BAB I
PENDAHULUAN

1.1. Latar Belakang

Sektor makanan dan minuman merupakan salah satu sektor strategis dalam perekonomian Indonesia karena berperan dalam memenuhi kebutuhan dasar masyarakat serta memberikan kontribusi signifikan terhadap Produk Domestik Bruto (PDB). Selain itu, sektor ini memiliki pertumbuhan yang relatif stabil dan termasuk dalam industri pengolahan non-migas yang mampu menyerap tenaga kerja dalam jumlah besar. Kondisi tersebut menjadikan sektor ini sebagai sektor defensif yang tetap menarik bagi investor di pasar modal [1].

Seiring dengan perkembangan pasar modal di Indonesia, sektor makanan dan minuman menjadi salah satu sektor yang aktif diperdagangkan di Bursa Efek Indonesia (BEI). Saat ini terdapat 26 perusahaan makanan dan minuman yang tercatat di BEI dengan 15 perusahaan memenuhi kriteria untuk dijadikan sampel penelitian [2]. Beberapa saham terpandang dalam sektor ini antara lain PT Indofood CBP Sukses Makmur Tbk (ICBP) yang merupakan salah satu saham blue chip dengan kapitalisasi pasar terbesar [3],PT Mayora Indah Tbk (MYOR) yang dikenal luas melalui produk biskuit dan permen [4]. Serta PT Garudafood Putra Putri Jaya Tbk (GOOD) yang memiliki portofolio produk makanan dan minuman yang beragam [5]. Dalam penelitian ini digunakan tiga emiten, yaitu PT Indofood CBP Sukses Makmur Tbk (ICBP), PT Mayora Indah Tbk (MYOR), dan PT Garudafood Putra Putri Jaya Tbk (GOOD). Pemilihan ketiga saham tersebut didasarkan pada kapitalisasi pasar yang besar, tingkat likuiditas yang tinggi, serta ketersediaan data historis yang kontinu. Selain itu, ketiga saham tersebut menunjukkan pola pergerakan harga yang fluktuatif namun relatif stabil dalam jangka panjang, sehingga sesuai digunakan dalam pemodelan *time series*. Di sisi lain, sektor ini juga menghadapi tantangan berupa fluktuasi harga saham yang tinggi, di mana pergerakan harga dapat berubah secara signifikan dalam waktu singkat akibat berbagai faktor eksternal seperti kebijakan pemerintah, kondisi ekonomi makro, dan sentimen pasar [6]. Penggunaan saham dalam sektor yang sama juga bertujuan

untuk menjaga konsistensi analisis tanpa dipengaruhi oleh perbedaan karakteristik antar sektor.

Meskipun sektor ini tergolong defensif, harga saham tetap menunjukkan dinamika yang kompleks. Karakteristik data saham yang bersifat nonlinear, dinamis, serta memiliki dependensi temporal menjadikan proses prediksi harga saham menjadi sulit dilakukan menggunakan pendekatan konvensional [7][8]. Metode statistik tradisional seperti ARIMA dan regresi linier memiliki keterbatasan dalam menangkap pola hubungan jangka panjang serta kompleksitas data *time series*.

Permasalahan dalam prediksi harga saham tidak hanya terletak pada metode pemodelan, tetapi juga pada pengelolaan data. Dalam praktiknya, proses akuisisi data, pembersihan, transformasi, hingga pemodelan sering dilakukan secara terpisah sehingga menyebabkan proses menjadi tidak efisien, sulit dimonitor, dan tidak dapat berjalan secara otomatis dalam skala besar. Kondisi ini menunjukkan perlunya suatu sistem yang mampu mengintegrasikan seluruh proses pengolahan data secara terstruktur dalam satu alur kerja yang berkelanjutan. Untuk mengatasi permasalahan tersebut, digunakan pendekatan data *pipeline* yang mengintegrasikan seluruh proses pengolahan data dalam satu alur kerja terstruktur [9]. Data *pipeline* mencakup tahapan data *ingestion*, data *lake*, *staging, preprocessing*, hingga penyediaan data siap pakai untuk pemodelan. Implementasi data *pipeline* dilaporkan mampu meningkatkan efisiensi pengolahan data hingga sekitar 40% serta meningkatkan *reproducibility* sistem secara signifikan karena seluruh proses terdokumentasi dan terotomatisasi [9].

Dalam arsitektur data pipeline tersebut, diperlukan komponen pemrosesan data yang mampu menangani data dalam jumlah besar secara efisien. Oleh karena itu, *Apache Spark* digunakan sebagai engine pemrosesan data karena kemampuannya dalam komputasi terdistribusi berbasis *in-memory*. Pendekatan ini dilaporkan mampu meningkatkan kecepatan pemrosesan data hingga 10–100 kali dibandingkan metode berbasis disk tradisional serta menurunkan waktu komputasi *preprocessing* lebih dari 60% pada dataset skala besar [10]. Selain itu, untuk memastikan seluruh tahapan dalam pipeline dapat berjalan secara otomatis dan terkoordinasi, diperlukan mekanisme orchestrasi yang mampu mengatur alur kerja

antar proses. Dalam hal ini, *Apache Airflow* digunakan sebagai *tools orchestrasi* untuk mengelola alur kerja data *pipeline* dalam bentuk *Directed Acyclic Graph* (DAG). *Airflow* memungkinkan setiap proses dijalankan secara otomatis dan terjadwal, serta menyediakan fitur *monitoring*, *logging*, dan pengelolaan dependensi antar *task*. Setiap proses dalam *DAG* direpresentasikan dalam bentuk *task* yang memiliki identitas unik berupa *task_id*. *Task_id* digunakan untuk membedakan setiap *task* dalam *workflow* sehingga *Airflow* dapat mengatur urutan eksekusi, mendeteksi keberhasilan maupun kegagalan proses, serta menampilkan status eksekusi pada dashboard *monitoring*. Selain itu, Airflow juga dapat menampilkan informasi durasi eksekusi pada setiap *task*, sehingga proses yang berjalan lambat dapat dianalisis untuk meningkatkan efisiensi *pipeline*. Dalam penelitian ini, *task_id* digunakan sebagai identitas pada setiap tahapan proses dalam *pipeline*. Penggunaan *Airflow* dilaporkan mampu meningkatkan keandalan *pipeline* dengan mengurangi kegagalan proses hingga sekitar 30% serta meningkatkan efisiensi monitoring sistem [11]. Pada tahap pemodelan, digunakan metode *Stacked Long Short-Term Memory* (*Stacked* LSTM) yang mampu menangkap pola jangka panjang dalam data *time series*. Model ini terbukti mampu menghasilkan performa prediksi yang lebih baik dibandingkan metode konvensional. Hasil penelitian menunjukkan bahwa model LSTM dapat menghasilkan nilai *Root Mean Square Error* (RMSE) sebesar 0,012 dibandingkan metode ARIMA sebesar 0,021 [12]. Selain itu, penggunaan arsitektur *Stacked* LSTM dilaporkan mampu menurunkan nilai *Mean Absolute Error* (MAE) hingga 15% dibandingkan model LSTM tunggal [13]. Studi lain juga menunjukkan bahwa pendekatan *deep learning* mampu meningkatkan akurasi prediksi secara signifikan dalam data finansial [14]. Meskipun model prediksi seperti LSTM telah menunjukkan performa yang baik, keberhasilan implementasinya sangat bergantung pada ketersediaan data yang terkelola dengan baik serta sistem yang mampu mendukung proses pengolahan data secara berkelanjutan. Oleh karena itu, selain aspek model, diperlukan pula sistem yang mampu mengintegrasikan seluruh proses pengolahan data secara efisien. Meskipun berbagai penelitian telah menunjukkan keberhasilan penggunaan LSTM dalam meningkatkan akurasi prediksi serta penggunaan data *pipeline, Apache Spark*, dan *Apache Airflow* dalam meningkatkan efisiensi pengolahan data,

sebagian besar penelitian tersebut masih dilakukan secara terpisah. Penelitian prediksi umumnya hanya berfokus pada model, sedangkan penelitian data pipeline lebih berfokus pada sistem tanpa mengintegrasikan model prediksi secara langsung. Hal ini menunjukkan adanya kesenjangan penelitian, yaitu belum banyak penelitian yang mengintegrasikan model prediksi berbasis Stacked LSTM dengan sistem data pipeline yang terstruktur menggunakan *Apache Spark* dan *Apache Airflow* dalam satu arsitektur terpadu. Oleh karena itu, penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem prediksi harga saham berbasis data *pipeline* yang terintegrasi menggunakan *Apache Spark* dan *Apache Airflow* dengan model *Stacked* LSTM. Sistem yang dibangun diharapkan mampu menghasilkan prediksi yang akurat sekaligus memiliki efisiensi, otomatisasi, dan skalabilitas dalam pengolahan data.

1.2. Rumusan Masalah

Rumusan masalah yang menjadi fokus utama pada penelitian ini adalah sebagai berikut:

1. Bagaimana perancangan dan implementasi arsitektur data *pipeline* berbasis *Apache Airflow* dan *Apache Spark* dalam pengolahan data saham sektor makanan dan minuman?
2. Bagaimana integrasi model *Stacked Long Short-Term Memory* dalam *pipeline* untuk menghasilkan prediksi harga saham?
3. Bagaimana hasil implementasi data *pipeline* dalam menjalankan proses prediksi secara otomatis ditinjau dari keberhasilan eksekusi task dan durasi proses pada *Airflow*?
4. Bagaimana evaluasi performa model *Stacked Long Short-Term Memory* dalam memprediksi harga saham berdasarkan nilai *RMSE*, *MAE*, dan *MAPE*?
5. Bagaimana perancangan dan implementasi antarmuka (GUI) berbasis web menggunakan Streamlit untuk visualisasi hasil prediksi harga saham?

1.3. Batasan Masalah

Batasan yang ditetapkan pada masalah yang diamati pada penelitian ini adalah sebagai berikut:

1. Penelitian ini berfokus pada saham sektor makanan dan minuman di BEI, yaitu ICBP, MYOR, dan GOOD.
2. Data yang digunakan merupakan data historis harga saham yang diperoleh dari Yahoo Finance.
3. Metode prediksi yang digunakan adalah *Stacked Long Short-Term Memory* (LSTM).
4. *Pipeline* data dibangun menggunakan *Apache Airflow* sebagai orches*tration tool* dan *Apache Spark* sebagai data *processing engine*.
5. Evaluasi model dilakukan menggunakan metrik MSE, RMSE, dan MAE tanpa membandingkan dengan metode lain.
6. Penelitian tidak membahas faktor eksternal seperti sentimen pasar, berita, atau analisis fundamental saham.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab rumusan masalah serta memberikan arahan yang lebih jelas bagi penelitian ini. Adapun tujuan penelitian ini adalah sebagai berikut:

1. Merancang dan mengimplementasikan arsitektur data pipeline berbasis *Apache Airflow* dan *Apache Spark* dalam pengolahan data saham sektor makanan dan minuman
2. Mengintegrasikan model *Stacked Long Short-Term Memory* dalam *pipeline* untuk menghasilkan prediksi harga saham
3. Melihat hasil implementasi data *pipeline* dalam menjalankan proses prediksi secara otomatis ditinjau dari keberhasilan eksekusi task dan durasi proses pada *Airflow*
4. Merancang dan mengimplementasikan antarmuka (GUI) berbasis web menggunakan Streamlit untuk visualisasi hasil prediksi harga saham

1.5. Manfaat Penelitian

Manfaat penelitian yang dapat diperoleh dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan kontribusi dalam pengembangan sistem prediksi harga saham berbasis *deep learning* dan *big data pipeline* secara terintegrasi.
2. Menjadi referensi dalam penerapan model *Stacked* LSTM untuk prediksi data *time series* di bidang pasar modal
3. Membantu investor atau pengguna dalam memperoleh gambaran prediksi harga saham sebagai bahan pertimbangan dalam pengambilan keputusan.
4. Menjadi dasar pengembangan sistem prediksi saham yang lebih kompleks dan siap digunakan pada lingkungan produksi