BAB II

TINJAUAN PUSTAKA

2.1 Justifikasi Solusi

Proses clustering pelanggan yang saat ini diterapkan di PT. XYZ dilakukan secara manual melalui tahapan ekstraksi data, pembersihan, pengelompokan berbasis aturan bisnis sederhana, hingga penyusunan laporan setiap bulan. Seluruh tahapan ini memakan waktu sekitar 1–3 hari kerja berdasarkan kuesioner yang diberikan kepada tim di *e-commerce* PT. XYZ, menjadikan perusahaan kurang responsif dalam menghadapi perubahan pasar e-commerce yang berlangsung cepat. Menghadapi tantangan tersebut, dibutuhkan metode yang mampu memangkas waktu proses clustering. Solusi ini harus menghasilkan kelompok pelanggan yang konsisten, *reproducible*, dan dapat menangkap pola-pola kompleks dalam data campuran numerik dan kategorikal secara efisien. Selain itu, solusi yang *scalable* sangat diperlukan agar proses tetap dapat berjalan optimal seiring bertambahnya volume data di masa mendatang. Adanya mekanisme validasi objektif juga harus diterapkan agar kualitas hasil clustering dapat diukur secara konsisten dan dapat dipertanggungjawabkan sebagai dasar pengambilan keputusan bisnis yang strategis.

2.1.1 *Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values* [6]

Penelitian tentang clustering untuk data campuran terus berkembang dan relevan seiring semakin kompleksnya data bisnis modern, termasuk pada sektor *e-commerce*. Salah satu kemajuan paling signifikan dalam bidang ini adalah diperkenalkannya algoritma K-Prototypes oleh Huang. Algoritma ini menjadi solusi matematis yang efektif untuk menangani data yang terdiri dari gabungan fitur numerik dan kategorikal secara bersamaan, sehingga tidak menghadapi hambatan utama yang ditemukan pada K-Means yang biasa digunakan khusus data numerik ataupun K-Modes yang biasa digunakan khusus data kategorikal. Dengan demikian, K-Prototypes telah menjadi acuan utama pada banyak

penelitian dan aplikasi *customer segmentation* dalam data asli dunia industri yang bersifat campuran.

Walaupun memiliki pondasi teoritis yang kuat, penerapan K-Prototypes masih memiliki keterbatasan. Dua masalah utama yang jarang teratasi dalam praktik adalah ketergantungan pada inisialisasi awal *cluster center* yang dapat menyebabkan hasil akhir berbeda pada setiap eksekusi. Metode yang dikembangkan dalam penelitian ini sangat relevan untuk diterapkan pada masalah clustering pelanggan *e-commerce* karena mampu mengakomodasi kompleksitas data pelanggan yang terdiri atas berbagai tipe fitur. Berdasarkan review literatur di atas, K-Prototypes menjadi pilihan algoritma yang paling sesuai untuk kebutuhan clustering pelanggan PT. XYZ karena dapat menangani data campuran secara *native* dan telah terbukti efektif dalam berbagai studi kasus *e-commerce*.

2.1.2 *Bagging to improve the accuracy of a clustering procedure* [4]

Penelitian yang dilakukan oleh Dudoit dan Fridlyand [4] menunjukkan bahwa metode *ensemble learning*, khususnya pendekatan *bagging*, mampu meningkatkan akurasi *clustering* secara signifikan pada data numerik, yaitu mencapai peningkatan 15-25% pada kasus *clustering* data *gene expression*. Melalui eksperimen yang komprehensif, mereka membuktikan bahwa dengan menggabungkan hasil *clustering* dari beberapa *bootstrap sample*, variansi hasil dapat dikurangi dan solusi *clustering* menjadi lebih stabil serta *robust* terhadap *noise data*. Namun, cakupan penelitian mereka masih terbatas pada data dengan tipe numerik saja dan belum menjangkau pengembangan ensemble learning untuk algoritma clustering yang secara khusus dirancang menangani data campuran seperti K-Prototypes.

Hal ini mengindikasikan adanya kesempatan untuk mengintegrasikan metode *ensemble learning* dan K-Prototypes *clustering* khususnya pada kebutuhan aplikasi modern seperti *e-commerc*e yang data pelanggannya hampir selalu berupa kombinasi fitur numerik dan kategorikal. Penerapan prinsip-prinsip *ensemble learning* yang telah terbukti efektif untuk numerik seperti bagging pada K-Means atau *hierarchical clustering*, sangat potensial untuk diadaptasi dan

dioptimalkan dalam konteks K-Prototypes yang menangani tipe data campuran. Pendekatan tersebut diharapkan dapat memberikan hasil *clustering* yang lebih konsisten, stabil, dan berkualitas tinggi, sehingga mendukung kebutuhan clustering pelanggan *e-commerce* berbasis data yang kompleks dan nyata.

2.1.3 *Intuitive-K-prototypes: A mixed data clustering algorithm with intuitionistic distribution centroid* [10]

Penelitian terbaru oleh Wang et al. [10] menghadirkan inovasi pada algoritma K-Prototypes melalui pengembangan Intuitive K-Prototypes. Algoritma ini menunjukkan peningkatan performa dalam mengelompokkan data campuran pada berbagai dataset UCI, dengan memfokuskan pada optimalisasi representasi *prototype* melalui strategi heuristik dan pembobotan atribut. Namun, meskipun telah sukses meningkatkan akurasi *clustering* secara statistik, metode ini belum menggabungkan pendekatan *ensemble* untuk menambah robustness dan stabilitas hasil clustering secara konsisten di lingkungan data yang bervariasi.

Berdasarkan hal tersebut, penelitian ini fokus pada pengembangan solusi yang lebih relevan untuk clustering pelanggan dengan data campuran asli, yaitu dengan menerapkan ensemble learning pada K-Prototypes. Pendekatan ini diharapkan tidak hanya meningkatkan akurasi dan kualitas clustering seperti yang telah dicapai pada penelitian-penelitian sebelumnya, tetapi juga memberikan stabilitas dan interpretabilitas yang lebih baik, tanpa harus melakukan transformasi data yang dapat mengorbankan aspek interpretasi penting dari data pelanggan *e-commerce*.

2.1.4 *Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost over Random Forest Algorithm* [7]

Penelitian Customer Churn Prediction [7] menunjukkan bahwa kombinasi algoritma *clustering* K-means dengan AdaBoost dapat memberikan hasil yang sangat baik dalam memprediksi potensi churn customer pada data e-commerce, tercermin dari akurasi klasifikasi yang tinggi, yaitu 95,55%. Temuan ini memperkuat peran AdaBoost sebagai classifier yang sangat efektif untuk memvalidasi kualitas clustering atau cluster yang terbentuk, karena hasil

clustering yang baik akan menghasilkan cluster yang jelas dan mudah dibedakan oleh classifier. Di sisi lain, penelitian tersebut juga mengungkapkan bahwa dengan melakukan clustering sebelum prediksi, model dapat mencapai akurasi yang lebih tinggi dibandingkan metode prediksi langsung.

Oleh karena itu, untuk konteks clustering pelanggan e-commerce yang menggunakan dataset dengan karakteristik data campuran, diperlukan strategi yang mampu mengatasi kelemahan ini. Penelitian ini membawa inovasi tersebut dengan mengadaptasi konsep validasi menggunakan AdaBoost—yang terbukti efektif di penelitian sebelumnya—untuk diaplikasikan pada hasil clustering dari K-Prototypes (bukan K-means). Dengan cara ini, kualitas cluster pada data campuran dapat divalidasi secara lebih objektif dan relevan, sehingga solusi clustering yang diperoleh tidak hanya lebih akurat, tetapi juga benar-benar sesuai dengan kebutuhan bisnis digital masa kini.

2.1.5 *Customer Segmentation and Classification Using K-Modes Clustering with Ensemble Learning* [8]

Niloy et al. [8] mengembangkan pendekatan *customer segmentation* menggunakan kombinasi *K-Modes clustering* dan AdaBoost *ensemble learning* pada data pelanggan asuransi. Penelitian ini menggunakan dataset sebanyak 53.503 *records* dari Kaggle yang mencakup informasi demografis, transaksional, *behavioral attributes*, dan *policy particulars* pelanggan dari Januari 2018 hingga Desember 2023. Algoritma K-Modes dipilih sebagai metode *clustering* utama karena kemampuannya menangani data kategorikal secara *native* dengan menggunakan *mode* sebagai *cluster centers* dan *Hamming distance* sebagai *dissimilarity measure*.

Metodologi penelitian tersebut melibatkan lima tahapan utama: *preprocessing data* (*handling missing values*, *outlier treatment*, dan *discretization*), clustering menggunakan K-Modes dengan jumlah *cluster* k = 5, evaluasi fitur menggunakan Gain Ratio untuk seleksi 15 fitur terbaik dari 19 fitur yang tersedia, klasifikasi menggunakan AdaBoost (dengan n_estimators=300), dan evaluasi model menggunakan confusion matrix. Hasil eksperimen

menunjukkan bahwa kombinasi K-Modes clustering dengan AdaBoost classification mencapai akurasi yang sangat tinggi, yaitu 94.68%, jauh melampaui metode baseline seperti Naive Bayes (26.98%), Decision Tree (26.82%), Random Forest (26.11%), dan Bagging (24.96%) ketika menggunakan cluster yang dihasilkan K-Modes dan selected features.

Relevansi penelitian Niloy et al. [8] terhadap penelitian ini terletak pada beberapa aspek. Pertama, penggunaan AdaBoost sebagai *validation mechanism* untuk kualitas cluster sejalan dengan pendekatan yang diterapkan dalam penelitian ini, di mana AdaBoost digunakan untuk mengukur *cluster separability* secara kuantitatif. Kedua, fokus pada data kategorikal atau data campuran menunjukkan pentingnya pemilihan algoritma clustering yang sesuai dengan karakteristik data. Ketiga, ensemble approach yang diterapkan Niloy et al. melalui AdaBoost untuk klasifikasi membuktikan bahwa *ensemble methods* dapat meningkatkan *robustness* dan akurasi hasil clustering.

2.2 Tinjauan Teori

Bagian ini menjelaskan teori-teori dan konsep-konsep fundamental yang menjadi dasar implementasi solusi untuk PT. XYZ. Pemahaman teori ini penting untuk memastikan solusi yang dikembangkan memiliki landasan ilmiah yang kuat dan dapat dipertanggungjawabkan secara akademis maupun praktis.

2.2.1 Customer Segmentation in E-Commerce

Customer segmentation adalah salah satu teknik *marketing* yang sangat krusial dalam upaya meningkatkan efektivitas strategi bisnis, khususnya di ranah e-commerce. Dengan melakukan clustering, perusahaan dapat membagi basis pelanggan menjadi kelompok-kelompok homogen yang memiliki karakteristik atau kebutuhan serupa. Segmentasi yang tepat memungkinkan perusahaan menjalankan kampanye pemasaran yang lebih terarah, sehingga pesan yang disampaikan menjadi lebih relevan dan ROI (Return on Investment) dari aktivitas pemasaran meningkat. Selain itu, clustering juga memudahkan implementasi rekomendasi produk yang dipersonalisasi, yang terbukti dapat meningkatkan tingkat konversi dan kepuasan pelanggan. Di samping itu, strategi clustering

memungkinkan perusahaan untuk mengoptimalkan skema harga berdasar karakteristik willingness-to-pay masing-masing segmen, sekaligus memperkuat loyalitas pelanggan melalui program-program yang disesuaikan secara khusus untuk kebutuhan tiap kelompok.

Seiring perkembangan teknologi *artificial intelligence* dan *machine learning*, penelitian-penelitian baru dalam bidang customer segmentation terus menampilkan kemajuan signifikan. Contohnya, Wang et al. [9] berhasil memperlihatkan bahwa integrasi metode *reinforcement learning* dengan K-means clustering dapat menghasilkan clustering pelanggan dengan akurasi klasifikasi lebih dari 95%. Selain itu, teknik dimensionality reduction seperti *Principal Component Analysis* (PCA) juga dilaporkan mampu meningkatkan kualitas clustering jika diimplementasikan sebelum proses clustering. Dalam hal validasi hasil clustering, penggunaan ensemble classifier seperti AdaBoost maupun *Artificial Neural Network* (ANN) semakin banyak digunakan karena bisa memberikan pengukuran obyektif mengenai seberapa mudah segmen-segmen yang dihasilkan bisa dipisahkan dan diidentifikasi dalam konteks penerapan bisnis nyata.

Namun demikian, mayoritas penelitian terdahulu umumnya hanya berfokus pada data berjenis homogen, baik numerik saja (seperti pada K-means) atau kategorikal saja (K-modes). Padahal, data asli pelanggan e-commerce sangat sering bersifat campuran—mengandung variabel numerik seperti usia, frekuensi transaksi, maupun variabel kategorikal seperti gender atau kategori produk. Pada kasus seperti ini, K-Prototypes menjadi pendekatan yang lebih relevan karena dapat mengakomodasi kedua tipe data secara optimal tanpa perlu transformasi yang menurunkan kualitas atau interpretasi data. Sayangnya, hingga saat ini, penelitian terkait penerapan K-Prototypes terutama dalam konteks ensemble learning dan perbaikan robustness melalui metode seperti bootstrap ensemble masih sangat terbatas. Hal ini membentuk landasan sekaligus urgensi bagi penelitian ini, yaitu mengembangkan framework ensemble K-Prototypes yang

dapat meningkatkan kualitas clustering pelanggan e-commerce dengan data campuran secara signifikan.

2.2.2 K-Prototypes Clustering Algorithm

K-Prototypes adalah algoritma clustering yang dikembangkan oleh Huang pada tahun 1997 [6] sebagai solusi atas keterbatasan algoritma clustering populer sebelumnya, yaitu K-Means dan K-Modes, yang hanya mampu menangani satu jenis data saja. K-Prototypes secara inovatif menggabungkan prinsip dasar K-Means yang efektif untuk data numerik dan K-Modes yang optimal untuk data kategorikal, sehingga dapat memproses dataset dengan tipe data campuran secara langsung. Algoritma ini menggunakan distance metric gabungan, yaitu kombinasi linear antara Euclidean distance untuk fitur numerik seperti usia dan matching dissimilarity untuk fitur kategorikal seperti gender serta kategori produk. Di dalam perhitungannya, terdapat parameter gamma (γ) sebagai penimbang yang mengatur kontribusi relatif kedua tipe fitur tersebut terhadap total jarak antar data dan pusat cluster. Untuk menyeimbangkan kedua kontribusi ini, Huang [6] menyarankan nilai gamma sebesar setengah kali standar deviasi fitur numerik, sehingga kombinasi jarak numerik dan kategorikal dapat menjadi adil dan proporsional secara matematis.

Pusat cluster pada K-Prototypes dinamakan prototype, yakni kombinasi rata-rata (mean) untuk atribut numerik dan modus (mode) untuk atribut kategorikal. Dengan pendekatan ini, K-Prototypes memiliki sejumlah keunggulan. Pertama, algoritma ini tidak mengalami information loss karena data numerik, seperti usia, tetap dipertahankan sebagai nilai kontinu tanpa perlu diskretisasi atau binning yang dapat mereduksi detail informasi. Kedua, K-Prototypes menggunakan metrik jarak yang paling sesuai bagi setiap tipe data, menjadikannya solusi secara matematis dan interpretatif lebih optimal untuk data asli e-commerce yang memang bersifat campuran. Ketiga, bukti menunjukkan bahwa hasil clustering menggunakan K-Prototypes lebih baik daripada pendekatan berbasis binning seperti K-Modes, terutama dilihat dari berbagai metric kualitas cluster [6].

Meskipun demikian, K-Prototypes tetap memiliki tantangan utama, yaitu sensitivitas terhadap pemilihan inisialisasi pusat cluster. Sama seperti K-Means dan K-Modes, proses clustering K-Prototypes bersifat iteratif dan hasil akhirnya dapat sangat tergantung pada posisi awal cluster center, sehingga dapat menghasilkan variasi hasil pada eksekusi yang berbeda-beda. Permasalahan ini menjadi penting, khususnya untuk aplikasi production yang membutuhkan hasil clustering yang konsisten, stabil, dan dapat direplikasi. Untuk mengatasi tantangan tersebut, pada penelitian ini diimplementasikan pendekatan ensemble berbasis bootstrap dan meta-clustering consensus, sehingga variabilitas yang diakibatkan oleh random initialization dapat diminimalisir. Perkembangan terbaru, seperti yang dikembangkan oleh Wang et al. [10] dalam Intuitive-K-Prototypes, juga telah memperkenalkan beberapa perbaikan, termasuk centroid berbasis distribusi intuitionistic untuk data kategorikal, selection heuristik untuk initial prototype, dan pembobotan atribut berdasar intra-cluster complexity. Namun, kontribusi penelitian ini berfokus pada penambahan ensemble strategy yang terbukti dapat lebih meningkatkan robustness dan stabilitas hasil clustering customer pada data e-commerce dengan tipe campuran.

2.2.3 Bootstrap Ensemble Principles

Bootstrap aggregating, yang sering disebut bagging, merupakan salah satu teknik ensemble learning yang bertujuan utama untuk meningkatkan akurasi serta stabilitas model prediksi melalui proses pembentukan berbagai model pembelajar (learners) yang berbeda-beda. Teknik ini dilakukan dengan cara melakukan pelatihan model pada sekumpulan bootstrap samples, yaitu subsets data yang diperoleh melalui proses pengambilan acak dengan pengembalian dari dataset asli. Hasil dari berbagai model yang dilatih pada subsets yang berbeda tersebut kemudian digabungkan (agregasi) baik dalam konteks prediksi maupun hasil clustering, sehingga memberikan hasil akhir yang lebih robust dan stabil. Walaupun bagging awalnya dikembangkan untuk supervised learning seperti klasifikasi dan regresi [3], penelitian-penelitian setelahnya menunjukkan bahwa prinsip ini juga sangat bermanfaat dan dapat diadaptasi dalam unsupervised learning, khususnya pada tugas clustering.

Implementasi bagging pada masalah clustering, sebagaimana dijabarkan oleh Dudoit dan Fridlyand [4], melibatkan beberapa tahapan yang sistematis. Pertama, proses ini dimulai dari pembuatan multiple bootstrap samples dengan mengambil data secara acak dari dataset asli dimana satu data dapat terambil lebih dari satu kali atau bahkan tidak terambil sama sekali dalam satu bootstrap. Tahap berikutnya, setiap sampel bootstrap tersebut dicluster secara independen menggunakan algoritma clustering yang konsisten (misalnya K-Prototypes) dengan inisialisasi berbeda, sehingga menghasilkan berbagai solusi clustering yang mengungkap struktur yang bisa jadi berbeda dari data. Pada tahap akhir, hasil clustering dari semua bootstrap samples tersebut diagregasi melalui mekanisme konsensus, seperti voting atau co-association matrix, sehingga data points yang sering kali tercluster bersama akan tetap berada dalam cluster yang sama pada hasil akhir, sementara data yang posisinya ambigu akan diputuskan berdasarkan kemunculan mayoritas atau kemiripan pola.

Pendekatan ensemble berbasis bootstrap ini terbukti mampu memberikan berbagai keuntungan nyata, salah satunya adalah mengurangi variansi hasil clustering serta membuat hasil cluster menjadi lebih stabil dan dapat direproduksi—khususnya penting untuk kebutuhan production di dunia nyata. Dudoit dan Fridlyand [4] melalui eksperimen menyeluruh pada beragam dataset, membuktikan bahwa bagging dapat meningkatkan akurasi clustering sebesar 15-25% jika dibandingkan dengan metode clustering single-run. Peningkatan ini terukur baik melalui evaluasi eksternal dengan membandingkan hasil dengan ground truth, maupun dengan evaluasi internal menggunakan metrik kualitas cluster. Selain itu, keuntungan bagging akan semakin nyata untuk data dengan noise tinggi, dimensi besar, atau struktur cluster yang kompleks—karakteristik yang umum ditemukan pada data pelanggan e-commerce. Walaupun penelitian ini awalnya masih berfokus pada data numerik dan hierarchical clustering, prinsip dasarnya bersifat umum dan sangat potensial untuk dikembangkan lebih lanjut, termasuk dalam implementasi K-Prototypes untuk data campuran seperti pada penelitian ini.

2.2.4 Meta-Clustering dan *Cluster Ensemble*

Meta-clustering merupakan metode ensemble clustering lanjutan yang dirancang untuk mengatasi kekurangan metode bootstrap aggregation konvensional dengan menggunakan kerangka dua tahap yang bersifat hierarchical. Tidak seperti bagging standar yang hanya menggabungkan hasil clustering berdasarkan voting sederhana atau rata-rata, meta-clustering menerapkan tahap tambahan, yakni melakukan clustering terhadap hasil clustering (local cluster centers) dari beberapa subset data yang dihasilkan melalui bootstrapping atau partitioning. Tahap pertama dimulai dengan melakukan clustering pada beberapa subset data sehingga setiap subset menghasilkan pusat cluster lokal (local cluster centers) yang merepresentasikan pola struktur lokal dalam subset tersebut. Pada tahap kedua, seluruh pusat cluster lokal dari semua subset tersebut kemudian digabungkan dan dilakukan proses clustering ulang guna menemukan meta-centers yang merepresentasikan pola konsensus secara global. Selanjutnya, meta-centers inilah yang digunakan sebagai pusat cluster final, dan seluruh data pada dataset asli di-assign ke meta-center terdekat untuk menentukan keanggotaan cluster akhir. Framework dua tahap ini menawarkan beberapa keunggulan penting, seperti mengeksplorasi solution space lebih luas, mereduksi sensitivitas terhadap inisialisasi pusat cluster melalui proses averaging, serta meningkatkan robustnes terhadap noise dan outlier.

Keberhasilan meta-clustering ini telah didemonstrasikan secara spesifik oleh Rashedi et al. [1] melalui pengembangan algoritma K-Metamodes yang menerapkan prinsip meta-clustering untuk data kategorikal. K-Metamodes menggunakan pendekatan boosting-based clusterer ensemble dengan berbagai weighting scheme pada data points, mirip prinsip dasar boosting pada supervised learning. Hasil-hasil clustering dari level bawah kemudian diagregasi secara hierarchical ke level lebih tinggi menjadi meta-modes yang mencerminkan solusi clustering multi-resolusi. Eksperimen yang mereka lakukan menunjukkan bahwa K-Metamodes mampu memberikan kualitas clustering jauh lebih baik dibandingkan metode single clustering, khususnya pada data kategorik yang kompleks dan banyak mengandung ketidakjelasan boundary antar cluster.

Keberhasilan K-Metamodes tersebut menjadi validasi penting bahwa meta-clustering dapat meningkatkan kualitas clustering, tidak hanya untuk data numerik, tetapi juga untuk data non-numerik, dan menjadi inspirasi utama bagi pengembangan penelitian ini.

Meskipun telah terbukti efektif untuk data kategorikal, konsep meta-clustering belum banyak dieksplorasi untuk data campuran seperti pada e-commerce. Dalam penelitian ini, prinsip meta-clustering diterapkan pada algoritma K-Prototypes yang mampu menangani data numerik dan kategorikal secara bersamaan, tentunya dengan beberapa penyesuaian penting—di antaranya, teknik sampling harus tetap menjaga distribusi fitur numerik dan kategorikal, distance metric meta-clustering tetap memanfaatkan formula jarak khas K-Prototypes, dan hasil meta-cluster harus tetap secara semantik representatif sebagai mean untuk data numerik dan mode untuk data kategorikal. Melalui adaptasi ini, penelitian ini bertujuan memperlihatkan bahwa keunggulan meta-clustering untuk data homogen juga dapat berlaku, bahkan memberikan manfaat lebih, pada clustering data campuran yang jauh lebih kompleks dalam clustering pelanggan e-commerce.

2.2.5 Adaboost Classification

AdaBoost atau Adaptive Boosting merupakan salah satu algoritma ensemble learning pada supervised classification yang sangat berpengaruh, yang dikembangkan oleh Freund dan Schapire [5]. Konsep dasar dari algoritma ini adalah melakukan proses pembelajaran secara berurutan dengan menggabungkan sejumlah weak classifiers—biasanya berupa tree sederhana atau decision stump—menjadi sebuah strong classifier melalui mekanisme voting berbobot. Proses ini diawali dengan pemberian bobot yang sama pada setiap data training. Selanjutnya, weak classifier dilatih pada data berbobot tersebut. Ketika ada data yang salah terklasifikasi, bobot data tersebut dinaikkan sehingga pada iterasi berikutnya, classifier berikutnya akan lebih fokus pada data sulit tersebut. Proses pembentukan classifier ini diulang selama sejumlah T iterasi, sehingga dihasilkan ensemble dari T weak classifiers dengan spesialisasi berbeda-beda. Pada akhirnya,

seluruh voting hasil dari weak classifiers digabungkan secara bobot (weighted majority voting), memberikan pengaruh lebih pada model yang terbukti lebih akurat pada training sebelumnya. Sifat adaptif pada AdaBoost ini sangat efektif untuk kasus data dengan boundary klasifikasi yang rumit atau kelas yang saling overlapping.

Walaupun awalnya AdaBoost dikembangkan untuk supervised learning, penelitian terbaru menunjukkan bahwa algoritma ini dapat digunakan secara efektif dalam validasi hasil clustering. Caranya adalah dengan memperlakukan label cluster dari hasil clustering unsupervised sebagai pseudo-label (target) pada supervised classification task. Selanjutnya, fitur dari data—misalnya usia, gender, dan kategori produk pada kasus e-commerce—digunakan sebagai input untuk melatih model AdaBoost agar dapat mengklasifikasikan data ke dalam masing-masing cluster. Nilai akurasi yang diperoleh menjadi indikator obyektif tingkat separabilitas antar cluster: jika akurasi tinggi (umumnya di atas 85%), berarti cluster terdefinisi dengan baik dan dapat dengan jelas dibedakan berdasarkan fitur-fiturnya. Sebaliknya, jika akurasi rendah (<70%), menunjukkan adanya overlap atau ketidakjelasan boundary antar cluster, sehingga membedakan data di masing-masing cluster menjadi sulit secara praktis.

Efektivitas penggunaan AdaBoost sebagai validator kualitas cluster telah dibuktikan dalam sejumlah penelitian. Misalnya, pada penelitian Customer Churn Prediction tahun 2022 [7], kombinasi K-means clustering dan AdaBoost mampu mencapai akurasi hingga 95,55%, yang lebih baik dibandingkan metode lain seperti neural network. Hal serupa juga dilaporkan oleh Wang et al. [9], di mana kombinasi clustering customer dan klasifikasi mampu mencapai akurasi di atas 95% dengan menggunakan ANN dan KSVM, serta AdaBoost tampil kompetitif pada perbandingan performa. Berdasarkan bukti tersebut, penelitian ini menerapkan AdaBoost baik pada pengelompokan cluster standar K-Prototypes maupun hasil ensemble K-Prototypes, agar dapat mengukur secara kuantitatif peningkatan separabilitas cluster yang diperoleh melalui pendekatan ensemble. Hipotesis yang diuji adalah bahwa ensemble K-Prototypes mampu menghasilkan

cluster yang lebih mudah dipisahkan oleh classifier, dan hal ini tercermin lewat kenaikan akurasi AdaBoost yang signifikan.