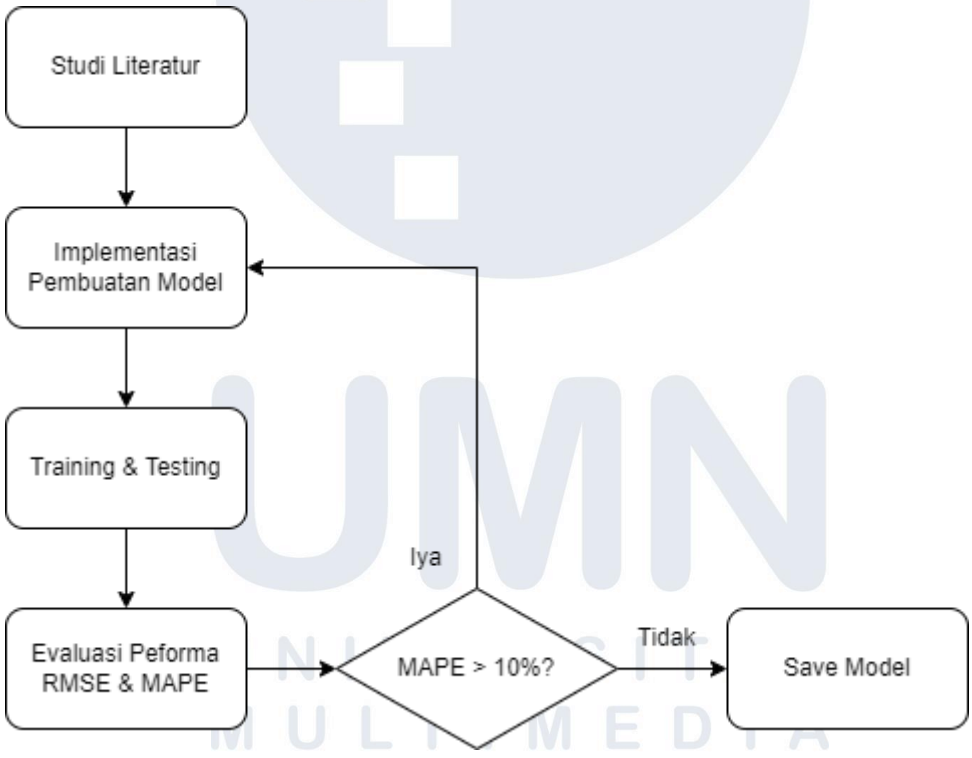
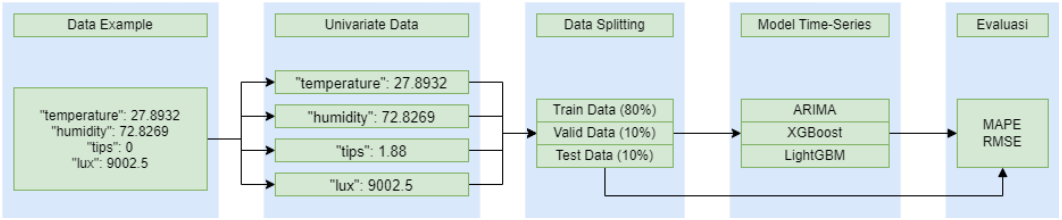
BAB III
ANALISIS DAN PERANCANGAN SISTEM

3.1 Metode Penelitian

Penelitian mengikuti alur seperti pada Gambar 3.1. Pertama adalah melakukan studi literatur, pada fase ini penulis lebih fokus terhadap pencarian informasi yang diperlukan. Lalu implementasi pembuatan model menggunakan data primer yang sudah di proses juga setelah melalui proses uji stasioner. Setelah fase *training* dan *testing*, akan terdapat pengujian performa model menggunakan metrik RMSE dan MAPE, jika performa model yang dibuat kurang baik maka akan mengulang kembali ke fase implementasi atau revisi model. Tetapi jika performa model dilihat sudah memberikan hasil yang baik dengan hasil metrik yang kecil, maka model akan disimpan untuk dipersiapkan digunakan dalam aplikasi.



Gambar 3.1 *Flowchart* alur penelitian



Gambar 3.2 *Data flow diagram* alur pengembangan model

Dari Gambar 3.2, dapat dilihat contoh dari data yang digunakan adalah temperatur, kelembapan, curah hujan (menggunakan *tipping bucket*), dan intensitas cahaya. Lalu data akan dibagi 80% untuk data pelatihan, 10% untuk data validasi dan 10% untuk data pengujian, akan ada tiga model kecerdasan buatan yang dibandingkan, yaitu ARIMA, XGBoost dan LightGBM. Perbandingan akan dilakukan dengan menggunakan metrik pengujian RMSE dan MAPE untuk melihat performa dari model kecerdasan buatan.

3.2 Pengumpulan dan Pemrosesan Data

Data primer diperoleh melalui sensor yang sudah dipasang di sekitar kebun salak di Paguyuban Mitra Turindo. Ada empat jenis variabel cuaca yang diambil, yaitu temperatur, kelembapan, curah hujan serta intensitas cahaya.

Data yang sudah didapatkan memiliki banyak celah, artinya ada data yang tidak tercatat pada hari tertentu, bahkan celah data lebih dari satu bulan. Kejernihan data sangat minimal, sehingga ada dua cara pemrosesan data yaitu:

1. Proses pelatihan hingga pengujian dilakukan menggunakan sebagian data yang jernih.
2. Proses pelatihan hingga pengujian dilakukan dengan menggunakan seluruh data yang sudah dibersihkan meskipun terdapat celah yang besar.

Data diambil dengan interval setiap jam, namun tidak menentu pada menit berapa. *Resampling* dilakukan pada data supaya interval waktunya terjadi pada jam yang sama pada menit yang sama. Terdapat juga *outlier* pada data, data dengan

deviasi yang terlalu besar perlu diproses, seperti nilai terlalu besar dan nilai yang negatif. Pengambilan data juga dilakukan oleh beberapa node berbeda yang menyebabkan beberapa data diambil pada waktu yang sama, pada data demikian dilakukan proses *averaging*, yaitu mengambil rata-rata untuk masing-masing data.

Untuk penelitian ini akan ada dua skenario yang diuji, berikut penjelasan mengenai kedua skenario:

1. Skenario 1, yang digunakan berasal dari node 10, yaitu sekitar tanggal 17 Juli hingga 22 Juli sekitar 129 data.
2. Skenario 2, menggunakan semua data yang tersedia mulai dari 17 Juli hingga 24 November sekitar 536 data.

Kedua skenario diatas dibuat untuk membandingkan performa model ketika dilatih dengan data yang berbeda, yaitu skenario 1 dilatih dengan menggunakan data tanpa celah atau semua data kontinyu tanpa adanya kerusakan. Skenario 2 dilatih dengan menggunakan jumlah data yang lebih banyak namun dengan celah yang besar. Dari kedua skenario yang sudah dibuat juga akan digunakan untuk memprediksi satu minggu kedepan dan dibandingkan dengan nilai aslinya untuk mengetahui performa model dalam melakukan prediksi.

3.3 Perancangan Model

Model ARIMA memiliki beberapa parameter yang perlu ditentukan. Ordo dari parameter p didapatkan dengan melakukan autokorelasi parsial terhadap dataset. Ordo dari parameter d dapat ditentukan dari hasil tes *Augmented Dickey-Fuller*. Lalu ordo untuk parameter q didapatkan dari autokorelasi terhadap dataset. Ordo parameter s yang menunjukan interval musiman diberikan nilai 24 karena data akan berulang setiap 24 jam. Untuk ordo parameter P, D, dan Q diberikan 1 untuk memberi informasi bahwa setiap 24 jam data akan mengalami musiman berulang.

Model XGBoost dan LightGBM dirancang serupa, dengan mengatur objektif untuk regresi, *n_estimators* atau iterasi pelatihan sebesar 50000, *learning_rate*

0.0001, dengan *booster decision tree*. Jumlah iterasi dan *learning_rate* diatur berkesinambungan, karena *learning_rate* yang kecil berarti proses pelatihan model lambat dan membutuhkan banyak iterasi supaya bisa memberikan performa yang tinggi, ditemukan dengan iterasi sebanyak 50000 model bisa memiliki performa yang baik tanpa terjadinya *overfitting*, yaitu kondisi ketika model terlalu mengenali data pelatihan sehingga untuk memprediksi data baru performanya kurang optimal. Pada model *ensemble* juga fitur data perlu diekstraksi untuk melakukan pelatihan, seperti mendapatkan hari, tanggal, jam, minggu, bulan.

3.4 Evaluasi Metrik

Ada dua metrik pengujian yang digunakan untuk menilai performa dari model yang dibuat yaitu *Root Mean Square Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE), RMSE merupakan metrik yang menjadi standar pengukuran performa model dalam penelitian iklim [23], sedangkan MAPE merupakan metrik yang independen dan mudah di interpretasi [24]. Namun MAPE memiliki kekurangan jika data bernilai nol, maka hasil MAPE akan sangat besar. Masing-masing metrik memiliki rumus sebagai berikut:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N-1}(yi-\hat{y}i)^2}{N}} \quad (1)$$

$$MAPE = \frac{1}{N}\sum_{i=0}^{N-1}\frac{|yi-\hat{y}i|}{\max(\varepsilon,|yi|)} \quad (2)$$

N merupakan jumlah data pada dataset, yi merupakan nilai sebenarnya, ŷi merupakan nilai prediksi, dan notasi max(ε,|yi|) merupakan sebuah fungsi agar ketika nilai yi nol tidak menghasilkan hasil MAPE yang tidak terdefinisi. Epsilon dalam Rumus 2 merupakan sebuah variabel yang bernilai sangat kecil, yaitu tepatnya bernilai 2^{-52} atau sekitar 2.2e-16.