BAB III

METODOLOGI PENELITIAN

3.1 Metode Penelitian

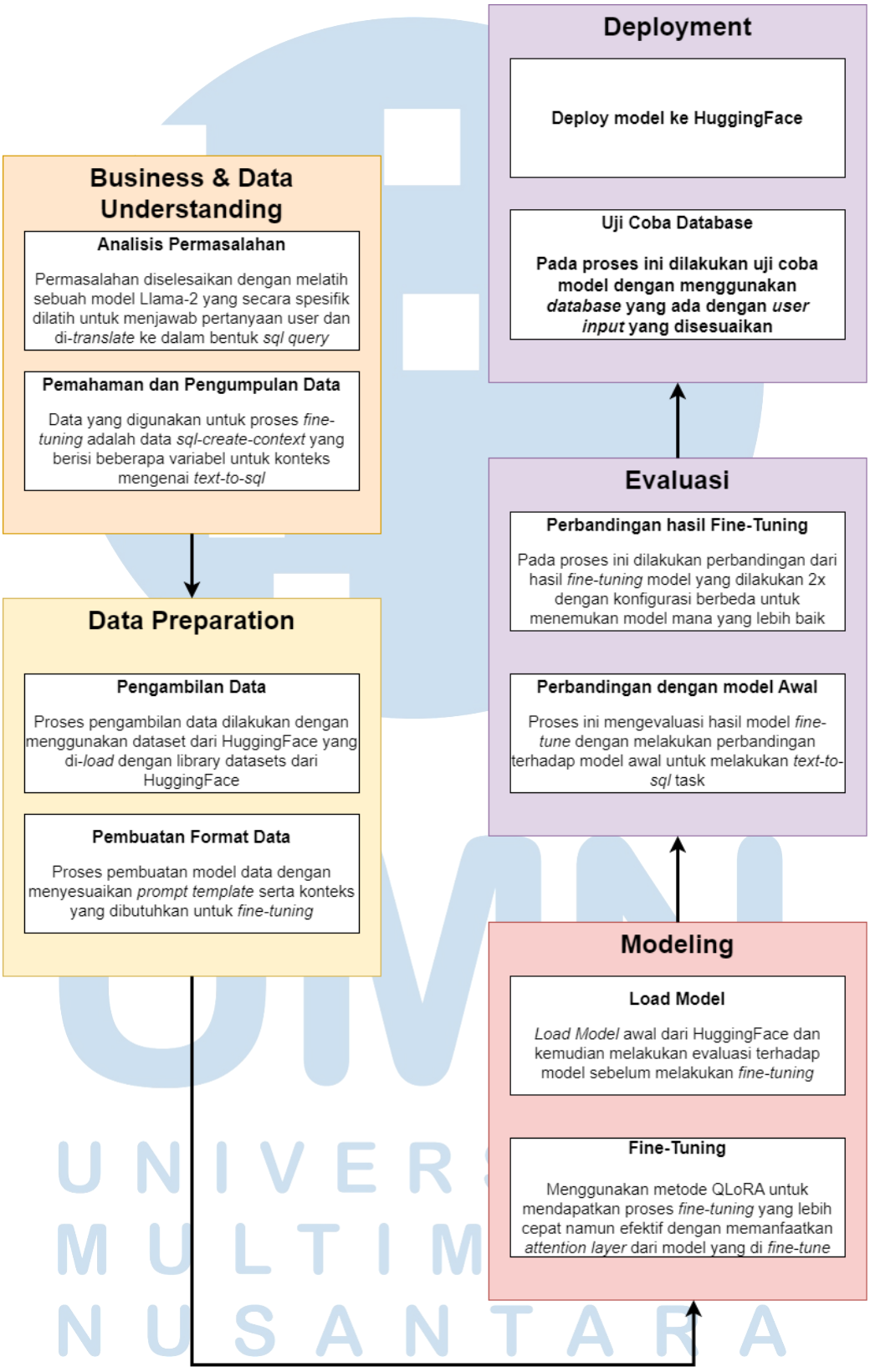
Penelitian ini dilakukan dengan menggunakan kerangka atau *framework* yang sesuai dengan tujuan penelitian. *Framework* yang paling sering digunakan untuk penelitian *data mining* dan *machine learning* antara lain adalah *CRISP-DM* yang merupakan sebuah metode penelitian *Data Mining* yang menjadi sebuah standar bagi semua proses *Data Mining* tersebut. Namun pada kenyataannya, *CRISP-DM* tidak memiliki performa yang sesuai ketika proses *Machine Learning* dilakukan sehingga hadir sebuah metode yaitu *CRISP-ML*. Metode *CRISP-ML* ini sudah dijelaskan pada bab 2 mengenai landasan teori. Dalam hal ini, perlu dilakukan perbandingan terhadap kedua metode sehingga terdapat alasan yang jelas mengenai mengapa *CRISP-ML* menjadi metode yang digunakan dalam penelitian ini. Berikut adalah perbandingan antara kedua model[37][50] :

Tabel 3. 1 Tabel perbandingan

	CRISP-DM	CRISP-ML
TUJUAN	Membantu proses *Data Mining* secara terstruktur yang sesuai dengan kebutuhan setiap industri	Menjadi metode yang mampu memberikan hasil implementasi *Machine Learning* yang sesuai dengan kebutuhan penelitian
PROSES	- Business Understanding - Data Understanding - Data Preparation - Modeling - Evaluation	- Business and Data Understanding - Data Preparation - Modeling - Evaluation - Deployment

	- Deployment	- Monitoring and Maintenance
KELEBIHAN	Memiliki struktur penelitian yang jelas dengan dokumentasi yang lengkap. Metode ini juga banyak digunakan oleh berbagai industri.	Di desain khusus untuk melakukan proses *Machine Learning* dengan memberikan struktur yang mampu memberikan hasil model yang terbaik
KEKURANGAN	Secara desain tidak memiliki proses khusus yang mempermudah proses *Machine Learning* dan tidak terlalu bergantung pada Evaluasi dan Penerapan Model	Membutuhkan *tools-tools* yang lebih spesifik terhadap *machine learning* dan juga membutuhkan *resources* yang lebih besar daripada metode CRISP-ML.

3.1.1 Alur Penelitian



Gambar 3. 1 Alur Penelitian

Alur Penelitian dirancang untuk memberikan gambaran mengenai proses penelitian yang akan dilakukan. Alur penelitian pada penelitian ini ditampilkan pada *flowchart* gambar 3.1 diatas dengan implementasi model CRISP-ML. Dalam beberapa tahun terakhir, telah terjadi pertumbuhan pesat dalam pengumpulan dan penghubungan data, yang diiringi dengan semakin beragamnya teknik-teknik AI berbasis data, termasuk machine learning (ML). Meskipun ada banyak peluang untuk proyek analitik data, kebutuhan akan regulasi dan akuntabilitas hasil proyek-proyek ini juga meningkat. Salah satu tantangan utama adalah bagaimana menjelaskan hasil dari algoritma ML kepada para praktisi dan pembuat kebijakan, khususnya dalam bidang kesehatan. Hal ini menghambat adopsi yang lebih luas dari pendekatan ilmu data di sektor ini[51].

CRISP-ML adalah metodologi yang dikembangkan untuk mengatasi masalah interpretabilitas dalam proyek-proyek AI. Metodologi ini dibangun di atas kekuatan CRISP-DM dan dirancang untuk menentukan tingkat interpretabilitas yang diperlukan oleh para pemangku kepentingan untuk solusi dunia nyata yang sukses, dan kemudian membantu mencapainya. Metodologi CRISP-ML telah berhasil diterapkan di berbagai industri dan bidang, termasuk risiko kredit, asuransi, utilitas, dan olahraga. Dengan demikian, penerapan CRISP-ML dalam penelitian ini diharapkan dapat memberikan struktur yang jelas dan komprehensif dalam mengembangkan model AI yang dapat membantu proses pengolahan data menggunakan query SQL secara otomatis.

Tahap pertama yang dilakukan adalah *Business & Data Understanding*. Pada tahap ini dilakukan analisis permasalahan yang ada agar dapat menemukan serta mengidentifikasi apa saja yang menjadi masalah yang terjadi. Perlu juga ada pengumpulan data mengenai konteks sebuah deskripsi *query* yang diinginkan dengan jawaban *sql query* yang diberikan. Hal ini juga merujuk kepada proses pemahaman data bagi banyak orang yang dinilai masih cukup sulit jika tidak memiliki kemampuan dasar dalam bidang data. Visualisasi data dengan kurang kemampuan dasar dalam pengolah data akan berakibat kurangnya informasi yang disampaikan.

Dengan permasalahan yang ada, dirumuskan sebuah solusi berupa *chatbot* yang dapat menampilkan hasil visualisasi berdasarkan deskripsi yang diinginkan *user* sehingga memudahkan *user* untuk mendapatkan informasi mengenai data yang dimiliki.

Setelah pemahaman bisnis dan data telah dilakukan, tahap berikutnya adalah *data preparation* dengan tujuan untuk membangun sebuah data yang sesuai dengan tujuan *fine-tuning* untuk mendapatkan hasil yang sesuai dengan tujuan penelitian. Pada tahapan ini dilakukan beberapa proses seperti konversi data yang ada dengan menyesuaikan ketentuan *fine-tuning* dan membagi data *training* dan data *test*. Setelah dilakukan pemisahan data *training* dan *test*, dilakukan tahap modeling yang merupakan implementasi *fine-tuning* pada model Llama-2 menggunakan dataset yang sudah diolah. Setelah implementasi *fine-tuning* telah dilakukan, proses berikutnya adalah melakukan evaluasi dengan membuat visualisasi singkat menggunakan hasil yang telah diberikan dan melakukan evaluasi dengan pihak yang bergerak pada pengolahan dan visualisasi data untuk menemukan visualisasi yang sesuai dengan deskripsi yang dihasilkan. Setelah evaluasi telah dilakukan, proses terakhir adalah implementasi model *fine-tuning* ke dalam *chatbot*.

3.1.2 Metode Pengembangan Sistem / Metode *Machine Learning*

Dengan merujuk pada gambar 3.1 diatas mengenai gambaran tentang alur penelitian yang dilakukan, penelitian dilakukan dengan melakukan proses CRISP-ML dengan berbagai tahapan yang diakhiri dengan tahap Deployment dari model yang dibuat. Pemilihan model CRISP-ML dipilih dengan mempertimbangkan implementasi model *Machine Learning* yang dinilai lebih cocok dengan model CRISP-ML dibandingkan model lain seperti CRISP-DM.

Dengan begitu, maka tahapan yang dimulai pada CRISP-ML adalah sebagai berikut :

3.1.2.1 Business & Data Understanding

Penelitian ini dilatarbelakangi oleh kesulitan yang sering dihadapi dalam pengolahan data berskala besar menggunakan SQL (Structured Query Language). Kompleksitas SQL seringkali menjadi hambatan bagi pengguna yang tidak memiliki keahlian teknis mendalam, sehingga menghambat produktivitas dan efisiensi. Untuk mengatasi masalah ini, penelitian ini mengusulkan pengembangan model bahasa besar (LLM) Llama-2 yang telah disempurnakan untuk tugas text-to-SQL. Model ini diharapkan dapat menerjemahkan perintah bahasa alami menjadi kueri SQL yang valid, sehingga menyederhanakan interaksi pengguna dengan data.

Tujuan utama penelitian ini adalah menciptakan model text-to-SQL yang akurat dan efisien yang dapat mengotomatiskan tugas-tugas pengolahan data. Dengan demikian, diharapkan dapat meningkatkan produktivitas pengguna, mengurangi hambatan teknis, dan membuka peluang baru dalam analisis bisnis, manajemen data, dan pengembangan perangkat lunak. Keberhasilan model ini akan memberikan dampak signifikan pada berbagai industri dan aplikasi yang bergantung pada pengolahan data yang efisien dan mudah diakses.

Pada penelitian ini, digunakan dataset "*SQL-Create-Context*" yang diperoleh dari HuggingFace.co. Dataset ini memuat data konteks yang relevan dengan tugas Text-to-SQL, terdiri dari tiga komponen utama: Question (pertanyaan dalam bahasa manusia), Answer (*Query* SQL yang sesuai), dan Context (informasi tabel yang relevan). Dataset ini akan digunakan sebagai landasan dalam proses fine-tuning model Llama-2. Dengan mempelajari pola dan hubungan antara pertanyaan, jawaban, dan konteks dalam dataset ini, model diharapkan dapat memahami tugas Text-to-SQL secara mendalam dan menghasilkan *Query* SQL yang akurat dan relevan berdasarkan pertanyaan pengguna.

3.1.2.2 Data Preparation

Pada tahap Data Preparation ini, dilakukan beberapa proses untuk mengubah format dari *dataset* yang dimiliki sehingga dapat digunakan sebagai data *training*. Untuk meminimalisir proses *training* yang akan dilakukan, data *training* akan difilter menjadi total 10000 data daripada 78000 data yang tersedia. Selanjutnya adalah membuat sebuah kolom baru berisikan *template* yang akan digunakan sebagai bahan latihan bagi model untuk memahami konteks *text-to-sql* yang ingin dicapai. Kolom tersebut akan menjadi kolom yang nantinya diberikan kepada model pada proses *training*.

3.1.2.3 Modeling

Pada proses modeling, dilakukan proses *fine-tuning* dengan melatih model Llama-2 menggunakan data *training* yang sudah dibuat dengan menggunakan berbagai konfigurasi seperti *BitsandBytes*, QloRA, dan beberapa konfigurasi lainnya yang difokuskan untuk membuat proses *training* berjalan dengan cepat dan tingkat efektivitas yang tinggi.

Pada proses modeling ini, pertama dilakukan *load* model dari Llama-2 awal. Setelah dilakukan *load* model, dilakukan percobaan terlebih dahulu terhadap model untuk melakukan pengujian awal dari *text-to-sql task*. Setelah dilakukan pengujian, selanjutnya adalah melakukan proses *training* dengan menggunakan konfigurasi QloRA untuk mengurangi penggunaan memori GPU pada proses *fine-tuning*.

3.1.2.4 Evaluation

Tahap Evaluasi dilakukan dengan melakukan perbandingan antara beberapa model yang dilatih untuk menemukan model mana yang memiliki tingkat kemiripan paling baik ketika dilakukan uji coba pada *test dataset*. Evaluasi yang dilakukan ada beberapa, yaitu uji Akurasi, *Precision, Recalling, F1-Score* dan *ROUGE Score*.

3.1.2.5 Deployment

Tahap terakhir yang dilakukan adalah melakukan *deployment* dari model ke dalam HuggingFace untuk dapat

digunakan. Selanjutnya proses yang dilakukan adalah uji coba model terhadap *database* yang ada dan menjalankan *sql query* dari hasil yang diberikan model untuk menentukan apakah model sesuai atau tidak.

3.2 Teknik Pengumpulan Data

Data yang digunakan untuk proses *Fine*-Tuning diambil dari data bernama *sql-create-context* yang merupakan sebuah dataset dengan memberikan konteks terkait *sql query* dan berasal dari HuggingFace.co yang merupakan sebuah *platform* penyedia berbagai bentuk dataset dan model LLM yang dapat digunakan untuk melakukan penelitian dan pengembangan.

Terdapat total 78600 baris konteks pertanyaan terkait pembuatan *sql query* yang memberikan jawaban bentuk *sql query* dengan mengikuti konteks yang diberikan. *Fine-tuning* dapat dilakukan pada model Llama-2 dengan membagi konteks ke dalam 3 kategori yaitu *context*, *question*, *answer*. Tabel 3.1 menampikan bentuk konteks, pertanyaan, dan jawaban yang diberikan.

Tabel 3. 2 Struktur dataset *sql-create-context*

Context	Pertanyaan	Jawaban
`CREATE TABLE head (age INTEGER)`	How many heads of the departments are older than 56 ?	`SELECT COUNT(*) FROM head WHERE age > 56`
`CREATE TABLE head (name VARCHAR, born_state VARCHAR, age VARCHAR)`	List the name, born state and age of the heads of departments ordered by age.	`SELECT name, born_state, age FROM head ORDER BY age`

CREATE TABLE department (creation VARCHAR, name VARCHAR, budget_in_billions VARCHAR)	List the creation year, name and budget of each department.	SELECT creation, name, budget_in_billions FROM department