BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian akan berfokus untuk meneliti data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta dengan rentang tahun 2010 – 2023. Rentang waktu yang lebih tepat pada *dataset* adalah mulai dari 1 Januari 2010 hingga 30 November 2023. Data tersebut memiliki beberapa *parameter* atau atribut, yaitu *tanggal, pm_10, pm_25, so2, co, o3, no2, max, critical, category, lokasi_spku*. Data tersebut diambil dari data Dinas Lingkungan Hidup DKI Jakarta melalui *website* https://satudata.jakarta.go.id/. Selain data ISPU, terdapat juga data meteorologi seperti cuaca, kelembapan, dan kecepatan angin dalam mendukung pembuatan model seperti yang direkomendasikan pada penelitian sebelumnya [16]. Namun, beberapa kolom akan diseleksi kembali dan mengutamakan pada parameter-parameter ISPU sebagai *features* untuk menentukan sebagai *input model* terutama dalam pembagian data *training* dan *testing*. Kemudian dari data tersebut akan dilakukan proses *clustering* terlebih dahulu menggunakan *K-Means* dan setelah itu akan dilakukan *re-clustering* dengan *K-Medoids*. Hasil akhir *clustering* akan dilakukan prediksi menggunakan *Long Short-Term Memory* (LSTM). Selain hasil *clustering*, terdapat variasi *dataset* untuk diprediksi, yaitu dengan menggunakan keseluruhan *dataset* dan per stasiun dengan pembagian *dataset* menjadi tahun 2010 – 2023 dan 2021 – 2023. Pembagian waktu tersebut dikarenakan pada tahun 2021 dan seterusnya terdapat parameter PM2.5 yang tidak ada di tahun-tahun sebelumnya. Hal tersebut juga memungkinkan untuk melihat performa *model* dalam menggunakan ukuran *dataset* yang berbeda-beda.

3.2 Metode Penelitian

Penelitian ini akan menggunakan sebuah kerangka atau *framework* sesuai dengan topik *data mining*. *Framework data mining* yang umum digunakan di antaranya adalah *KDD*, *SEMMA*, dan *CRISP-DM*. Berikut ini adalah perbandingan ketiga *framework* tersebut[60][61][62]:

Tabel 3. 1 Perbandingan *Framework Data Mining*

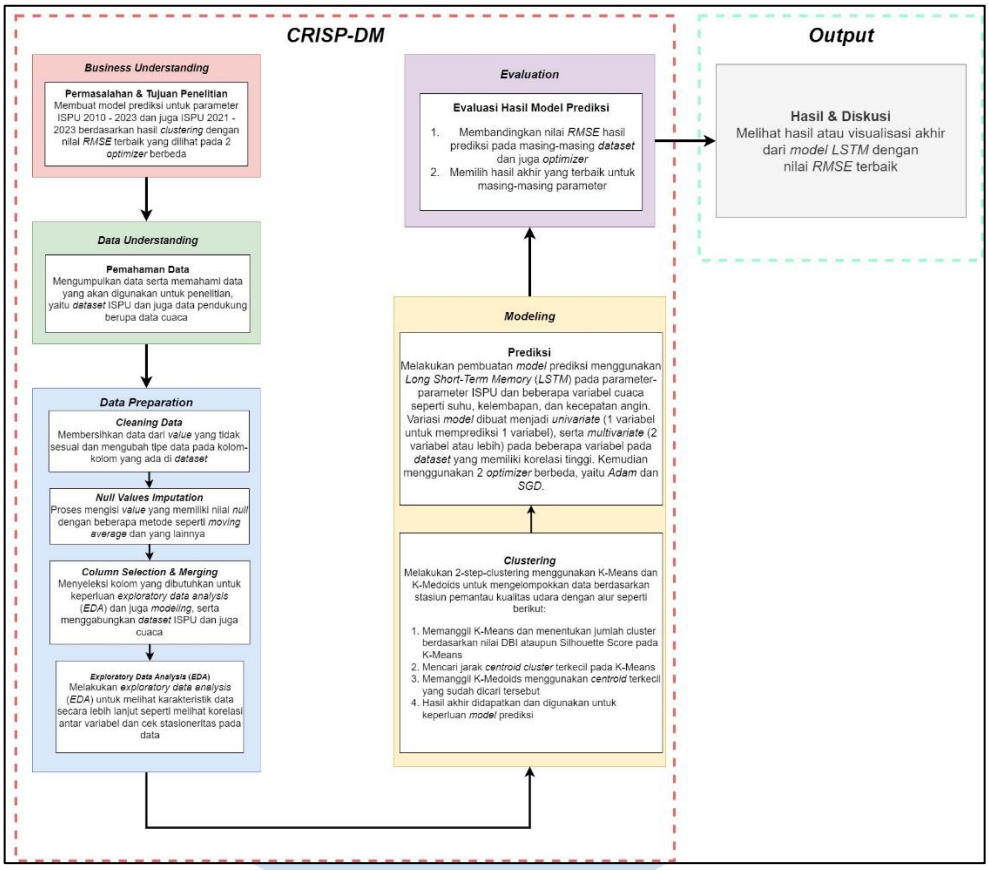
	KDD	SEMMA	CRISP-DM
Tujuan	Untuk menemukan informasi yang berguna dari data terutama pada *database*	Bertujuan untuk membantu *tools* SAS pada perusahaan SAS Institute dalam melakukan *data mining*	Untuk menjadi proses umum dalam melakukan *data mining* terutama pada hampir keseluruhan industri
Proses	1. *Selection* 2. *Data Preprocessing* 3. *Data Transformation* 4. *Data Mining* 5. *Data Interpretation / Evaluation*	1. *Sample* 2. *Explore* 3. *Modify* 4. *Model* 5. *Assess*	1. *Business Understanding* 2. *Data Understanding* 3. *Data Preparation* 4. *Modeling* 5. *Evaluation* 6. *Deployment*
Kelebihan	Menjadi pelopor bagi metode *data mining* lainnya, serta menghadirkan cara melakukan *knowledge discovery* pada data	Menjadi panduan terstruktur pada *software* SAS dalam melakukan *data mining*, dan juga bersifat dinamis karena model yang telah dibuat dapat diperbaiki dengan proses yang iteratif	Proses yang dilakukan terstruktur dan terdokumentasi sehingga dapat diterapkan pada banyak industri, serta bersifat fleksibel dengan adanya proses yang iteratif
Kekurangan	Kurang fokus terhadap tujuan bisnis karena tidak disebutkan secara jelas dalam proses yang dilakukan	Kurang fleksibel karena penerapannya yang spesifik hanya untuk *software* SAS	Proses yang dilakukan harus disesuaikan dengan tujuan dan pemahaman terhadap data yang dapat berubah-ubah

Tabel 3.1 merupakan perbandingan *framework data mining* untuk menjadi acuan dalam menetapkan *framework* yang akan digunakan pada penelitian ini. Penelitian ini akan menggunakan *framework CRISP-DM* dalam melakukan proses *data mining*. Hal tersebut dikarenakan proses yang fleksibel dan iteratif pada prosesnya, serta dapat diterapkan pada banyak topik atau industri.

3.2.1 *CRISP-DM*

Pada penelitian ini akan menggunakan salah satu metode *data mining*, yaitu CRISP-DM. Penggunaan CRISP-DM sendiri adalah karena cocok dalam proyek *data science* terutama jika proyek tersebut menggunakan konsep *goal-directed* dan juga *process-driven*, sehingga CRISP-DM masih sangat

memungkinkan untuk diterapkan [63]. Tahapan pada CRISP-DM memiliki 6 proses yang keenam proses tersebut menjadi sebuah siklus atau proses dapat dilakukan secara berulang. Tahapan pertama adalah *business understanding* yang dilakukan dengan tujuan membuat *model* prediksi yang menggunakan tahapan *hybrid clustering* terlebih dahulu. Kemudian, tahapan kedua adalah *data understanding* yang dilakukan dengan tujuan memahami data yang didapatkan dari proses pengambilan data. Tahapan ketiga, yaitu *data preparation* yang bertujuan untuk menyiapkan data sehingga dapat diproses ke tahap selanjutnya, dimana pada tahap ini data dapat dibersihkan (*data cleaning*) dari *error* atau permasalahan lainnya. Tahapan selanjutnya yang keempat, yaitu *modeling* yang dilakukan dengan tujuan untuk membuat model *hybrid* sesuai dengan algoritma *clustering* yang dipilih, *K-Means* dan *K-Medoids*, dimana *clustering* pertama dilakukan menggunakan *K-Means* dan selanjutnya dilakukan *re-clustering* menggunakan *K-Medoids*. Setelah dilakukan *re-clustering* dan didapatkan hasil tersebut, maka selanjutnya adalah membuat model prediksi menggunakan *Long Short-Term Memory* (LSTM) untuk memprediksi parameter-parameter ISPU dan juga beberapa data pendukung, seperti suhu, kelembapan udara, dan juga kecepatan angin. Tahapan kelima, yaitu *evaluation* sesuai dengan *metric* yang digunakan, yaitu *root mean square error* (*RMSE*), serta tujuan melakukan evaluasi terhadap *model* yang telah dibuat untuk melihat performa dan kinerja *model*. Tahapan terakhir atau tahapan keenam, yaitu berupa *output* atau hasil dan diskusi yang dilakukan dengan tujuan dapat memberikan informasi berdasarkan hasil *modeling*, dimana pada penelitian ini memilih *model* antara *optimizer* dengan nilai evaluasi terbaik, serta mengambil kesimpulan berdasarkan hasil *model* yang telah dihasilkan untuk masing-masing *input* atau parameter-parameter pada *dataset*.



Gambar 3. 1 Alur Penelitian Berdasarkan *CRISP-DM*

Berikut adalah penjelasan tahapan-tahapan pada *CRISP-DM* berdasarkan Gambar 3.1 yang akan digunakan pada penelitian ini:

1. ***Business Understanding***

   Tahapan pertama *CRISP-DM* adalah *business understanding* untuk membentuk pemahaman mengenai tujuan dilakukan penelitian. Melalui penelitian ini, tujuan yang ingin dicapai adalah memprediksi konsentrasi polutan menggunakan data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta dari tahun 2010 – 2023. Kemudian, terdapat variasi *dataset* tersebut seperti dibagi menjadi tahun 2010 – 2023 dan juga tahun 2021 – 2023. Tujuan lainnya adalah melakukan visualisasi dari *model* terbaik berdasarkan nilai *evaluation metrics* yang digunakan antara 2 *optimizer* yang berbeda juga. Selanjutnya, penelitian ini juga dapat menjadi informasi mengenai bagaimana penerapan *hybrid model K-Means* dan *K-Medoids* dengan LSTM untuk memprediksi data ISPU DKI Jakarta tahun

2010 – 2023 dengan tambahan data cuaca sebagai pendukung, yaitu suhu, kelembapan udara, dan juga kecepatan angin.

2. ***Data Understanding***

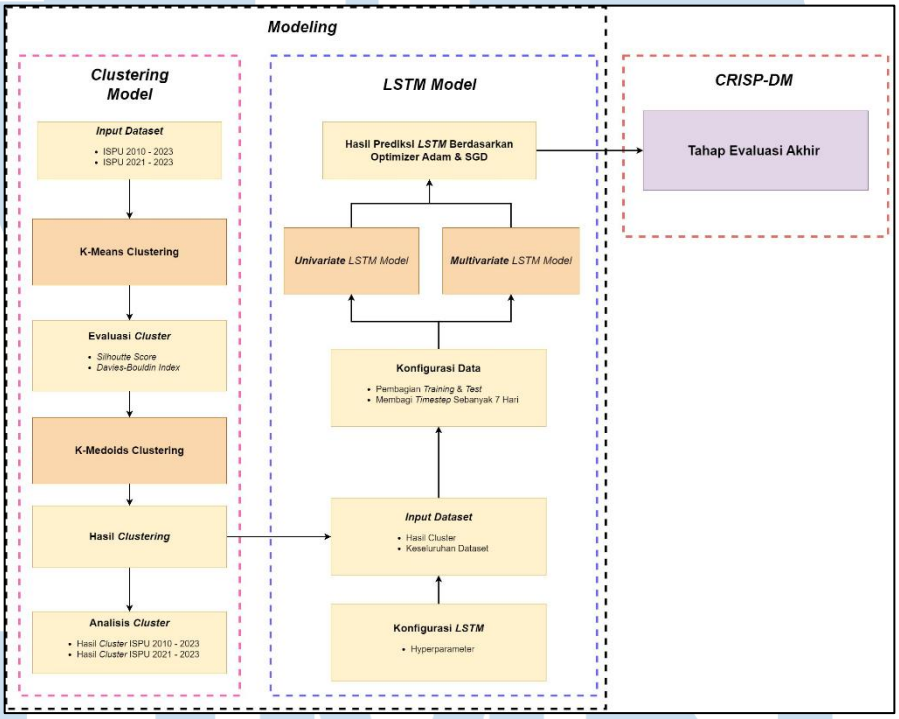
Tahapan selanjutnya adalah melakukan *data understanding* untuk memahami data agar tujuan penelitian dapat diproses lebih lanjut dan tercapai nantinya. Pada penelitian ini, data yang akan dipahami adalah data ataupun atribut-atribut yang terdapat pada *dataset* Indeks Standar Pencemar Udara (ISPU) DKI Jakarta tahun 2010 – 2023. Selain itu, terdapat beberapa data pendukung dari *dataset* cuaca, yaitu suhu, kelembapan dan juga kecepatan angin. Data yang dipilih tersebut memiliki rentang waktu dari 1 Januari 2010 hingga 30 November 2023. Pengambilan data berasal dari *website* https://satudata.jakarta.go.id/, serta data pendukungnya diambil dari *website* https://www.visualcrossing.com/weather-data.

3. ***Data Preparation***

Pada tahapan *data preparation* dilakukan pengecekan dan pemeriksaan terhadap data yang akan digunakan. *Dataset* mentah yang telah diambil akan dibersihkan terlebih dahulu dari *values* yang tidak sesuai pada data. Kemudian dilakukan imputasi atau mengisi *null values* dengan menggunakan metode tertentu seperti *moving average*. Selanjutnya, beberapa kolom dipilih dan *dataset* tersebut di-*merge* sesuai dengan kebutuhan penelitian terutama sebagai *input* pada tahap *modeling*. Secara lebih detil, proses *merge dataset* bertujuan untuk memudahkan dan mempersingkat waktu *import dataset* ketika *modeling*, serta untuk tahapan *exploratory data analysis* (*EDA*) digunakan untuk mempermudah melihat variabel yang memiliki korelasi tinggi antara variabel pada *dataset* ISPU ataupun cuaca yang berupa variabel suhu, kelembapan udara, dan kecepatan angin. Proses terakhir adalah melakukan *exploratory data analysis* (*EDA*) untuk melihat karakteristik variabel secara lebih lanjut seperti cek stasioneritas data dan pemilihan variabel dengan korelasi tertinggi agar dapat dibuat *model multivariate* menggunakan *LSTM*.

4. *Modeling*

Tahapan *modeling* ini adalah dengan membuat model berdasarkan algoritma yang dipilih. Penelitian ini akan menggunakan model *hybrid* dari gabungan *K-Means* dan *K-Medoids* sesuai dengan acuan penelitian [15] yang kemudian akan dilakukan prediksi menggunakan *Long Short-Term Memory* (LSTM) seperti pada penelitian [16]. Berikut ini tahapan *modeling* secara lebih detil:



Gambar 3. 2 Alur Tahapan *Modeling* Penelitian Berdasarkan *CRISP-DM*

Gambar 3.2 merupakan alur *modeling* yang dimulai dari pembuatan *model cluster* kemudian diprediksi menggunakan *LSTM* hingga hasil prediksi digunakan untuk tahapan evaluasi akhir pada alur penelitian utama berdasarkan *CRISP-DM* sebelumnya yang ada di Gambar 3.1. Untuk tahapan ini juga disesuaikan dengan kebutuhan penelitian seperti melakukan *clustering* dan hasil *clustering* tersebut akan diprediksi dalam memprediksi parameter-parameter ISPU. Secara lebih detil, proses *clustering* yang dilakukan adalah dengan menggunakan *K-Means* sebagai proses awal dan juga untuk menentukan jumlah *cluster* yang

optimal berdasarkan nilai evaluasi *Davies Bouldin Index* ataupun *Silhouette Score*. Kemudian dilanjutkan dengan memilih jarak *centroid* terkecil dari hasil *K-Means* tersebut dan menggunakan *K-Medoids* berdasarkan jarak *centroid* terkecil yang sudah dicari. Hasil akhir dari *clustering* akan digunakan untuk keperluan *model* prediksi dan juga analisis *cluster*. Selanjutnya, untuk *model* prediksi dengan *LSTM* akan menggunakan beberapa variasi dari *dataset*. Variasi lainnya adalah pada *model LSTM* dilakukan secara *univariate* dan juga *multivariate*. *Univariate* sendiri adalah menggunakan 1 variabel untuk memprediksi variabel tersebut saja, sedangkan *multivariate* adalah menggunakan 2 variabel untuk memprediksi variabel pertama atau kedua. *Optimizer* yang akan digunakan adalah *Adam* dan *SGD*.

5. ***Evaluation***

Tahapan selanjutnya ini adalah melakukan *evaluation* terhadap hasil yang diberikan pada model yang telah dibuat. Pada penelitian ini, hasil dari *hybrid model* yang telah dibuat akan dievaluasi berdasarkan *model* terakhir, yaitu prediksi menggunakan *LSTM*. Pengukuran evaluasi yang digunakan adalah menggunakan nilai *root mean square error* (RMSE). Kemudian, pada masing-masing *dataset* dan juga *optimizer*, dibandingkan dan kemudia dilihat mana yang memiliki nilai *RMSE* terkecil sebagai kriteria hasil *model* terbaik. Hasil terbaik tersebut adalah pada masing-masing parameter ISPU ataupun data pendukung seperti suhu, kelembapan udara, dan juga kecepatan angin.

6. ***Ouput* (Hasil & Diskusi)**

Tahapan terakhir adalah hasil dan diskusi agar dapat memberikan informasi yang berguna dari proses yang sudah dilakukan. Pada penelitian ini tahap akhir berupa *output* merupakan tahapan hasil dan diskusi, serta hasil akhir dari prediksi akan dilakukan visualisasi dari *model* terbaik berdasarkan nilai evaluasi menggunakan *root mean square error* (*RMSE*) untuk *optimizer* yang digunakan. Selain itu, terdapat *forecasting* selama 14 hari ke depan pada masing-masing visualisasi akhir tersebut.

3.3 Teknik Pengumpulan Data

Dataset yang akan digunakan dalam penelitian ini berasal dari *website* https://satudata.jakarta.go.id/ yang berupa data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta mulai dari tahun 2010 – 2023 dan 2021 – 2023 dengan menggunakan variabel-variabel pada data tersebut. Kemudian untuk data meteorologi atau cuaca yang digunakan untuk mendukung data ISPU pada pembuatan model nantinya diambil melalui *website* https://www.visualcrossing.com/.

Kemudian, pengambilan data dengan rentang waktu lebih dari 10 tahun dilakukan untuk keperluan *model* prediksi menggunakan *LSTM*. Data tersebut dapat dikatakan berjumlah banyak karena terdapat lebih dari 20 ribu dan bertujuan untuk meningkatkan kompleksitas *model LSTM* yang dimana algoritma tersebut dapat mempelajari data jangka panjang secara efisien [64]. Pada salah satu penelitian prediksi penyakit *influenza* atau penyakit pernapasan menggunakan *LSTM* [65], *model LSTM* memiliki performa akurasi yang baik dalam memprediksi data jangka panjang dengan rentang waktu 10 tahun. Selain itu, terdapat variasi *dataset* antara tahun 2010 – 2023 dan 2021 – 2023 dikarenakan variabel ISPU berupa PM2.5 hanya terdapat setelah tahun 2020.

3.4 Teknik Pengambilan Sampel

Pada penelitian ini, pengambilan sampel yang dimaksud adalah memilih beberapa hasil dari *model* untuk dilakukan visualisasi akhir. Pemilihan sampel *model* yang diambil adalah mengacu pada hasil prediksi dengan menggunakan *dataset* hasil *clustering*. Selain itu, untuk hasil *model* prediksi secara *multivariate* akan dibandingkan pada keseluruhan *dataset* dan per *cluster*, serta variabel yang dipilih dari *dataset* cuaca digunakan pada keseluruhan. Namun, semua itu juga bergantung pada nilai evaluasi yang dihasilkan.

3.5 Variabel Penelitian

Variabel penelitian ini akan menggunakan data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta di tahun 2010 – 2023 dan 2021 – 2023. Pada

penelitian ini variabel dibagi menjadi variabel dependen dan variabel independen.

1. **Variabel Dependen**

   Variabel dependen sendiri adalah variabel yang dipengaruhi oleh variabel lainnya, yaitu variabel independen. Pada penelitian ini, yang menjadi variabel dependen adalah hasil *clustering* ataupun hasil prediksi dari *hybrid model LSTM*.

2. **Variabel Independen**

   Variabel independen merupakan variabel yang tidak dipengaruhi oleh variabel lainnya. Pada penelitian ini, variabel independen yang digunakan adalah variabel atau kolom dari data Indeks Standar Pencemar Udara (ISPU) DKI Jakarta pada tahun 2010 – 2023 ataupun 2021 – 2023 seperti, *pm_10*, *pm_25*, *so2*, *co*, *o3*, *no2*, *stasiun*, serta data cuaca sebagai data pendukung seperti suhu, kelembapan dan kecepatan angin. Variabel independen tersebut digunakan sebagai data yang di-*input* pada *model clustering* dan juga prediksi.

3.6 Teknik Analisis Data

Pada penelitian ini, teknik analisis data yang akan digunakan adalah menggunakan *framework* CRISP-DM dengan melakukan *clustering* berdasarkan *hybrid model K-Means* dan *K-Medoids* yang kemudian diprediksi menggunakan *Long Short-Term Memory* (LSTM). *Tools* yang akan digunakan dalam membantu analisis data ini, yaitu bahasa pemrograman *Python* melalui *platform Jupyter Notebook*. Tahap pemodelan prediksi tersebut dilakukan pada hasil *cluster*, keseluruhan *dataset*, dan per stasiun pemantau kualitas udara yang ada di *dataset* ISPU 2010 – 2023 dan 2021 – 2023 dengan menggunakan 2 *optimizer*, yaitu *Adam* dan juga *SGD*. Selanjutnya, setelah *model* dibuat dan diperoleh hasil prediksi berdasarkan 7 hari (*timestep*) yang lalu, maka dilakukan evaluasi dari nilai *root mean square error* (*RMSE*) untuk menentukan *optimizer* yang memiliki nilai terbaik. Kemudian, tahap akhir yang merupakan visualisasi adalah menampilkan perbandingan antara data aktual dengan prediksi, serta menambahkan *forecast* selama 14 hari ke depan.