

MODUL PERKULIAHAN

PENGANTAR DATA MINING

Penerapan Studi Kasus untuk Tekhnik Data Mining CRISP-DM

**Fakultas Program Studi Tatap Muka Kode MK Disusun Oleh**

Ilmu Komputer Tekhnik Informatika 14  **W151720006** Saruni Dwiasnati,ST.MM

**Abstract Kompetensi**

Memahami dan mampu menjelaskan bagaimana penerapan studi kasus pada Tekhnik CRISP-DM.Mampu memahami Konsep dari Studi Kasus CRISP-DM.

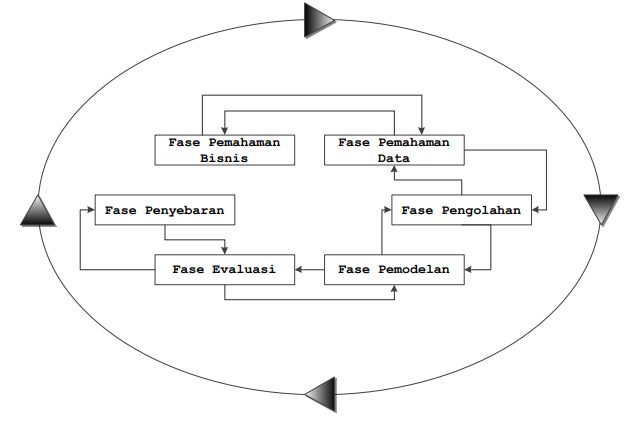
Pengantar Data Mining

* 1. **Tekhnik Data Mining CRISP-DM**

CRIPS-DM merupakan siklus hidup proyek data mining yang digunakan untuk penambangan data yang dapat diterapkan pada strategi pemecahan masalah umum untuk bisnis atau unit penelitian lainnya. Tekhnik ini terbagi dalam 6 fase pengembangan data mining, yaitu:

1. Fase Pemahaman Bisnis (Business Undersanting Phase)
2. Fase Pemahaman Data (Data Understanding Phase)
3. Fase Pengolahan Data (Data Preparation Phase)
4. Fase Pemodelan (Modeling Phase)
5. Fase Evaluasi (Evaluation Phase)
6. Fase Penyebaran (Deployment Phase)

Seperti terlihat pada gambar di bawah ini:



Gambar 1. Proses Data Mining menurut CRIPS-DM

Pengerjaan setiap fase bersifat adaptif dan saling ketergantungan, dimana setiap fase berikutnya bergantung pada hasil dari fase sebelumnya. Fase yang sudah dikerjakan dan belum menemukan hasil yang maksiman mungkin bisa kembali pada fase sebelumnya apabila diperlukan evaluasi atau kesalahan pada fase sebelumnya agar mendapatkan hasil yang dibutuhkan.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 2 | **Pengantar Data Mining**  Saruni Dwiasnati,ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

Berikut penjelasan dari setiap fase CRIPS-DM (Larose, 2005):

1. Fase Pemahaman Bisnis (Business Undersanting Phase)

a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.

b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.

c. Menyiapkan strategi awal untuk mencapai tujuan.

2. Fase Pemahaman Data (Data Understanding Phase)

a. Mengumpulkan data.

b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.

c. Mengevaluasi kualitas data.

d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengandung pola dari permasalahan.

3. Fase Pengolahan Data (Data Preparation Phase)

a. Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.

b. Pilih kasus dan variable yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.

c. Lakukan perubahan pada beberapa variable jika dibutuhkan.

d. Siapkan data awal sehingga siap untuk perangkat pemodelan.

4. Fase Pemodelan (Modeling Phase)

a. Pilih dan aplikasi teknik pemodelan yang sesuai dengan tipe data yang dimiliki.

b. Kalibrasi aturan model untuk mengoptimalkan hasil.

c. Perlu diperhatikan bahwa beberapa teknik mungkin bisa digunakan pada permasalahan data mining yang sama.

d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk mendapatkan hasil data ke dalam bentuk yang sesuai dengan klasifikasi yang dibutuhkan pada teknik data mining tertentu.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 3 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

5. Fase Evaluasi (Evaluation Phase)

a. Mengevaluasi satu atau lebih model yang didapat untuk digunakan dalam fase pemodelan

agar mendapatkan kualitas dan efektivitas sebelum hasilnya tersebut disebarkan untuk digunakan.

b. Menetapkan apakah terdapat model yang memenuhi syarat pada fase awal agar mendapatkan hasil yang maksimal.

c. Menentukan apakah memiliki permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.

d. Mengambil keputusan yang berkaitan dengan penggunaan hasil dari data mining yang diolah.

6. Fase Penyebaran (Deployment Phase)

a. Menggunakan model yang dihasilkan agar terbentuknya model yang menandakan telah terselesaikannya proyek.

b. Pembuatan hasil laporan dari .proyek atau penelitian yang kita kerjakan menggunakan data mining tersebut.

c. Penerapan proses hasil data mining secara parallel pada departemen lain atau objek penelitian lain.

Selain CRIPS-DM terdapat langkah sistematis lainnya yang kurang lebih intinya sama. Berikut tiga langkah utama dalam proses data mining (Gonunescu, 2011):

1. Ekplorasi/pemrosesan awal data Ekploasi/pemrosesan awal data terdiri dari normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.
2. Membangun model dan melakukan validasi terhadapnya Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis cluster, deteksi anomali, analisis asosiasi, analisis pola sekuensial, dan sebagainya. Dalam beberapa refrensi, deteksi anomali juga masuk dalam langkah ekplorasi. Akan tetapi, deteksi anomali juga dapat digunakan sebagai algoritma utama, terutama untuk mencari data-data yang special.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **2018** | 4 |  |  | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |  |  |

1. Penerapan, Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

Metode yang tepat perlu digunakan untuk melakukan tahapan penelitian ini. CRISP-DM menyediakan standar proses standar untuk data mining yang dapat diterapkan ke dalam strategi pemecahan masalah umum pada bisnis atau pada unit peneltian. CRISP-DM membandingkan metodologi data mining lain lebih lengkap dan terdokumentasi dengan baik. Setiap fase terstruktur dan terdefinisi dengan jelas sehingga mudah diaplikasikan bahkan bagi pemula sekalipun. Penggunaan metodologi CRISP-DM pada data mining sudah banyak diterapkan dalam penelitian, salah satunya pada jurnal “Business and Data Understanding dalam rangka pembentukan model tata letak dan tata ruang pasar tradisional dalam rangka meningkatkan laba penjualan menggunakan metode association rule dan decision tree” jurnal ini membahas mengenai masalah pemahaman tujuan bisnis dan data yang akan digunakan pada proses data mining untuk menentukan tata letak strategis dan tata ruang pasar tradisional dalam rangka meningkatkan laba penjualan. Penelitian ini menggunakan metodologi CRISPDM sebagai tahapan penelitian. Kasus yang dapat diterapkan pada tekhnik CRIP-DM antara lain:

* Memprediksi pola kelulusan studi mahasiswa dalam menempuh mata kuliah, pola ini dihasilkan dari pohon keputusan yang diharapkan dapat menjadi acuan untuk kelulusan studi mahasiswa menempuh mata kuliah pada Universitas XYZ.
* Metodologi CRISP-DM juga digunakan dalam jurnal “Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM” . Jurnal ini membahas tentang bagaimana prediksi kinerja penjualan karya musik dengan memanfaatkan proses data mining yang mengikuti langkahlangkah dalam framework CRISP-DM. CRISP-DM dapat membantu analisis kinerja secara lebih menyeluruh dan terstruktur.
* Penelitian lain yang menggunakan CRISP- DM sebagai metodologi dan proses clustering dengan menggunakan algoritma K-Means adalah penelitian “Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan Tridharma” [2]. Penelitian ini menggunakan tahapan CRISP-DM sebagai metodologi dan proses clustering dengan menggunakan algoritma K-Means.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 5 | **Pengantar Data Mining**  Saruni Dwiasnati,ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

Dan untuk contoh nya mengambil tema Memprediksi pola kelulusan studi mahasiswa dalam menempuh mata kuliah pada Universitas XYZ. Tahapan yang digunakan mengacu pada penelitian yang berjudul ‘Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan tridharma’. Pemahaman tahapan mengacu pada penelitian tersebut akan tetapi terdapat perbedaan pada objek serta variabel penelitian yang dilakukan.

Terdapat tiga tahapan penelitian, pertama merupakan tahapan pengumpulan data. Pada tahapan ini menjelaskan mengenai cara pengumpulan data yang dilakukan dalam penelitian. Pengumpulan data dilakukan selain dengan melakukan wawancara dan kuesioner, data juga dikumpulkan terkait dengan suasana akademik mahasiswa pada universitas XYZ. Pengumpulan data tersebut dilakukan dengan melakukan observasi terkait dengan suasana akademik pada Universitas XYZ, sumber data tersebut didapatkan dari sistem informasi akademik yang telah diimplementasikan pada Universitas XYZ.

Tahapan kedua adalah tahapan pemahaman bisnis dan pemahaman data. Proses pada tahapan ini merupakan bagian dari Metodologi CRISP-DM. Proses pada tahapan ini yaitu, pemahaman bisnis, pemahaman data, dan proses persiapan data. Tahapan ketiga merupakan tahapan pemodelan dan evaluasi. Proses pada tahapan ini juga merupakan bagian dari metodologi CRISP-DM. Proses pada tahapan ini yaitu, proses pemodelan, proses evaluasi dan proses penyebaran.

1. Tahapan pengumpulan data awal:

a. Studi Pustaka

b. Observasi

c. Wawancara

d. Kuesioner

2. Tahapan Pemahaman bisnis dan pemahaman data:

a. Pemahaman bisnis Pada tahapan pemahaman bisnis ada beberapa hal yang dilakukan antara lain, menentukan tujuan bisnis, menilai situasi, dan menentukan tujuan data mining.

b. Pemahaman data Pada tahapan pemahaman data, ada beberapa hal yang dilakukan antara lain, mengumpulkan data awal, mendeskripsikan data, mengeksplorasi data, dan memverifikasi kualitas data.

c. Persiapan Data Pada tahapan persiapan data, ada beberapa hal yang harus dilakukan antara lain, deskripsikan dataset yang ada, pemilihan data yang dibutuhkan, membangun data, mengintegrasikan data satu dengan data yang liannya dan membersihkan data.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 6 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

Tahapan persiapan data yang dilakukan berdasarkan penelitian yang berjudul ‘Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan tridharma’.

3. Tahap Pemodelan dan Evaluasi:

a. Pemodelan Pada penelitian yang berjudul’ Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan tridharma’, didefinisikan ‘tahapan pemodelan merupakan tahapan yang secara langsung melibatkan data mining. Pemilihan teknik data mining, algoritma dan menentukan parameter dengan nilai yang optimal’. Pada tahapan pemodelan, ada beberapa hal yang dilakukan antara lain, memilih teknik pemodelan, membangun model, dan menilai model. Tahapan yang dilakukan mengacu pada penelitian tersebut.

b. Evaluasi Pada penelitian yang berjudul’ Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan tridharma’, didefinisikan tahapan evaluasi adalah 'fase interpretasi terhadap hasil data mining. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap modelling sesuai dengan sasaran yang ingin dicapai dalam tahap business understanding’. Pada tahapan evaluasi, ada beberapa hal yang dilakukan antara lain:

1) Mengevaluasi hasil (Evaluate Results)

2) Meninjau Proses (Review Process)

3) Menentukan Langkah selanjutnya (Determine Next Steps).

Pada penelitian ini memanfaatkan proses yang terdapat pada metode CRISP-DM, akan tetapi tidak melakukan proses terakhir yakni proses implementasi. Tahapan yang dilakukan untuk melakukan penerapan metode CRISP-DM mengacu pada penelitian yang berjudul ‘Data Clustering menggunakan metodologi CRISP-DM untuk pengenalan pola proporsi pelaksanaan tridharma’.

4.Pemodelan (Modeling) Pemilihan teknik data mining, algoritma dan menentukan parameter dengan nilai yang optimal. Pada tahapan pemodelan, ada beberapa hal yang dilakukan antara lain, memilih teknik pemodelan, membangun model, dan menilai model:

1. Memilih Teknik Pemodelan (Select Modelling Technique) Teknik data mining yang dipilih adalah decision tree dengan menggunakan algoritma C4.5. Decision tree dan algortima C4.5 sangat tepat digunakan untuk mencapai tujuan awal penelitian ini yaitu menggali pengetahuan tentang prediksi mahasiswa menempuh matakuliah.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 7 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

Pemodelan data mining diawali dengan membuat rule untuk pembentukan pohon keputusan.

2. Membangun Model (Build Model) Sesuai ketentuan umum yang telah diterapkan, kriteria penilaian mahasiswa menjadi salah satu tolak ukur kelulusan mahasiswa dalam setiap mata kuliah.

5. Evaluasi Evaluasi pada penelitian ini lebih difokuskan pada model atau pola yang dihasilkan oleh algoritma C4.5. Model yang dihasilkan dianalisa untuk mengetahui apakah pola yang dihasilkan sudah sesuai dengan standar yang terdapat di universitas XYZ. Jika pola yang dihasilkan belum sesuai, maka analisa lebih lanjut terhadap pola yang dihasilkan perlu dilakukan sehingga dapat menghasilkan rekomendasi perbaikan dalam suasana akademik, yang diharapkan dapat membantu keberhasilan studi mahasiswa dalam menempuh mata kuliah.

6. Penyebaran, berdasarkan buku algoritma dan data mining tahapan penyebaran merupakan pembuatan laporan, dan tahapan penyebaran juga bisa didefinisikan sebagai penerapan proses data mining secara paralel pada departemen lain, dalam hal ini tahapan penyebaran data mining dapat dilakukan pada seluruh program studi yang ada di Universitas XYZ. Pada penelitian ini tahapan deployment atau tahap penyebaran tidak dilakukan.

Karakteristik cross validation:

- Sumber data dipisahkan menjadi beberapa folder untuk cross validation.

- Seleksi fitur-fitur dilakukan pada masing-masing folder training data secara terpisah.

- Beragam tree diperoleh dari beberapa kasus.

Adapun parameter pengukuran yang digunakan adalah :

* Accuracy adalah persentase dari total data yang benar diidentifikasi.
* Precision adalah perbandingan jumlah data relevan yang diambil dengan jumlah seluruh dokumen yang diambil oleh sistem (Mandala, 2006).
* Recall merupakan salah satu pengukuran untuk permasalahan dalam klasifikasi.Recall adalah perbandingan antara jumlah data relevan yang diambil dengan jumlah data relevan yang berada di koleksi dokumen (database).
* F Measures merupakan sebuah parameter pengukuran yang membandingkan secara seimbang antara precision dan Recall.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 8 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

Metode yang digunakan untuk mengatasi permasalahan imbalance class yang terjadi adalah databoost IM yang memadukan boosting dengan data generation,data generation adalah membuat data sintetis untuk ditambahkan ke data asli (data sintetik yang dibuat sama dengan data asli).Sehingga akurasi prediksi terhadap kelas mayor dan kelas minor dapat diperbaiki.

Format data hanya dilakukan pada penamaan atribut tanpa mengubah makna aslinya, seperti atribut A1 untuk kategori cluster yang didefinisikan oleh carrier, A2 untuk jenis layanan yang dipakai pelanggan, A3 untuk jenis kelamin, dan seterusnya.

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 9 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |

DAFTAR PUSTAKA

* Turban, E, 2005, ***Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia Jilid 1***. Andi: Yogyakarta.
* Larose, Daniel T. 2005. ***Discovering Knowledge in Data : An Introduction to Data Mining***. John Willey & Sons, Inc.
* ayyad, Usama. 1996. ***Advances in Knowledge Discovery and Data Mining***. MIT Press.
* *Decision Support and Expert Systems; Management Support Systems* (E. Turban) Prentice Hall. New Jersey.USA.
* <https://informatikalogi.com/algoritma-naive-bayes/>
* <https://www.ilmuskripsi.com/2016/07/algoritma-c45.html>
* <https://yudiagusta.wordpress.com/k-means/>
* <https://www.researchgate.net/publication/282479259_Survey_on_Anomaly_Detection_using_Data_Mining_Techniques>

|  |  |  |  |
| --- | --- | --- | --- |
| **2018** | 10 | **Pengantar Data Mining**  Saruni Dwiasnati, ST.MM | **Pusat Bahan Ajar dan eLearning**  <http://www.mercubuana.ac.id> |
|  |