Esai Use Case Metode CRISP-DM

Repository Github: https://github.com/rizps/k-means_in-sales

Sumber jurnal: http://stmik-budidarma.ac.id/ejurnal/index.php/jurikom/article/view/4486/2896

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

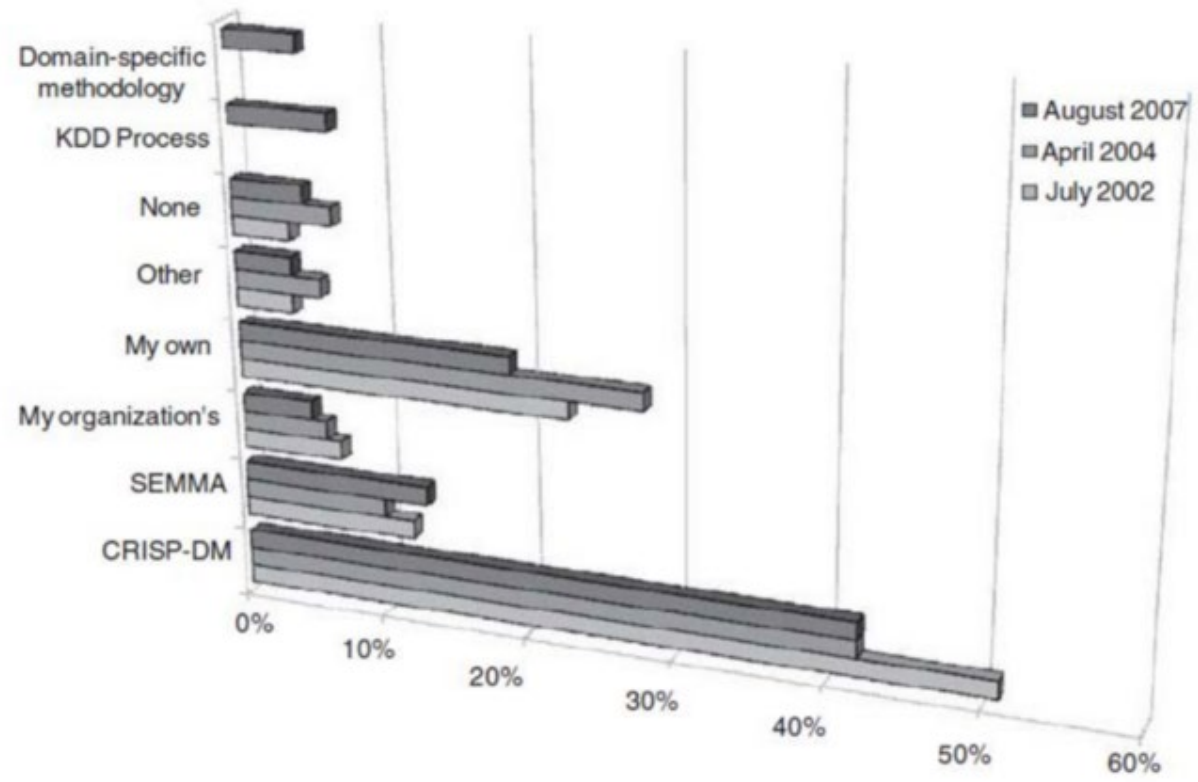
Penulis: **Yefta Christian* dan Katherine Oktaviani Yap Rui Qi**

Pengertian

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah salah satu model proses datamining (*datamining framework*) yang awalnya (1996) dibangun oleh 5 perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA. Framework ini kemudian dikembangan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan *methodology standard non-proprietary* bagi *data mining*. Versi pertama dari methodologi ini dipresentasikan pada 4th CRISP-DM SIG Workshop di Brussels pada bulan Maret 1999 (Pete Chapman, 1999); dan langkah langkah proses datamining berdasarkan model ini di publikasikan pada tahun berikutnya (Pete Chapman,2000).

Antara tahun 2006 dan 2008 terbentuklah grup CRISP-DM 2.0 SIG yang berkeinginan untuk mengupdate CRISP-DM process model (Colin Shearer, 2006). Namun produk akhir dari inisiatip ini tidak diketahui.

Banyak hasil penelitian yang mengungkapkan bahwa CRISP-DM adalah datamining model yang masih digunakan secara luas di kalangan industry, sebahagian dikarenakan keunggulannya dalam menyelesaikan banyak persoalan dalam proyek proyek data mining.

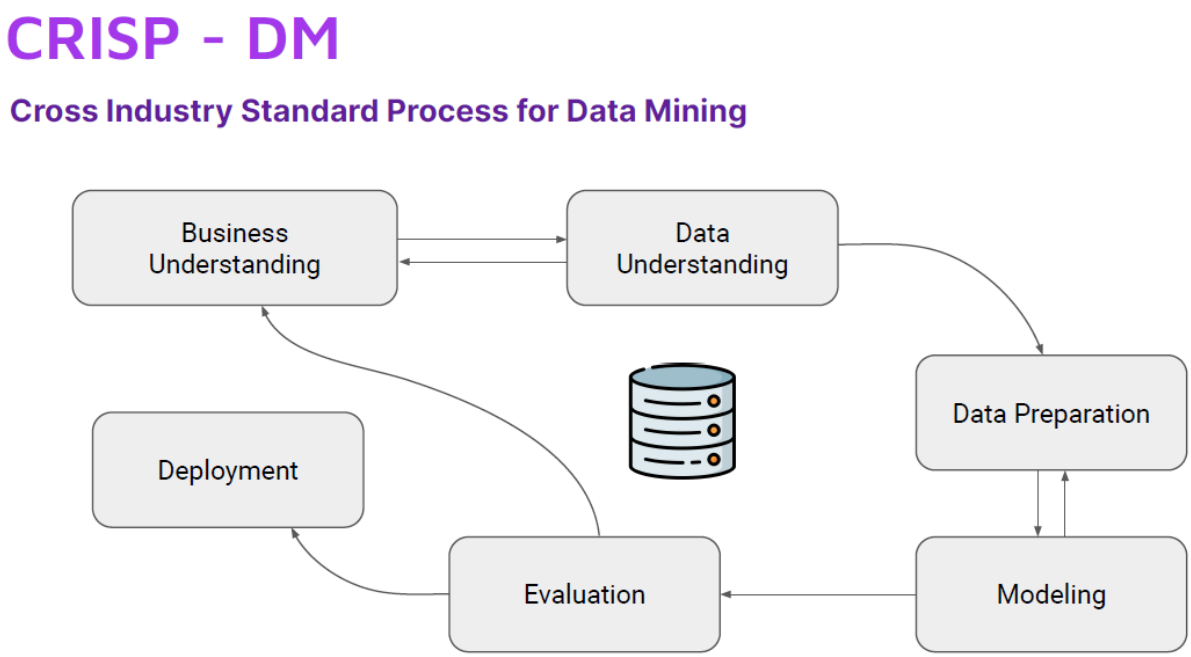


Gambar 1. Survei Penggunaan Metodologi Data Mining (Mariscal, Marban, and Fernandez 2010)

Mariscal, Marba dan Fernandez (Mariscal, Marban, and Fernandez 2010) menyatakan CRISP-DM sebagai *defacto* menjadi standar untuk pengembangan proyek *data mining* dan *knowledge discovery* karena paling banyak digunakan dalam pengembangan data mining. Hal tersebut dapat terlihat dari survei yang ditunjukkan pada Gambar 1 yang dilakukan terhadap penggunaan metodologi dalam proyek *data mining*.

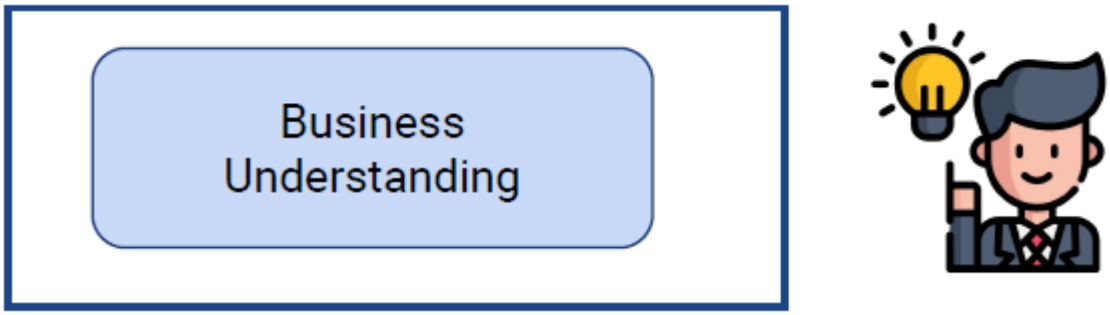
Hasil survei "Penggunaan Metodologi dalam Proyek *Data Mining*", memperlihatkan pengguna CRISP-DM di tahun 2002 mencapai 51%, kemudian menurun menuju 41% di tahun 2004. Meskipun persentasi penggunaan CRISP-DM menurun 10%, jumlah pengguna metodologi ini masih terbilang lebih banyak daripada pengguna metodologi lain.

Model proses CRISP-DM memberikan gambaran tentang siklus hidup proyek data mining. CRISP-DM memiliki 6 tahapan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* seperti ditunjukkan pada Gambar 2.



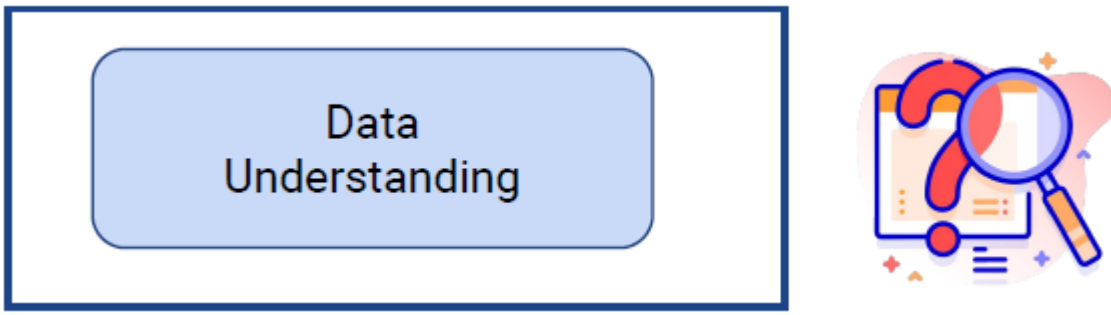
Gambar 2. Tahapan CRISP-DM

1. Business Understanding



Tahapan ini memerlukan pemahaman tentang tujuan dan persyaratan proyek dari sudut pandang bisnis. Perspektif bisnis seperti itu digunakan untuk mencari tahu masalah bisnis apa yang harus dipecahkan melalui penggunaan data mining.

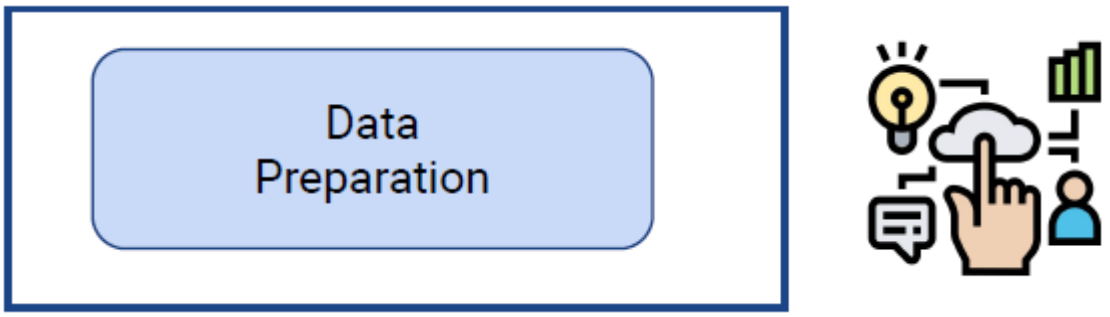
2. Data Understanding



Tahap ini memungkinkan kita untuk membiasakan diri dengan data dan ini melibatkan melakukan analisis data eksplorasi. Eksplorasi data awal tersebut memungkinkan kita untuk

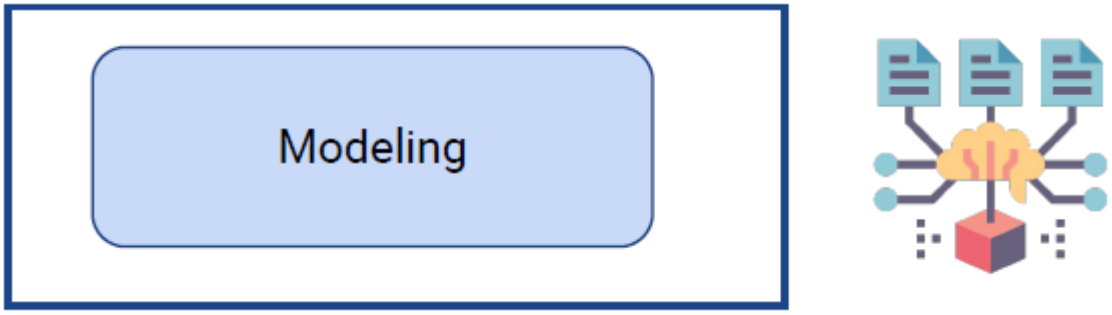
mengetahui subset data mana yang akan digunakan untuk pemodelan lebih lanjut serta membantu dalam menghasilkan hipotesis untuk dijelajahi.

3. Data Preparation



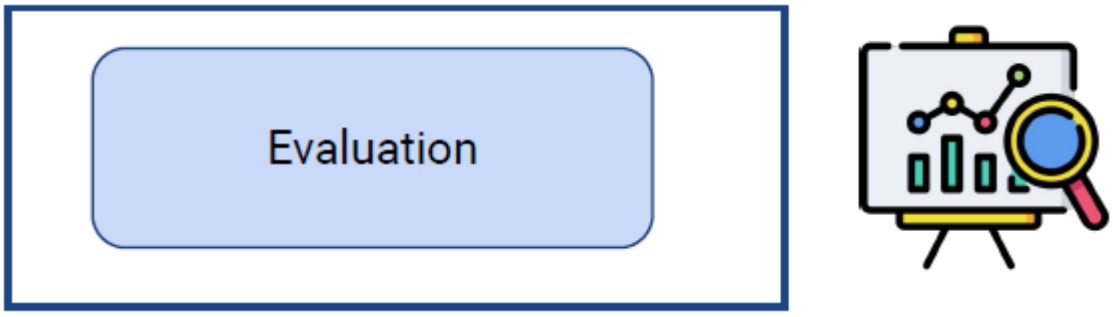
Tahapan ini dapat dianggap sebagai fase yang paling memakan waktu dari proses data mining karena melibatkan pembersihan data yang ketat dan pra-pemrosesan serta penanganan data yang hilang.

4. Modelling



Data preparation digunakan untuk membangun model di mana algoritma pembelajaran digunakan untuk melakukan analisis multivariat. Ulangi pembuatan dan penilaian model sampai Anda sangat yakin bahwa Anda telah menemukan model terbaik.

5. Evaluation

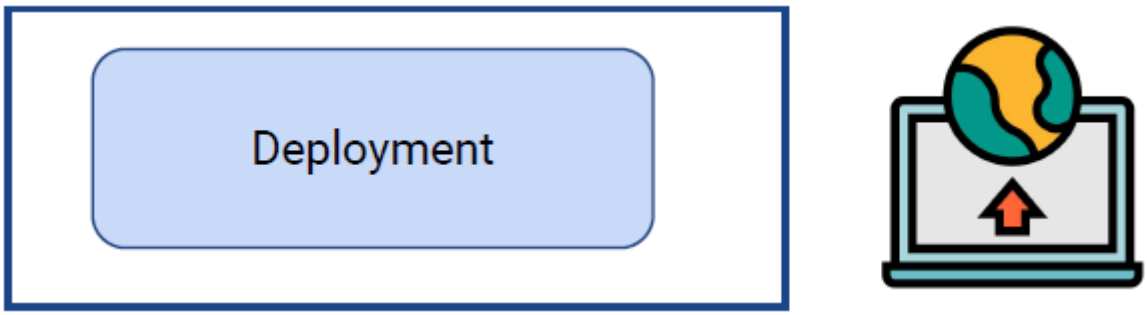


Penting untuk mengevaluasi hasil model dan meninjau proses yang dilakukan untuk menentukan apakah tujuan bisnis yang ditetapkan semula terpenuhi atau tidak.

Jika dianggap tepat, beberapa langkah mungkin perlu dilakukan lagi. Bilas dan ulangi. Setelah dirasa hasil dan prosesnya memuaskan maka kita siap untuk pindah ke deployment.

Selain itu, dalam tahap evaluasi ini, beberapa temuan dapat memicu ide-ide proyek baru untuk dieksplorasi.

6. Deployment



Setelah model memiliki kualitas yang memuaskan, model tersebut kemudian disebarkan, yang dapat berupa laporan sederhana, API yang dapat diakses melalui panggilan terprogram, aplikasi web, dll.

Pembahasan use case pada jurnal

Pendahuluan

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 metode CRISP-DM.

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*.

Perolehan data

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.

Ada 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.

Perancangan kegiatan

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. Telah ada beberapa segmen yang diperoleh dari survei sebelumnya.

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.

Data Preparation

Data wawancara riset pemasaran disimpan dalam bentuk excel. Berikut temuan yang diperoleh dari evaluasi dan eksplorasi terhadap data tersebut.

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*

Dalam tahap ini penulis menemukan bahwa 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. Oleh karena target startup difokuskan pada atribut no 5, chi-square test dilakukan untuk membandingkan atribut lain dengan atribut no 5.

Chi-square test yang dilakukan menggunakan nilai alpha = .05 dan ditemukan adanya korelasi antara kolom no 6 dengan kolom no 5, kolom no 9 dengan kolom no 5, kolom no 10 dengan kolom no 5, kolom no 7 dengan kolom no 5, serta kolom no 8 dengan kolom no 5.

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.

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.

Pada tahap ini penulis melakukan beberapa hal:

Membersihkan *null value* pada kolom no 10.

Menghapus kolom “unnamed”, “nama”, “umur”, “tempat_tinggal”, dan “pendidikan_terakhir” yang tidak diperlukan untuk melakukan data modelling.

Mengurutkan nomor indeks yang teracak pada data karena proses pembersihan null value

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
```

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.

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. 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.

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.

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*.

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

Dibandingankan dengan metode elbow, ternyata silhoutte lebih unggul, berikut adalah perbandingannya.

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

Dikarenakan keunggulan metode silhoutte dibandingkan metode elbow, maka 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*.

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*.

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.