BAB 2
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Pengumpulan informasi pada penelitian ini diawali dengan melakukan studi literatur yang digunakan sebagai salah satu cara untuk melengkapi informasi terkini dan mempertajam masalah yang dibahas. Beberapa hasil penelitian yang ditulis dalam bentuk jurnal dipilih sesuai dengan topik dan temanya.

Tidak banyak penelitian yang ditemukan terkait prediksi popularitas gim indi pernah dilakukan sebelumnya. Penelitian [6], menggunakan metode *logistic regression* dan *random forest* untuk prediksi popularitas gim indi pada data gim *steam* yang diambil dari *Kaggle,* menggunakan berbagai tipe fitur mulai dari numerikal, kategorikal, dan teks didalamnya. Menunjukan model *logistic regression* menghasilkan akurasi 80.2 % dan *random forest* menghasilkan akurasi 77.9%. Walaupun memiliki performa model yang masih relatif rendah dari kedua algoritma yang digunakan meskipun sudah diterapkan *hyperparameter tuning.*

Penelitian terkait popularitas gim indi [12] dengan melibatkan berbagai jenis fitur pada dataset gim indi *steam platform* dan menggunakan algoritma *neural network* menyimpulkan bahwa model yang menggunakan berbagai jenis fitur akan menghasilkan performa yang lebih menjanjikan. Pada penelitian ini menyimpulkan bahwa dengan menggunakan lebih dari satu tipe fitur menghasilkan performa model yang lebih unggul.

Selain penelitian [6], [12], peneliti tidak menemukan penelitian lain terkait dengan prediksi popularitas gim indi. Dari penelitian [6], [12], juga menunjukan belum banyak algoritma yang diteliti performanya pada prediksi gim indi. Menurut penelitian [13], algoritma *boosting* lebih baik dalam memprediksi. Penelitian [11], membandingkan algoritma *boosting* dengan *logistic regression* dalam memprediksi dataset risiko kredit. Penelitian ini menyimpulkan bahwa algoritma *CatBoost* memiliki performa

yang lebih baik daripada algoritma lainnya. Model *CatBoost* pada klasifikasi risiko kredit jerman mencapai akurasi 81% dan nilai *AUC* 73%. Sedangkan model *CatBoost* memiliki akurasi 94% dan *AUC* 92% pada pinjaman ekuitas rumah USA. Sedangkan model *logistic regression* memiliki akurasi dan *AUC* yang lebih rendah dibawah algoritma *CatBoost* dan *boosting* lainnya.

Kinerja *CatBoost* tersebut yang lebih baik daripada algoritma yang lain juga dibuktikan pada penelitian [8], yang membandingkan algoritma *CatBoost* dengan berbagai macam algoritma *machine learning* lainnya seperti *random forest, logistic regression, decission tree, neural netkwork,* dan algoritma *boosting* seperti *XGBoost,* dan *gradient boosting* untuk prediksi persetujuan pinjaman jangka panjang dan promosi staf. Hasil penelitian ini menunjukan bahwa algoritma *CatBoost* memiliki performa yang lebih baik pada persetujuan pinjaman jangka panjang, dengan akurasi 72% dan nilai *AUC* sebesar 78%. Begitu pula pada dataset promosi staf model mencapai akurasi 94% dan nilai *AUC* sebesar 82%.

Selain model prediksi diperlukan juga pemahaman tentang fitur fitur yang berpengaruh pada hasil model prediksi. Menurut penelitian [9], *Shapley Additive Explanations* (*SHAP)* memungkinkan untuk menginterpretasikan model prediksi yang bersifat *blackbox* atau sulit dipahami. Saran penggunaan *SHAP* juga disampaikan penelitian [14] karena kemampuannya yang lebih baik dalam menginterpretasikan model daripada metode lainnya.

Penggunaan metode *SHAP* pada model prediksi pernah dilakukan sebelumnya. Penelitian [15], membuat model deteksi secara *real-time* kecelakaan jalan raya menggunakan XGBoost, menunjukkan bahwa lalu lintas harian rata-rata yang lebih tinggi menghasilkan probabilitas kecelakaan yang lebih tinggi ditunjukkan oleh hasil interpretasi dari *SHAP*.

Penggunaan *SHAP* pada penelitian [16], untuk model prediksi pengendapan akibat pengeboran *shield tunneling* menggunakan algoritma boruta. Hasil interpretasi yang dibuat *SHAP* menunjukan bahwa fitur *soil*

type (ST) merupakan fitur paling penting diantara fitur lainnya. Fitur *soil type* yang memiliki dua instansi didalamnya, yaitu lempung berlumpur (titik biru) dan pasir berlumpur (titik merah). Lempung berlumpur memiliki nilai *shapley* positif yang lebih tinggi memberikan kontribusi lebih besar pada keluaran model daripada pasir berlumpur.

Penggunaan algoritma *CatBoost* dan *Shapley Additive Explanations* (*SHAP*) secara bersamaan pernah dilakukan pada penelitian [10] dalam analisis stabilitas seismik yang efisien pada lereng tanggul yang mengalami perubahan muka air. Menunjukkan bahwa semua nilai koefisien determinasi (R2) *CatBoost* lebih besar dari 0,90 untuk dataset pelatihan dan dataset pengujian. Kemudian pada fitur yang berpengaruh menggunakan *SHAP* menunjukan bahwa di antara empat fitur, fitur yang mempengaruhi, faktor sudut gesek memiliki pengaruh paling signifikan terhadap prediksi, diikuti oleh koefisien seismik horizontal, kohesi, dan permeabilitas jenuh. Berbeda dengan parameter kekuatan geser yang memiliki pengaruh positif berpengaruh positif terhadap stabilitas lereng timbunan, peningkatan koefisien seismik horizontal dan permeabilitas jenuh cenderung mendestabilisasi kestabilan lereng timbunan.

Penelitian yang menggunakan metode *CatBoost* dan *Shapley Additive Explanations* (*SHAP*) juga pernah dilakukan pada data Diabetes Mellitus [11]. Model *CatBoost* disimpulkan memiliki kinerja yang baik untuk mengklasifikasikan apakah seorang pasien menderita diabetes melitus atau tidak dengan nilai *AUC* validasi sebesar 86,86% menggunakan classifier *CatBoost*. Dari model tersebut dengan penggunaan SHAP menunjukkan angka d1_glucose_max yang tinggi memiliki nilai *shapley* semakin tinggi pada prediksi Diabetes Mellitus.

Kemampuan *SHAP* menginterpretasikan model *CatBoost* pernah dibuktikan pada penelitian [17]. Model *CatBoost* yang mencapai kinerja optimal dalam prediksi risiko T2DKD, *SHAP* menunjukkan pada model *CatBoost* yang memilili AUC 0,84. Fitur yang memiliki kontribusi positif terdiri dari tekanan darah sistolik, kreatinin, lama rawat inap, waktu

trombin, usia, waktu protrombin, rasio sel trombosit, albumin, glukosa, fibrinogen, lebar distribusi sel darah merah-simpangan standar, serta hemoglobin A1C.

Didukung penelitian [18] yang Dimana model *CatBoost* menghasilkan *AUC* tertinggi (0,933) untuk mortalitas dibandingkan dengan SAPS3 dan *SOFA* (0,860 dan 0,867). *SHAP* mengungkapkan bahwa peningkatan *DNI* pada hari ke-3, syok septik, penggunaan terapi norepinefrin, dan peningkatan rasio normalisasi internasional pada hari ke-3 memiliki dampak terbesar pada prediksi model kematian.

Terdapat beberapa perbedaan utama dibandingkan dengan penelitian sebelumnya. Pertama, penelitian ini menggunakan algoritma *CatBoost* sebagai metode pengklasifikasian utama, sementara beberapa penelitian sebelumnya lebih cenderung menggunakan metode *logistic regression, random forest*, dan *neural network*. Kedua, penelitian ini menggunakan *Shapley Additive Explanations* untuk menginterpretasikan model prediksi, memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang berpengaruh dalam menentukan popularitas gim indi berdasarkan data dari *steam platform*.

Tabel 2.1 menunjukkan penelitian terkait prediksi popularitas gim indi pada *steam platform* menggunakan algoritma *catboost* dan *shapley additive explanations (SHAP) :*

Tabel 2.1 Studi Literatur

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
Ziyang Jiang (2021)	*Predicting the Popularity of Independent Video Games on the Steam Platform*	Membuat gim mereka menonjol di pasar yang meningkat secara eksponensial adalah tantangan besar bagi setiap pengembang independent. Maka perlu dibuat model prediksi	Membuat, menguji, dan memvalidasi model *linear regression* dan *random forest* untuk prediksi popularitas gim indi.	*Linear regression* dan *random forest* diberi *hyper parameter tuning*	Dataset *game steam* Kaggle.	Kedua model yang digunakan untuk membuat model prediksi popularitas gim indi menggunakan semua tipe data yang ada dan *hyperparameter tuning* maksimal 80% berhasil dibuat
Huang Y, Chu W (2022)	*Indie Games Popularity Prediction by Considering Multimodal Features*	Sistem prediksi popularitas untuk gim komputer independen sebelumnya tidak mempertimbangkan informasi visual, teks, dan metadata secara bersamaan.	Mengetahui performa prediksi popularitas untuk gim komputer independen dengan mempertimbangkan informasi visual, teks, dan metadata secara bersamaan.	*neural network*	Data s*team platform*	Penggunaan lebih dari satu jenis fitur akan meningkatkan performa model
Linn´ea Machado, David Holmer (2022)	*Credit risk modelling and prediction: Logistic regression versus machine*	Ingin mengetahui efektivitas dan ketepatan tiga algoritma *machine learning*, yaitu *XGBoost, CatBoost,* dan *logistic regression,*	Membandingkan performa tiga algoritma *machine learning*, yaitu *XGBoost, CatBoost,* dan *logistic regression*, dalam memprediksi risiko kredit	*XGBoost, CatBoost,* dan *logistic regression*	Risiko kredit Jerman dan USA	Algoritma *CatBoost* memiliki kemampuan yang lebih baik dalam melakukan prediksi daripada *logistic regression* dengan

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
	learning boosting algorithms	dalam tugas klasifikasi penilaian risiko kredit.				nilai akurasi 81% dan *AUC* 73% pada risiko kredit jerman, serta nilai akurasi 94% dan *AUC* 92% pada pinjaman ekuitas rumah USA
Abdullahi A. Ibrahim, Raheem L.Ridwan , Muhammed M. Muhammed, Rabiat O. Abdulaziz, Ganiyu A. Saheed (2020)	*Comparison of the CatBoost Classifier with other Machine Learning Methods*	Berdasarkan performa algoritma *CatBoost* dalam kedua analisis, apakah algoritma ini dapat direkomendasikan untuk prediksi yang lebih baik dalam persetujuan pinjaman dan promosi staff	Membandingkan algoritma *CatBoost* dengan berbagai macam algoritma *machine learning* lainnya seperti *random forest, logistic regression, decission tree, neural netkwork,* dan algoritma *boosting* seperti *XGBoost,* dan *gradient boosting* dalam memprediksi persetujuan pinjaman jangka panjang, dan promosi staff	*CatBoost, Random Forest, Logistic Regression, Decission Tree, Neural Netkwork,* dan algoritma *boosting* seperti *XGBoost,* dan *gradient boosting*	*Data of mortgage approvals USA government data,* dan *staff promotion*	*CatBoost* memiliki performa yang lebih baik daripada algoritma *random forest,* dan *logistic regression* dengan akurasi 72% dan nilai *AUC* sebesar 78% pada persetujuan pinjaman jangka panjang dan pada promosi staff dengan akurasi 94% dan nilai *AUC* sebesar 82%.
Bahador Parsa A, Movahedi A, Taghipour H, Derrible S, Mohammadadian A (2019)	*Toward Safer Highways, Application of XGBoost and SHAP for Real-Time*	Terjadinya kecelakaan lalu lintas merupakan masalah utama di berbagai negara di seluruh dunia. Dengan pesatnya peningkatan	Membuat model prediksi atau deteksi secara *real-time kecelakaan* lalu lintas dan mengetahui apa yang mempengaruhi	*CatBoost* dan *Shapley Additive Explanations*	Data kecelakaan dari *Illinois Department of Transportation (IDOT)*	Penggunaan *Shapley Additive Explanations* dapat menunjukan pengaruh fitur terhadap model

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
	Accident Detection and Feature Analysis	jumlah jalan raya dan kendaraan bermotor di sebagian besar negara mengakibatkan jumlah total kecelakaan telah meningkat secara substansial di dunia. Diperlukan model yang memdeteksi kecelakaan secara *real-time*	nya menggunakan *XGBoost* dan *Shapley Additive Explanations*			dengan menunjukkan bahwa lalu lintas harian rata-rata yang lebih tinggi menghasilkan probabilitas kecelakaan yang lebih tinggi
K.K. Pabodha M. Kannangara a, Wanhuan Zhou a, Zhi Ding b, Zhehao Hong a (2022)	*Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method*	Meskipun algoritma *Machine Learning (ML)* dapat digunakan untuk memprediksi penurunan yang disebabkan oleh terowongan sebelumnya, model yang memiliki performa yang baik biasanya kurang dapat diinterpretasikan. Artinya, meskipun model memberikan hasil yang akurat, tidak jelas bagaimana model tersebut membuat keputusan berdasarkan fitur-fitur inputnya.	Menggunakan metode *Shapley additive explanations (SHAP)* untuk mengeksplorasi bagaimana fitur-fitur input berkontribusi terhadap hasil dari model yang kompleks.	*Boruta algorithm*	Dataset *tunnel geometry, geological conditions and shield operational parameters*	Penggunaan *Shapley Additive Explanations* dapat menunjukan pengaruh fitur terhadap model dengan menunjukkan parameter kekuatan geser memiliki pengaruh positif terhadap stabilitas lereng timbunan, peningkatan koefisien seismik horizontal dan permeabilitas jenuh cenderung mendestabilisasi

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
						kestabilan lereng timbunan
Luqi Wang Jiahao Wu, Wengang Zhan, Lin Wang, Wei Cui (2021)	*Efficient Seismic Stability Analysis of Embankment Slopes Subjected to Water Level Changes Using Gradient Boosting Algorithms*	Pengaplikasian algoritma *machine learning* ini untuk meningkatkan akurasi dan efisiensi dalam analisis stabilitas lereng seismik yang seringkali terpengaruh oleh perubahan tingkat air	Membandingkan kinerja *CatBoost*, LightGBM, dan XGBoost dalam memprediksi faktor keselamatan (Factor of Safety, FS) dari lereng embankmen seismik berdasarkan empat faktor yang mempengaruhi, yaitu kohesi, sudut gesekan, koefisien seismik horizontal, dan permeabilitas jenuh. dan mencari kontribusi dari fitur diberi peringkat menggunakan metode *Shapley additive explanations*	*CatBoost, LightGBM, XGBoost,* dan *Shapley Additive Explanations*	Dataset dari database stabilitas lereng seismic	Penggunaan algoritma *boosting* seperti *CatBoost* memiliki performa yang lebih bagus daripada algoritma yang lain dengan koef determinasi 0.90, serta penggunaan *Shapley Additive Explanation*s mampu memberikan informasi kontribusi fitur dengan jelas dengan menunjukkan jika Friction angel memiliki kontribusi secara positif yang lebih besar daripada fitur lainnya
Novia Permatasari, Shafiyah Asy Syahidah, Aldo Leofiro Irfiansyah	*Predicting Diabetes Mellitus Using CatBoost Classifier And*	Diabetes melitus sebagai penyakit metabolik yang ditandai dengan hiperglikemia dapat	Melakukan deteksi dini pasien diabetes melitus menggunakan pendekatan machine learning dan dengan	*CatBoost* dan *Shapley Additive Explanations*	Data GOSSIS (*Global Open Source Severity of Illness Score*) milik MIT	Model algoritma *CatBoost* memiliki performa bagus dengan nilai *AUC* validasi sebesar

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
, M. Ghozy Al-Haqqoni (2022)	*Shapley Additive Explanations (Shap) Approach*	berbahaya jika tidak dapat ditangani dengan baik. Deteksi dini terhadap gejala yang ada dapat mengurangi dampak keterlambatan dalam pengobatan	menggunakan *Shapley Additive Explanations* yang memungkinkan penentuan prioritas fitur yang menentukan klasifikasi majemuk			86.86%., serta penggunaan *Shapley Additive Explanation*s mampu memberikan informasi kontribusi fitur dengan menunjukkan d1_glucose_max atau konsentrasi glukosa tertinggi pasien dalam serum atau plasma selama 24 jam pertama masa rawat inap memiliki pengaruh paling tinggi untuk diabetes
Xiao zhu Liu, Minjie Duan, Hao dong Huang, Yang Zhang, Tian yu Xiang, Wu ceng Niu, Bei Zhou, Hao lin Wang, Ting ting Zhang (2023)	*Predicting diabetic kidney disease for type 2 diabetes mellitus by machine learning in the real world: a multicenter retrospective study*	Penyakit ginjal diabetic telah dilaporkan sebagai komplikasi mikrovaskuler utama diabetes melitus. Meskipun biopsi ginjal mampu membedakannya dari penyakit ginjal Non-Diabetes, belum ada standar baku yang	membangun model diagnosis tambahan untuk penyakit ginjal diabetes tipe 2 (T2DKD) berdasarkan algoritme pembelajaran mesin.	*CatBoost, LGBM, XGBoost, Extra Trees Classifier, Gradient Boosting Classifier, Random Forest, Linear Discriminant*	Data klinis pada 3624 orang dengan diabetes tipe 2 (T2DM) dikumpulkan dari 1 Januari 2019 hingga 31 Desember 2019 menggunakan basis data retrospektif	Model algoritma *CatBoost* memiliki performa bagus dengan AUC 0,86, serta penggunaan *Shapley Additive Explanation*s mampu memberikan informasi kontribusi fitur dengan menunjukkan

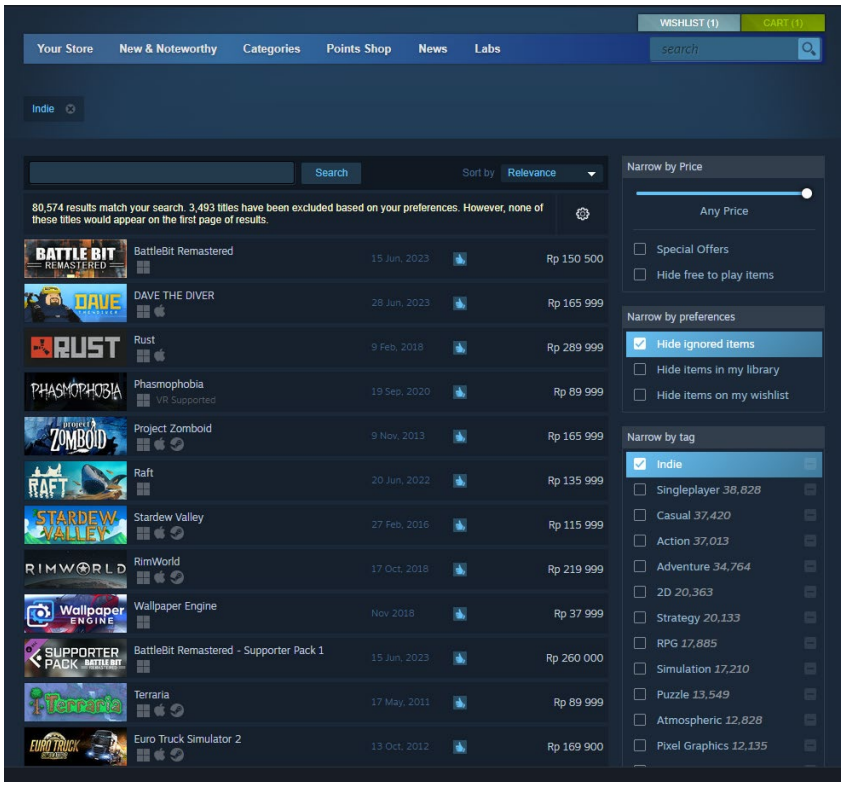
Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
		divalidasi untuk menilai perkembangannya		*Analysis, Logistic Regression, Quadratic Discriminant Analysis, AdaBoost, Naïve Bayes, KNN, DTree, SVM, Ridge*	multi-pusat	ekanan darah sistolik, kreatin, lama rawat inap, waktu trombin, usia, waktu protrombin, rasio sel trombosit, albumin, glukosa, fibrinogen, lebar distribusi sel darah merah-simpangan standar, serta hemoglobin A1C berkontribusi positif terhadap model
Seung Hee Lim, Min Jeong Kim, Won Hyuk Choi, Jin Cheol Cheong, Jong Wan Kim, Kyung Joo Lee, Jun Ho Park (2023)	*Explainable machine learning using perioperative serial laboratory results to predict postoperative mortality in patients with peritonitis-induced sepsis*	Sepsis adalah salah satu penyebab kematian paling umum setelah pembedahan. Beberapa sistem penilaian konvensional telah dikembangkan untuk memprediksi hasil dari sepsis; namun, kekuatan prediktifnya tidak memadai.	Penelitian ini menerapkan algoritma pembelajaran mesin yang dapat dijelaskan untuk meningkatkan akurasi prediksi kematian pasca operasi pada pasien dengan sepsis yang disebabkan oleh peritonitis.	*CatBoost, SAPS3, SOFA*	Pelaksanaan analisis retrospektif terhadap data dari analisis demografi, klinis, dan laboratorium, termasuk delta neutrofil index (DNI), jumlah WBC dan neutrofil, dan tingkat CRP. Data	Model algoritma *CatBoost* memiliki performa bagus AUC tertinggi 0,93, serta penggunaan *Shapley Additive Explanation*s mampu memberikan informasi kontribusi fitur dengan menunjukkan jika peningkatan DNI pada hari ke-3, syok septik, penggunaan terapi norepinefrin,

Studi Literatur		Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan / Hasil Temuan
Penulis	Judul					
					laboratorium diukur sebelum operasi, 12-36 jam setelah pembedahan, dan 60-84 jam setelah pembedahan.	dan peningkatan rasio normalisasi internasional pada hari ke-3 memiliki dampak terbesar pada prediksi model kematian.

2.2 Landasan Teori

2.2.1 Gim Indi

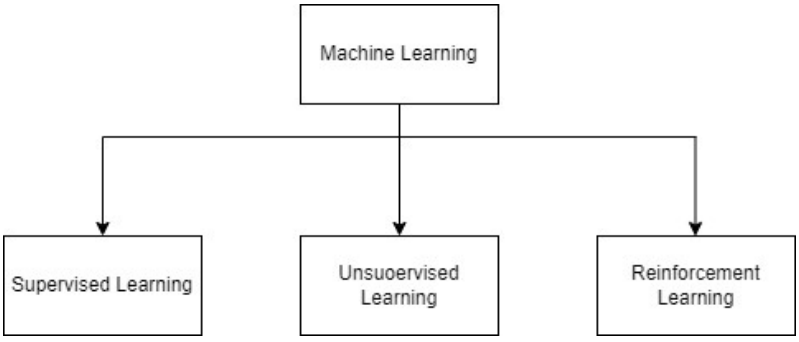
Belum ada studi yang definisi jelas tentang gim indi, namun kata "indi" dalam industri gim dunia sudah menjadi penanda umum untuk jenis-jenis gim digital dan pengembang tertentu. *Independent Games Festival,* pertemuan tahunan terbesar di sektor gim indi, menyatakan aturan resmi dari kandidat gim indi terbaik tahun ini haruslah gim yang yang dibuat dengan "semangat indi" oleh pengembang gim independen [19]. Jurnal [20], menganggap gim indi sebagai ideologi yang mengadu domba perbedaan pendapat secara politis-ekonomi dengan penerbit gim besar. Jurnal [20] juga menyatakan bahwa yang Namanya gim indi adalah gim yang diproduksi dengan biaya yang murah. Karena gim 2 dimensi lebih murah untuk diproduksi daripada 3 dimensi, maka gim indi lebih sering berupa ke 2 dimensi. Pada *steam platform* [21] sendiri sudah disediakan tag khusus yang menandakan jika suatu gim termasuk gim indi. Tampilan *tag* gim indi pada web *steam* ditunjukan pada Gambar 2.1 :



Gambar 2.1 Tampilan *steam* [21]

2.2.2 *Machine Learning*

Menurut Arthur Samuel, *machine learning* didefinisikan sebagai bidang studi yang memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit [22]. Menurut Taeho Jo [23], *machine learning* adalah suatu program komputer yang memiliki kemampuan memecahkan suatu permasalahan berdasarkan data sebelumnya. Fokus utama dari *Machine Learning* adalah membangun sebuah program komputer yang mampu mempelajari data dan mampu membuat model yang siap digunakan untuk memecahkan sebuah kasus atau masalah yang ada [24]. *Machine learning* terdiri dari tiga bagian yang terlihat pada Gambar 2.2, yaitu *supervised learning, unsupervised learning,* dan *reinforcement learning*.



Gambar 2.2 *Machine Learning*

Supervised Learning merupakan teknik pembelajaran yang bertujuan meminimalisir selisih target hasil prediksi dengan target data aktual [23]. *KNN, naïve Bayes,* dan *decission tree* merupakan bagian dari algoritma *supervised learning.* Contoh penerapan *supervised learning* adalah regresi dan klasifikasi

Klasifikasi merupakan proses pengelompokan objek yang mempunyai kesamaan karakteristik untuk beberapa kelas [23]. Proses klasifikasi dengan menerapkan *machine learning* dimulai dari inputan data berlabel yang akan diberikan pembelajaran menggunakan mesin akan mendeteksi data baru ke dalam label atau kelas tertentu

Unsupervised learning merupakan teknik pembelajaran pada machine learning yang digunakan untuk mengolah data tidak berlabel dan

bergantung pada kesamaan karakteristik pada data [23]. Contoh penerapannya *unsupervised learning* adalah clustering. Teknik pembelajaran yang terakhir adalah *reinforcement learning* yang prosesnya tanpa menggunakan inputan data, pada pembelajaran ini terdapat *agent* yang dapat mempelajari, memilih dan melakukan tindakan dan mendapatkan hasil untuk proses pembelajaran dan tindakan selanjutnya [23]

2.2.3 *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) merupakan bagian penting dalam analisis pada domain *data science*. *EDA* adalah pendekatan untuk menganalisis data yang bertujuan untuk membuat ringkasan dan memahami data dengan lebih baik. Tujuan utama *EDA* adalah untuk mengidentifikasi distribusi data, *outliers*, dan tren yang dapat membantu dalam pengujian hipotesis yang telah dibuat[25]

Exploratory Data Analysis (EDA) mencakup penggunaan grafik, statistik deskriptif, dan berbagai metode analisis lainnya untuk mengungkap wawasan yang membantu dalam pemahaman dataset secara lebih mendalam. EDA memiliki peran penting dalam mengidentifikasi dan mempersiapkan data sebelum digunakan dalam model analisis.

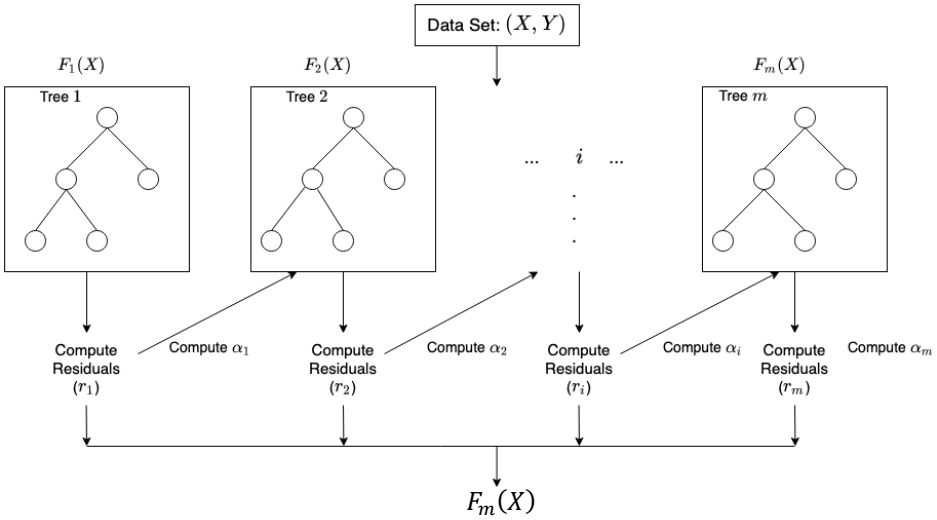
2.2.4 *CatBoost (Categorical Boosting)*

CatBoost (Categorical Boosting) adalah salah satu *gradient boosting algorithm* yang dibuat menggunakan konsep *decision tree.* Sama dengan *gradient boosting* lainnya seperti *XGBoost*, dan *LightGBM*, *CatBoost* juga mampu melakukan regresi dan klasifikasi. Cara kerja yang mendasari *CatBoost* sama dengan algoritma *boosting* lainnya, yaitu mempelajari banyak pengklasifikasi yang lemah dan menggabungkannya menjadi pengklasifikasi yang lebih kuat.

CatBoost mengimplementasikan algoritma *Gradient Boosting Decision Tree (GBDT*) konvensional dengan penambahan dua algoritmik penting yaitu Implementasi memerintahkan meningkatkan, alternatif

permutasi-driven untuk algoritma klasik, dan Algoritma inovatif untuk memproses fitur kategoris. Kedua teknik diciptakan untuk melawan pergeseran prediksi yang disebabkan *data leakage* yang ada di semua implementasi algoritma peningkatan gradien yang ada saat ini.

Gradient Boosting Decision Tree (GBDT) sendiri adalah salah satu teknik dalam *machine learning* yang digunakan untuk membangun model prediktif yang kuat dengan menggabungkan beberapa model lemah atau base learner *decision tree* secara adaptif. Teknik ini termasuk dalam kategori ensemble learning, yang berarti model akhir dibentuk dengan menggabungkan beberapa model individual dengan tujuan untuk mengoptimalkan *loss function* [26]. Proses iterasi pembuatan model dilakukan dengan hasil *residual* dari model sebelumnya. Ilustrasi cara kerja *Gradien Boosting* sendiri ditunjukan pada Gambar 2.3 berikut :



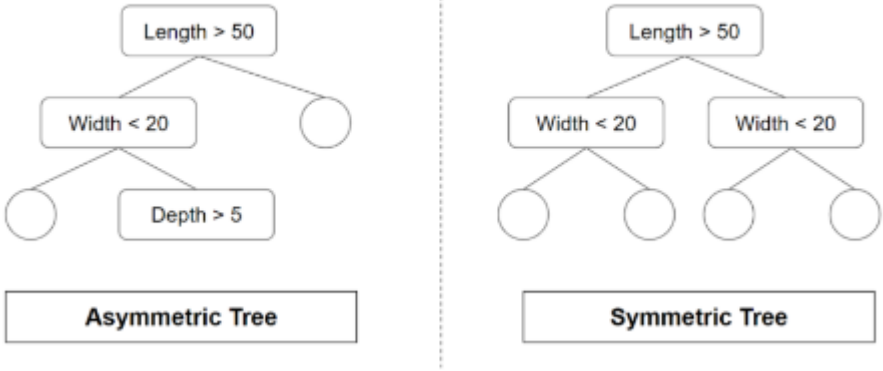
Gambar 2.3 Cara Kerja *Gradien Boosting* [27]

Pada Gambar 2.3, $F_m(X)$ merupakan prediksi model final, prediksi final didapatkan menggunakan persamaan 2.1:

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) \tag{2.1}$$

α_i adalah parameter *learning_rate,* dan r_{m-i} adalah residual dari model sebelumnya dan h_i model residual ke i.

Berbeda dengan jenis *Gradient boosting* lain seperti *XGBoost* dan *LightGBM* yang menggunakan pohon yang tidak simetris, *CatBoost* menggunakan pohon simetris untuk kecepatan prediksi yang baik. Hal tersebut membuatnya memiliki beberapa kelebihan yang sangat berguna seperti kemampuan mengurangi kemungkinan terjadinya *overfitting*, waktu latih yang lebih singkat, dan peningkatan efisiensi pemakaian *CPU* [28]. Berbagai keuntungan diatas terjadi karena *CatBoost* menggunakan kondisi yang sama di setiap nodenya dalam membuat pohon keputusan, sehingga setiap bagian pohon keputusan yang dibuat menggunakan algoritma lebih efisien dalam mengurangi error dibandingkan dengan bagian sebelumnya [7]. ditunjukan pada Gambar 2.4 :



Gambar 2.4 Perbedaan pohon simetris dan tidak simetris [7]

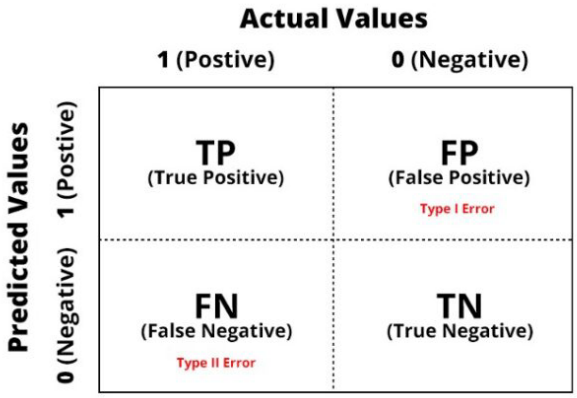
Selain kelebihan yang sudah disebutkan sebelumnya, kelebihan utama dari algoritma ini adalah kemampuannya dalam mengatasi berbagai tipe data secara otomatis baik data numerik, teks, dan khususnya kategorikal [7]. Hanya dengan mendefinisikan fitur apa saja yang bertipe kategorikal, *CatBoost* secara *default* akan memberikan perlakuan untuk data pada fitur kategorikal tersebut menjadi numerik, dengan menggunakan persamaan 2.2:

$$ctr_i = \frac{countInClass+prior}{totalCount+1} \quad (2.2)$$

ctr_i adalah data ke i pada fitur kategorikal. $countInClass$ adalah berapa kali nilai label melebihi i untuk objek dengan nilai fitur kategorikal saat ini. Fungsi ini hanya menghitung objek yang telah memiliki nilai ini, perhitungan dilakukan berdasarkan urutan objek setelah pengocokan. $totalCount$ adalah jumlah total objek yang memiliki nilai fitur yang cocok dengan nilai fitur saat ini. $prior$ adalah sebuah angka konstanta yang ditentukan oleh parameter awal.

2.2.5 Matriks Konfusi

Menilai performa suatu sistem klasifikasi menjadi hal yang penting karena hal tersebut dapat mengindikasikan sejauh mana kemampuan sistem dalam mengklasifikasikan data [29]. Salah satu metode yang umum digunakan untuk mengukur kinerja klasifikasi adalah Matriks Konfusi. Matriks Konfusi adalah tabel yang mencatat hasil dari proses klasifikasi. Berikut adalah contoh tabel Matriks Konfusi untuk sistem klasifikasi dengan dua kelas yang ditunjukan pada Gambar 2.5:



Gambar 2.5 Matriks Konfusi [30]

Akurasi, Presisi, *Recall,* dan *F1-score* merupakan beberapa matrik evaluasi yang digunakan untuk mengukur keakuratan model membedakan kelas dalam melakukan prediksi atau klasifikasi. Merupakan bagian dari

matriks konfusi yang secara perhitungan ditunjukkan persamaan 2.3, 2.4, 2.5, 2.6 :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

$$Presisi = \frac{TP}{TP + FP} \quad (2.4)$$

$$Recalll = \frac{TP}{TP + FN} \quad (2.5)$$

$$F1 - score = 2 \times \frac{Presisi + Recall}{Presisi + Recall} \quad (2.6)$$

dimana nilai TP, TN, FP, dan FN pada persamaan 2.3, 2.4, 2.5, 2.6 adalah :

a. TP (*True Positive*) merupakan jumlah observasi yang benar – benar termasuk ke dalam kelas positif dan diprediksi benar oleh model.

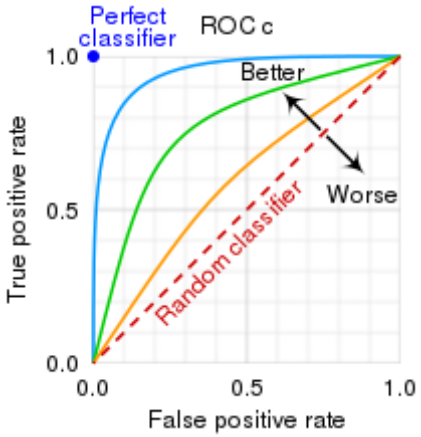
b. TN (*True Negative*) merupakan jumlah observasi yang benar – benar termasuk ke dalam kelas negatif dan diprediksi dengan benar oleh model.

c. FP (*False Positive*) merupakan jumlah observasi yang diprediksi sebagai bagian dari kelas positif oleh model, tetapi sebenarnya observasi tersebut termasuk dalam kelas negatif.

d. FN (*False Negative*) merupakan jumlah observasi yang diprediksi sebagai bagian dari kelas negatif oleh model, tetapi sebenarnya observasi tersebut termasuk dalam kelas positif.

2.2.6 Kurva *AUC (Area Under the Curve)*

Kurva *AUC* (Area Under the Curve) adalah alat evaluasi yang digunakan dalam pengukuran kinerja model klasifikasi. Kurva *AUC* menggambarkan hubungan antara tingkat *true positive* (*sensitivity*) dan tingkat *false positive (1-specificity)* pada berbagai *threshold* pengklasifikasi. Nilai *AUC* memiliki nilai berkisar antara 0 hingga 1. Semakin tinggi nilai *AUC*, semakin baik model klasifikasi dalam membedakan antara kelas positif dan negatif. Dengan menganalisis kurva *AUC* [31], maka dapat diperoleh wawasan tentang kinerja model dalam membedakan antar kelas.



Gambar 2.6 Kurva *AUC* [32]

Pada Gambar 2.6 menunjukan nilai *AUC* pada model. Apabila kurva ke diagonal semakin mendekati 45 derajat, maka semakin tidak akurat model tersebut. Namun jika kurva diagonal semakin mendekati 90 derajat, maka semakin akurat model tersebut [33].

2.2.7 *Shapley Additive Explanations (SHAP)*

Pendekatan *Shapley Additive Explanations (SHAP)* adalah metode yang memungkinkan untuk menginterpretasikan model prediksi pembelajaran mesin yang bersifat *blackbox* atau sulit dipahami. Tujuan dari *SHAP* adalah untuk menjelaskan prediksi dari fitur x dengan menghitung kontribusi dari setiap fitur terhadap prediksi. Metode *SHAP* menghitung nilai *Shapley* mirip dengan teori permainan koalisi. Nilai-nilai fitur dari sebuah contoh data bertindak sebagai pemain dalam sebuah koalisi. Nilai *Shapley* memberitahu kita cara mendistribusikan kontribusi prediksi secara adil di antara fitur-fitur. Seorang pemain dapat berupa nilai fitur individual [9]. Salah satu inovasi yang dibawa oleh pendekatan *Shapley Additive Explanations (SHAP)* adalah penjelasan nilai *Shapley* direpresentasikan sebagai metode atribusi fitur aditif, sebuah model linier. Pandangan tersebut menghubungkan nilai *LIME* dan *Shapley* [34]. pendekatan *Shapley Additive*

Explanations (SHAP) menggunakan rumus yang ditunjukan pada Persamaan 2.7:

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \frac{|S|!(p-|S|-1)!}{p!} \times [Val(S \cup \{i\}) - Val(S)] \qquad (2.7)$$

ϕ_i adalah nilai *shapley* dari instansi suatu fitur terhadap hasil prediksi. $Val(S)$ adalah output dari model ML yang akan dijelaskan menggunakan satu set fitur S, dan p adalah jumlah keseluruhan dari semua fitur.

Kontribusi akhir atau nilai *Shapley* dari fitur $i(\phi i)$ didefinisikan juga sebagai rata-rata dari *marginal* kontribusinya di seluruh permutasi yang mungkin dari set fitur . Proses ini menggunakan perhitungan nilai *Shapley* dengan mengukur bagaimana kontribusi setiap fitur berubah saat fitur-fitur lain berubah.