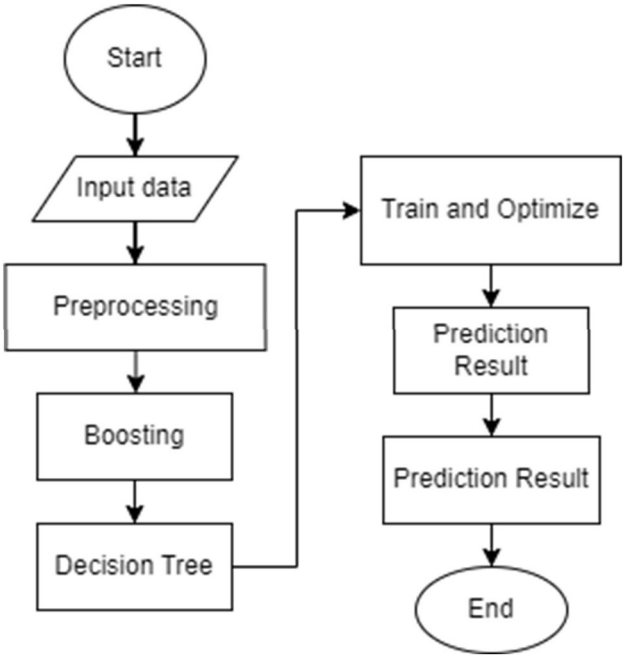
$$w_j^* = -\frac{g_j}{h_j+\lambda} \quad (12)$$

dan

$$-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda} + \lambda T \quad (13)$$

Dimana g_i dan h_i merupakan gradient orde pertama dan kedua dari loss function Loss function yang dapat digunakan sebagai skor kualitas struktur pohon q(Swetha & Dayananda, 2020).

Berikut adalah diagram proses algoritma *xgboost* pengaplikasiannya dalam *fraud detection*:



Gambar 11 Diagram proses xgboost

Berdasarkan gambar 11, dijelaskan proses dimulai dengan mengambil data set sebagai input. Pada tahap awal, data mungkin perlu diproses lebih lanjut melalui pembersihan dan transformasi untuk mempersiapkan proses boosting.

Boosting adalah teknik yang menggabungkan prediksi dari banyak model sederhana untuk membuat model yang lebih kuat. Dalam konteks XGBoost, model sederhana ini biasanya adalah pohon keputusan. Pohon-pohon ini dibuat dan ditambahkan ke model secara berurutan, dengan setiap pohon mencoba mengoreksi kesalahan yang dibuat oleh pohon sebelumnya. Ini berarti bahwa setiap pohon baru menambahkan informasi ke model, membantu mengidentifikasi pola yang mungkin terlewatkan oleh pohon sebelumnya.

Setelah pohon-pohon ini dibuat, model dilatih menggunakan algoritma optimasi seperti penurunan gradien. Tujuannya adalah untuk menemukan

parameter terbaik yang akan mengurangi kesalahan prediksi sebanyak mungkin. Proses ini bisa sangat kompleks dan memerlukan penyesuaian parameter yang cermat untuk mendapatkan hasil terbaik.

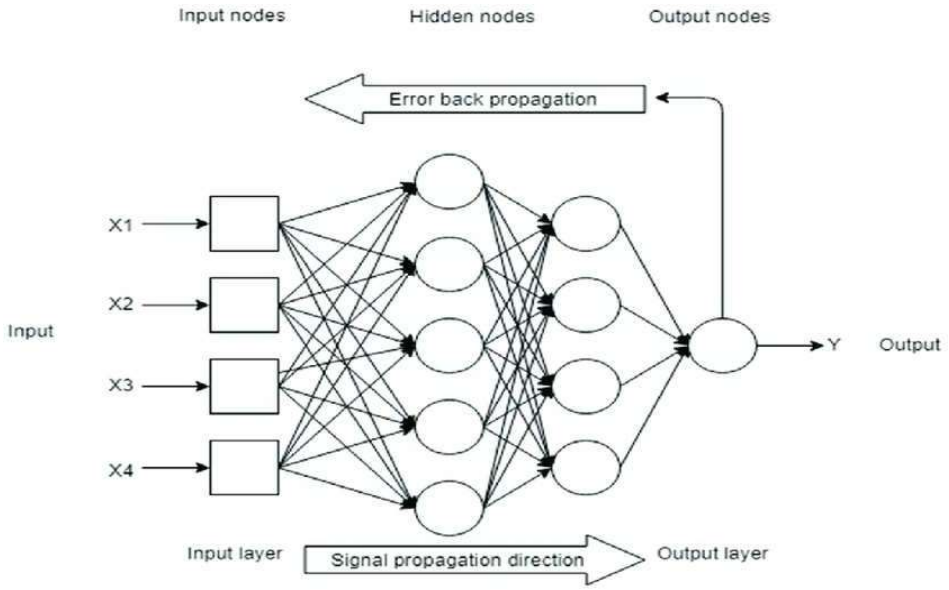
Akhirnya, setelah semua pohon keputusan dikombinasikan dan dioptimalkan, hasil prediksi akhir diperoleh. Proses ini selesai, dan hasil prediksi siap untuk digunakan dalam aplikasi yang diperlukan, seperti klasifikasi atau regresi.

XGBoost adalah algoritma yang sangat kuat dan fleksibel, dengan banyak pengaturan yang dapat disesuaikan untuk memenuhi kebutuhan spesifik dari setiap tugas pembelajaran mesin. Ini telah menjadi salah satu algoritma yang paling populer dan efektif dalam kompetisi pembelajaran mesin dan digunakan secara luas dalam industri.

2.4.4 Algoritma *Backpropagation Neural Network (BPNN)*

Arsitektur *BackPropagation* memiliki sejarah yang panjang, pertama kali diperkenalkan oleh Werbos pada tahun 1974. Model arsitekturnya sendiri terdiri dari baris unit pemrosesan yang saling terhubung yang disebut dengan *node*, pada struktur arsitektur neural network, node saling terhubung membentuk grup yang disebut dengan *layers* (Shinde & Shah, 2018).

Algoritma Backpropagation Neural Network berbasis *Multilayer perceptron* yang apabila dilihat berdasarkan data yang ada memiliki akurasi yang cukup baik untuk pengaplikasiannya di bidang klasifikasi, yaitu menghitung seberapa kecil perubahan yang terjadi yang bisa menimbulkan efek kepada jaringan dengan menggunakan *chain rule* kalkulus. Jaringan syaraf tiruan yang merupakan konsep dasar dari *backpropagation neural network* menggunakan metode komputasi yang meniru jaringan syaraf biologis. Dengan menggunakan metode komputasi yang meniru jaringan saraf biologis, Metode ini menggunakan perhitungan non linier dasar yang disebut neuron, yang saling terhubung dengan cara yang mirip dengan jaringan saraf manusia. Jaringan syaraf tiruan dibuat untuk memecahkan masalah pengenalan pola atau klasifikasi (Akhmad Hizham et al., 2018).



Gambar 12 Visualisasi backpropagation neural network

Backpropagation Neural Network(BPNN) menggunakan nilai output error untuk mengubah nilai bobotnya dalam arah mundur *(backward)*. Dimana untuk mendapatkan error ini harus melalui tahap perambatan maju *(forward propagation)* harus dikerjakan terlebih dahulu (Andrijasa, 2010).

Seperti disebutkan sebelumnya, proses *chain rule* yang terlibat adalah *forward propagation* dan *backward*. Feedforward biasanya dilakukan dengan memprediksi data dengan memasukkan variabel yang ada ke dalam output kelas untuk memprediksi, menjaga kinerja dan kesalahan model yang dibuat. Setelah itu parameter model disesuaikan dan dilakukan backpropagation. Proses ini berjalan terus menerus untuk mencapai performa model terbaik.

Berdasarkan proses kerja feed forward dan backpropagation yang dijelaskan sebelumnya, dapat disimpulkan bahwa algoritma *backpropagation neural network* memiliki waktu running paling lama dibandingkan dengan algoritma lain dalam penelitian kali ini. Namun, ketika mengevaluasi data dengan jaringan saraf dan parameter yang optimal, hasil yang diperoleh diharapkan dapat menjadi yang paling optimal dibandingkan dengan algoritma lain karena menggunakan metode trial and error.

Adapun langkah langkah untuk penerapan Backpropagation neural network (BPNN) adalah sebagai berikut:

1. Inisialisasi semua bobot dengan bilangan acak kecil.
2. Jika kondisi penghentian belum terpenuhi, lakukan langkah selanjutnya.

Fase 1 : Prograsi Maju

3. Tiap unit masukan menerima sinyal dan meneruskan ke unit tersembunyi.
4. Hitung semua keluaran di unit tersembunyi (Z_j):

$$zv_{0j} + \sum_{i=1}^{n} x_i v_{ij} \quad (14)$$

$$z_j = f \quad (15)$$

Dimana :

j = nilai unit tersembunyi

z_{0j} = Bobot layer input bias ke unit tersembunyi ke-j

x_i = unit input ke-i

v_{ij} = bobot unit input ke-i ke layer tersembunyi ke-j

z_j = nilai unit tersembunyi ke-j menggunakan fungsi aktivasi sigmoid

5. Hitung semua jaringan di unit keluaran (y_k):

$$y_n et_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk} \quad (16)$$

$$y_k = f(y_n et_k) = \frac{1}{1+e^{-y_n etk}} \quad (17)$$

Dimana:

$y_n e\ t_k$ = nilai unit *output* ke-k

w_{0k} = bobot unit tersembunyi bias ke unit *output* ke-k

w_{jk} = bobot unit tersembunyi ke-j unit *output* ke-k

y_k = nilai unit *output* ke-k menggunakan fungsi aktivasi sigmoid.

Fase 2: Propagasi mundur

6. Hitung faktor δ unit keluaran berdasarkan kesalahan di setiap unit keluaran:

$$\delta_k = (t_k - y_k) f'(y_n e\ t_k) = (t_k - y_k) y_k \qquad (18)$$

δk merupakan kesalahan yang akan dipakai dalam perubahan bobot layer dibawahnya (langkah 7), kemudian hitung suku perubahan bobot w_{jk} (yang akan dipakai nanti untuk merubah bobot w_{jk}) dengan learning rate α. Learning rate merupakan salah satu parameter yang digunakan untuk menghitung nilai perubahan bobot dengan range antara 0 sampai dengan 1.

$$\Delta wj\ k = \alpha\ \delta_k z_j \qquad (19)$$

Dimana :

δ_k = nilai *error* unit *output*

z_j = nilai target *output*

α = *learning rate*

$\Delta wj\ k$ = perubahan bobot unit tersembunyi ke-j ke unit output ke-k

7. Hitung faktor unit tersembunyi berdasarkan kesalahan di setiap unit tersembunyi z_j

$$net_j = \sum_{k=1}^{m} k w_{jk} \qquad (20)$$

Faktor kesalahan unit tersembunyi

$$\delta_j = \delta_n et_j f'(z_n et_j) = \delta_n et_j \; z_j \; (1 - z_j) \qquad (21)$$

Hitung suku perubahan bobot v_{ij} (yang akan digunakan untuk merubah v_{ij})

$$V_{ij} = \alpha \, \delta_j \, x_i \qquad (22)$$

Dimana:

δ_j = nilai *error* unit tersembunyi

Δv_{ij} = perubahan bobot unit input ke-i ke unit tersembunyi ke-j

Fase 3 Modifikasi bobot

8. Hitung semua perubahan bobot

   Perubahan bobot garis yang menuju ke unit keluaran :

$$W_{jk}\,(baru) = w_{jk}\,(lama) + \Delta w_{jk} \qquad (23)$$

   Perubahan bobot garis yang menuju ke unit tersembunyi :

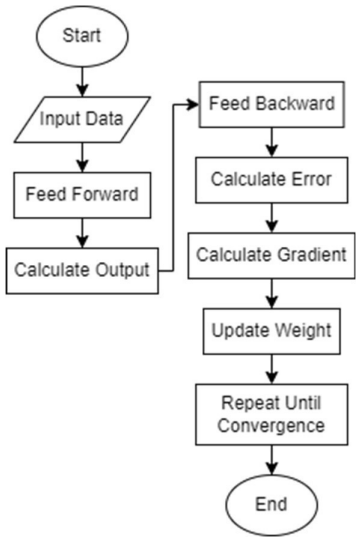
$$V_{ij}\,(baru) = V_{ij}\,(lama) + \Delta V_{ij} \qquad (24)$$

9. Selesai

Setelah pelatihan selesai dilakukan maka dapat digunakan untuk pengenalan pola. Dalam hal ini, hanya Propagasi Maju (langah 4 dan 5) saja yang dipakai untuk menentukan keluaran. Apabila fungsi aktivasi yang dipakai bukan Sigmoid biner, maka langkah 4 dan 5 harus disesuaikan. Demikian juga turunannya pada langkah 6 dan 7. Kondisi penghentian terpenuhi jika besarnya iterasi lebih besar dari besarnya iterasi maksimum yang telah ditetapkan. Iterasi merupakan rangkaian langkah dalam pembelajaran jaringan syaraf tiruan. Satu

iterasi diartikan sebagai satu kali pembelajaran yang dilakukan pada langkah 2 sampai 8 (Akhmad Hizham et al., 2018).

Berikut adalah diagram proses algoritma *backpropagation neural network* pengaplikasiannya dalam *fraud detection*:



Gambar 13 Diagram proses *Backpropagation Neural Network*

Proses backpropagation neural network dalam sistem deteksi fraud dimulai dengan memasukkan data dan dilanjutkan pada proses feed forward melalui jaringan saraf. Setiap lapisan jaringan menerapkan bobot dan fungsi aktivasi untuk menghitung output, yang kemudian dibandingkan dengan nilai target untuk menghitung kesalahan. Gradien dihitung untuk setiap bobot, yang menunjukkan seberapa besar perubahan bobot akan mempengaruhi kesalahan, dan bobot diperbarui sesuai.

Langkah-langkah ini diulangi berulang kali dalam proses iteratif sampai model konvergen, yaitu sampai perubahan dalam bobot menjadi sangat kecil atau kesalahan mencapai tingkat yang dapat diterima. Dalam konteks deteksi penipuan, proses ini memungkinkan model untuk 'belajar' dari data, menyesuaikan bobot internalnya untuk menjadi lebih efisien dalam mengidentifikasi transaksi yang mencurigakan, sehingga meningkatkan keakuratan dalam mendeteksi penipuan.

2.5 Evaluasi Performa *Supervised Machine Learning*

2.5.1 Confusion Matrix

Confusion Matrix merupakan konsep dari *machine learning* yang berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh sistem klasifikasi (Deng et al., 2016). *Confusion matrix* memiliki dua dimensi, satu dimensi berdasarkan data yang diambil dari *raw data* dan dimensi lainnya yaitu hasil klasifikasi setelah proses *data testing* kemudian dibagi berdasarkan empat jenis karakteristik diantaranya *True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN)*. Pada penelitian ini karakteristik dikategorikan *fraud* dikategorikan sebagai positif dan *non-fraud* dikategorikan sebagai negatif . *True Positive* (TP) yang berarti data pelanggan terdeteksi *fraud* dan dideteksi oleh model sebagai *fraud*, disisi lain *False Positive* (FP) saat data menunjukkan *non-fraud* namun hasil klasifikasi menunjukkan hasil positif atau *fraud*. Sedangkan *False Negative* (FN) saat data menunjukkan *fraud* namun tidak terdeteksi oleh model algoritma yang dibangun dan diklasifikasikan menjadi pelanggan yang tidak melakukan *fraud*, sedangkan *True Negative* (TN) menunjukkan data yang tidak melakukan *fraud* dan model algoritma mendeteksinya sebagai pelanggan yang tidak melakukan *fraud*. Dalam penelitian ini, *False Positive* (FP) dan *False Negative (FN)* merupakan jenis error yang ingin dihindari.

Tabel 1 *Confusion matrix*

	Data Prediksi	
Data Asli		
	Fraud	Bukan *Fraud*
Fraud	*True Positive* (TP)	*False Negative* (FN)
Bukan *Fraud*	*False Positive* (FP)	*True Negative* (TN)

Dengan adanya *confusion matrix*, dapat dihitung indikator performa klasifikasi yang mencerminkan kualitas dari model klasifikasi yang telah dibuat, indikator-indikator yang paling umum digunakan adalah *precision (TP / (TP + FP)), sensitivity(TP / (TP + FN)), specificity (TN / (TN + FP)), dan accuracy ((TP + TN) / (TP + TN + FP + FN))* (Ruuska, 2018). Dimana dalam penelitian ini output yang diinginkan yaitu model dengan performa klasifikasi dalam mendeteksi *fraud* terbaik.

2.5.2 Performa Klasifikasi

Setelah mendapatkan *confusion matrix,* dilakukan pengukuran performa untuk masing-masing model dengan tujuan mendapatkan perbandingan hasil dari model algoritma yang digunakan dengan prioritas mendapatkan nilai akurasi dan *recall* namun nilai *recall* masih paling menjadi tujuan utama dikarenakan nilai *recall* merupakan nilai yang menunjukkan seberapa banyak model sistem dapat mendeteksi *fraud* dibandingkan dengan jumlah total *fraud* yang ada.

- Akurasi

  Adalah pengukuran performa klasifikasi yang menunjukkan performa model dalam mengklasifikasi data yang benar.

$$\text{Akurasi} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{F}\ )} \tag{25}$$

- *Recall*

  Atau biasa disebut sensitivitas merupakan performa yang menunjukkan kemampuan model algoritma dalam mendeteksi positif yang menjadiprioritas.

$$Recall = \frac{(\text{TP})}{(\text{TP}+\text{FN})} \tag{26}$$

Dalam konteks *fraud detection,* nilai *recall* sering dianggap lebih penting daripada nilai akurasi, dikarenakan *recall* mengukur seberapa baik model kita dalam mengidentifikasi semua kasus penipuan yang sebenarnya. Dengan kata lain, ini adalah rasio antara jumlah penipuan yang benar-benar dideteksi oleh model terhadap jumlah total penipuan yang sebenarnya ada. Nilai recall yang tinggi berarti model tersebut lebih akurat dalam mendeteksi kasus penipuan. Sedangkan nilai

akurasi Mengukur seberapa sering model membuat prediksi yang benar, baik itu penipuan atau bukan penipuan. Namun, dalam konteks fraud detection, kasus penipuan biasanya jauh lebih jarang daripada kasus non-penipuan, sehingga model yang hanya memprediksi "bukan penipuan" untuk semua kasus bisa mendapatkan accuracy yang tinggi.

2.5.3 ROC Curve

Kurva the Receiver Operating Characteristic (ROC) dikembangkan oleh para insinyur selama perang dunia ke-2 untuk mendeteksi keberadaan musuh di medan perang (Gonçalves et al., 2014). Disiplin ilmu ini terus berkembang dan mulai merambat ke bidang lainnya dengan sangat cepat, dalam beberapa tahun saja bidang ilmu ini telah banyak diterapkan di berbagai bidang lainnya termasuk ilmu atmosfer, biosains, psikologi eksperimental, keuangan, geosains, dan sosiologi (Krzanowski & Hand, 2009).

Analisis ROC pun mulai dikembangkan dan banyak digunakan di *machine learning* dan *data mining* sebagai metode yang sederhana namun efektif digunakan untuk membandingkan akurasi variable (Streiner & CAIRNEY, 2013). ROC yang dimaksud dalam penelitian ini merupakan kurva yang membandingkan interaksi antara *True Positive Rate (True Positive/(True Positive+False Negative))* atau *sensitivity* dan *False Positive Rate(False Positive/(False Positive+True Negative))* atau *specificity.*

2.5.4 Execution Time

Mengetahui *execution time* dari sebuah *code* sangat membantu untuk proses pengoptimalan sebuah *code* (Stewart, 2006). Dalam penelitian kali ini execute time digunakan sebagai perbandingan seberapa efisien algoritma yang digunakan dalam mendeteksi *fraud.* Gambar 10 dan 11 menunjukkan *code* yang digunakan dalam algoritma untuk menghitung *executing*, dan gambar 12 menunjukkan *flowchart execute time.* Pengukuran menggunakan time.process_time yang mengimpor modul time dan kemudian memanggil fungsi tersebut. Ini memberikan waktu CPU yang

telah dihabiskan oleh proses sejak dimulai, yang lebih tepat untuk mengukur waktu pemrosesan karena tidak termasuk waktu saat sistem atau aplikasi sedang tidur.

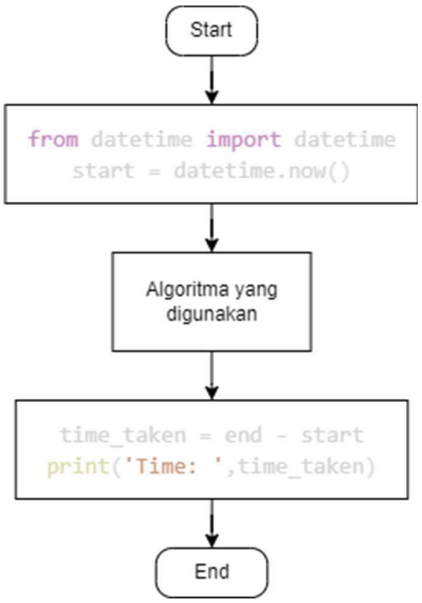
```
import time

start_time = time.process_time()
```

Gambar 14 Start menghitung execute time

```
end_time = time.process_time()
execution_time = end_time - start_time
print(f"Waktu eksekusi: {execution_time:.6f} detik")
```

Gambar 15 End menghitung execute time

Gambar 14 adalah code yang digunakan yang dituliskan tepat sebelum algoritma di eksekusi dan gambar 15 dituliskan tepat setelah algoritma selesai berjalan dengan tujuan menampilkan hasil berapa lama waktu eksekusi yang dibutuhkan suatu algoritma.



Gambar 16 Flowchart execution time