JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM

Yefta Christian*, Katherine Oktaviani Yap Rui Qi

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Internasional Batam, Batam, Indonesia
Email: [1,*]yefta@uib.ac.id, [2] 1931153.katherine@uib.edu
Email Penulis Korespondensi: yefta@uib.ac.id
Submitted **17-07-2022**; Accepted **16-08-2022**; Published **30-08-2022**

Abstrak

Startup *early stage* akan melakukan ideasi, *problem solving*, dan riset pasar. Salah satu tahapan dalam riset pasar adalah segmentasi pasar. Pada umumnya, startup *early stage* tidak memiliki sumber daya yang mumpuni sehingga banyak tahapan yang dilakukan secara manual. Hal ini mendorong terjadinya ketidakakuratan dan berkurangnya objektifitas dalam menilai situasi pasar yang sebenarnya penting bagi pertumbuhan startup *early stage*. Penelitian ini fokus mengembangkan sebuah aplikasi berbasis *machine learning* untuk segmentasi pasar. Pengembangan aplikasi dilakukan menggunakan kerangka kerja CRISP-DM. Kerangka kerja ini digunakan dalam data mining dengan menerapkan enam fase, yaitu *business understanding*, *data understanding*, *data preparation*, *modelling*, *deployment*, dan evaluasi, dalam metodenya untuk mengidentifikasi input dan output dalam suatu proses. *Data model* yang digunakan dalam aplikasi ini adalah K-Means, yaitu algoritma yang sering digunakan untuk *clustering* atau pengelompokkan suatu kumpulan data menjadi beberapa kelompok berdasarkan atribut yang dimiliki. Aplikasi yang dihasilkan telah mampu memberikan hasil dalam bentuk visualisasi dan pembacaan segmentasi dalam bentuk excel. Dengan begitu, startup early stage ini dapat terbantu dalam pengolahan data untuk identifikasi segmen dalam pasar. Untuk penelitian lebih lanjut, aplikasi ini dapat ditingkatkan untuk efisiensi dan keakuratan model. Oleh karena itu, sistem dapat dikembangkan dari sisi desain tampilan, algoritma, pemilihan jumlah cluster, dan analisa terhadap data, serta menambahkan fitur *decision-making* atau *predictive analysis*.

Kata Kunci: Pembelajaran Mesin; Riset Pemasaran; Startup Early Stage; K-Means; CRISP-DM

Abstract

Early stage startup would conduct ideation, problem solving, and market research. One of the stages in market research is market segmentation. Normally, early stage startups do not have enough resources, thus many processes are done manually. This encourages inacurracies and lessen the objectiveness in evaluating market situation which is important for the growth of early stage startups. This research focuses on developing a machine learning based application for market segmentation. The framework used here is CRISP-DM, which is a framework used in data mining. This framework has six phases, which consists of business understanding, data understanding, data preparation, modelling, deployment, and evaluation, to identify the input and output of a process. The data model used in this application is K-Means, which is a common algorithm used for clustering or dividing a set of data into groups according to their attributes. The application developed is able to output results in the form of visualization and the segmentation in excel format. With this, the early stage startup is able to process their data to identify segments in the market. For future research, this application could be improved in the efficiency and accuracy of the model. The application could be improved in the aspects of UI/UX design, the algorithm used, the number of clusters, and the analysis of the dataset, as well as adding a predictive analysis feature.

Keywords: Machine learning; Market Research; Early Stage Startup; K-Means; CRISP-DM

1. PENDAHULUAN

Kewirausahaan di Indonesia memerlukan peningkatan. Walau begitu, data *Global Entrepreneurship Index* telah menunjukkan peningkatan dari tahun 2018 ke 2019. Tahun 2018 menunjukkan Indonesia berada di posisi 94 dari 137 negara [1]. Tahun 2019 menunjukkan Indonesia di posisi 75 dari 137 negara [2]. Namun dalam sebuah laporan yang diterbitkan *Global Entrepreneurship Monitor* untuk periode 2020/2021 menunjukkan aktivitas wirausaha di Indonesia yang cukup rendah pada tahun 2020. Aktivitas wirausaha tahap awal tidak mencapai 10% dari jumlah penduduk berumur 18-64 tahun dan jumlah kepemilikan usaha juga tidak mencapai 15% [3].

Keberadaan kewirausahaan sebenarnya memberi dampak signifikan bagi pembangunan ekonomi, penyediaan lapangan kerja, dan perkembangan potensi untuk masyarakat [4]. Oleh karena itu, kewirausahaan di Indonesia juga harus terus ditingkatkan. Melalui program-program kependidikan baru, seperti Kampus Merdeka dan Merdeka Belajar di institusi dari sekolah hingga pendidikan tinggi, pemerintah telah memperlebar jalan bagi calon wirausaha dan calon inovator untuk mengembangkan usahanya sejak dini. Hal ini cukup membantu untuk mendorong dan memotivasi peningkatan kewirausahaan di Indonesia. Kewirausahaan sendiri memiliki beberapa model, seperti konvensional dan startup atau perusahaan rintisan yang banyak kita kenal dalam beberapa tahun terakhir.

Startup, khususnya startup digital, melewati tiga tahap perkembangan, yaitu *startup* atau *early stage*, *stabilization*, dan *growth*. *Startup* atau *early stage* adalah tahapan dimana *startup* membuat produk terlebih dahulu, sebelum menjual. *Stabilization stage* adalah tahapan di antara mendapatkan penjualan pertama dan memperoleh produk yang stabil. *Growth stage* adalah tahapan di mana distribusi produk dapat terjadi tanpa membuat halangan terhadap pengembangan produk [5].

Pada tahap *early stage*, startup melakukan ideasi dan *problem solving* untuk mengetahui permasalahan yang dihadapi dan solusinya sehingga bisa terbentuk produk yang akan dipasarkan. Selanjutnya startup melakukan riset pasar untuk mengetahui umpan balik dan situasi pasar.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Penelitian ini berpusat pada tahap riset pasar. Segmentasi pasar adalah metode yang menggolongkan pasar menjadi beberapa segmen berdasarkan variabel seperti demografis, geografis, psikografis, perilaku, dan faktor yang berkaitan dengan produk [6]. Segmentasi pasar telah mengintegrasikan metode *data-driven*. Sebuah studi kasus di Taipei menunjukkan bagaimana metode *data-driven* mempengaruhi proses segmentasi data. Metode yang *data-driven* dianggap lebih objektif dan akurat jika data dianalisa menggunakan tools statistis [7].

E. Y. L. Nandapala dan K. P. N. Jayasena melakukan penelitian untuk melakukan segmentasi pasar secara efektif dan efisien [8]. Dalam penelitian ini, digunakan algoritma K-Means. Penelitian ini dimulai dengan melakukan data *preprocessing*, menentukan jumlah *cluster*, menjalankan algoritma K-Means dan menganalisa hasil yang diperoleh. Untuk melakukan identifikasi jumlah *cluster*, digunakan perbandingan dari tiga metode, yaitu *elbow method*, *silhouette method*, dan *gap statistic method*. Penelitian ini membuktikan bahwa jumlah *cluster* yang sesuai untuk *use case* ini adalah enam *cluste*r. Penelitian ini pun berhasil mengembangkan *framework clustering* yang dapat membantu perusahaan dalam membuat keputusan. Berdasarkan penelitian, digunakan jumlah sama dengan tiga dalam proses analisa yang diperlukan.

Penelitian yang dilakukan oleh I Gusti Ayu Widiantari Putri dan Ida Bagus Gede Dwidasmara fokus melakukan segmentasi pada data transaksi produk toko kerajinan tangan "Kreasi Slaka Bali" [9]. Dengan begitu, dapat mengidentifikasi *cluster* minat beli *consumer* dengan atribut-atribut yang membentuk *cluster* tersebut. Penelitian ini menggunakan K-Means untuk melakukan *clustering* dan *elbow method* untuk mengidentifikasi jumlah *cluster* yang diperlukan. *Clustering* ini berguna bagi usaha untuk memfokuskan area produksi mereka sehingga dapat menghasilkan pendapatan yang lebih maksimal.

Penelitian yang dilakukan oleh Salman Kimiagari, Samira Keivanpour, dan Matti Haverila, bertujuan untuk menyediakan solusi bagi perusahaan-perusahaan dalam melakukan segmentasi pasar sehingga mampu memperoleh wawasan lebih dari target pasar [10]. Penelitian ini menggunakan data pengguna ponsel yang dikumpulkan melalui sebuah survei terhadap siswa SLTA dan mahasiswa di Finland, RRC, Kanada, dan Selandia Baru. Data yang dikumpulkan selanjutnya melalui empat tahap, yaitu visualisasi data, clustering, analisa perbandingan, dan menghubungkan hasil dengan pendapat ahli. Penelitian ini membandingkan penggunaan K-Means, K-Medoids, SOM, dan *fuzzy* c-means. Penelitian ini berhasil mengembangkan suatu *framework clustering* dan keberadaan *intermarket segments* dalam pasar telepon seluler.

Azad Abdulhafedh telah melakukan penelitian untuk menerapkan algoritma *clustering unsupervised* pada segmentasi pasar untuk menentukan strategi pemasaran bagi sebuah perusahaan kartu kredit [11]. Penelitian ini menggunakan data akun pengguna kartu kredit, seperti saldo, pembelian, *credit scores*, dan lain-lain. Untuk mengolah dan *clustering*, digunakan tiga algoritma, yaitu *Hierarchical clustering*, K-Means, dan *Principal Componen Analysis* (PCA). Penelitian menemukan bahwa K-Means lebih cocok dengan dataset yang digunakan. Sementara PCA dapat digunakan untuk mereduksi dimensionalitas dan visualisasi data.

Penelitian yang dilakukan Xiaochuan Pu, Ning Qi, dan Junli Huang menerapkan K-Means untuk merancang metode segmentasi pelanggan yang dapat mengidentifikasi pelanggan tingkat atas [12]. Penelitian ini menggunakan data konsumsi pelanggan yang diperoleh dari UCI Repository. Dataset ini diterapkan *preprocessing*, eksplorasi, dan analisa data. Pengolahan data dilakukan pada KNIMES dan penelitian ini menghasilkan metode segmentasi pasar yang mampu memetakan pelanggan dan potensialnya sehingga perusahaan mampu memprioritaskan pasar yang akan dicapai.

Penelitian-penelitian relevan di atas menunjukkan penggunaan K-Means dan *data processing* untuk melakukan segmentasi pasar pada beberapa bidang bisnis dapat dilakukan dengan baik. Oleh karena itu, penulis memutuskan untuk menggunakan K-Means sebagai *clustering algorithm* dengan penggunaan CRISP-DM sebagai pedoman proses analisa data hingga *deployment*. Namun dalam penelitian ini, untuk mempermudah penggunaan K-Means dalam segmentasi pasar secara jangka panjang, penulis mengembangkan aplikasi *web* yang mampu memroses data dan memberikan output yang mudah dipahami. Penelitian relevan di atas juga menunjukkan segmentasi pasar *data-driven* pada perusahaan atau bidang bisnis yang sudah berdiri dan berjalan. Adapun penelitian ini fokus pada analisa pasar yang akan dituju suatu perusahaan (startup) yang baru berdiri. Penelitian ini diharapkan dapat meningkatkan penggunaan kecerdasan buatan pada startup baru. Berikut ini tujuan dari penelitian ini:

a. Mengembangkan *tools* analisa data calon pelanggan bagi *startup*, khususnya yang termasuk dalam *early stage*.
b. Memberikan kontribusi pada penerapan pembelajaran mesin pada *startup early stage*.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan 51 data dari sebuah *startup early stage* di Batam, yang berisi hasil wawancara terhadap calon pelanggan. Melalui wawancara, diperoleh data demografis dan data lainnya yang berkaitan dengan interest, passion, dan latar belakang dari responden. Penelitian ini menerapkan kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) dengan algoritma *machine learning* berupa K-Means *clustering* untuk melakukan segmentasi pasar pada data di atas. Kerangka kerja CRISP-DM adalah kerangka kerja data mining yang menerapkan enam fase dalam metodenya untuk mengidentifikasi input dan output dalam suatu proses. Keenam fase tersebut adalah *business understanding*, *data understanding*, *data preparation*, *modelling*, *deployment*, dan evaluasi [13].

2.1 *Business Understanding*

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tahap ini bertujuan untuk menilai kondisi suatu bisnis sehinga mendapatkan gambaran dari sumber daya yang tersedia dan dibutuhkan. Penentuan target *data mining* adalah aspek penting dari tahap ini [14]. Penulis melakukan identifikasi dan penelaahan terhadap masalah yang dihadapi, objektif bisnis, target yang hendak dicapai, serta menghasilkan rencana pengembangan proyek. Tahap ini dilakukan dengan mengadakan diskusi dengan pihak *startup early stage* untuk mengetahui rencana dan kebutuhannya.

2.2 *Data Understanding*

Penulis melakukan pengumpulan data yang diperlukan, mengevaluasi dan mengeksplorasi data, serta memastikan kualitas data. Melalui tahap ini, didapatkan kesimpulan terkait data yang akan digunakan dan strategi yang harus diambil untuk mengatasi data tersebut. Pengumpulan data dilakukan dengan menggunakan dokumen *excel* yang diperoleh dari *startup early stage*. Penulis juga melakukan *exploratory data analysis* (EDA) yang merupakan proses dimana representasi dalam bentuk numerik, tabel ataupun grafik dibuat untuk memberi rangkuman serta menerangkan aspek dan fitur utama dari suatu data [15].

2.3 *Data Preparation*

Data yang sudah dieksplorasi pada tahap sebelumnya harus dilakukan *preprocessing*, dibersihkan dan dinilai dimensionalitas datanya [16]. Tahap pembersihan yaitu dengan menghapus atau mengganti *null value* atau nilai data yang *error*. Selanjutnya, dengan menilai dimensionalitas data, dapat ditentukan atribut data yang perlu dan tidak perlu dihilangkan.

2.4 *Data Modelling*

Melalui tahap ini, penulis menentukan dan mengaplikasikan model yang akan digunakan. Algoritma K-Means menggunakan metode non-hierarkikal yang mengelompokkan data menggunakan suatu sistem pembagi. K-Means menggunakan data numerik dan sering digunakan karena kemudahannya dalam melakukan implementasi [9]. Pengembangan *data model* dibantu dengan *tools* Jupyter *Notebook*.

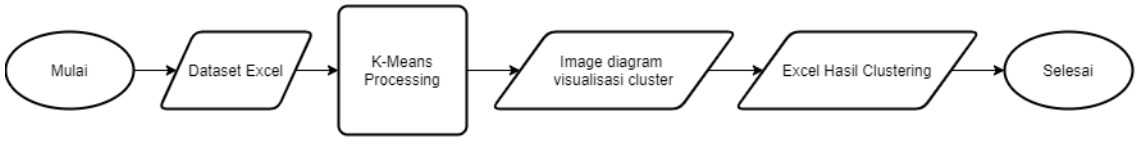
2.5 *Evaluation*

Pada tahap ini, penulis akan melakukan evaluasi terhadap performa data model dan memeriksa kembali seluruh proses yang sudah dijalankan, untuk memastikan tidak adanya data dan tahapan yang terlewatkan.

2.6 *Deployment*

Setelah tahap *Evaluation* berhasil, produk dapat diserahterima dan digunakan oleh *startup*. Produk dilakukan *deployment* melalui sebuah *web app* yang dikembangkan menggunakan *microservice* Flask dengan bahasa pemrograman utama Python. Pengembangan *web app* dilakukan dengan tahapan sebagai berikut.

a. Desain alur *web app*.
   Desain alur dilakukan untuk memberi gambaran terkait cara kerja atau *use case web app* dan pengembangannya.



Gambar 1. Bagan Alur *Web App*

b. Desain tampilan *web app*.
   Desain tampilan UI dilakukan dengan menggunakan Figma. Desain ini menjadi landasan untuk *frontend web app*.

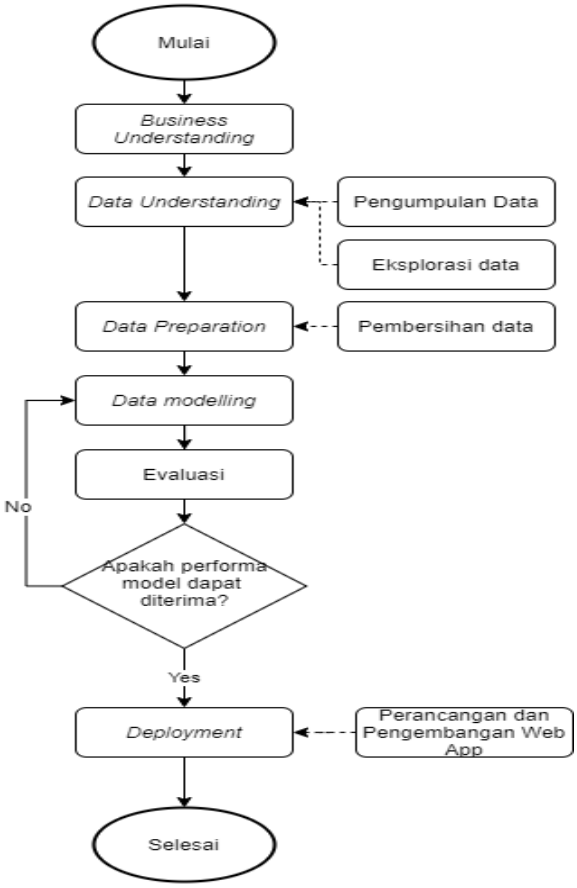
Gambar 2. Desain Awal Tampilan *Web App*

c. Pengembangan *backend* dan *frontend web app* sesuai desain.
   Pengembangan dilakukan menggunakan *microservice* Flask dan Python menggunakan Pycharm.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

d. Penyatuan *script data model* Python ke *web app*.
   Script data model diambil dari file Jupyter Notebook yang sudah dikembangkan dan digabungkan dalam *web app* pada Pycharm.
e. Review dan *enhancement web app* sesuai *request customer.*

Berhubung produk merupakan *web app*, *Startup* akan menentukan apakah produk digunakan secara lokal atau akan di-*host* secara *live*.



Gambar 3. Bagan Alur Pengembangan Proyek

3. HASIL DAN PEMBAHASAN

3.1 Perancangan Luaran Kegiatan

3.1.1 *Business Understanding*

Pada tahap ini, penulis melakukan pemahaman terhadap kebutuhan bisnis yang dimiliki dan target yang hendak dicapai. *Startup early stage* yang dipilih sebagai studi kasus ini memiliki permasalahan dalam melakukan identifikasi segmen dalam pasar. *Startup* ini sebelumnya telah melakukan wawancara kepada 51 responden dari beberapa segmen pasar yang hendak dicapai. Segmen demografis dibagi menjadi SMA/SMK, SMP, dan mahasiswa. Adapun segmen psikografis dibagi menjadi dua, yaitu memiliki ketertarikan ke bidang IT dan tidak memiliki ketertarikan di bidang IT. Dengan mengetahui segmen-segmennya, startup mampu memilih segmen yang lebih menguntungkan untuk disasar. Analisa terhadap hasil wawancara sebelumnya dilakukan secara manual. Namun untuk kepentingan jangka panjang, di mana *startup* akan melakukan penyebaran kuesioner secara rutin, diharapkan adanya suatu tools yang dapat membantu dalam analisa segmen pasar. Sebagai startup di bidang pendidikan IT, startup hendak mengetahui bagaimana segmentasi pasar yang tertarik mempelajari IT. Target ini yang akan dijadikan patokan dalam proses pengembangan tools.

3.1.2 *Data Understanding*

Data yang digunakan adalah jawaban wawancara terhadap 51 responden yang sebelumnya sudah diperoleh startup melalui wawancara riset pemasaran. Data ini disimpan dalam bentuk excel. Berikut temuan yang diperoleh dari evaluasi dan eksplorasi terhadap data tersebut.

a. Untuk melakukan eksplorasi, penulis mengganti nama *header* sehingga lebih mudah diakses. Penamaan *header* dapat dilihat pada Tabel 1.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 1. Penamaan *Header*

Nama *Header*	Penjelasan
Unnamed	Penomoran terhadap baris yang dimiliki data
nama	Data kategoris yang berisi nama responden
umur	Data kategoris yang berisi umur responden
tempat_tinggal	Data kategoris yang berisi tempat tinggal responden
pendidikan_terakhir	Data kategoris yang berisi pendidikan terakhir yang ditempuh responden
tertarik_belajar_it	Data kategoris yang berisi ketertarikan responden untuk mempelajari IT (Ya jika tertarik, Tidak jika tidak tertarik, Mungkin jika ragu tertarik atau tidak tertarik)
pernah_belajar_it	Data kategoris yang berisi pengalaman mempelajari IT yang dimiliki responden (Ya jika pernah mempelajari, Tidak jika tidak pernah mempelajari)
jurusan	Data kategoris yang berisi jurusan yang diambil responden ketika menempuh pendidikan terakhir
tertarik_bidang_it	Data kategoris yang berisi apakah bidang ketertarikan responden berhubungan dengan IT (Ya jika berhubungan, Tidak jika tidak berhubungan, Mungkin jika ragu berhubungan atau tidak berhubungan)
literasi_digital	Data kategoris yang berisi apakah responden pernah memiliki literasi digital atau tidak (Ya jika ada, Tidak jika tidak ada)
ketertarikan_jurusan_it	Data kategoris yang berisi ketertarikan responden untuk mempelajari IT secara formal melalui jurusan IT (Ya jika tertarik, Tidak jika tidak tertarik)

b. Data memiliki dimensi 51 baris dan 11 kolom dengan rincian sebagai berikut.

Tabel 2. Informasi *Dataset*

No	Kolom	Jumlah *Non-Null Value*	Tipe Data
0	Unnamed: 0	51	Int64
1	nama	51	*Object*
2	umur	51	*Object*
3	tempat_tinggal	51	*Object*
4	pendidikan_terakhir	51	*Object*
5	tertarik_belajar_it	51	*Object*
6	pernah_belajar_it	51	*Object*
7	jurusan	51	*Object*
8	tertarik_bidang_it	51	*Object*
9	literasi_digital	51	*Object*
10	ketertarikan_jurusan_it	49	*Object*

c. Terdapat dua *null value* pada atribut "ketertarikan_jurusan_it".
d. Data responden memiliki atribut yang tidak diperlukan dalam analisa ini, yaitu kolom nama dan kolom indeks yang berisi nomor urut. Kedua kolom ini tidak memiliki keterkaitan terhadap analisa data sehingga dihapus.
e. Atribut data responden mengandung data kategoris. Untuk melakukan analisa korelasi, digunakan metode *chi-square test of independence* yang digunakan untuk menganalisa frekuensi antara dua variabel yang memiliki satu kategori atau lebih. *Chi-square test of independence* mampu menentukan apakah kedua variabel tersebut independen satu sama lain [15]. Oleh karena target *startup* difokuskan pada atribut "tertarik_belajar_it", *chi-square test* dilakukan untuk membandingkan atribut lain dengan atribut "tertarik_belajar_it".
f. *Chi-square test* yang dilakukan menggunakan nilai alpha = .05 dan ditemukan adanya korelasi antara kolom "pernah_belajar_it" dengan "tertarik_belajar_it", "literasi_digital" dengan "tertarik_belajar_it", "ketertarikan_jurusan_it" dengan "tertarik_belajar_it", "jurusan" dengan "tertarik_belajar_it", serta "tertarik_bidang_it" dengan "tertarik_belajar_it".

Berdasarkan eksplorasi yang telah dilakukan, didapatkan kesimpulan hasil analisa sebagai berikut.

a. Ketertarikan untuk mempelajari IT memiliki korelasi dengan pengalaman pembelajaran IT melalui ada-tidaknya literasi digital ataupun pembelajaran IT secara formal di institusi.
b. Ketertarikan untuk mempelajari IT memiliki korelasi dengan ketertarikan terhadap jurusan spesifik di bidang IT ataupun bidang IT secara keseluruhan.
c. Ketertarikan untuk mempelajari IT juga memiliki korelasi dengan jurusan yang diambil oleh responden.
d. Pendidikan terakhir dan umur yang telah diambil oleh responden tidak memiliki korelasi terhadap ketertarikan untuk mempelajari IT.
e. Terdapat dua *null value* pada kolom "ketertarikan_belajar_it".
f. Terdapat dua kolom yang tidak relevan, yaitu kolom "nama" dan kolom pertama yang berisikan nomor indeks.

Berdasarkan kesimpulan di atas, berikut strategi yang akan diterapkan penulis pada tahap-tahap selanjutnya.

a. Penulis harus membersihkan *null value* pada kolom "tertarik_belajar_it".
b. Penulis harus menghapus kolom "nama" dan kolom pertama yang berisikan nomor indeks.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

c. Penulis akan memunculkan *cluster* berdasarkan dua aspek, yaitu tertarik atau tidak tertarik untuk mempelajari IT dengan atribut pengalaman pembelajaran IT, ketertarikan terhadap jurusan IT, ketertarikan terhadap bidang-bidang IT, dan jurusan dari pendidikan yang sedang diambil saat ini.

3.1.3 *Data Preparation*

Data yang sudah dieksplorasi pada tahap sebelumnya dibersihkan dan dinilai dimensionalitas datanya. Tahap pembersihan yaitu dengan menghapus atau mengganti *null value* atau nilai data yang *error*. Selanjutnya, dengan menilai dimensionalitas data, dapat ditentukan atribut data yang perlu dan tidak perlu dihilangkan. Berikut tahapan yang dilakukan berdasarkan hasil eksplorasi data.

a. Menghapus kolom "nama" dan kolom pertama yang berisikan nomor indeks;
b. Menghapus kolom "umur", "tempat_tinggal", dan "pendidikan_terakhir" yang tidak diperlukan untuk melakukan data modelling. Kolom "tertarik_belajar_it", "pernah_belajar_it", "jurusan", "literasi_digital", dan "ketertarikan_jurusan_it" akan digunakan dalam pemodelan dikarenakan korelasi yang dimiliki terhadap segmen yang dicapai;
c. Membersihkan *null value* pada kolom "ketertarikan_belajar_it";
d. Mengurutkan nomor indeks yang teracak pada data karena proses pembersihan null value;
e. Melakukan *encoding* pada seluruh data kategoris sehingga dapat diproses oleh algoritma K-Means. *Encoding* dilakukan dengan menggunakan fungsi LabelEncoder() pada gambar berikut.

```
mk = LabelEncoder()
df['tertarik_belajar_it'] = mk.fit_transform(df['tertarik_belajar_it'])
df['pernah_belajar_it'] = mk.fit_transform(df['pernah_belajar_it'])
df['jurusan'] = mk.fit_transform(df['jurusan'])
df['tertarik_bidang_it'] = mk.fit_transform(df['tertarik_bidang_it'])
df['literasi_digital'] = mk.fit_transform(df['literasi_digital'])
df['ketertarikan_jurusan_it'] = mk.fit_transform(df['ketertarikan_jurusan_it'])
df
```

Gambar 4. Fungsi LabelEncoder()

Encoding ini dilakukan untuk mengubah data kategoris menjadi data numerik yang mampu diproses algoritma K-Means. Data kategoris akan ditandai dengan bilangan real sesuai dengan *unique values* yang ada.

3.1.4 *Data Modelling*

Identifikasi segmen pasar dilakukan dengan proses clustering menggunakan algoritma K-Means. *Data model* dirancang dan dimodifikasi dari *Notebook* Kaggle oleh Luiz Bueno dengan akun juniorbueno [17]. Data yang digunakan adalah data riset pasar yang dilakukan sebuah perusahaan mobil. Dalam menjalankan model, *Notebook* menggunakan *elbow method* untuk menentukan jumlah *cluster*. *Elbow method* memetakan jumlah *cluster*. Grafik akan menurun seiring bertambahnya jumlah *cluster* hingga hasil K-Means dengan jumlah *cluster* tertentu cenderung stabil. Hal ini membentuk suatu siku pada suatu angka jumlah *cluster*. Angka ini digunakan untuk menentukan nilai K atau jumlah *cluster* yang paling optimal [18]. Seluruh parameter lainnya dibiarkan *default* dan *Notebook* hanya mendefinisikan jumlah *cluster*. Sayangnya, *Notebook* ini tidak mencantumkan tahap evaluasi kinerja model.

Dengan pertimbangan seperti di atas, penulis melakukan modifikasi terhadap *Notebook* dalam penelitian ini sebagai berikut.

a. Menambahkan function *convert* excel ke CSV untuk mengakomodasi input excel dari pengguna. Hal ini dilakukan sebagai respon terhadap kebutuhan *startup early stage*.
b. Menyesuaikan metode *cleaning data* dengan dataset yang ada dan melakukan pengurutan terhadap *indexing* pada dataset sehingga mempermudah untuk menghasilkan hasil analisa dalam bentuk excel.
c. Membuat duplikat *dataset* setelah dilakukan *cleaning data*. Duplikat *dataset* digunakan untuk di-output sebagai hasil analisa dalam bentuk excel.

3.1.5 *Evaluation*

Evaluasi dilakukan terhadap performa data model dengan menggunakan *silhouette coefficient* atau *silhouette score*. *Silhouette coefficient* dinilai berdasarkan *score* yang didapatkan, dari rentang nilai -1 sampai 1. Jika nilai rata-rata mendekati nilai 1, maka *clustering* dianggap semakin baik. Jika nilai rata-rata mendekati -1, maka *clustering* dianggap tidak baik. Berikut kriteria *silhouette coefficient* [19].

Tabel 3. Kriteria *Silhouette Coefficient*

Silhouette Coefficient	**Kriteria Penilaian**
0.7 <SC<=1.0	Strong Structure
0.5 <SC<=0.7	Medium Structure
0.25 <SC<=0.5	Weak Structure
SC<=0.25	No Structure

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 4, Agustus 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i4.4486
Hal 966−973
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Dengan menggunakan *elbow method* untuk penentuan *cluster* dalam *data model*, ditemukan bahwa nilai *silhouette coefficient* cenderung lebih rendah dari 0.4. Oleh karena itu, metode penentuan *cluster* dengan menggunakan *silhouette score method*. Berikut hasil perbandingan *silhouette coefficient* dengan dua metode di atas.

Tabel 4. Perbandingan Metode Penentuan Jumlah *Cluster*

Metode Penentuan Jumlah *Cluster*	Jumlah *Cluster* Optimal	*Silhouette Coefficient Value*
Silhouette Method	9	0.47
Elbow Method	3	0.33

Berdasarkan penentuan *cluster* yang dilakukan di atas, didapatkan bahwa jumlah *cluster* optimal adalah 9, dengan nilai *silhouette coefficient* paling tinggi, yaitu 0.46 dalam proses penentuan dan 0.47 pada tahap evaluasi. Berdasarkan kriteria penilaian, 0.46-0.47 menandakan struktur yang lemah pada *cluster*. Hal ini akan menjadi catatan untuk pengembangan selanjutnya. Untuk penelitian kali ini, metode penentuan *cluster* untuk model diubah dengan menggunakan *silhouette coefficient*. Penulis juga menambahkan variabel yang menampung jumlah *cluster* yang ditentukan dari *silhouette coefficient* sehingga mampu menyesuaikan seiring adanya penambahan data dari pengguna.

Setelah melakukan perubahan pada alur model, selanjutnya penulis memastikan kembali seluruh proses sudah benar dan tidak ada proses yang belum terkoneksi ataupun *error*. Dengan memastikan proses sudah lancar dan dapat digunakan, model dapat dilanjutkan ke tahap *Deployment*.

3.1.6 *Deployment*

Tahap *deployment* dilakukan menggunakan *web app*. Pengembangan dilakukan menggunakan *Python* dengan *microservice Flask*. *Web app* bertujuan untuk menerima *input dataset* dari pengguna dan memberi *output* hasil *clustering* dan bagannya.

Pengembangan *web app* dimulai dengan melakukan desain tampilan menggunakan Figma. *Web app* tergolong cukup sederhana, yaitu hanya mencakup tiga halaman serta fitur *create*, *read*, dan *update*. *Web app* tidak didesain untuk menyimpan data pemrosesan dalam bentuk basis data, namun tetap menyimpan *file* Excel dan CSV yang di-*input*. *Web app* juga menyediakan fitur untuk mengunduh *image pie chart* serta Excel hasil *clustering* yang dihasilkan. Dengan begitu, *startup early stage* dapat mempelajari pengelompokan dan atribut dari *cluster* yang ditargetkan. Penulis melakukan pengembangan *front-end* dan *back-end* menggunakan *Pycharm*. Dalam proses pengembangan, terdapat penyesuaian pada desain awal menjadi tampilan yang lebih sederhana. Eksekusi dan *testing web app* dilakukan secara lokal menggunakan *Pycharm*.

3.2 Implementasi

Setelah melakukan deployment dan menyelesaikan *tools* secara keseluruhan, penulis melakukan implementasi luaran. Proses implementasi dimulai dengan melakukan demo hasil analisa data, *web app* yang dihasilkan, serta rencana pengembangan lebih lanjut ke depannya kepada pihak *startup*. *Improvement* lebih lanjut akan dilakukan pada iterasi selanjutnya.

4. KESIMPULAN

Untuk membantu *startup early stage* dalam melakukan segmentasi pasar, penulis melakukan *data analytics* terhadap data wawancara yang dimiliki, serta merancang dan mengembangkan aplikasi berbasis web yang menggunakan algoritma K-Means. Pengembangan aplikasi menggunakan kerangka kerja CRISP-DM, yaitu: (1) *Business Understanding*; (2) *Data Understanding*; (3) *Data Preparation*; (4) *Data Modelling*; (5) *Evaluation*; dan (6) *Deployment*. Aplikasi ini mampu menerima input *dataset* dari Excel dan memberi *output* berupa *image* berisi visualisasi distribusi frekuensi segmen dalam diagram *pie* dan excel yang berisi segmentasi beserta atributnya. *Data model* yang diterapkan menggunakan algoritma K-Means dan penentuan jumlah *cluster* dilakukan dengan *silhouette method*. Adapun evaluasi performa *data model* dilakukan dengan menggunakan *silhouette coefficient* dan didapatkan nilai sebesar 0.47, dibanding penggunaan *elbow method* yang menhasilkan nilai sebesar 0.33. Dengan aplikasi ini, *startup* mampu mengetahui segmen dalam data calon pelanggan serta dataset lain yang diinput sesuai *template* yang tersedia. *Startup* mampu menganalisa atribut yang membentuk segmen tersebut. Penulis menyadari masih banyak aspek yang dapat diperbaiki dan dikembangkan. Oleh karena itu, penulis menyarankan agar penelitian selanjutnya dapat meningkatkan efisiensi dan keakuratan model. Sistem dapat dikembangkan dari sisi desain tampilan, algoritma, pemilihan jumlah cluster, dan analisa terhadap data, serta menambahkan fitur decision-making atau *predictive analysis*. Penulis juga menyarankan agar perusahaan *startup early stage* dapat mempertimbangkan penggunaan *machine learning* untuk mendukung perkembangannya, misalnya pada riset pemasaran hingga tugas administrasi.

REFERENCES

[1] Z. Acs, L. Szerb, and A. Lloyd, “The Global Enterpreneuship Index 2018,” Washington D.C., 2018. [Online]. Available: http://thegedi.org/wp-content/uploads/dlm_uploads/2017/11/GEI-2018-1.pdf.

[2] Z. J. Ács, L. Szerb, E. Lafuente, and G. Márkus, “The Global Entrepreneurship Index 2019,” Washington D.C., 2019. doi: 10.13140/RG.2.2.17692.64641.

[3] N. Bosma, S. Hill, A. Ionescu-Somers, D. Kelley, M. Guerrero, and T. Schott, "2020/2021 Global Report," London, 2021. [Online]. Available: https://www.gemconsortium.org/report/gem-20202021-global-report.
[4] R. T. P. B. Santoso, I. W. R. Junaedi, S. H. Priyanto, and D. S. S. Santoso, "Creating a startup at a University by using Shane's theory and the entrepreneural learning model: a narrative method," *J. Innov. Entrep.*, vol. 10, no. 1, 2021, doi: 10.1186/s13731-021-00162-8.
[5] R. A. M. Kencanasari and W. Dhewanto, "Digital Startups Fundamental Capabilities in New Product Development: Multiple Case Studies in Bandung, Indonesia," *J. Manaj. Indones.*, vol. 22, no. 1, p. 62, 2022, doi: 10.25124/jmi.v22i1.3286.
[6] M. A. Camilleri, "Market Segmentation, Targeting and Positioning," *Travel Mark. Tour. Econ. Airl. Prod.*, no. 4, pp. 69–83, 2018, doi: 10.1108/978-1-78635-746-520161006.
[7] J.-H. Chen, T. Ji, M.-C. Su, H.-H. Wei, V. T. Azzizi, and S.-C. Hsu, "Swarm-inspired data-driven approach for housing market.pdf," *J. Hous. Built Environ.*, vol. 36, pp. 1787–1811, 2021, doi: https://doi.org/10.1007/s10901-021-09824-1.
[8] E. Y. L. Nandapala and K. P. N. Jayasena, "The practical approach in Customers segmentation by using the K-Means Algorithm," *2020 IEEE 15th Int. Conf. Ind. Inf. Syst. ICIIS 2020 - Proc.*, no. 978, pp. 344–349, 2020, doi: 10.1109/ICIIS51140.2020.9342639.
[9] I. G. A. W. Putri and I. B. G. Dwidasmara, "Application of the K-Means Algorithm to Segmentation of Consumer Interest in Silver Craft 'Kreasi Slaka Bali,'" *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 4, p. 541, 2021, doi: 10.24843/jlk.2021.v09.i04.p12.
[10] S. Kimiagari, S. Keivanpour, and M. Haverila, "Developing a high-performance clustering framework for global market segmentation and strategic profiling," *J. Strateg. Mark.*, vol. 29, no. 2, pp. 93–116, 2021, doi: 10.1080/0965254X.2019.1628099.
[11] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *J. City Dev.*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.
[12] X. Pu, N. Qi, and J. Huang, "Data Analysis and Application of Retail Enterprises Based on Knime," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 782, no. 5, 2020, doi: 10.1088/1757-899X/782/5/052030.
[13] E. Bakhshizadeh, H. Aliasghari, R. Noorossana, and R. Ghousi, "Customer Clustering Based on Factors of Customer Lifetime Value with Data Mining Technique (Case Study: Software Industry)," *Int. J. Ind. Eng. Prod. Res.*, vol. 33, no. 1, pp. 1–16, 2022, doi: 10.22068/ijiepr.33.1.1.
[14] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
[15] K. Black *et al.*, *Business Analytics and Statistics*. 2019.
[16] D. Astuti, "Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 60–72, 2019, doi: 10.20895/inista.v1i2.71.
[17] Luiz Bueno, "Customer/K-Means/Hierarchical Grouping/DBSCAN | Kaggle." Kaggle.com, 2022, Accessed: Jun. 24, 2022. [Online]. Available: https://www.kaggle.com/code/juniorbueno/customer-k-means-hierarchical-grouping-dbscan.
[18] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.
[19] R. A. Farissa, R. Mayasari, and Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient," vol. 5, no. 2, pp. 109–116, 2021.