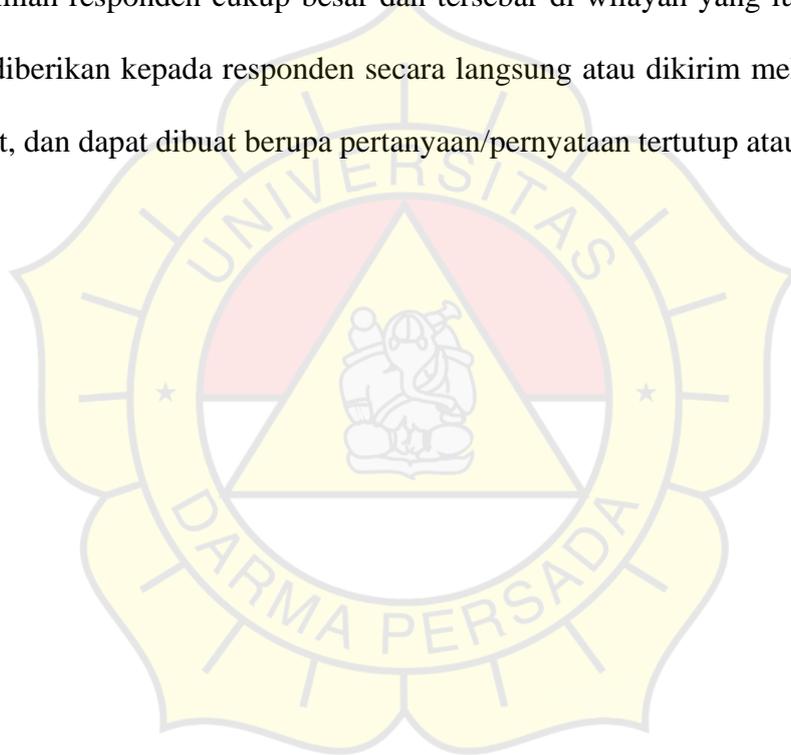


BAB II

LANDASAN TEORI

2.1 Kuesioner

Menurut (Dr. Sulaiman Saat, M.Pd., dkk., 2020) Kuesioner (angket) merupakan teknik pengumpulan data dengan cara memberikan seperangkat pertanyaan atau pernyataan tertulis kepada responden untuk dijawab. Kuesioner cocok digunakan bila jumlah responden cukup besar dan tersebar di wilayah yang luas. Kuesioner dapat diberikan kepada responden secara langsung atau dikirim melalui pos, atau internet, dan dapat dibuat berupa pertanyaan/pernyataan tertutup atau terbuka.



Hal yang perlu diingat oleh peneliti adalah bahwa angket hanya bisa digunakan untuk memperoleh data berupa pendapat atau persepsi responden tentang suatu masalah, dan penggunaannya tidak dengan berhadapan langsung (face to face) dengan responden. Pembuatannya, dapat berupa angket tertutup, terbuka atau semi tertutup terbuka (jawaban sudahkan disiapkan oleh peneliti, tapi responden diberi kesempatan untuk menambahkan jawaban lain).

Angket tertutup adalah yang sudah disiapkan jawabannya oleh peneliti dan tidak diberi kemungkinan atau kesempatan kepada responden untuk memberikan jawaban selain yang sudah disediakan. Angket terbuka adalah pertanyaan/ Pernyataan yang tidak disiapkan jawabannya, dan memberi peluang kepada responden untuk memberikan jawaban secara bebas dalam bentuk uraian tentang suatu hal.

2.2 Text Mining

Menurut Feldman dan James Sanger (2007) dalam buku *The Text mining Handbook*, text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools anylisis yang merupakan komponen-komponen data mining yang salah satunya kategorisasi. (Feldman & Sanger, 2007) Text mining memberikan solusi dari permasalahan

seperti pemrosesan, pengorganisasian/pengelompokan dan menganalisa unstructured text dalam jumlah besar. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen dan mendukung proses knowledge discovery pada koleksi dokumen yang besar, jadi sumber data yang digunakan pada text mining adalah kumpulan text yang memiliki format yang tidak teratur atau minimal semi teratur. Tahapan-tahapan dalam text mining secara umum adalah text preprocessing. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen, jadi sumber data yang digunakan dalam text mining adalah sekumpulan text yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining adalah pengkategorisasian teks dan pengelompokan teks (Feldman, R., & Sanger, J. 2007).

Hotho (2005) juga mengatakan bahwa pada penelitian text mining diperlukan tahapan text pre-processing pada koleksi dokumen dan menyimpan informasi tersebut dalam struktur data. Pendekatan text mining didasarkan pada pemikiran bahwa dokumen teks dapat diwakili oleh satu set kata-kata, yaitu 12 dokumen teks digambarkan berdasarkan pada set kata-kata yang terkandung di dalamnya.

2.2.1 Text Processing

Merupakan tahapan pemrosesan data agar menjadi data yang siap untuk dianalisis. Setelah data terstruktur maka dapat diolah lebih lanjut. Beberapa proses yang dilakukan pada tahap ini adalah sebagai berikut :

i. Case folding

Merupakan proses perubahan semua huruf pada dokumen kuesioner menjadi huruf kecil. Hanya huruf a sampai z diproses. Karakter selain huruf akan dibiarkan. Contoh : Sb menambah wangi segar !! Menjadi : sb menambah wangi segar.

ii. Tokenizing

Pada tahap ini, kalimat dipotong atau dipecah berdasarkan tiap kata yang menyusunnya. Bekerja untuk mengidentifikasi kata-kata dalam teks menjadi beberapa urutan yang terpotong oleh spasi atau karakter special Berikut contoh penerapan tokenizing.

Tabel 2. 1 Penerapan Stemming (Jurnal , Wilianto, Fajri rahmat umbara, 2017).

Data Latih	
Input	Output
Menginspirasi	Inspirasi

iii. Stopword Removal

Merupakan tahap pembuangan kata-kata yang dianggap tidak penting. Langkah ini dilakukan supaya perhitungan lebih berfokus pada katakata yang jauh lebih penting. Berguna untuk membuang kata-kata yang sering muncul dan bersifat umum, kurang menunjukkan relevansinya dengan teks. Kata-kata yang akan dibuang tersebut didefinisikan dalam stopword list. Contoh beberapa kata yang

iv. Stemming

Tahap stemming adalah tahap mencari root (bentuk dasar) dari tiap kata. Pada tahap ini, dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tabel 2.2 Penerapan Stemming (Jurna, Wilianto, Fajri rahmat umbara, 2017).

2.2.2 Pembobotan Kata

Hal yang perlu diperhatikan dalam pencarian informasi dari koleksi dokumen yang heterogen adalah pembobotan term. Karena setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, maka untuk setiap kata tersebut diberikan sebuah indikator, yaitu term weight (Zafikri, 2008). Menurut Zafikri (2008) term weighting atau pembobotan term sangat dipengaruhi oleh hal-hal berikut ini :

i. Document Frequency (df)

Metode document frequency (df) merupakan salah satu metode pembobotan dalam bentuk sebuah metode yang merupakan perhitungan dokumen yang mengandung suatu term tertentu. Tiap term akan dihitung nilai document frequency (df)

ii. Term Frequency (TF)

Term frequency (TF) yaitu factor yang menentukan bobot term pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Nilai jumlah kemunculan suatu kata (term frequency) diperhitungkan dalam pemberian bobot terhadap suatu kata. Semakin besar jumlah kemunculan suatu term (tf tinggi) dalam dokumen, semakin besar pula bobotnya dalam dokumen atau akan memberikan

nilai kesesuaian yang semakin besar. Berikut rumus Term Frequency (TF) yaitu : $tf = 1 + \log(tf)$

Keterangan :

tf : term frequency

$\log t.f$: jumlah kemunculan kata/term t di dalam dokumen d

iii. Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) yaitu pengurangan dominasi term yang sering muncul diberbagai dokumen. Hal ini diperlukan karena term yang banyak muncul diberbagai dokumen dapat dianggap sebagai term umum sehingga tidak penting nilainya. Sebaliknya, factor kejarang muncul kata dalam kumpulan dokumen harus diperhatikan dalam pemberian bobot. Menurut wittern dan zafikri (2010), kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting dari pada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan factor kebalikan frekuensi dokumen yang mengandung suatu kata (inverse document frequency). Berikut rumus idf :

$$idf_j = \log(D/df_j)$$

Keterangan :

D : adalah jumlah semua dokumen dalam koleksi

df_j : adalah jumlah dokumen yang mengandung term t_j

Lakukan penghitungan bobot kata (term) dari data ada menggunakan metode TFIDF (Term Frequency-Inverse Document Frequency). Term frequency menyatakan frekuensi (tingkat keseringan) munculnya suatu term dalam suatu

dokumen. Sedangkan document frequency adalah banyaknya jumlah dokumen dimana sebuah term itu muncul.

2.2.3 Sentimen Analisis

Berdasarkan penelitian yang dilakukan oleh (Sipayung, Maharani, & Zefanya, 2016) memaparkan bahwa sentimen analisis merupakan suatu metode yang berasal dari gabungan proses penambangan data (data mining), dimana dapat dijadikan suatu metode yang digunakan dengan tujuan untuk memproses berbagai macam pendapat dari pelanggan mengenai suatu layanan, barang komersial, ataupun suatu badan penyedia layanan. Langkah ini dapat dimanfaatkan untuk mengidentifikasi sentimen dalam suatu pendapat atau komentar yang disampaikan pelanggan dari berbagai macam sarana komunikasi. (Darmawan, 2022)

Menurut (Liu, 2012) sentimen analisis merupakan suatu metode pembelajaran secara komputasional yang berkaitan dengan pendapat, sentimen, serta perasaan orang lain yang direpresentasikan dalam bentuk teks, dimana metode ini dapat menentukan suatu kelompok, polaritas didalam teks yang terdapat pada suatu kalimat atau berkas, sehingga setelah melalui proses tersebut dapat diketahui apakah opini dari suatu kalimat atau dokumen tergolong kedalam sentimen positif atau negatif.

Oleh karena itu hasil dari suatu metode ini dapat digunakan sebagai bahan pertimbangan yang berfungsi untuk memperbaiki pengadaan layanan maupun kualitas dari suatu barang agar menjadi lebih baik. Sangat tepat digunakan untuk menyelesaikan masalah-masalah yang sulit serta bergantung pada waktu.

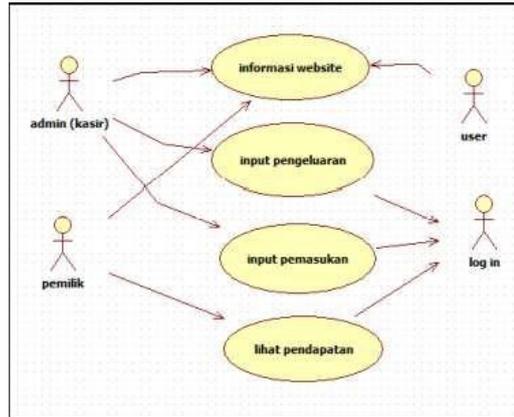
2.3 UML Diagram

Unified Modelling Language (UML) adalah keluarga notasi grafis yang didukung oleh meta-model tunggal, yang membantu pendeskripsian dan desain system perangkat lunak, khususnya sistem yang dibangun menggunakan pemrograman berorientasi objek (OO).

UML merupakan standar yang relative terbuka yang dikontrol oleh object Management group (OMG), sebuah konsorsium terbuka yang terdiri dari banyak perusahaan. OMG mungkin lebih dikenal dengan standar-standar CORBA (Common Object Request Broker Architecture). UML lahir dari penggabungan banyak bahasa pemodelan grafis berorientasi objek yang berkembang pesat pada akhir 1980-an dan awal 1990-an.

2.3.1 Use Case Diagram

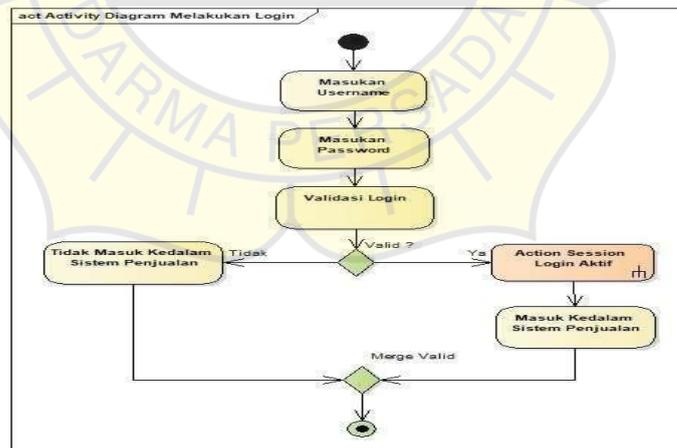
Use case adalah teknik untuk merekam persyaratan fungsional sebuah sistem. Use case mendeskripsikan interaksi tipikal antara para pengguna sistem dengan sistem itu sendiri, dengan memberi sebuah narasi tentang bagaimana sistem tersebut digunakan. Dalam bahasa Use case, para pengguna disebut sebagai actor. Aktor merupakan sebuah peran yang dimainkan seorang pengguna dalam kaitannya dengan sistem. Setiap use case memiliki actor utama yang meminta sistem untuk memberi sebuah layanan. Setiap langkah dalam use case adalah sebuah elemen dalam interaksi antara actor dan sistem. Langkah tersebut harus menunjukkan tujuan actor, bukan mekanisme yang harus dilakukan actor.



Gambar 2.1 Use Case Diagram

2.3.2 Activity Diagram

Activity diagram adalah teknik untuk menggambarkan logika procedural, proses bisnis dan jalur kerja. Dalam beberapa hal, diagram ini memainkan peran mirip sebuah diagram alir, tetapi perbedaan prinsip antara diagram ini dan notasi diagram alir adalah diagram ini mendukung behavior parallel.



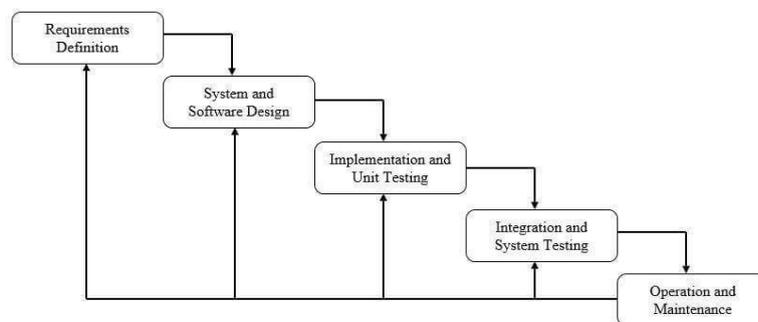
Gambar 2.2 Activity Diagram

Activity diagram telah mengalami beberapa perubahan paling besar selama perkembangan versi – versi UML, jadi tidaklah mengejutkan jika activity diagram telah dikembangkan secara signifikan dan diubah lagi dalam UML.

2.3.3 Sequence Diagram

Interaction diagram menunjukkan bagaimana kelompok – kelompok objek saling berkolaborasi dalam beberapa behavior. UML memiliki beberapa bentuk interaction diagram dan yang paling umum digunakan adalah sequence diagram. Sebuah sequence diagram, secara khusus menjabarkan behavior sebuah cenario tunggal. Diagram tersebut menunjukkan sejumlah objek contoh dan pesan – pesan yang melewati objek – objek ini didalam use case. Sequence diagram sangat bagus untuk memperlihatkan kolaborasi antar objek, tetapi mereka tidak terlalu bagus pada definisi yang rinci tentang behavior.

Saat ini model waterfall merupakan model pengembangan perangkat lunak yang sering digunakan. Model pengembangan ini melakukan pendekatan secara sistematis dan berurutan. Disebut 22 waterfall karena tahap demi tahap yang dilalui harus menunggu selesainya tahap sebelumnya dan berjalan berurutan. Model pengembangan ini bersifat linear dari tahap awal pengembangan sistem yaitu tahap perencanaan sampai tahap akhir pengembangan sistem yaitu tahap pemeliharaan. Tahapan berikutnya tidak akan dilaksanakan sebelum tahapan sebelumnya selesai dilaksanakan dan tidak bisa kembali atau mengulang ke tahap sebelumnya.



Gambar 2.3 Tahapan Metode Waterfall

1. Requirement

Metode pengumpulan informasi ini dapat diperoleh dengan berbagai macam cara diantaranya, diskusi, observasi, survei, wawancara, dan sebagainya. Informasi yang diperoleh kemudian diolah dan dianalisa sehingga didapatkan data atau informasi yang lengkap mengenai spesifikasi kebutuhan pengguna akan perangkat lunak yang akan dikembangkan.

2. System dan Software Design

Perancangan desain dilakukan dengan tujuan membantu memberikan gambaran lengkap mengenai apa yang harus dikerjakan. Tahap ini juga akan membantu pengembang untuk menyiapkan kebutuhan *hardware* dalam pembuatan arsitektur sistem perangkat lunak yang akan dibuat secara keseluruhan.

3. Implementation dan Unit Testing

Perancangan desain dilakukan dengan tujuan membantu memberikan gambaran lengkap mengenai apa yang harus dikerjakan. Tahap ini juga akan membantu pengembang untuk menyiapkan kebutuhan *hardware* dalam pembuatan arsitektur sistem perangkat lunak yang akan dibuat secara keseluruhan.

4. Integration dan System Testing

Setelah proses integrasi selesai, selanjutnya dilakukan pemeriksaan dan pengujian sistem secara keseluruhan untuk mengidentifikasi kemungkinan adanya kegagalan dan kesalahan sistem.

5. Operation dan Maintenance

Pada tahap terakhir dalam Metode Waterfall, perangkat lunak yang sudah jadi dioperasikan pengguna dan dilakukan pemeliharaan. Pemeliharaan memungkinkan pengembang untuk melakukan perbaikan atas kesalahan yang tidak terdeteksi pada tahap-tahap sebelumnya. Pemeliharaan meliputi perbaikan kesalahan, perbaikan implementasi unit sistem, dan peningkatan dan penyesuaian sistem sesuai dengan kebutuhan.

2.4 Algoritma

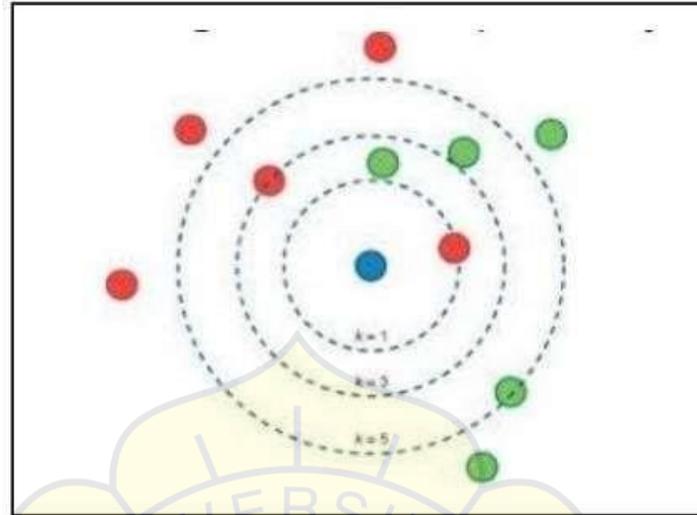
Algoritma merupakan suatu prosedur berupa urutan langkah-langkah yang saling berhubungan, atau suatu metode yang digunakan untuk menyelesaikan suatu permasalahan yang benar-benar terjadi di dunia nyata. Kesimpulannya algoritma adalah susunan langkah-langkah untuk membenahi suatu persoalan dalam pemrosesan kerja computer secara logis, efisien dan terstruktur sehingga dapat memberi solusi yang tepat untuk suatu masalah dengan tepat.

Algoritma terbaik akan menghasilkan output yang benar, tepat guna (efektif) dalam waktu yang relatif singkat dan penggunaan penyimpanan yang relatif sedikit.

2.4.1 Metode Algoritma (KNN)

Metode K-Nearest Neighbor (K-NN) merupakan salah satu dalam top 10 metode data mining yang paling banyak digunakan. Metode ini melakukan klasifikasi berdasarkan kemiripan suatu data dengan data yang lain. Prinsip sederhana metode ini adalah “Jika suatu hewan berjalan seperti bebek, bersuara

kwek-kwek seperti bebek, dan penampilannya seperti bebek, maka hewan itu mungkin bebek”.(Rangga et al., 2019)



Gambar 2. 4 ilustrasi K-Nearest Neighbor

Ilustrasi di atas menjelaskan metode klasifikasi. Pada gambar ini, titik biru akan dijadikan objek diprediksi kelasnya. Untuk $k = 1$ maka kemungkinan objek masuk ke kelas merah. Selanjutnya $k = 3$, objek diprediksi masuk kelas merah, dengan perhitungan 2-1 lebih banyak atas kelas hijau. Untuk $k = 5$, maka akan diprediksi masuk kelas hijau dengan perhitungan 3-2 lebih banyak dari kelas merah(Mitchell B.O., 2014). K-Nearest Neighbor digunakan untuk mengklasifikasi data yang tidak dilabeli. Karakteristik data didapatkan dari training set dan test set(Zhang, 2016). Langkah-langkah klasifikasi data menggunakan K-Nearest Neighbor adalah sebagai berikut:

1. Tentukan nilai K
2. Hitung jarak antara data baru ke setiap label data

3. Tentukan k label data yang mempunyai jarak yang paling minimal
4. Klasifikasikan data baru ke dalam label data yang mayoritas K-NN dipilih berdasarkan metrik jarak.

2.4.2 Metode Algoritma SVM

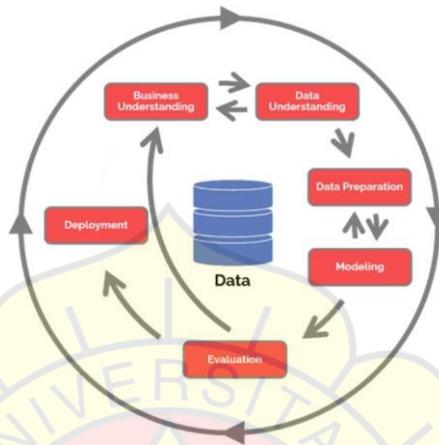
Support Vector Machine (SVM) merupakan salah satu teknik pembelajaran supervised dengan kualitas dan akurasi baik yang membuatnya menjadi algoritma populer diantara algoritma lainnya. Klasifikasi SVM berupaya untuk mempartisi ruang data dengan menggunakan penggambaran linier atau non-linier antara kelas yang berbeda. Dalam geometri, SVM dapat dilihat sebagai hyperlane pada fitur ruang yang memisahkan titik-titik yang mewakili kategori dari hal positif dan dari titik-titik yang mewakili hal negative. (Analisis Dan Penerapan et al., 2019)

Prinsip utama SVM adalah untuk menentukan ruang pemisah di ruang pencarian yang dapat memisahkan kelas-kelas yang berbeda. Proses SVM dimulai dari mengubah data text kedalam bentuk vector data dan dikombinasikan dengan nilai TF-IDF untuk pembobotan. Fungsi deskriptif dari SVM dapat dilihat sebagai berikut : Dimana X merupakan fitur vector, z merupakan vector dari bobot yang berbeda, ϕ merupakan fungsi pemetaan non-linier, dan c adalah vector bias, z dan c diperoleh secara otomatis dari dataset training.

2.5 Pemodelan CRISP-DM

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah salah satu model proses datamining (datamining framework) yang awalnya dibangun oleh 5 perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR

Corporation dan OHRA. Metodologi ini terdiri dari enam tahapan yaitu Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment. Proses metodologi ini terdiri dari 6 tahapan yang dapat dijelaskan sebagai berikut :



Gambar 2. 5 Alur Crisp-dm

1. Business Understanding (Pemahaman Bisnis)

Ini adalah tahap pertama dalam CRISP-DM dan termasuk bagian yang cukup vital. Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokkan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun. Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut. Dalam penelitian ini, penulis menggunakan data kuesioner dari responden mahasiswa unsada

2. Data Understanding (Pemahaman Data)

Secara garis besar untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan (summary) dan mengidentifikasi potensi masalah dalam data. Tahap ini juga harus dilakukan secara cermat dan tidak terburu-buru, seperti pada visualisasi data, yang terkadang insight-nya sangat sulit didapat jika dihubungkan dengan summary data nya. Jika ada masalah pada tahap ini yang belum terjawab, maka akan mengganggu pada tahap modeling. Ringkasan atau summary dari data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan, atau mengungkapkan penyimpangan tak terduga yang perlu ditangani pada tahap selanjutnya, yaitu Data Preparation. Masalah dalam data biasanya seperti nilai-nilai yang hilang, outlier, berdistribusi spike, berdistribusi bimodal harus diidentifikasi dan diukur sehingga dapat diperbaiki dalam Data Preparation. Data kuesioner yang diperoleh dengan memanfaatkan kuesioner

3. Data Preparation (Persiapan Data)

Secara garis besar untuk memperbaiki masalah dalam data, kemudian membuat variabel derived. Tahap ini jelas membutuhkan pemikiran yang cukup matang dan usaha yang cukup tinggi untuk memastikan data tepat untuk algoritma yang digunakan. Bukan berarti saat Data Preparation pertama kali dimana masalah- masalah pada data sudah diselesaikan, data sudah dapat digunakan hingga tahap terakhir. Tahap ini merupakan tahap yang sering ditinjau kembali saat menemukan masalah pada saat pembangunan model. Sehingga dilakukan iterasi sampai menemukan hal

yang cocok dengan data. Tahap sampling dapat dilakukan disini dan data secara umum dibagi menjadi dua, data training dan data testing. Kegiatan yang dilakukan antara lain: memilih kasus dan parameter yang akan dianalisis (Select Data), melakukan transformasi terhadap parameter tertentu (Transformation), dan melakukan pembersihan data agar data siap untuk tahap modeling (Cleaning). Data tweet yang telah diperoleh kemudian diolah dan kemudian dipersiapkan sebagai data set.

4. Modeling (Pemodelan)

Secara garis besar untuk membuat model prediktif atau deskriptif. Pada tahap ini dilakukan metode statistika dan Machine Learning untuk penentuan terhadap teknik data mining, alat bantu data mining, dan algoritma data mining yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap data preparation. Algoritma yang digunakan yaitu SVM dan KNN.

5. Evaluation (Evaluasi)

Melakukan interpretasi terhadap hasil dari data mining yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

6. Deployment (Penyebaran)

Tahap deployment atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk Deployment dimulai selama Business Understanding dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengkonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional. Pada akhirnya, rencana sistem Deployment mengakui bahwa tidak ada model yang statis. Model tersebut dibangun dari data yang diwakili data pada waktu tertentu, sehingga perubahan waktu dapat menyebabkan berubahnya karakteristik data. Model pun harus dipantau dan mungkin diganti dengan model yang sudah diperbaiki.

