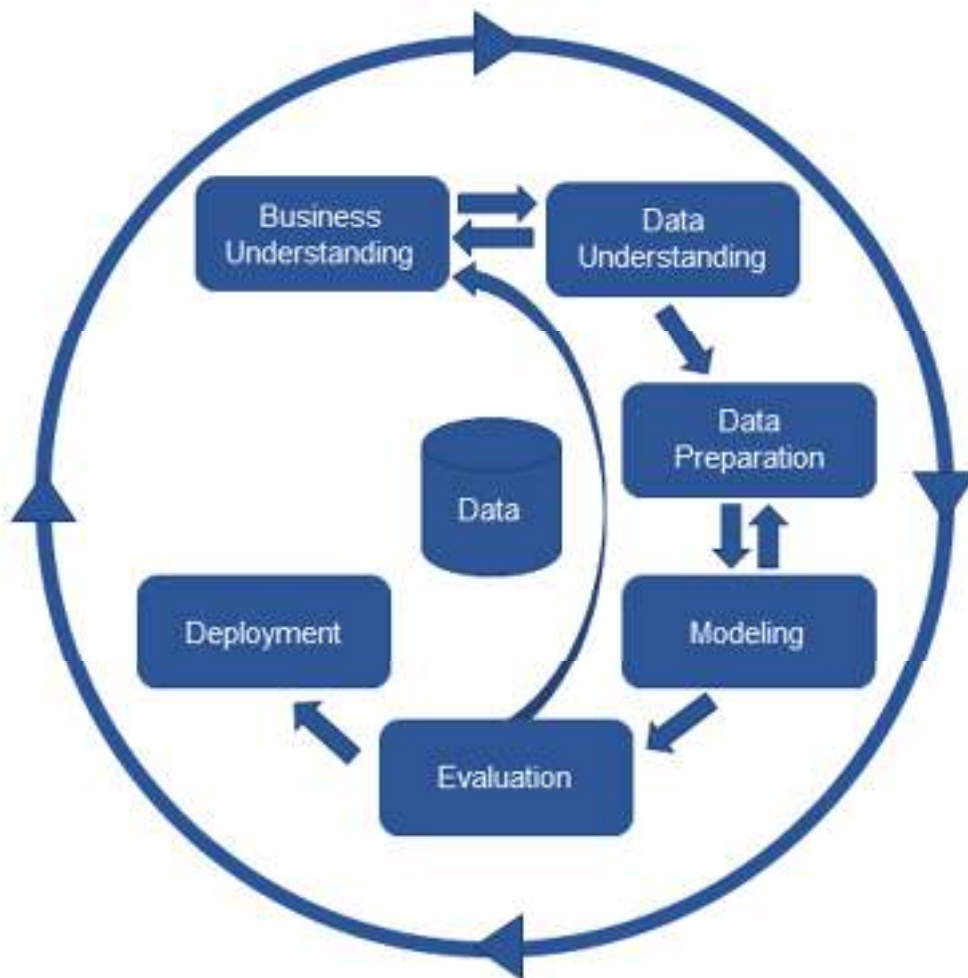


BAB III

METODE PENELITIAN

3.1 Metode Penelitian

Metode Penelitian yang digunakan pada penelitian Tugas Akhir adalah metode CRISP-DM atau *Cross Industry Standard Process for Data Mining*. Gambar 3.1 dibawah ini merupakan gambaran langkah-langkah yang akan dilakukan.



Gambar 3.1 Proses CRISP-DM (Larose, 2005)

Berdasarkan Gambar 3.1 diatas, proses penelitian dapat dipaparkan dalam beberapa fase dibawah ini:

1. Fase *Business Understanding* atau Pemahaman Bisnis
 - a. Menetapkan tujuan penelitian serta menentukan kebutuhan penelitian secara rinci.

Tujuan penelitian ini adalah melakukan analisis secara komprehensif terhadap kasus pandemi *Covid-19* di sebagian besar negara yang terjangkiti. Penelitian akan mengklasifikasikan kasus kedalam beberapa kelas berdasarkan tingkat kasus yang dilaporkan antara tanggal 22 Januari 2020 sampai tanggal 23 September 2020.

b. Membuat strategi untuk mencapai tujuan penelitian.

Langkah yang dilakukan adalah dengan mencari *dataset* melalui beberapa *online repository*. Ada beberapa yang peneliti akses pada tanggal 5 november 2020, diantaranya *UCI Machine Learning* (<https://www.ucilearning.com>), *OpenML* (<https://www.openml.com>), dan *Kaggle* (<https://www.kaggle.com>). Kemudian peneliti memutuskan untuk menggunakan *dataset* dari *kaggle.com* karena pertimbangan kelengkapan dan kerapihan. *Dataset* kasus *Covid-19* yang diperoleh dari <https://www.kaggle.comhttps://archive.ics.uci.edu/ml/datasets/Student+Performance> yang terdiri dari 116805 *records* dan 7 variabel seperti tanggal kasus dilaporkan atau *ObservationDate*, *State/Province* atau provinsi maupun negara bagian tempat kejadian, negara tempat terjadi kasus atau *Country/Region*, waktu kasus dilaporkan atau *LastUpdate*, jumlah terkonfirmasi positif atau *confirmed*, jumlah kematian atau *deaths*, dan jumlah sembuh atau *recovered*.

2. Fase *Data Understanding* atau Pemahaman Data

Dataset kasus pandemi *Covid-19* yang didapatkan dari *online repository* *www.kaggle.com* adalah *dataset* berbentuk *spreadsheet* yang kemudian dikonversi menjadi *csv*. Beberapa proses dalam fase ini adalah:

a. *Data Collecting* atau Mengumpulkan data.

Dataset yang digunakan adalah *dataset* kasus pandemi *Covid-19* disebagian besar negara yang terjangkit pandemi virus ini.

b. Melakukan analisis lebih lanjut untuk mengenali data.

Dataset kasus pandemi *Covid-19* pada sebagian besar negara ini terdiri dari 7 variabel utama yaitu *ObservationDate*, *State/Province*, *Country/Region*, *LastUpdate*, *Confirmed*, *Deaths*, dan *Recovered*.

c. Melakukan evaluasi terhadap kualitas data.

Evaluasi dilakukan untuk mengetahui apakah terdapat data dengan nilai *null* atau tidak untuk menghindari *missing value* dan menciptakan *dataset* yang baik untuk proses mining.

Dengan dataset yang baik maka diharapkan tujuan data mining yang telah ditetapkan dapat dicapai.

Selanjutnya peneliti melakukan proses *data preprocessing* dan *data cleansing* untuk membersihkan data dari ketidaklengkapan dan menghilangkan *outlier*. Tabel 3.1 berikut adalah contoh dataset kasus covid-19 yang akan digunakan.

Tabel 3.1 Contoh dataset kasus covid-19

ObsaservationDate	State/Province	Country/Region	LastUpdate	confirmed	deaths	recovered
1/27/2020	Taiwan	Taiwan	1/27/2020 23:59	5	0	0
1/27/2020	Xinjiang	Mainland	1/27/2020 23:59	5	0	0
1/27/2020	Washington	China	1/27/2020 23:59	1	0	0
1/27/2020	Illinois	US	1/27/2020 23:59	1	0	0
1/27/2020	California	US	1/27/2020 23:59	2	0	0
1/27/2020	Arizona	US	1/27/2020 23:59	1	0	0
1/27/2020	Japan	Japan	1/27/2020 23:59	4	0	1
1/27/2020	Thailand	Thailand	1/27/2020 23:59	8	0	2
1/27/2020	South Korea	South Korea	1/27/2020 23:59	4	0	0
1/27/2020	Singapore	Singapore	1/27/2020 23:59	5	0	0
1/27/2020	Vietnam	Vietnam	1/27/2020 23:59	2	0	0
1/27/2020	France	France	1/27/2020 23:59	3	0	0
1/27/2020	Nepal	Nepal	1/27/2020 23:59	1	0	0
1/27/2020	Malaysia	Malaysia	1/27/2020 23:59	4	0	0
1/27/2020	Ontario	Canada	1/27/2020 23:59	1	0	0
1/27/2020	Cambodia	Cambodia	1/27/2020 23:59	1	0	0

3. Fase *Data Preparation* atau Pengolahan Data

a. Menyiapkan data untuk proses pengolahan.

Dataset kasus pandemi *Covid-19* yang terjadi disebagian besar negara di dunia ini disiapkan untuk proses *data mining*.

b. Memilih bagian yang akan dianalisis.

Variabel yang menjadi acuan adalah *confirmed*, *deaths*, dan *recovered* yang selanjutnya ketiga kategori ini juga dilabelkan tingkat kasusnya menjadi *tinggi* untuk jumlah kasus diatas rata-rata seluruh kasus, dan *rendah* untuk jumlah kasus dibawah rata-rata seluruh kasus.

4. Fase *Modeling* atau Pemodelan

Pada fase keempat ini peneliti menentukan teknik *data mining* yang akan digunakan beserta algoritmanya. Teknik yang digunakan dalam penelitian ini adalah teknik klasifikasi

menggunakan algoritma C4.5. Kemudian menggunakan model fungsi *logistic* untuk memprediksi tren kasus covid-19 di Indonesia dengan menggunakan bahasa pemrograman Python.

5. Fase *Evaluation* atau Evaluasi

Pada fase ini peneliti akan memastikan bahwa:

- a. Semua model yang dihasilkan sesuai dengan tujuan yang diharapkan.
- b. Tidak ada permasalahan yang terlewatkan atau luput dari penanganan; serta peneliti akan melakukan:
- c. Pengambilan keputusan berdasarkan informasi yang didapatkan dari hasil *data mining*.

6. Fase *Deployment* atau Penyebaran

Penyebaran dilakukan dengan membuat laporan jurnal ilmiah tentang penelitian *data mining* ini.

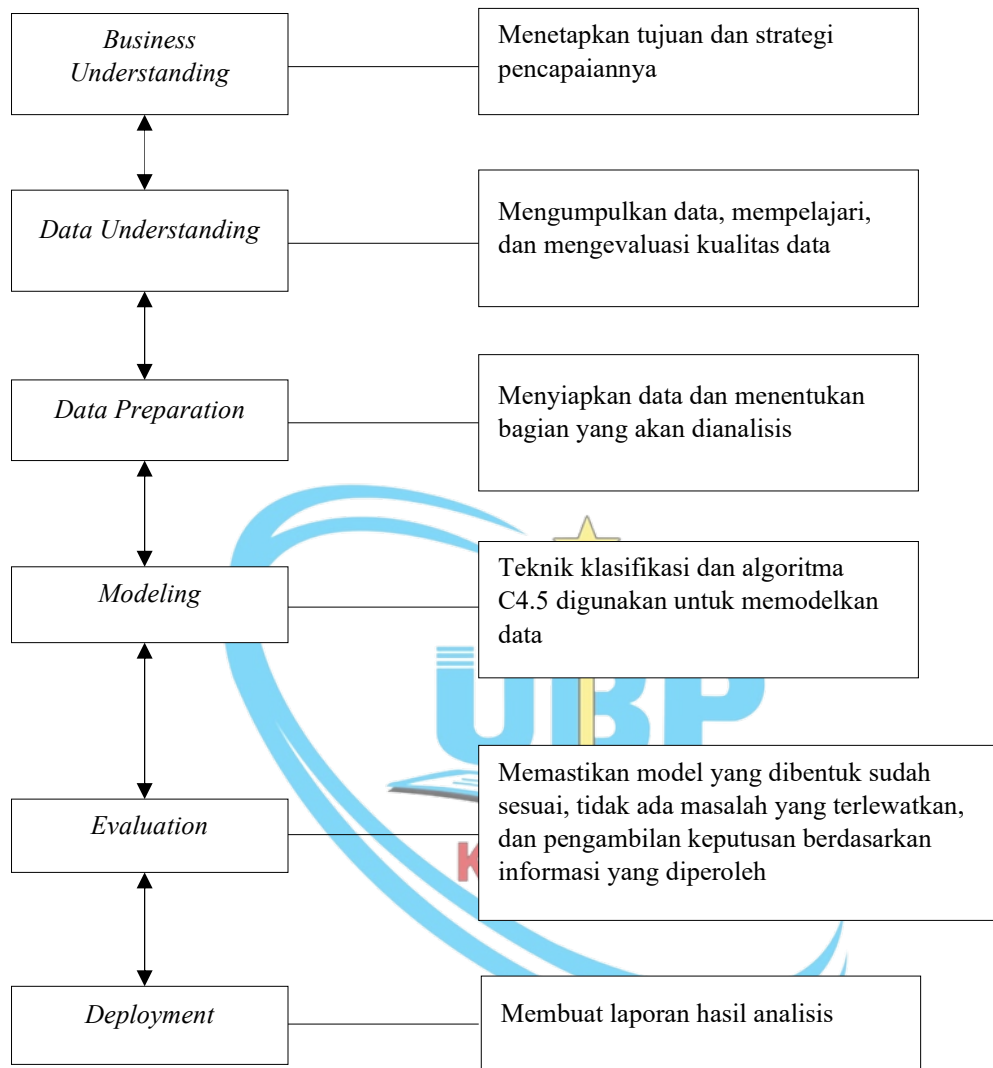
Selain karena *detail*-nya setiap proses yang dijalankan, metode CRISP-DM juga digunakan karena metode ini termasuk metode standar yang dapat digunakan pada sebagian besar industri. Metode lain yang perlu juga dipertimbangkan adalah *Team Data Science Process* (TDSP) yang dikembangkan oleh Microsoft. Komponen utama TDSP adalah *data science lifecycle definition* yang menyediakan siklus pengembangan *data science task*. Proses dalam komponen ini termasuk *business understanding*, *data acquisition and understanding*, *modeling*, dan *deployment*. Proses *evaluation* dilakukan pada tahap *acceptance* atau tingkat penerimaan hasil pemodelan. Pada dasarnya metode lain seperti SEMMA (*Sample, Explore, Modify, Model, Access*), TDSP, IBM *Data Science Methodology* dalam prosesnya adalah sama akan tetapi dengan istilah berbeda. Contohnya adalah pada metode SEMMA, fase *access* adalah serupa pada fase *evaluation* di CRISP-DM.

3.2 Objek Penelitian

Objek dalam penelitian Tugas Akhir ini adalah dataset kasus pandemi *Covid-19* yang menjangkiti hampir sebagian besar negara yang dilaporkan dari tanggal 22 Januari sampai 23 September 2020, dengan 116805 *records* dan 7 variabel, berasal dari www.kaggle.com.

3.3 Prosedur Penelitian

Alur penelitian Tugas Akhir yang akan dilakukan dapat digambarkan sebagai berikut:



Gambar 3.2 Alur penelitian

Penelitian Tugas Akhir ini dilakukan sesuai standar CRISP-DM (Larose, 2005) dengan 6 tahap penelitian.

Berdasarkan Gambar 3.2 di atas, penelitian dimulai dengan penetapan tujuan penelitian dan strategi untuk mencapai tujuan penelitian tersebut, keduanya masuk dalam fase pertama penelitian yaitu *Business Understanding*. Kemudian penelitian dilanjutkan dengan mengumpulkan data melalui *online repository* <https://www.kaggle.com>. Di kaggle.com ini, peneliti mendapatkan data kasus pandemi Covid-19 di sebagian besar negara di dunia. Setelah didapatkan, data akan dipelajari lebih lanjut untuk mengenali data dan dilakukan evaluasi dengan

menghilangkan *missing value* untuk mendapatkan data yang baik untuk proses *mining*. Proses-proses ini masuk dalam fase kedua penelitian yaitu *Data Understanding*.

Setelah didapatkan data yang baik untuk proses *mining*, langkah selanjutnya adalah menyiapkannya dan menentukan bagian data mana yang akan dilakukan analisis. Menentukan variabel yang menjadi acuan, yaitu *confirmed*, *deaths*, dan *recovered*, kemudian nilai ketiga variabel tersebut dikategorikan menjadi *tinggi*, untuk nilai yang berada diatas rata-rata kasus, dan *rendah* untuk nilai yang berada dibawah rata-rata kasus. Fase ini adalah fase ketiga penelitian yang disebut *Data Preparation*.

Fase penelitian keempat adalah fase *Modeling*, dimana pada fase ini peneliti menentukan teknik dan algoritma apa yang akan digunakan. Peneliti memutuskan untuk menggunakan teknik klasifikasi dan algoritma C4.5 pada proses data mining ini dan menggunakan model fungsi logistic untuk memprediksi tren pandemic di Indonesia. Pada fase *Evaluation*, hasil pemodelan dengan algoritma C4.5 dan hasil prediksi dengan model fungsi *logistic* kemudian dievaluasi untuk mengetahui kategori klasifikasi dan kategori modelnya. *Deployment* adalah fase terakhir dari penelitian. Fase ini berupa pembuatan laporan hasil penelitian berupa karya tulis atau jurnal ilmiah.

Keenam fase penelitian tersebut dilakukan secara berkesinambungan, antara fase yang satu saling berkaitan dengan fase yang lain, sehingga jika suatu fase penelitian belum selesai maka fase yang lain tidak dapat dilakukan.

