SKRIPSI

PREDIKSI GAGAL BAYAR CALON KREDITUR MENGGUNAKAN METODE DATA MINING

Disusun dan diajukan oleh:

ANDI SUNGKURUWIRA BATARA UNRU
D121171527

DEPARTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

LEMBAR PENGESAHAN SKRIPSI

PREDIKSI GAGAL BAYAR CALON KREDITUR MENGGUNAKAN METODE DATA MINING

Disusun dan diajukan oleh:

ANDI SUNGKURUWIRA BATARA UNRU
D121171527

Telah dipertahankan di depan Panitia Ujian yang dibentuk
dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika
Fakultas Teknik Universitas Hasanuddin pada Tanggal 31 Juli 2024
Dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing I	**Pembimbing II**
Elly Warni, S.T., M.T.	**Mukarramah Yusuf, B.Sc., M.Sc., Ph.D**
NIP. 198202162008122001	NIP. 198310082012122003

Ketua Departemen
Teknik Informatika,

Prof. Dr. Ir. Indrabayu, S.T., M.T. M.Bus.Sys., IPM, ASEAN.
NIP. 197507162002121004

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : **Andi Sungkuruwira Batara Unru**

NIM : **D121171527**

Program Studi : **Teknik Informatika**

Jenjang : **S1**

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Prediksi Gagal Bayar Calon Kreditur Menggunakan Metode Data Mining

Adalah karya tulisan sendiri dan bukan merupakan pengambil 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 diberikan 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, 31 Juli 2024

Menyatakan

Andi Sungkuruwira Batara Unru

ABSTRAK

ANDI SUNGKURUWIRA BATARA UNRU. *Prediksi Gagal Bayar Calon Kreditur Menggunakan Metode Data Mining* (Dibimbing oleh Elly Warni dan Mukarramah Yusuf).

Gagal bayar adalah ketidakmampuan debitur untuk memenuhi kewajiban pembayaran pinjaman tepat waktu. Dalam konteks kredit, gagal bayar dapat berdampak signifikan terhadap ekonomi baik pada skala makro maupun mikro. Pada tingkat makro, tingginya tingkat gagal bayar dapat mengganggu stabilitas keuangan nasional dan merusak kepercayaan investor. Pada tingkat mikro, gagal bayar dapat menyebabkan kerugian besar bagi lembaga keuangan dan penurunan kualitas hidup bagi individu yang mengalaminya. Dampak negatif ini menjadikan pencegahan gagal bayar sebagai prioritas dalam pengelolaan kredit.

Penelitian ini bertujuan mengembangkan sistem prediksi gagal bayar calon kreditur dan mengidentifikasi pola hubungan antar variabel yang mempengaruhi gagal bayar. Dengan sistem prediksi yang akurat, lembaga keuangan dapat mengurangi risiko gagal bayar serta meningkatkan efisiensi pemberian kredit. Penelitian ini juga berupaya memahami karakteristik calon kreditur yang berpotensi gagal bayar, sehingga dapat diambil tindakan pencegahan yang tepat.

Metode yang digunakan meliputi analisis asosiasi dengan algoritma *FP-Growth* menggunakan *minimum support* 0.3 dan *minimum confidence* 0.5, serta prediksi menggunakan algoritma *Extreme Gradient Boosting* (*XGBoost*) dengan *learning rate* 0.3. Analisis asosiasi bertujuan menemukan pola hubungan antar variabel, sedangkan metode prediksi mengklasifikasikan calon kreditur ke dalam kategori resiko gagal bayar.

Hasil penelitian menunjukkan bahwa calon kreditur dengan karakteristik berpendidikan kelas menengah, harga barang yang diajukan untuk pinjaman bernilai rendah, tidak memiliki mobil, dan sumber pendapatan dari bekerja memiliki potensi tinggi untuk gagal bayar. Sedangkan calon kreditur dengan karakteristik memiliki rumah atau apartemen, kepemilikan properti, berjenis kelamin perempuan, memiliki keluarga kecil, telah menikah, dan tidak memiliki anak memiliki potensi besar untuk melunasi kredit tanpa gagal bayar, sehingga aplikasi kredit mereka dapat diterima. Karakteristik lainnya seperti rendahnya anuitas, peminjaman tunai, dan jumlah kredit tinggi juga muncul, namun kurang signifikan dalam mempengaruhi pengambilan kebijakan kredit.

Kata Kunci: Gagal Bayar, *Credit Scoring*, Asosiasi, Prediksi, *FP-Growth*, *Machine Learning*

ABSTRACT

ANDI SUNGKURUWIRA BATARA UNRU. Predicting Loan Default Using Data Mining Methods (Supervised by Elly Warni and Mukarramah Yusuf).

Default is the inability of a debtor to fulfill their loan repayment obligations on time. In the context of credit, default can significantly impact both macro and microeconomic scales. At the macro level, high default rates can disrupt national financial stability and undermine investor confidence. At the micro level, default can cause substantial losses for financial institutions and diminish the quality of life for individuals experiencing it. These Negative impacts make default prevention a priority in credit management.

This study aims to develop a system for predicting loan default among prospective borrowers and to identify the relationships between variables that influence default. With an accurate prediction system, financial institutions can reduce the risk of default and improve the efficiency of credit provision. This study also seeks to understand the characteristics of potential defaulters, allowing for appropriate preventive measures to be taken.

The methods used include association analysis with the FP-Growth algorithm, employing a minimum support of 0.3 and a minimum confidence of 0.5, as well as prediction using Extreme Gradient Boosting (XGBoost) with 0.3 learning rate. The association analysis aims to discover Patterns between variables, while the prediction methods classify prospective borrowers into default risk categories.

The research findings indicate that loan applicants with characteristics such as middle-class education, low value of goods requested for the loan, not owning a car, and deriving income from employment have a high potential for default. On the other hand, applicants who own a house or apartment, possess property, are female, have a small family size, are married, and do not have children show a strong potential for successfully repaying loans without default, thereby increasing the likelihood of their loan applications being approved. Other characteristics such as low annuities, cash loans, and high loan amounts also emerge but are less significant in influencing credit policy decisions.

Keywords: Default, Credit Scoring, Association, Prediction, FP-Growth, Machine Learning

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
ξ	*Slack*
$h_t(x)$	*Weak / Basic Classifier*
α_t	*Learning Rate*
θ	*Parameter XGBoost*
y_i	Nilai Aktual
$\hat{y_i}$	Hasil Prediksi XGBoost
n	Jumlah Iterasi XGBoost
$H(x)$	*Final Classifier*
$y(x)$	Nilai Target
$F_0(x)$	*Model Baseline*
AIMP	*Artificial Intelligence and Multimedia Processing*
CSV	*Comma Separated Value*
DTC	*Decision Tree*
FP	*Frequent Pattern*
FP	*False Positive*
FN	*False Negative*
KDD	*Knowledge Discovery In Databases*
RAM	*Random Access Memory*
TID	*Transaction ID*
TP	*True Positive*
TN	*True Negative*
TPR	*True Positive Rates*
TNR	*True Negative Rates*
XGB	*Extreme Gradient Boosting*

DAFTAR LAMPIRAN

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa, yang telah memberikan rahmat dan karunia-Nya, sehingga dapat menyelesaikan tugas akhir dengan judul "Prediksi Gagal Bayar Calon Kreditur Menggunakan Metode *Data Mining*" sebagai salah satu syarat dalam menyelesaikan jenjang Strata-1 di Departemen Teknik Informatika, Fakultas Teknik Universitas Hasanuddin.

Penelitian ini bertujuan untuk mengembangkan model prediksi gagal bayar calon kreditur menggunakan metode *Data Mining* guna memberikan kontribusi signifikan terhadap stabilitas ekonomi makro dan mikro. Dalam konteks ini, penggunaan *machine learning* diharapkan dapat mengidentifikasi faktor-faktor kunci yang mempengaruhi gagal bayar dan membangun model prediktif yang efisien. Dengan demikian, penelitian ini diharapkan dapat mengurangi risiko gagal bayar, mendorong pertumbuhan ekonomi melalui peningkatan kepercayaan dalam sistem kredit, serta mendukung kesehatan finansial baik secara makro maupun mikro.

Penulis menyadari bahwa penyusunan dan penulisan tugas akhir ini tidak dapat terselesaikan dengan baik tanpa adanya bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih banyak kepada :

1. Keluarga penulis, H. Andi Fahry Makkasau Krg Unjung dan Andi Ramlah Pg Tayu yang telah menjadi sosok orangtua yang tak pernah lelah dalam mendidik, memberi semangat, dan mendoakan penulis. Istri tercinta penulis, Jumrah Nurfadila Krg Layu yang senantiasa memberikan kasih dan mendampingi menjadi *support system* terbaik yang penulis dapatkan. Juga untuk Tanteku Andi Mutmainnah Krg Puji, dan tak lupa pula untuk Adikku dan suami, Andi Pancaitana Bunga Walie Krg Sompa dan Andi Muh. Riang Rezky Krg Laja.
2. Ibu Elly Warni, S.T., M.T. selaku pembimbing I dan Ibu Mukarramah Yusuf, B.Sc., M.Sc., Ph.D. selaku pembimbing II, yang senantiasa menyediakan waktu, tenaga,

pikiran, dan perhatian yang luar biasa dalam mengarahkan penulis untuk menyelesaikan tugas akhir.

3. Bapak Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng, selaku Ketua Departemen Teknik Informatika Universitas Hasanuddin atas segala bimbingan dan dukungan selama masa perkuliahan.
4. Bapak Andi Ais Prayogi Alimuddin, S.T., M.Eng. selaku Pembimbing Akademik penulis, segenap dosen, dan staf Departemen Teknik Informatika Universitas Hasanuddin yang telah banyak memberikan ilmu dan pengalaman, serta bantuan kepada penulis selama menuntut ilmu di kampus tercinta ini.
5. Saudaraku *RECOGN17ER* atas dukungan, bantuan, semangat, dan solidaritasnya.
6. Saudara dalam baktiku di Dewan Kerja Daerah Gerakan Pramuka Sulawesi Selatan, Kakak Ketua Mursaha, Kakak Wakil Ketua Eky Sri Septiana, Virdhalya, Zulfahmi, Mantik, Yogie, Akhsan, Heni, Murod, Azhar, Shafira, Kak Eka, Riska, Nunu, dan Fitrah yang senantiasa menyemangati dan membersamai penulis.
7. Saudaraku AMPG Maros, Pak Ketua Andi Alif Maulana, Bu Bendum Pratiwi Putri Rosidin, Dzuljalalii Ikraam, Andi Aidil, Andi Sri Mulyani, Alvito yang senantiasa menyemangati dan membersamai penulis.
8. Pihak lainnya yang tanpa penulis sadari telah banyak memberikan doa, bantuan, dan semangat kepada penulis dalam menyelesaikan tugas akhir ini.

Penulis berharap semoga Tuhan Yang Maha Esa berkenan membalas segala kebaikan serta jasa dari semua pihak yang telah banyak membantu penulis dalam menyelesaikan tugas akhir ini. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna. Oleh karena itu penulis mengharapkan segala bentuk saran serta masukan yang membagun dari berbagai pihak. Semoga tugas akhir ini dapat memberi manfaat bagi para pembaca dan semua pihak. Aamiin.

Gowa, Juni 2024
Penulis,

Andi Sungkuruwira Batara Unru

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penilaian kredit atau *Credit Scoring* adalah suatu aspek integral dalam aktivitas lembaga keuangan modern. Dalam upaya mengamankan pinjaman dan mengurangi risiko gagal bayar, lembaga keuangan melakukan evaluasi mendalam terhadap calon kreditur (Xia et al, 2017). Proses penilaian ini melibatkan analisis kompleks atas berbagai variabel, termasuk karakteristik pribadi, informasi keuangan, dan perilaku calon peminjam. Model penilaian kredit berkembang melalui pendekatan statistik, matematika, dan bahkan *Machine learning* (Moradi et al, 2019), memungkinkan lembaga keuangan membuat keputusan yang lebih cerdas dan risiko yang lebih terkendali.

Gagal bayar oleh pelanggan memiliki dampak signifikan, tidak hanya bagi perusahaan pemberi pinjaman, tetapi juga bagi pelanggan dan bahkan ekonomi negara. Bagi perusahaan, risiko gagal bayar dapat mengakibatkan kerugian finansial yang serius. Selain kehilangan jumlah pokok pinjaman, perusahaan juga harus menanggung biaya pengumpulan hutang, biaya hukum, dan bahkan penurunan kepercayaan dari investor dan pemegang saham. Di sisi lain, bagi pelanggan, gagal bayar dapat merusak sejarah kredit mereka, membuatnya sulit mendapatkan pinjaman di masa mendatang, dan menghadapi tekanan finansial yang lebih besar. Selain itu, tingkat gagal bayar yang tinggi dalam populasi dapat menciptakan tekanan ekonomi makro. Ini dapat mengurangi kepercayaan investor, meningkatkan suku bunga secara keseluruhan, dan merusak stabilitas ekonomi, mengakibatkan konsekuensi negatif bagi pertumbuhan dan kesempatan kerja di negara tersebut (Giesecke et al, 2011). Oleh karena itu, pemahaman mendalam mengenai faktor-faktor yang mempengaruhi gagal bayar melalui penilaian kredit yang cermat sangat penting untuk mencegah risiko-risiko ini dan memastikan keberlanjutan ekonomi yang stabil (Khemais et al, 2016).

Data Mining, sebagai cabang penting dalam analisis data, membuka peluang besar untuk mendapatkan wawasan berharga dari kumpulan data yang besar dan

kompleks. Dalam konteks penilaian kredit, teknik *Data Mining* menjadi kunci dalam meningkatkan akurasi dan efisiensi model-model penilaian kredit (Maharjan, 2022). Dengan *Data Mining*, kita dapat melakukan pengklasifikasian, memprediksi, memperkirakan, dan mendapatkan informasi lain yang bermanfaat dari kumpulan data dalam jumlah yang besar (Mardi, 2017).

Salah satu lembaga keuangan non-bank yang berperan aktif dalam memberikan pinjaman kepada individu dengan sejarah kredit yang terbatas atau bahkan tidak ada, adalah *Home Credit*. Dengan jangkauannya yang mencakup 10 negara di Eropa, Asia, dan Amerika, *Home Credit* membanggakan basis pelanggan yang luas, mencapai lebih dari 29 juta orang. Dalam menjalankan misinya, *Home Credit* berkomitmen untuk memperluas inklusi keuangan dengan memberikan pinjaman yang mudah dimengerti, transparan, dan bertanggung jawab kepada pelanggan-pelanggannya.

Penelitian ini secara khusus bertujuan untuk menyelidiki data historis peminjaman calon kreditur *Home Credit* menggunakan teknik *Data Mining*. Dalam pendekatan studi kasus, penelitian ini akan menerapkan metode asosiasi dan prediksi pada data yang telah terkumpul. Tujuan utama adalah mengidentifikasi faktor-faktor kunci yang mempengaruhi kinerja kredit dan risiko gagal bayar peminjam. Melalui analisis mendalam ini, penelitian diharapkan dapat memberikan pemahaman yang lebih baik terkait praktek penilaian kredit, sekaligus memberikan kontribusi penting dalam pengembangan strategi risiko dan kebijakan kredit di lembaga keuangan, terutama dalam konteks *Data Mining*.

Dengan merangkum, penelitian ini tidak hanya mendalami proses penilaian kredit dan teknik *Data Mining* yang digunakan, tetapi juga membawa dampak yang potensial pada perumusan kebijakan kredit di *Home Credit*. Dengan memahami faktor-faktor yang secara signifikan memengaruhi kinerja kredit, lembaga keuangan dapat mengambil keputusan yang lebih cerdas dan berbasis data. Sebagai hasilnya, penelitian ini bukan hanya relevan bagi *Home Credit* tetapi juga memiliki implikasi yang lebih luas dalam konteks industri keuangan secara keseluruhan.

1.2 Rumusan Masalah

1. Bagaimana mengembangkan sistem prediksi calon kreditur gagal bayar dengan menggunakan algoritma *Data Mining* berbasis data historis?
2. Bagaimana unjuk kerja sistem prediksi calon kreditur gagal bayar dengan menggunakan algoritma *Data Mining* berbasis data historis?

1.3 Tujuan

Tujuan dari penelitian ini adalah:

1. Mengembangkan prediksi calon kreditur gagal bayar dengan algoritma *Data Mining* berbasis data historis.
2. Mengetahui prediksi calon kreditur gagal bayar dengan menggunakan algoritma *Data Mining* berbasis data historis.

1.4 Manfaat

Manfaat dari penelitian ini adalah:

1. Memberikan informasi terhadap bidang ilmu pengetahuan yang relevan sesuai dengan tema tugas akhir.
2. Mengidentifikasi pola dan hubungan antar variabel dalam data histori peminjaman.
3. Mengembangkan model prediktif yang dapat mengklasifikasikan calon kreditur ke dalam kelompok risiko yang berbeda.

1.5 Batasan Masalah

1. Data yang digunakan berupa historis peminjaman dari *Home Credit* dan keuangan sejenis yang telah dilaporkan ke Biro Kredit.
2. Penelitian ini akan menggunakan algoritma asosiasi yang dilanjutkan dengan algoritma prediksi.
3. Output akhir dari penelitian ini akan memprediksi setiap kostumer yang dikemudian hari akan berpotensi untuk gagal bayar.

BAB II
TINJAUAN PUSTAKA

2.1 *Credit Scoring*

Credit Scoring melibatkan penggunaan berbagai data dan teknik statistik untuk menilai risiko kredit seorang peminjam. Tujuannya adalah untuk memprediksi apakah seorang individu atau entitas cenderung gagal membayar pinjaman atau kewajiban keuangannya. Proses ini melibatkan analisis mendalam terhadap data historis seperti riwayat pembayaran pinjaman, jumlah hutang, riwayat kredit, dan faktor-faktor lain yang relevan. Dengan menggunakan model *Credit Scoring*, lembaga keuangan dapat membuat keputusan kredit yang lebih terinformasi dan mengelola risiko dengan lebih efektif (Lyn et al, 2010).

Metode *Credit Scoring* sangat penting dalam industri keuangan karena membantu dalam memprediksi perilaku pembayaran calon peminjam secara lebih akurat. Melalui teknik-teknik seperti pengklasifikasi data, analisis statistik, dan model prediktif seperti regresi logistik atau jaringan saraf tiruan, *Credit Scoring* memungkinkan penilaian risiko kredit yang lebih terinci dan objektif. Hasil dari proses ini memberikan informasi berharga bagi pemberi pinjaman untuk menentukan tingkat bunga, syarat pinjaman, atau bahkan keputusan untuk menolak aplikasi kredit (Lyn et al, 2010).

2.2 Gagal Bayar

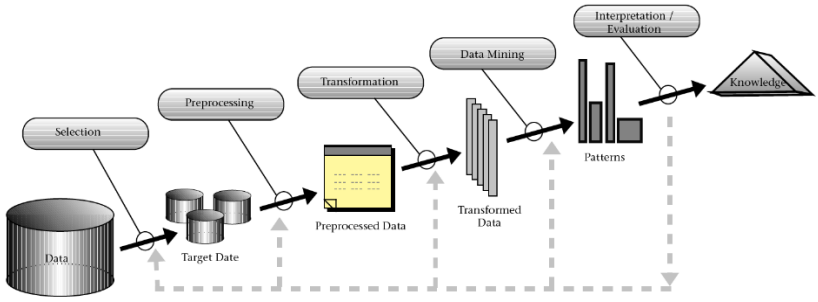
Gagal bayar terjadi ketika peminjam tidak mampu membayar kembali kewajiban kreditnya sesuai dengan persyaratan yang telah disepakati, yang biasanya mencakup pembayaran pokok dan bunga pinjaman (Altman et al, 1998).

Ketidakpastian kualitas dan asimetri informasi di pasar dapat menyebabkan risiko gagal bayar yang lebih tinggi karena pemberi pinjaman tidak memiliki informasi yang cukup untuk membedakan antara peminjam berkualitas tinggi dan rendah. Hal ini

menyebabkan seleksi yang merugikan, di mana peminjam yang berisiko tinggi lebih mungkin mendapatkan kredit daripada mereka yang berisiko rendah (Akerlof, 1970).

2.3 *Knowledge Discovery in Database* (*KDD*)

*Kno**wledge Discovery in Database* (*KDD*) adalah proses mengidentifikasi pola yang valid, baru, berpotensi, berguna, dan pada akhirnya dapat dipahami di dalam suatu data (Fayyad et al., 1996). *KDD* dalam prosesnya melingkupi antara lain *data cleaning*, *data integration*, *data selection*, *data transformation*, *Data Mining*, *Pattern evaluation*, *dan k*no*wledge presentation* (Han et al., 2012).



Gambar 2. 1 Proses *Knowledge Discovery in Database* (Fayyad et al, 1996)

Adapun penjelasan dari setiap tahapan proses dari *Kno**wledge Discovery in Database* (*KDD*) antara lain sebagai berikut:

2.3.1 *Data Cleaning*

Data Cleaning atau Pembersihan Data adalah langkah awal dalam proses *KDD* yang melibatkan penghapusan data yang berisik dan tidak konsisten dari kumpulan data. Langkah ini sangat penting karena kualitas data yang rendah dapat mengarah pada hasil analisis yang menyesatkan. Data yang berisi kesalahan, duplikasi, atau nilai yang hilang harus diidentifikasi dan diperbaiki atau dihapus untuk memastikan bahwa data yang akan digunakan dalam tahap berikutnya valid dan reliabel (Han et al, 2012).

2.3.2 *Data Integration*

Data Integration atau Integrasi Data adalah proses menggabungkan data dari berbagai sumber ke dalam satu penyimpanan data yang koheren. Ini melibatkan penyatuan data yang mungkin berasal dari basis data yang berbeda, sistem informasi, atau file yang terpisah. Tantangan utama dalam tahap ini adalah menangani perbedaan format, skema, dan struktur data dari berbagai sumber. Dengan mengintegrasikan data secara efektif, analisis selanjutnya dapat dilakukan dengan lebih efisien dan akurat (Han et al, 2012).

2.3.3 *Data Selection*

Data Selection atau Seleksi Data adalah sebuah proses memilih subset data yang relevan untuk analisis dari basis data yang lebih besar. Tidak semua data yang tersedia akan berguna untuk tujuan analisis tertentu, sehingga penting untuk memilih data yang memiliki keterkaitan langsung dengan masalah yang sedang diteliti. Seleksi data yang baik membantu dalam mengurangi kompleksitas dan meningkatkan fokus analisis. (Nofriansyah, 2014).

2.3.4 *Data transformation*

Data transformation atau Transformasi Data adalah tahap di mana data yang dipilih diubah menjadi bentuk yang sesuai untuk analisis lebih lanjut. Ini bisa mencakup normalisasi, agregasi, dan konstruksi fitur baru. Transformasi ini diperlukan untuk memastikan bahwa data berada dalam format yang optimal untuk metode penambangan data yang akan diterapkan. Proses ini memastikan bahwa data dapat dianalisis dengan cara yang paling efektif dan efisien (Han et al, 2012).

2.3.5 *Data Mining*

Data Mining atau Penambangan Data adalah inti dari proses *KDD*, di mana algoritma khusus diterapkan untuk mengekstrak pola dari data. Teknik penambangan data dapat mencakup prediksi, klasifikasi, klastering, estimasi, dan asosiasi. Tujuannya

adalah menemukan pola yang tersembunyi dalam data yang dapat memberikan wawasan baru atau mendukung pengambilan keputusan (Fayyad et al., 1996).

2.3.6 *Pattern Evaluation*

Pattern Evaluation atau Evaluasi Pola adalah tahap di mana pola yang dihasilkan dari penambangan data dinilai berdasarkan kriteria tertentu untuk menentukan apakah mereka menarik dan berguna. Ini melibatkan penggunaan ukuran keterarikan untuk menilai relevansi dan utilitas pola yang ditemukan. Pola yang tidak memenuhi kriteria ini dapat diabaikan atau disesuaikan untuk analisis lebih lanjut (Fayyad et al., 1996).

2.3.7 *Knowledge Presentation*

Knowledge Presentation atau Presentasi Pengetahuan adalah tahap akhir dalam proses *KDD*, di mana hasil dari penambangan data disajikan kepada pengguna melalui teknik visualisasi dan presentasi. Tujuannya adalah untuk membuat pengetahuan yang dihasilkan dapat dipahami dan berguna bagi pengguna akhir. Teknik presentasi yang baik dapat membantu dalam interpretasi dan pengambilan keputusan yang lebih baik (Fayyad et al., 1996).

2.4 *Data Mining*

Seperti yang dijelaskan diatas, bahwasanya *Data Mining* adalah proses inti dari *Knowledge Discovery in Database* atau *KDD*, adapun beberapa metode pengelompokan yang digunakan dalam *Data Mining* dibagi menjadi 5 yaitu estimasi, prediksi, asosiasi, klasifikasi, dan klasterisasi (Suntoro, 2018).

- **Estimasi** hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori, model dibangun dengan menggunakan *record* lengkap yang menyediakan nilai dari variabel prediksi (Nofriansyah, 2014).
- **Prediksi** hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi, nilai dari hasil akan ada di masa mendatang (Nofriansyah, 2014).

- **Asosiasi** dalam *Data Mining* adalah menemukan atribut yang muncul dalam satu waktu (Nofriansyah, 2014).
- **Klasifikasi** terdapat target variabel kategori, sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam 3 kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah (Nofriansyah, 2014).
- **Klasterisasi** atau pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan satu dengan yang lainnya, dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain (Nofriansyah, 2014).

2.5 *Association Rules*

Dalam bidang keilmuan *Data Mining*, terdapat suatu metode yang dinamakan *association rule*. *Association rule* mining adalah suatu prosedur untuk mencari hubungan antar *item* dalam suatu *dataset* yang ditentukan. Asosiasi dikenal sebagai salah satu teknik *Data Mining* yang menjadi dasar dari salah satu teknik *Data Mining* lainnya (Fauzy et al., 2016). Metode ini sering juga dinamakan dengan *market basket analysis*. Contoh aturan asosiasi dari analisis pembelian di suatu pasar swalayan adalah dapat diketahuinya berapa besar kemungkinan seorang konsumen membeli roti bersama dengan susu. Aturan osisasi memberikan informasi dalam bentuk hubungan "*if-then*" atau "jika-maka" dan memeriksa semua kemungkinan hubungan *if-then* antar *item* dan memilih hanya yang paling mungkin (*most likely*) sebagai indikator dari hubungan ketergantungan antar *item*. Istilah *antecedent* untuk mewakili bagian "jika" dan *consequent* untuk mewakili bagian "maka" (Listriani et al., 2018). Metodologi dasar aturan asosiasi dijelaskan sebagai berikut (Listriani et al., 2018):

2.5.1 Pembentukan Pola Frekuensi Tinggi (*Frequent Itemset*)

Tahap ini mencari kombinasi *item* yang memenuhi syarat minimum dari nilai *support* dalam suatu *database*. *Support* yaitu seberapa banyak suatu *item* yang muncul dari keseluruhan transaksi. Nilai *support* adalah nilai penunjang atau persentase kombinasi sebuah *item* bersamaan dalam suatu *database*. Semakin besar nilai *support*

menandakan semakin banyak data pendukung yang ditemukan dalam *database*. Nilai *support* sebuah *item* diperoleh dari persamaan (1) atau persamaan (2):

$$Support\ (A) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A}{Total\ Transaksi} \quad (1)$$

Cara mencari nilai *support* dari 2 *item*:

$$Support\ (A \cup B) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A\ dan\ B}{Total\ Transaksi} \quad (2)$$

2.5.2 Pembentukan Aturan Asosiasi (*Association Rules*)

Setelah seluruh pola frekuensi tinggi ditemukan , maka tahap selanjutnya adalah membentuk aturan asosiasi dengan melihat kombinasi *item* yang memenuhi syarat minimum dari nilai *confidence*. Nilai *confidence* adalah nilai keyakinan berupa kuatnya hubungan antar *item* yang didapatkan. Semakin besar nilai *confidence* menandakan semakin besar kemungkinan kombinasi *item* muncul secara bersamaan. Nilai *confidence* sebuah *item* diperoleh dari persamaan (3).

$$Confidence\ P\ (B|A) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A\ dan\ B}{Jumlah\ transaksi\ yang\ mengandung\ A} \quad (3)$$

Kedua parameter diatas digunakan untuk menentukan kekuatan suatu pola dan menemukan pola yang memenuhi syarat minimum untuk *support* dan syarat minimum untuk *confidence* (Priyana et al, 2015).

2.5.3 Rasio Peningkatan (*Lift Ratio*)

Lift Ratio merupakan nilai yang menunjukkan keabsahan aturan yang terbentuk dalam proses transaksi dan memberikan informasi apakah benar produk A dibeli bersamaan dengan produk B. *Lift ratio* mengukur seberapa penting aturan yang telah terbentuk berdasarkan nilai *support* dan *confidence* yang telah didapatkan sebelumnya. Jika nilai *lift ratio* kurang dari atau sama dengan (<=) 1, maka hubungan sebab-akibat yang terjadi bersifat saling lepas satu sama lain. Sedangkan, jika nilai *lift ratio* lebih dari (>) 1, maka hubungan sebab-akibat yang terjadi bersifat saling berhubungan satu

sama lain dan dapat dikatakan kejadian tersebut bukan kebetulan dan akan berulang. Nilai *lift ratio* diperoleh dari persamaan (4):

$$Lift\ Ratio = \frac{Confidence\ Antecedent}{Support\ Consequence} \quad (4)$$

2.6 Algoritma *FP-Growth*

Algoritma *Frequent Pattern-Growth* atau yang biasa disebut dengan *FP-Growth* merupakan pengembangan dari algoritma *apriori*, sehingga pada algoritma *FP-Growth*, segala kekurangan dalam algoritma *apriori* telah diperbaiki. *Frequent Pattern Growth* (*FP-Growth*) merupakan algoritma yang dapat digunakan untuk menentukan himpunan data yang paling sering muncul (*Frequent Itemset*) dalam sebuah kumpulan data. Algoritma *FP-Growth* hanya memerlukan 2 kali *scanning database* untuk menentukan *Frequent Itemset*. Struktur data yang digunakan untuk mencari *Frequent Itemset* dengan algoritma *FP-Growth* adalah perluasan dari penggunaan sebuah pohon prefix, yang biasa disebut adalah *FP-Tree*. Dengan menggunakan *FP-Tree*, algoritma *FP-Growth* dapat langsung mengekstrak *Frequent Itemset* dari *FP-Tree* yang telah terbentuk dengan menggunakan prinsip *divide and conquer* (Samuel, 2008).

2.6.1 Pembangunan *FP-Tree*

FP-Tree dibangun dengan memetakan setiap data transaksi ke dalam setiap lintasan tertentu dalam *FP-Tree*. Setiap transaksi yang dipetakan, mungkin ada transaksi yang memiliki *item* yang sama, maka lintasannya memungkinkan untuk saling menimpa. Semakin banyak data transaksi yang memiliki *item* yang sama, maka proses pemampatan dengan struktur data *FP-Tree* semakin efektif. Kelebihan dari *FP-Tree* adalah hanya memerlukan dua kali pemindaian data transaksi yang terbukti sangat efisien (Samuel, 2008).

Misal I = {a1, a2, …, an} adalah kumpulan dari *item* dan basis data transaksi DB = {T1, T2, …, Tn}, di mana Ti (i € [1..n]) adalah sekumpulan transaksi yang mengandung *item* di I. Sedangkan *support* adalah penghitung (*counter*) frekuensi kemunculan transaksi yang mengandung suatu pola. Suatu pola dikatakan sering

muncul (*Frequent Pattern*) apabila *support* dari pola tersebut tidak kurang dari suatu konstanta ξ (batas ambang *minimum support*) yang telah didefinisikan sebelumnya. Permasalahan mencari pola *Frequent* dengan batas ambang *minimum support count* (ξ) inilah yang dicoba untuk dipecahkan oleh *FP-Growth* dengan bantuan struktur *FP-Tree* (Samuel, 2008). Adapun *FP- Tree* adalah sebuah pohon dengan definisi sebagai berikut:

- *FP-Tree* dibentuk oleh sebuah akar yang diberi label *Null*, sekumpulan pohon yang beranggotakan *item-item* tertentu, dan sebuah tabel *Frequent header*.
- Setiap simpul dalam *FP-Tree* mengandung tiga informasi penting, yaitu label *item*, menginformasikan jenis *item* yang direpresentasikan simpul tersebut, *support count*, merepresentasikan jumlah lintasan transaksi yang melalui simpul tersebut, dan *pointer* penghubung yang menghubungkan simpul-simpul dengan label *item* sama antar lintasan, ditandai dengan garis panah putus-putus (Samuel, 2008). Sebagai contoh, tabel 2.6 merupakan data transaksi dengan *minimum support count* ξ=2.

Tabel 2. 1 Data transaksi mentah

No	Transaksi
1	a,b
2	b,c,d,g,h
3	a,c,d,e,f
4	a,d,e
5	a,b,z,c
6	a,b,c,d
7	a,r
8	a,b,c
9	a,b,d
10	b,c,e

Sumber: Samuel, 2008

Kemudian dilakukan perhitungan frekuensi kemunculan tiap *item* dari semua *item* yang ada. Frekuensi kemunculan tiap *item* dapat dilihat pada tabel 2.2 berikut:

Tabel 2. 2 Frekuensi kemunculan tiap karakter

Item	**Frekuensi**
a	8
b	7
c	6
d	5
e	3
f	1
g	1
h	1
i	1
j	1

Sumber: Samuel, 2008

Setelah dilakukan pemindaian pertama di dapat *item* yang memiliki frekuensi di atas *support count* ξ=2 adalah a,b,c,d, dan e. Kelima *item* inilah yang akan berpengaruh dan akan dimasukkan ke dalam *FP-Tree*, selebihnya (r, z, g, dan h) dapat dibuang karena tidak berpengaruh signifikan. Tabel 2.3 menunjukkan kemunculan *item* yang *Frequent* dalam setiap transaksi, diurut berdasarkan frekuensinya yang paling tinggi (Samuel, 2008).

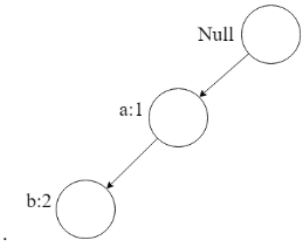
Tabel 2. 3 Data transaksi

TID	***Item***
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Sumber: Samuel, 2008

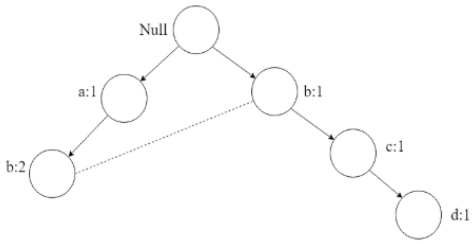
Adapun hasil-hasil dari pembacaan *TID* di atas diilustrasikan pada gambar-gambar berikut:

- Hasil pembentukan *FP-Tree* pada pembacaan *TID* 1



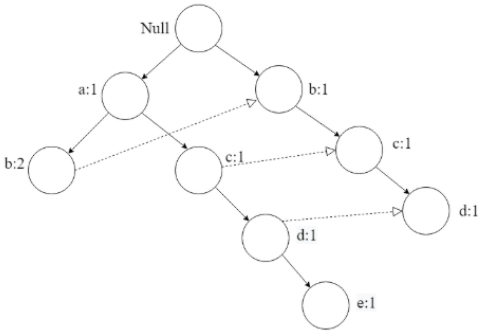
Gambar 2. 2 Hasil Pembentukan FP-Tree TID 1 (Samuel, 2008)

- Hasil pembentukan *FP-Tree* pada pembacaan *TID* 2



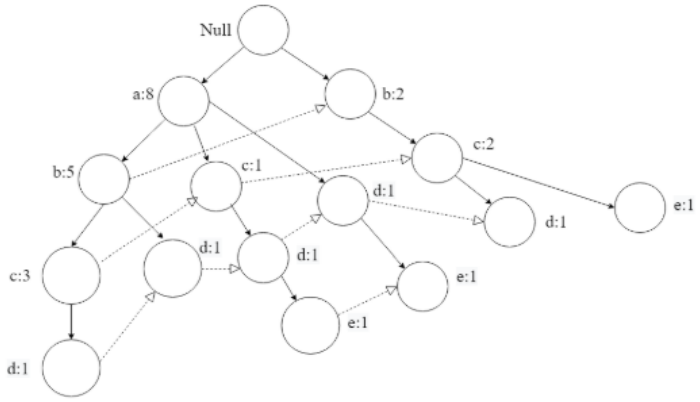
Gambar 2. 3 Hasil Pembentukan FP-Tree TID 2 (Samuel, 2008)

- Hasil Pembentukan *FP-Tree* pada pembacaan *TID* 3



Gambar 2. 4 Hasil Pembentukan FP-Tree *TID* 3 (Samuel, 2008)

- Hasil pembentukan *FP-Tree* pada pembacaan *TID* 10



Gambar 2. 5 Hasil Pembentukan FP-Tree TID 10 (Samuel, 2008)

Diberikan 10 data transaksi dengan 5 jenis *item* seperti pada tabel 2.3. Gambar 2.2, Gambar 2.3, Gambar 2.4 dan Gambar 2.5 menunjukkan proses terbentuknya *FP-Tree* setiap *TID* dibaca. Setiap simpul pada *FP-Tree* mengandung nama sebuah *item* dan *counter support* yang berfungsi untuk menghitung frekuensi kemunculan *item* tersebut dalam tiap lintasan transaksi (Samuel, 2008).

2.6.2 Penerapan Algoritma *FP-Growth*

Setelah tahap pembangunan *FP-Tree* dari sekumpulan data transaksi, akan diterapkan algoritma *FP-Growth* untuk mencari *Frequent Itemset* yang signifikan (Samuel, 2008). Algoritma *FP-Growth* dibagi menjadi tiga langkah utama, yaitu :

a. Pembangkitan *Conditional Pattern Base*

Conditional Pattern Base merupakan *subdatabase* yang berisi *prefix path* (lintasan awal) dan *suffix Pattern* (pola akhiran). Pembangkitan *conditional Pattern base* didapatkan melalui *FP-Tree* yang telah dibangun sebelumnya.

b. Pembangkitan *Conditional FP-Tree*

Pada tahap ini, *support count* dari setiap *item* pada setiap *conditional Pattern base* dijumlahkan, lalu setiap *item* yang memiliki jumlah *support count* lebih besar sama dengan *minimum support count* ξ akan dibangkitkan dengan *conditional FP-Tree.*

c. Pencarian *Frequent Itemset*

Apabila *conditional FP-Tree* merupakan lintasan tunggal (*single path*), maka didapatkan *Frequent Itemset* dengan melakukan kombinasi *item* untuk setiap *conditional FP-Tree*. Jika bukan lintasan tunggal, maka dilakukan pembangkitan *FP-Growth* secara rekursif.

2.7 *Machine Learning*

Machine learning (pembelajaran mesin) mengeksplorasi bagaimana komputer dapat meningkatkan kinerjanya berdasarkan data yang diberikan. Program komputer secara otomatis belajar untuk mengenali pola yang kompleks dan membuat keputusan cerdas berdasarkan data tersebut. Sebagai contoh, pembelajaran mesin memungkinkan komputer untuk secara otomatis mengenali kode pos tulisan tangan pada surat setelah mempelajari sejumlah contoh. Pembelajaran mesin adalah bidang yang berkembang dengan cepat. (Han et al., 2012). Berikut merupakan beberapa pendekatannya :

2.7.1 *Supervised Learning*

Supervised learning adalah pendekatan yang digunakan untuk data yang memiliki informasi kelas atau label. Label adalah variabel yang mengidentifikasi setiap data dalam kumpulan data. Tujuan dari pendekatan ini adalah untuk mengklasifikasikan suatu data ke dalam kategori yang telah ada. (Han et al., 2012).

2.7.2 *Unsupervised Learning*

Unsupervised learning adalah proses pembelajaran yang tidak diawasi karena data yang digunakan tidak memiliki label. Akibatnya, hasil dari *unsupervised learning* tidak bisa diprediksi sebelumnya karena model ini belajar secara mandiri untuk

menemukan pola dalam kumpulan data yang diberikan. Salah satu aplikasi dari *unsupervised learning* adalah *association rules* (Han et al., 2012).

2.7.3 *Semi-supervised Learning*

Semi-supervised learning adalah gabungan antara *supervised learning* dan *unsupervised learning*. Dalam pendekatan ini, data yang digunakan terdiri dari kombinasi data yang berlabel dan tidak berlabel (Han et al., 2012).

2.7.4 *Active Learning*

Active learning adalah kasus khusus dalam *machine learning* di mana sistem secara interaktif meminta peneliti atau pengguna untuk memberikan label pada data dengan keluaran yang diinginkan (Han et al., 2012).

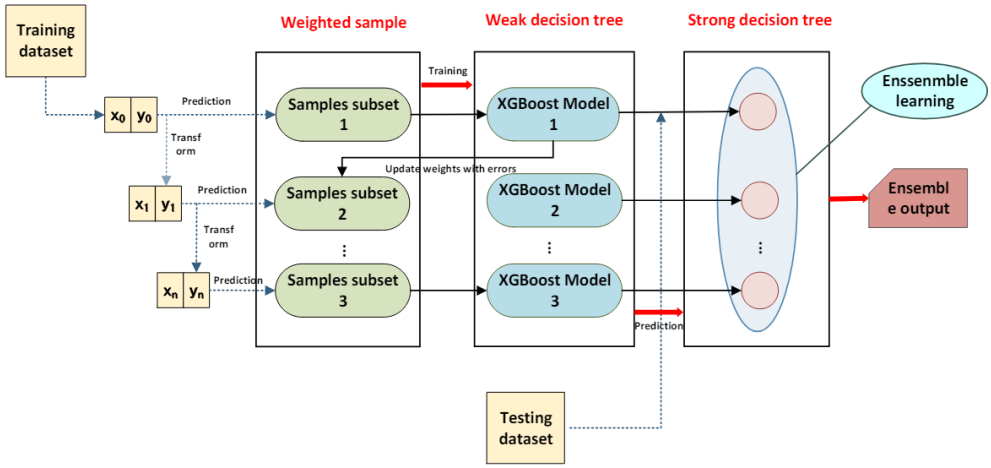
2.8 Prediksi

Dalam konteks *machine learning*, prediksi adalah proses menggunakan data yang diamati untuk membuat estimasi atau pernyataan tentang data yang belum diamati atau data baru. Tujuan utama dari prediksi ini adalah untuk menghasilkan model atau algoritma yang dapat mempelajari pola-pola yang ada dalam data historis, sehingga ketika diberikan data baru, model tersebut dapat memberikan estimasi yang paling akurat atau probabilitas yang tinggi terhadap hasil yang diharapkan. Proses ini mendasarkan diri pada analisis statistik dan teknik-teknik pembelajaran mesin yang digunakan untuk mengklasifikasikan atau mengidentifikasi pola yang relevan dalam data yang tersedia (Kevin, 2012).

2.9 Algoritma *Extreme Gradient Boost* (*XGBoost*)

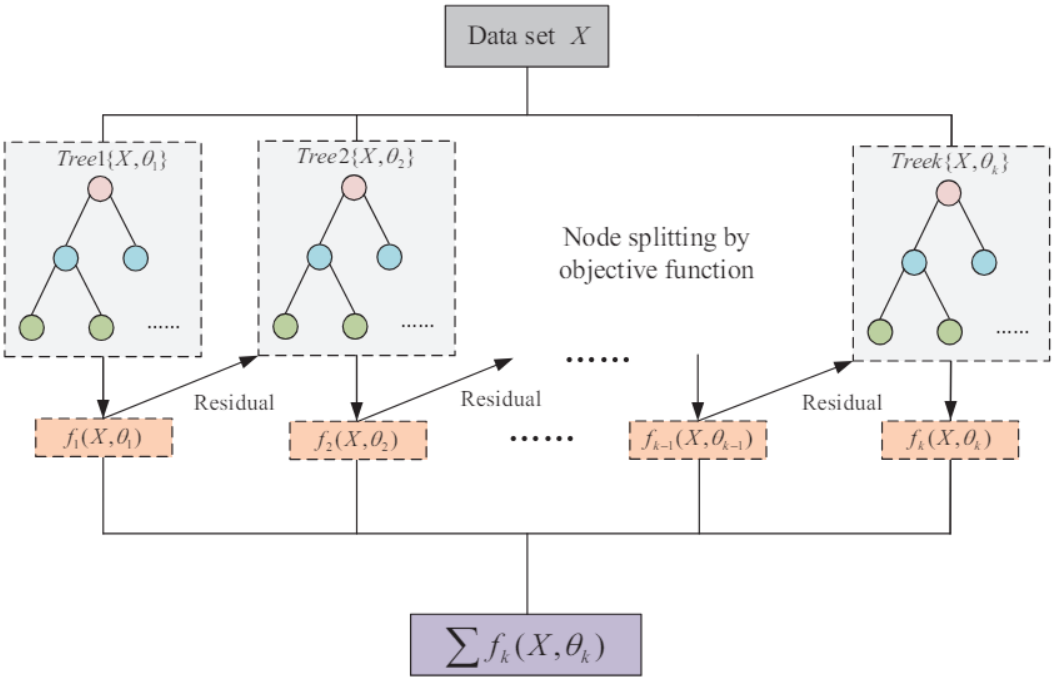
Algoritma *XGBoost* saat ini merupakan algoritma dengan jenis *Decision Tree* tercepat dan paling terintegrasi dengan baik. Algoritma ini menggunakan *CART* sebagai klasifier dasar, dan diputuskan secara bersama-sama oleh beberapa *Decision Tree* terkait; di mana sampel *input* dari *Decision Tree* berikutnya akan terkait dengan

hasil pelatihan dan prediksi dari *Decision Tree* sebelumnya. *XGBoost* adalah alat yang sangat fleksibel dan serbaguna yang dapat menyelesaikan sebagian besar masalah regresi dan klasifikasi, serta fungsi tujuan yang dibuat pengguna (Ma et al, 2021).



Gambar 2. 6 Struktur Algoritma *XGBoost* (Ma et al, 2021)

Selama proses pelatihan, model menghitung kerugian pada simpul-simpul secara berkelanjutan untuk memilih simpul daun dengan peningkatan kerugian terbesar. *XGBoost* menambahkan pohon-pohon baru dengan terus membagi fitur-fitur. Setiap penambahan pohon pada setiap iterasi sebenarnya adalah pembelajaran fungsi baru $fk(X,\theta k)$ untuk menyesuaikan sisa prediksi sebelumnya. Setelah pelatihan dan mendapatkan *K* pohon, atribut-atribut prediksi pada sampel akan sesuai dengan simpul daun tertentu di setiap pohon, dan masing-masing simpul daun ini sesuai dengan sebuah skor. Pada akhirnya, skor-skornya dari setiap pohon dijumlahkan untuk mendapatkan nilai prediksi yang menggambarkan sampel tersebut (Guo et al, 2020). Adapun desain dari *flowchart* algoritma *XGBoost* seperti yang ditampilkan pada gambar 2.7.



Gambar 2. 7 *Flowchart* algoritma *XGBoost* (Guo et al, 2020)

2.10 Evaluasi Sistem

Evaluasi sistem adalah proses sistematis yang bertujuan untuk menilai nilai, manfaat, dan signifikansi suatu sistem dengan menggunakan kriteria yang diatur oleh standar tertentu. Proses evaluasi ini sering kali melibatkan pengukuran melalui observasi langsung atau metodologi eksperimental untuk memastikan bahwa sistem berfungsi sesuai dengan tujuan yang diharapkan. Dalam konteks *machine learning*, evaluasi dilakukan menggunakan berbagai metrik seperti presisi, *recall*, spesifitas, dan akurasi untuk menilai kinerja model.

a. **Akurasi** adalah proporsi hasil yang benar (baik positif benar maupun negatif benar) dalam populasi. Ini mengukur keseluruhan ketepatan model dengan menunjukkan berapa banyak sampel yang diklasifikasikan dengan benar dari semua contoh (Powers, 2011)

b. **Presisi (*Precision*)** adalah fraksi dari sampel yang diambil yang relevan. Ini adalah ukuran dari akurasi hasil yang diambil, yang mewakili proporsi hasil positif benar dalam semua prediksi positif (Manning et al, 2008).
c. ***Recall*** juga dikenal sebagai sensitivitas, adalah fraksi dari sampel relevan yang telah diambil dari jumlah total sampel relevan. Dalam *machine learning*, *recall* mengukur kemampuan model untuk menangkap semua contoh yang relevan (Manning et al, 2008).
d. **Spesifitas (*Spesificity*)** adalah proporsi negatif yang benar-benar diidentifikasi sebagai negatif oleh model. Ini mengukur kemampuan model untuk mengidentifikasi dengan benar contoh-negatif, yaitu, proporsi dari contoh negatif yang diprediksi negatif oleh model (Fawcett, 2006).

2.11 *Confusion Matrix*

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja dari suatu model klasifikasi. Tabel ini memiliki empat sel utama yang mewakili jumlah prediksi yang benar dan salah yang dilakukan oleh model terhadap data uji. Secara umum, *confusion matrix* terdiri dari empat bagian utama:

1. ***True Positive* (*TP*)**: Jumlah data positif yang diprediksi benar sebagai positif.
2. ***False Positive* (*FP*)**: Jumlah data negatif yang salah diprediksi sebagai positif.
3. ***True Negative* (*TN*)**: Jumlah data negatif yang diprediksi benar sebagai negatif.
4. ***False Negative* (*FN*)**: Jumlah data positif yang salah diprediksi sebagai negatif.

Confusion matrix membantu dalam menghitung berbagai metrik evaluasi klasifikasi seperti akurasi, presisi, *recall* (sempurna), dan nilai F1 (Han et al, 2012).