ANALISA EFISIENSI PENDEKATAN DATA-DRIVEN DALAM PROSES SEGMENTASI PASAR DENGAN STUDI KASUS STARTUP

Yefta Christian[1], Katherine Oktaviani Yap Rui Qi[2]
[1,2]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Internasional Batam
Email: [1]yefta@uib.ac.id, [2]1931153.katherine@uib.edu

ABSTRAK
Segmentasi pasar mengacu pada karakteristik yang digunakan untuk mengkategorikan customer dalam segmen. Teknokasi Edutech merupakan perusahaan rintisan yang diluncurkan pada awal Juni 2021. Teknokasi fokus pada education technology (ed-tech), terutama di bidang Teknologi Informasi. Perusahaan baru membutuhkan metode segmentasi pasar yang relevan dan akurat karena perusahaan harus memahami pasar sebelum membuat dan memasarkan produknya. Penelitian ini bertujuan untuk menyelidiki efisiensi algoritma pembelajaran mesin yang digunakan dalam segmentasi pasar, terutama untuk perusahaan early startup seperti Teknokasi. Penelitian dilakukan dengan pengembangan tiga model (K-Means dengan AHP-TOPSIS, SVM dengan AHP, dan Decision Tree dengan AHP) dengan tahapan data gathering, data preprocessing dan training, data modelling, evaluasi model, serta komparasi dan seleksi ketiga model. Berdasarkan analisa, didapatkan bahwa metode supervised learning lebih aplikatif, dengan SVM menunjukkan tingkat keakuratan yang relatif lebih tinggi. Penelitian ini membuktikan SVM beserta AHP mampu melakukan prediksi segmen pasar dan dengan begitu membantu decision-making berdasar label yang ditentukan.

Kata Kunci: AHP-TOPSIS, K-Means, SVM, Decision Tree, pembelajaran mesin

1. PENDAHULUAN

1.1. Latar Belakang

Riset pasar dan validasi adalah masalah umum yang dihadapi oleh banyak perusahaan. Kegagalan untuk memahami masalah tersebut berpotensi mengganggu pertumbuhan dan bahkan menyebabkan kebangkrutan. Bahkan pada tahap kreasi dan perancangan, produk telah gagal untuk menyesuaikan dengan target pasar mereka dan dengan demikian gagal untuk memenuhi kebutuhan pasar. Target pasar adalah aset penting yang perlu diteliti oleh perusahaan secara menyeluruh sebelum merilis produk mereka.

Segmentasi pasar mengacu pada karakteristik tertentu yang digunakan untuk mengkategorikan pelanggan ke dalam segmen. Variabel yang umum digunakan adalah faktor demografi, geografis, psikografis, perilaku, dan *product-related* (Camilleri, 2018). Segmentasi pasar bukanlah proses yang pasti dan 100% akurat, namun segmentasi pasar sangat membantu perusahaan dalam menekan biaya sekaligus memaksimalkan penjualan. Dengan biaya produksi minimum, perusahaan dapat mengelola upaya mereka ke segmen pertumbuhan lainnya, meningkatkan jumlah penjualan dan laba secara signifikan, serta mengeksplorasi potensi inovatif lainnya.

Menghadapi pertumbuhan kebutuhan perusahaan dalam pemasaran, individu yang terlibat terus mengejar efektivitas dan merampingkan proses untuk memudahkan dan mempercepat proses identifikasi pasar. Saat ini, banyak produk di pasaran yang mencoba untuk mengedepankan nilai keberagaman, keunikan dan adanya berbagai fitur dan manfaat yang ditawarkan. Namun, pada kenyataannya, produk-produk tersebut gagal mendapatkan penilaian positif di mata publik. Terdapat masalah dalam bagaimana bisnis menganalisa pasar sehingga gagal menentukan target pasar dan preferensinya. Kemampuan manusia tidak dapat mengikuti analisis data yang dihasilkan saat ini. Untuk menghadapi *big data*, kemampuan analisis yang lebih kuat harus dimanfaatkan (Miklosik et al., 2019).

Kecerdasan buatan adalah teknologi yang memungkinkan pelacakan data secara *real-time* untuk melakukan analisis data, seperti dengan melatih komputer untuk memahami pola data. Oleh karena itu, komputer mampu melakukan tugas dan memecahkan masalah seperti halnya manusia, bahkan dalam skala yang lebih besar dan pengurangan latensi akibat kesalahan manusia (Wisetsri et al., 2021). Kecerdasan buatan telah berhasil berkembang dan telah dipelajari secara ketat dalam dekade terakhir (Ma & Sun, 2020). Penerapan kecerdasan buatan dalam kehidupan modern sudah berkembang pesat, terutama dalam bidang pemasaran.

Bersama dengan pertumbuhan potensi *digital marketing* yang menguntungkan dan perluasan pengembangan kecerdasan buatan, keberadaan *marketing intelligence* dan AI-*powered decision-making tools* lainnya terbukti menjadi lebih signifikan bagi para pemimpin bisnis di seluruh dunia (Saura, 2021). Aspek digitalisasi, seperti kecerdasan buatan, tidak hanya mempercepat proses yang berulang tetapi telah berhasil meningkatkan akurasi dalam perhitungan dan kecepatan dalam pengambilan keputusan, terutama di bidang pemasaran. Oleh karena itu, setiap perusahaan harus

memanfaatkan kecerdasan buatan dalam proses pemasarannya, salah satunya adalah *machine learning*.

Proses pembelajaran mesin yang terlibat dalam penelitian ini adalah pendekatan *clustering* seperti algoritma K-Means, SVM, dan *decision tree*. *Clustering* melibatkan pembagian kumpulan data, atau lebih khusus lagi polanya, ke dalam kelompok-kelompok sejenisnya. Setiap algoritma memiliki kelebihan dan kekurangan tersendiri. Misalnya, algoritma K-Means, yang lebih umum digunakan, terbukti lebih efektif dalam menciptakan hasil pengelompokan untuk banyak aplikasi praktis. Namun algoritma K-Means tidak efisien dari segi waktu dan membutuhkan inisialisasi dimana peneliti harus menentukan jumlah cluster sebelum menjalankan algoritma (Sinaga & Yang, 2020). Oleh karena itu, algoritma K-Means tidak dapat diimplementasikan begitu saja dan harus dikonfigurasi lebih lanjut serta dipoles untuk menciptakan hasil yang lebih efisien.

Teknokasi Edutech merupakan perusahaan rintisan yang diluncurkan pada awal Juni 2021. Teknokasi lebih dikenal dengan fokus pada penyediaan teknologi pendidikan (*ed-tech*), terutama di bidang Teknologi Informasi. Teknokasi saat ini sedang dalam tahap awal inkubasi. Oleh karena itu, Teknokasi sedang meneliti nilai-nilai inti, nilai-nilai unik, dan target pasarnya untuk memberikan solusi yang lebih besar dan cocok bagi publik. Atas dasar inilah peneliti menyadari bahwa perusahaan yang baru diluncurkan membutuhkan riset dan segmentasi pasar yang relevan dan akurat.

Penelitian ini bertujuan untuk menyelidiki efisiensi algoritma pembelajaran mesin yang digunakan dalam *marketing intelligence*, terutama untuk perusahaan *early* startup seperti Teknokasi. Melalui penelitian ini, diharapkan perusahaan startup dapat menumbuhkan kesadaran dan meningkatkan penerapan kecerdasan buatan dalam proses bisnis mereka, seperti segmentasi pasar. Karena segmentasi pasar membutuhkan pemrosesan data dalam jumlah besar dan memiliki banyak proses berulang, *analytic tools* (salah satu aplikasi pembelajaran mesin) akan terbukti sangat berguna bagi perusahaan pemula.

Berdasarkan latar belakang di atas, rumusan penelitian inilah yang akan peneliti jawab.

a. Bagaimana efisiensi algoritma SVM dan *decision tree* dibandingkan dengan K-Means dalam proses segmentasi pasar?
b. Bagaimana segmentasi pasar berbasis data dapat mendukung pengambilan keputusan di perusahaan?

1.2. Tinjauan Pustaka

Segmentasi pasar merupakan proses penting yang harus dilakukan oleh perusahaan, bahkan sebelum memproduksi produknya. Segmentasi pasar membantu perusahaan untuk menemukan pasar yang cocok untuk ditargetkan oleh bisnis mereka. Oleh karena itu, perusahaan perlu memastikan kualitas proses segmentasi pasarnya. Untuk melakukannya, metode yang efisien dan akurat harus digunakan, seperti memanfaatkan *marketing intelligence*. *Marketing intelligence* membantu dalam mengekstraksi dan mengidentifikasi pola dari kumpulan *big data*, membuat keputusan dalam pemasaran, dan bahkan memprediksi perilaku pelanggan (Lies, 2019).

Berdasarkan kebutuhan tersebut, banyak peneliti telah melakukan penelitian dan laporan tertulis untuk membantu memecahkan masalah yang dihadapi. Sebuah penelitian oleh (Christian & Qi, 2022) bertujuan untuk mengembangkan aplikasi berbasis pembelajaran mesin untuk melakukan segmentasi potensial pasar dari startup tahap awal. Penelitian ini menggunakan data pasar potensial yang telah dikumpulkan sebelumnya yang kemudian dianalisis dan disegmentasi menggunakan K-Means melalui framework CRISP-DM. Modul K-Means kemudian diintegrasikan ke dalam aplikasi berbasis web yang dibangun menggunakan Python dan Flask. Aplikasi menerima input melalui file excel, dan output berupa visualisasi analisis dan excel data segmentasi pasar.

Penelitian lain yang dilakukan oleh (Chi-Hsien & Nagasawa, 2019) bertujuan untuk menganalisis perilaku pembelian pelanggan dengan bantuan model pembelajaran mesin, bukan intervensi manusia. Peneliti menggunakan pendekatan tanya jawab untuk mengumpulkan data dari subjek, menyaring data dan akhirnya menempatkan data ke dalam model statistik. Peneliti mengusulkan penggunaan lima model pembelajaran mesin yang paling umum seperti *logistic regression*, *nearest neighbours*, *decision tree*, *random forests*, dan *support vector machine* (SVM). Selain kelima model tersebut, peneliti juga menggunakan *Principle Component Analysis* pada tahap akhir. Data yang digunakan dalam penelitian ini adalah 760 data survei pembelian produk mewah yang sebenarnya oleh pelanggan. Melalui penelitian ini, peneliti berhasil mengimplementasikan semua model yang disebutkan dengan akurasi lebih dari 90%.

Penelitian yang dilakukan oleh (Saura, 2021) berusaha untuk menguraikan metode utama analisis, penggunaan, dan metrik kinerja yang paling baik digunakan untuk ilmu data di bidang pemasaran digital. Penelitian ini memberikan rekomendasi topik penelitian lebih lanjut terkait *data science* dan *digital marketing*. Penelitian ini dilakukan dengan menggunakan metode *Systematic Literature Review* dengan data yang berasal dari berbagai publikasi ilmiah. Kerangka konseptual yang dihasilkan memberikan wawasan tentang metrik yang harus diperhitungkan untuk membuat metode *data science* yang sukses, konsep yang harus dipertimbangkan

saat menggunakan intelijen pemasaran, serta sembilan topik untuk penelitian masa depan di bidang *data science* dan *digital marketing*.

Penelitian lain oleh (Raharja et al., 2022) mencoba mengelompokkan pelanggan salah satu toko kerajinan di Bali untuk menentukan strategi pemasaran. Penelitian ini menggunakan K-Means dan AHP-TOPSIS. Kriteria tersebut menggunakan perhitungan *Analytic Hierarchy Process* (AHP) dengan saran dari ahli. K-Means kemudian digunakan untuk mengelompokkan data. Terakhir, hasil pengelompokan diberi peringkat menggunakan *Technique for Order of Preference by Similarity to Ideal Solution* (TOPSIS). Proses ini memungkinkan pengguna untuk memilih *cluster*/segmen terbaik. Dengan menggunakan metode ini, peneliti dapat menyorot dua *cluster* pelanggan beserta atributnya.

Pada kasus penggunaan yang berbeda, salah satu penelitian (Taufiqurrahman et al., 2018) mengimplementasikan proses *fuzzy analytic hierarchy process* (F-AHP) yang dikombinasikan dengan K-Means untuk memberi peringkat dan mengelompokkan hasil rekrutmen staf. Hal ini dilakukan di Bagian Tata Usaha SMKN 7 Samarinda. Divisi Administrasi telah menerima banyak staf potensial tetapi mengalami kesulitan memilih yang terbaik. F-AHP kemudian digunakan untuk mengurutkan dan mengelompokkan hasil seleksi dengan menerapkan bobot atau nilai kriteria. K-Means juga digunakan untuk mengelompokkan hasil peringkat untuk menentukan pilihan terbaik.

Segmentasi pasar yang dilakukan pada perusahaan terkadang masih membutuhkan proses yang manual, berulang, dan tidak efisien. Selain itu, pengambilan keputusan manual tentang segmen pasar mana yang akan ditargetkan meningkatkan risiko subjektivitas dalam pilihan. Penelitian serupa sebelumnya (Christian & Qi, 2022) telah dilakukan pada data survei pasar menggunakan K-Means dan CRISP-DM untuk melakukan segmentasi data. Tetapi penelitian tersebut belum memberikan dukungan untuk pengambilan keputusan dan berada pada tahap awal proses pembelajaran mesin. Oleh karena itu, penelitian ini mencoba untuk meningkatkan kerangka segmentasi pasar menggunakan pembelajaran mesin dan mengembangkan sistem pendukung keputusan seperti yang ditunjukkan oleh (Raharja et al., 2022). Untuk mendapatkan hasil yang lebih baik dalam *clustering*, penelitian ini juga akan melibatkan analisis komparatif algoritma K-Means, SVM, dan Decision Tree yang mirip dengan penelitian oleh (Chi-Hsien & Nagasawa, 2019). Hasil kerangka pembelajaran mesin akan dianalisis menggunakan metrik akurasi dan skalabilitas yang terinspirasi oleh metrik kinerja dalam kerangka yang dikonsep oleh (Chi-Hsien & Nagasawa, 2019; Saura, 2021). Sedangkan untuk sistem pendukung keputusan, peneliti ini akan menggunakan AHP, dilanjutkan dengan proses clustering, dan terakhir TOPSIS seperti yang ditunjukkan dalam penelitian oleh (Raharja et al., 2022; Taufiqurrahman et al., 2018).

1.3. Metodologi Penelitian

Masalah yang diidentifikasi dalam penelitian ini adalah masalah validasi pasar. Validasi pasar dilakukan untuk menyempurnakan proses *go-to market* yang akan dilakukan oleh perusahaan. Data validasi pasar dikumpulkan dengan menggunakan kuesioner yang disebarkan ke berbagai strata yang berpotensi menjadi bagian dari pasar perusahaan. Data ini kemudian dikumpulkan dan diekstraksi sebagai dataset untuk diproses lebih lanjut.

Data dinormalisasi dan dibersihkan sesuai dengan kebutuhan data yang akan digunakan pada tahap pengolahan. Karena penelitian ini akan mengembangkan tiga model yang akan dibandingkan, langkah-langkah dari *preprocessing* data akan dipisahkan sesuai dengan kebutuhan model. Inilah ketiga model tersebut.

a. *Decision Tree*

Decision Tree adalah algoritma *supervised learning* yang menggunakan struktur percabangan untuk membagi dan mengklasifikasikan berdasarkan aturan (Chi-Hsien & Nagasawa, 2019).

a). *Data Preprocessing dan Training*

Data tersebut akan disaring untuk menghilangkan nilai dan atribut yang tidak relevan atau terkait dengan penelitian ini. Kriteria/subkriteria dari dataset akan diterapkan *Analytical Hierarchy Process* (AHP) untuk menimbang kriteria/subkriteria tersebut. AHP merupakan suatu metode yang digunakan dalam pengambilan keputusan, dimulai dari menentukan persyaratan keputusan untuk membuat dan membuat hierarki kriteria tersebut (Taufiqurrahman et al., 2018). Untuk label, "ketertarikan_jurusan_it" dipilih dan disimpan sebagai data kategoris.

Data tersebut akan dilanjutkan ke proses *training*. Proses *training* melibatkan pemisahan dataset menjadi dua kelompok sesuai dengan rasio yang ditetapkan sebelumnya. Kelompok terbesar akan digunakan sebagai *training data*, sedangkan kelompok terkecil akan digunakan sebagai *testing data*. Kedua kelompok ini akan melanjutkan proses *modeling*.

b). *Data Modeling*

Proses pemodelan mengacu pada proses pelatihan mesin untuk dapat memprediksi label yang sesuai berdasarkan atribut *training data* (Koehrsen, n.d.). *Training*

data akan digunakan sebagai dasar atau "panduan studi" untuk algoritma yang akan digunakan.

c). Model Evaluation

Model akan dievaluasi apakah pengambilan keputusan akhir sejalan dengan tujuan perusahaan, yaitu menemukan segmen pasar yang tertarik untuk belajar dan memasuki karir IT. Kita akan melihat bagaimana prediksi dibandingkan dengan label.

b. SVM

SVM (*Support Vector Machine*) adalah algoritma *supervised learning* lainnya yang memeriksa kumpulan data untuk mengidentifikasi fitur utama dan perilaku serupa, kemudian mengklasifikasikan nilai *instance* yang berbeda (Saura, 2021). Mirip dengan *decision tree*, ini adalah langkah-langkah untuk menyelesaikan model ini.

a). Data Preprocessing dan Training

Data tersebut akan disaring untuk menghilangkan nilai dan atribut yang tidak relevan atau terkait dengan penelitian ini. Kriteria/subkriteria dari dataset akan diterapkan *Analytical Hierarchy Process* (AHP) untuk menimbang kriteria/subkriteria tersebut. AHP merupakan suatu metode yang digunakan dalam pengambilan keputusan, dimulai dari menentukan persyaratan keputusan untuk membuat hierarki kriteria tersebut (Taufiqurrahman et al., 2018). Untuk label, "ketertarikan_jurusan_it" dipilih dan disimpan sebagai data kategoris.

Data tersebut akan dilanjutkan ke proses *training*. Proses *training* melibatkan pemisahan dataset menjadi dua kelompok sesuai dengan rasio yang ditetapkan sebelumnya. Kelompok terbesar akan digunakan sebagai *training* data, sedangkan kelompok terkecil akan digunakan sebagai *testing* data. Kedua kelompok ini akan melanjutkan proses *modeling*.

b). Data Modelling

Proses pemodelan mengacu pada proses pelatihan mesin untuk dapat memprediksi label yang sesuai berdasarkan atribut data pelatihan (Koehrsen, n.d.). Peneliti akan membagi dataset untuk ini.

c). Model Evaluation

Model tersebut akan dievaluasi apakah pengambilan keputusan akhir sejalan dengan tujuan perusahaan, yaitu menemukan segmen pasar yang tertarik untuk belajar dan memasuki karir IT.

c. K-Means dengan AHP-TOPSIS

a). Data Preprocessing dan Training

Sama seperti *decision tree* dan model SVM, model ini juga akan menyaring nilai dan atribut yang hilang berdasarkan relevansinya dengan penelitian. Kriteria/subkriteria dari dataset akan diterapkan *Analytical Hierarchy Process* (AHP) untuk menimbang kriteria/subkriteria tersebut. AHP merupakan suatu metode yang digunakan dalam pengambilan keputusan, dimulai dari menentukan persyaratan keputusan untuk membuat dan membuat hierarki kriteria tersebut (Taufiqurrahman et al., 2018).

b). Data Modelling

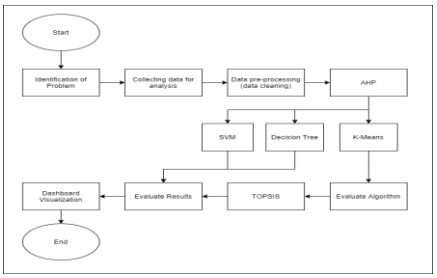
Dataset berbobot tersebut akan dilanjutkan untuk diklasterisasi menggunakan algoritma *machine learning* yaitu K-Means *clustering*. Hasil *clustering* akan dievaluasi menggunakan *Silhouette Coefficient*.

c). TOPSIS (*Technique for Order of Preference by Similarity to Ideal Solution*)

TOPSIS adalah metode pengambilan keputusan berdasarkan beberapa kriteria, yang menggunakan prinsip bahwa alternatif yang dipilih memiliki jarak terdekat dari solusi ideal positif dan terjauh dari solusi ideal negatif. Setelah hasil klasterisasi K-Means dievaluasi, TOPSIS akan diterapkan untuk menentukan klaster mana yang lebih baik bagi perusahaan untuk memasuki pasar.

d). Model Evaluation

Setelah TOPSIS, peneliti akan mengevaluasi apakah keputusan akhir yang dibuat selaras dengan tujuan perusahaan.



Gambar 1. Alur Kerja Penelitian

Hasil akhir harus dianalisis sesuai dengan kriteria di bawah ini. Penelitian ini membatasi istilah "efisiensi" menjadi dua kategori, *unsupervised learning* dan *supervised learning*, masing-masing dengan metrik di bawah ini.

a. Supervised Learning

Decision Tree dan SVM adalah algoritma *supervised learning* dan tidak akan memiliki metrik yang sama dengan *unsupervised learning*. Ini adalah metrik yang digunakan.

a). Ketepatan (*Accuracy*)
Akurasi mengacu pada rasio jumlah total prediksi yang benar dengan jumlah total dalam data. Semakin tinggi rasionya, semakin akurat modelnya. Persamaannya adalah seperti itu (Mishra, n.d.).

$$Akurasi = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

b. *Precision*
Precision mengacu pada rasio jumlah hasil positif yang benar dengan jumlah semua hasil positif oleh model. Persamaannya adalah seperti itu (Mishra, n.d.).

$$\text{Pr}ecision = \frac{TP}{(TP + FP)} \quad (2)$$

c. *Recall*
Recall mengacu pada jumlah hasil yang benar dibagi dengan jumlah nilai yang dibuang (Saura, 2021). Persamaannya seperti di bawah.

$$\text{Re}call = \frac{TP}{(TP + FN)} \quad (3)$$

d. *Unnsupervised learning*
K-Means menggunakan *Silhouette coefficient* sebagai metode evaluasi. *Silhouette coefficient* dinilai berdasarkan skor yang dicapai, yaitu pada rentang -1 hingga 1. Jika skor rata-rata mendekati 1, maka pengelompokan dianggap lebih baik. Jika skor rata-rata mendekati -1, maka clustering dianggap kurang baik. Ini adalah kriteria *silhouette coefficient* (Farissa et al., 2021)

Terlepas dari pengukuran metrik di atas, karena penelitian ini memberikan dukungan pengambilan keputusan, model akan dievaluasi berdasarkan keputusan yang dibuat. Peneliti akan mengevaluasi apakah keputusan yang diambil sudah sesuai dengan tujuan Teknokasi.

2. PEMBAHASAN

2.1. Data Gathering

Pada tahap ini peneliti mengumpulkan data yang dibutuhkan melalui Teknokasi Edutech. Untuk penelitian ini peneliti menggunakan data survei pasar yang disebarluaskan oleh Teknokasi Edutech sebagai sarana untuk memahami pasar dan melihat apakah perusahaan mampu memasuki pasar. Ini juga membantu perusahaan merancang strategi masuk pasar (*go-to market*).

Tabel 1. Atribut data pada dataset

Atribut Data	Tipe	Keterangan
Unnamed	Non-kategoris	Indeks urutan data
nama	Kategoris	Nama Responden
umur	Kategoris	Umur responden
tempat_tinggal	Kategoris	Domisili responden
pendidikan_terakhir	Kategoris	Pendidikan terakhir responden
tertarik_belajar_it	Kategoris (Ya/Tidak)	Ketertarikan responden dalam belajar IT
pernah_belajar_it	Kategoris (Ya/Tidak)	Pengalaman belajar IT responden
jurusan	Kategoris	Bidang yang dipelajari responden pada pendidikan terakhir
tertarik_bidang_it	Kategoris (Ya/Tidak)	Kaitan bidang studi responden dengan IT
literasi_digital	Kategoris (Ya/Tidak)	Pengalaman literasi digital
ketertarikan_jurusan_it	Kategoris (Ya/Tidak)	Ketertarikan dalam mempelajari IT secara formal melalui edukasi IT
previous_knowledge	Kategoris (Ya/Tidak)	Pengetahuan tentang Teknokasi
previous_experience	Kategoris (Ya/Tidak)	Pengalaman dengan produk Teknokasi (*webinars*, *trial courses*, atau *paid course*)
trust	Non-kategoris (Skala Likert)	Level kepercayaan atau keterbukaan responden terhadap Teknokasi
curiosity	Kategoris (Ya/Tidak)	Ketertarikan tentang Teknokasi

Survei tersebut mengumpulkan 224 tanggapan. Tanggapan ini perlu dibersihkan dan disaring sebelum dapat diproses oleh model data.

2.2. AHP: Pembobotan Kriteria dan Subkriteria

Pada tahap ini peneliti dengan saran dari perusahaan membandingkan dan menimbang masing-masing kriteria dan subkriteria. Ini menghasilkan dataset berbobot yang digunakan dalam proses pemodelan. Tolok ukur kepentingan didasarkan pada seberapa besar unsur tersebut mempengaruhi tujuan secara positif, dibandingkan dengan yang lain. Tujuan dari perbandingan ini adalah untuk mengidentifikasi pasar yang lebih memungkinkan untuk terhubung dan membeli dari Teknokasi.

a. Memberi Bobot pada Kriteria
Berdasarkan hasil perhitungan didapatkan rasio konsistensi yang dicapai sebesar 0,085 yang lebih rendah dari 0,1 sehingga penilaian ini

dapat diterima. Di bawah ini adalah bobot akhir.

Tabel 2. Bobot pada Kriteria

Atribut Data	Bobot	Atribut Data	Bobot
umur	0,01753	literasi_digital	0,033738
tempat_tinggal	0,015619	ketertarikan_jurusan_it	0,142853
pendidikan_terakhir	0,020933	previous_knowledge	0,103027
tertarik_belajar_it	0,092393	previous_experience	0,113769
pernah_belajar_it	0,040294	trust	0,208641
jurusan	0,031818	curiosity	0,093415
tertarik_bidang_it	0,085969		

b. Memberi Bobot pada Sub-Kriteria

Selanjutnya, dilakukan pembobotan untuk setiap sub kriteria. Hal ini dilakukan dengan membandingkan antar alternatif dalam masing-masing sub kriteria.

Tabel 3. Rasio Konsistensi Bobot pada Subkriteria

Subkriteria	Rasio Konsistensi	Catatan
umur	0,064	Konsisten
tempat_tinggal	0,0518	Konsisten
pendidikan_terakhir	0,088	Konsisten
tertarik_belajar_it	0,074	Konsisten
pernah_belajar_it	0	Konsisten
jurusan	0,087	Konsisten
tertarik_bidang_it	0,074	Konsisten
literasi_digital	0	Konsisten
ketertarikan_jurusan_it	0	Konsisten
previous_knowledge	0,074	Konsisten
previous_experience	0	Konsisten
trust	0,055	Konsisten
curiosity	0,074	Konsisten

c. Dataset yang Sudah Berbobot

Setelah bobot kriteria dan setiap subkriteria diputuskan dan dievaluasi, data dalam dataset dikonversi sesuai dengan bobot yang terkait dengan kategori dan atribut data. Dataset yang dihasilkan menjadi dataset berbobot dan akan digunakan pada tahap penelitian selanjutnya.

2.3. Supervised Learning Model

Supervised learning adalah pendekatan pembelajaran yang membutuhkan *expert advice* atau kurasi manual untuk memberi label pada kumpulan data. Algoritma kemudian akan menganalisis hubungan dan pola antara fitur data dan label. Dari proses ini, algoritma akan memprediksi label yang sesuai untuk data yang lebih baru, terpisah dari *training* dataset (Myszczynska et al., 2020).

a. *Data Preprocessing* dan *Training*

Setelah data melalui AHP, dataset tersebut kemudian dilanjutkan ke tahap kedua yaitu *preprocessing*. Karena "ketertarikan_jurusan_it" dipilih sebagai label, atribut data ini tidak diubah dengan nilai bobot, tetapi disimpan dalam bentuk kategorisnya. Data lainnya diubah dengan nilai bobot terkait. *Preprocessing* kemudian dilanjutkan dengan pembersihan *missing value* dan atribut data yang tidak diperlukan.

Dataset yang sudah bersih dibagi untuk *data features* (df_data) dan label (df_target). *Data features* akan menjadi data yang dipelajari oleh *classifier* dan kemudian dibandingkan dengan labelnya. Peneliti lalu membagi df_data dan df_target menjadi *training* dan *test* data menggunakan fungsi train_test_split dari Sci-Kit Learn. Peneliti akan mengevaluasi hasil model dengan dua rasio split, yaitu 80-20 dan 70-30.

b. Data Modelling

a). Decision Tree

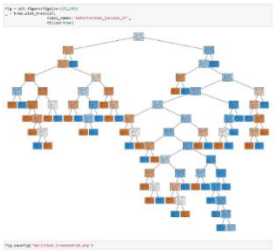
Dalam penelitian ini, untuk menjalankan algoritma *decision tree*, peneliti menggunakan Sci-Kit Learn. *Decision tree* pertama-tama memasukkan *training data* ke dalam model, kemudian menjalankan prediksi pada *test data*.

```
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
```

Gambar 2. Menjalankan Decision Tree

Decision tree yang dihasilkan di-output dan disimpan.



Gambar 3. Output Decision Tree

b). SVM

Untuk menjalankan model SVM, peneliti menggunakan library Sci-Kit Learn. SVM mempelajari *training data* dan menjalankan prediksi pada *test data*.

c. Evaluasi Model

Peneliti mengevaluasi bagaimana setiap model bereaksi terhadap dua rasio split data. Peneliti juga melakukan iterasi dari split hingga modelling untuk mendapatkan hasil rata-rata. Tabel di bawah ini menunjukkan hasil model SVM.

Tabel 4. Hasil Evaluasi SVM

Data Training Ratio (train-test)	Rata-Rata Hasil Evaluasi		
	Accuracy	*Precision*	*Recall*
70-30	0,627	0,627	0,627
80-20	0,665	0,665	0,665

Tabel di bawah ini menunjukkan hasil dari model *decision tree*.

Tabel 5. Hasil Evaluasi Decision Tree

Data Training Ratio (train-test)	Rata-Rata Hasil Evaluasi		
	Accuracy	*Precision*	*Recall*
70-30	0,638	0,638	0,638
80-20	0,65	0,65	0,65

2.4. Unsupervised Learning: K-Means dengan AHP-TOPSIS

Pada tahap ini, peneliti menggunakan K-Means untuk mengelompokkan dataset berbobot. K-Means adalah algoritma yang menggunakan centroid yang ditetapkan secara acak di setiap *cluster* dan melakukan iterasi pada mean grup (*cluster*) yang dihitung ulang dan menetapkan kembali unit ke dalam cluster ini (Farissa et al., 2021; Ma & Sun, 2020).

a. *Data Preprocessing* dan *Training*

Untuk model ini, peneliti menggunakan dataset berbobot dimana semua atribut data ditimbang dan dataset dikonversi sesuai dengan bobotnya. Selanjutnya, dataset berbobot dibersihkan dari *missing values* dan menghilangkan fitur data yang tidak diperlukan dalam model ini.

Karena ini adalah *unsupervised learning*, tidak diperlukan label data dan kumpulan data juga tidak dibagi. Data tersebut kemudian dimasukkan ke dalam model.

b. *Data Modelling*

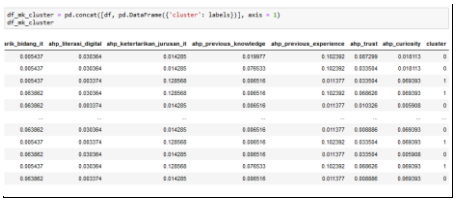
Dalam menjalankan model ini, pertama-tama perlu diputuskan berapa banyak *cluster* yang akan digunakan. Untuk melakukan ini, peneliti membandingkan antara Metode *Elbow* dan Metode *Silhouette*. Di bawah ini adalah jumlah *cluster* optimal yang ditentukan.

Tabel 6. Jumlah Cluster Optimal Berdasar Metode Identifikasi Cluster

Metode Identifikasi Jumlah *Cluster*	Jumlah *Cluster* Optimal
Elbow Method	3
Silhouette Method	2

Dengan masing-masing metode, peneliti kemudian menjalankan algoritma K-Means dengan dataset berbobot dan dilakukan pencatatan pusat *cluster* yang akan digunakan untuk proses TOPSIS untuk menentukan peringkat *cluster*.

Hasil pengelompokan kemudian digabungkan ke dataset dan output ke file Excel untuk mendokumentasikan hasil.



Gambar 4. Hasil Segmentasi

Sebelum melanjutkan ke TOPSIS, hasil dari kedua metode dievaluasi. Dengan menggunakan kriteria *Silhouette Coefficient*, peneliti dapat memutuskan metode penentuan jumlah cluster mana yang terbaik untuk dataset ini.

Tabel 7. Kriteria Evaluasi Nilai *Silhouette Coefficient*

Silhouette Coefficient	Kriteria Evaluasi
0.7 <SC<=1.0	Struktur Kuat
0.5 <SC<=0.7	Struktur Medium
0.25 <SC<=0.5	Struktur Lemah
SC<=0.25	Tidak Berstruktur

Ini adalah *Silhouette Coefficient* dari kedua metode.

Tabel 8. Nilai *Silhouette Coefficient* Berdasar Metode Identifikasi Cluster

Metode Identifikasi Jumlah *Cluster*	No. *Cluster* Optimal	*Silhouette Coefficient Value*	Keterangan
Silhouette Method	2	0.90	Struktur yang Kuat
Elbow Method	3	0.91	Struktur yang Kuat

Clustering dari *Elbow Method* digunakan pada versi final dari model ini dan dilanjutkan ke metode TOPSIS untuk diranking.

c. TOPSIS (*Technique for Orders Preference by Similarity to Ideal Solution*)

Pada tahap TOPSIS, dilakukan normalisasi dan pembobotan pada matriks keputusan. Dalam penelitian ini, matriks keputusan didasarkan pada pusat-pusat *cluster* yang dikumpulkan dari *clustering*. Nilai setiap kriteria untuk setiap alternatif dijumlahkan dan nilai setiap kriteria akan dibagi dengan jumlah tersebut (Siregar, 2017).

Untuk melakukan pembobotan, digunakan tabel kriteria urgensi di bawah ini.

Tabel 9. Nilai Kriteria Kepentingan

Nilai Kepentingan	Keterangan	Nilai Kepentingan	Keterangan
1	Sangat tidak penting	4	Penting
2	Tidak penting	5	Sangat penting
3	Cukup penting		

Peneliti kemudian menghitung bobot untuk matriks yang dinormalisasi.

Selanjutnya, peneliti menentukan matriks solusi ideal positif dan negatif, serta jarak antara setiap alternatif ke matriks solusi positif dan negatif.

Tabel 10. Jarak ke Solusi Ideal Positif dan Negatif

Jarak ke Solusi Ideal Positif		Jarak ke Solusi Ideal Negatif	
D_1^+	1,917390044	D_1^-	42,05420689
D_2^+	36,97372769	D_2^-	9,447067289
D_3^+	26,68957236	D_3^-	32,61126947

Terakhir, peneliti menghitung nilai preferensi sesuai rumus (4) untuk setiap alternatif. Ini akan diurutkan dan nilai yang lebih besar menunjukkan alternatif mana yang lebih disarankan.

$$V_i = \frac{D_i^-}{(D_i^- + D_i^+)} \quad (4)$$

Tabel di bawah ini menunjukkan nilai preferensi dan hasil pemeringkatan.

Tabel 11. Ranking pada Segmen K-Means

Alternatif	Nilai Preferensi	Nilai	Ranking
Segmen_1	V1	0,956394805	1
Segmen_2	V2	0,203509382	3
Segmen_3	V3	0,549929284	2

Berdasarkan peringkat di atas, Segmen_1 direkomendasikan sebagai pasar yang akan dimasuki Teknokasi.

d. Evaluasi Model

Dari keputusan yang dibuat oleh TOPSIS, kami kemudian mengevaluasi atribut dari segmen yang dipilih (Segmen_1). Di bawah ini adalah data untuk Segmen_1.

Tabel 12. Dataset Segmen_1

No.	umur	tempat_tinggal	…	trust	curiosity
1	0,002862333	0,00087557	…	0,010326269	0,005908405
2	0,009243446	0,000728907	…	0,03350378	0,018113449
…	…	…	…	…	…
67	0,002862333	0,001577613	…	0,03350378	0,005908405
68	0,002862333	0,000869822	…	0,008885603	0,069393146

Dari atribut di atas, segmen yang dipilih TOPSIS terdiri dari :

a). Calon pelanggan yang tidak tertarik mengikuti jurusan IT dari kriteria "tertarik_jurusan_it" yang justru menjadi penentu dalam keputusan apakah segmen dapat didekati.

b). 50% calon pelanggan berminat belajar IT.

c). 83,82% calon pelanggan belum mengetahui tentang Teknokasi sebelumnya.

d). Calon pelanggan tidak memiliki pengalaman sebelumnya dengan Teknokasi.

e). 80,88% calon pelanggan ingin tahu tentang Teknokasi.

f). 64,71% calon pelanggan memilih nilai 4 dan 3 dari 5 untuk tingkat kepercayaan yang mereka berikan Teknokasi.

2.5. Evaluasi Final dan Seleksi Data Model

Melalui model *supervised learning*, proses pengambilan keputusan lebih sederhana. Pemangku kepentingan dapat memutuskan atribut mana yang dapat menjadi tujuan, seperti "ketertarikan_jurusan_it" yang dipilih dalam penelitian ini. Selanjutnya dari rata-rata akurasi, presisi, dan *recall*, kita melihat bahwa algoritma SVM dengan data split 80-20 adalah pilihan yang lebih baik. Ini memiliki akurasi, presisi, dan *recall* 66,5%.

Dibandingkan dengan penelitian sebelumnya oleh (Christian & Qi, 2022), kita dapat melihat bahwa pengelompokan K-Means telah meningkat secara signifikan dengan menggabungkan AHP dalam dataset. Dataset berbobot terdiri dari nilai numerik dengan bobot relatif yang lebih spesifik daripada nilai kategoris. Tampak dari *silhouette coefficient* sebesar 0,9.

Karena sifat tak terduga pada segmen yang dipilih, peneliti melihat bahwa studi lebih lanjut dengan TOPSIS dapat dilakukan untuk meningkatkan efisiensi proses pengambilan keputusan. Tolok ukur dalam penelitian ini adalah

"ketertarikan_jurusan_it". Namun, segmen yang dipilih menunjukkan semua responden yang tidak tertarik mempelajari pendidikan IT. K-Means dengan AHP-TOPSIS cukup baik tetapi belum dapat diterapkan pada dataset ini. Peneliti harus mengeksplorasi masalah baik dalam TOPSIS atau validasi dataset.

Sebagai keputusan akhir, SVM dengan rasio pemisahan data 80-20 akan dipilih untuk didigitalkan dalam aplikasi berbasis web, sehingga dapat diakses oleh pengguna. Proses AHP juga akan diintegrasikan ke dalam model. Pengembangan solusi digital akan dilakukan pada penelitian selanjutnya.

3. ALGORITMA ATAU PROGRAM

Data modelling yang dilakukan dirangkai menggunakan Bahasa Python dengan Jupyter Notebook. Sesuai dengan jumlah data model, berikut algoritma yang digunakan.

a. SVM

```
df_target =
df['ketertarikan_jurusan_it'].copy()
df_data = df.iloc[:,0:8]
df_data = pd.concat([df_data,
df.iloc[:,9:]], axis = 1)
df_data = df_data.values.tolist()
df_target = df_target.values.tolist()
X, y = df_data, df_target
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3)
clf = svm.SVC()
clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score
(y_test, y_pred))
```

Gambar 5. Algoritma Data Model Berbasis SVM

b. Decision Tree

```
df_target =
df['ketertarikan_jurusan_it'].copy()
df_data = df.iloc[:,0:8]
df_data = pd.concat([df_data,
df.iloc[:,9:]], axis = 1)
df_data = df_data.values.tolist()
df_target = df_target.values.tolist()
X, y = df_data, df_target
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3)
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score
(y_test, y_pred))
```

Gambar 6. Algoritma Data Model Decision Tree

c. K-Means

```
from kneed import KneeLocator
clusters = []
for k in range(1, 10, 1):
    # train clustering with the
specified K
    model = KMeans(n_clusters=k)
    model.fit(df_original)# append model
to cluster list
    clusters.append(model)
wcss_1 = []
range_values = range(1, 10)
for i in range_values:
  kmeans = KMeans(n_clusters=i)
  kmeans.fit(df_original)
  wcss_1.append(kmeans.inertia_)
```

Gambar 7. Algoritma Data Model K-Means

4. KESIMPULAN

Proses segmentasi pasar yang dilakukan dengan *data-driven approach* memiliki potensi yang luas dan dapat dieksplorasi lebih jauh. Melalui penelitian dan analisa yang dilakukan untuk memahami efisiensi *data-driven approach* untuk segmentasi pasar, didapatkan kesimpulan sebagai berikut.

a. Pendekatan *supervised learning* mampu memberikan keputusan secara langsung tanpa melalui proses TOPSIS karena SVM dan *decision tree* digunakan untuk melakukan prediksi. Dengan begitu, pengguna mampu mendapatkan apakah calon pelanggan termasuk dalam segmen pasar yang diinginkan, sesuai dengan label yang telah diberikan. Namun validitas dan reliabilitas dari *data modelling* harus dipastikan sudah sesuai terlebih dahulu. Berdasarkan metrik penilaian, SVM memberikan hasil lebih baik daripada *decision tree*.
b. Pendekatan *unsupervised learning* berupa K-Means clustering menunjukkan pembagian cluster pada data yang sudah ada, namun belum menunjukkan keputusan jelas terkait cluster/segmen yang dapat ditargetkan. Oleh karena itu, K-Means harus dikombinasikan dengan metode lain seperti AHP-TOPSIS untuk memberikan informasi yang berguna dan membantu pembuatan keputusan.

Berdasar penelitian yang sudah dilakukan, peneliti menyarankan startup tahap awal dan startup yang sudah berjalan untuk memaksimalkan penggunaan *data-driven approach* dalam proses bisnis yang dilakukan, terutama segmentasi pasar. Dengan begitu, proses segmentasi pasar memiliki dukungan data, minim campur tangan manusia, dan metode analisa menjadi terstruktur. Untuk pengembangan lebih lanjut, dapat dilakukan analisa lebih dalam terkait data model dan *data analytics*, serta dilakukan *deployment* pada aplikasi.

PUSTAKA

Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. *Travel Marketing, Tourism Economics and the Airline Product*,

4, 69–83. https://doi.org/10.1108/978-1-78635-746-520161006

Chi-Hsien, K., & Nagasawa, S. (2019). Applying machine learning to market analysis: Knowing your luxury consumer. *Journal of Management Analytics*, *6*(4), 404–419. https://doi.org/10.1080/23270012.2019.1692254

Christian, Y., & Qi, K. O. Y. R. (2022). Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Early Stage Startup dengan Menggunakan CRISP-DM. *JURIKOM (Jurnal Riset Komputer)*, *9*(4), 966–973. https://doi.org/10.30865/jurikom.v9i4.4486

Farissa, R. A., Mayasari, R., & Umaidah, Y. (2021). *Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient*. *5*(2), 109–116.

Koehrsen, W. (n.d.). *Modeling: Teaching a Machine Learning Algorithm to Deliver Business Value*. Towards Data Science. Retrieved December 15, 2021, from https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86

Lies, J. (2019). Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(5), 134. https://doi.org/10.9781/ijimai.2019.05.002

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Miklosik, A., Kuchta, M., Evans, N., & Zak, S. (2019). Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing. *IEEE Access*, *7*, 85705–85718. https://doi.org/10.1109/ACCESS.2019.2924425

Mishra, A. (n.d.). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science. Retrieved December 16, 2021, from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), 440–456. https://doi.org/10.1038/s41582-020-0377-8

Raharja, M. A., Surya, I. K. A., & Mogi, I. K. A. (2022). CLUSTERING CUSTOMER FOR DETERMINE MARKET STRATEGY USING K-MEANS AND TOPSIS : CASE STUDY. *International Proceeding Conference on Information Technology, Multimedia, Architecture, Design, and E-Business (IMADE)*, *2*(August), 61–71.

Saura, J. R. (2021). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. *Journal of Innovation and Knowledge*, *6*(2), 92–102. https://doi.org/10.1016/j.jik.2020.08.001

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Siregar, R. A. (2017). Seleksi Penyerang Utama Menggunakan K-Means Clustering Dan Sistem Pendukung Keputusan Metode Topsis. *Technomedia Journal*, *2*(1), 37–48. https://doi.org/10.33050/tmj.v2i1.314

Taufiqurrahman, T., Malani, R., & Najib, A. (2018). Pengurutan dan Pengelompokan Divisi Hasil Penerimaan Calon Karyawan Menggunakan Metode F-AHP dan K-Means. *Prosiding SAKTI (Seminar Ilmu Komputer Dan Teknologi Informasi)*, *3*(1), 115–123.

Wisetsri, W., S, R. T., Julie Aarthy, C. C., Thakur, V., Pandey, D., & Gulati, K. (2021). Systematic Analysis and Future Research Directions in Artificial Intelligence for Marketing. *Turkish Journal of Computer and Mathematics Education*, *12*(11), 43–55.