BAB 2
LANDASAN TEORI

Beberapa literatur yang telah dikaji untuk melakukan penelitian ini yang berisi tentang *Hidden Markov Model*, *Akaike Information Criterion*, dan *Bayesian Information Criterion*.

2.1 Hidden Markov Model

HMM adalah model yang lebih luas yang memungkinkan hubungan yang lebih fleksibel antara data observasi (*Observation*) dan urutan keadaan tersembunyi (*Hidden State*). Hubungan tersebut dapat disajikan dengan fungsi probabilitas observasi yang sesuai dengan setiap keadaan tersembunyi. Fungsi-fungsi umum yang yang umum digunakan dalam model HMM adalah fungsi densitas normal, normal campuran, atau eksponensial [17].

Inisialisasi probabilitas keadaan awal, pada rumus ini dijelaskan proses untuk melatih model GHMM terhadap data harga penutupan (*close*):

$$P(q_1 = i) = \pi_i \tag{2.1}$$

Keterangan:

- P: Probabilitas terjadinya suatu peristiwa.
- q_1: Keadaan tersembunyi pada langkah pertama.
- i: Indeks yang menunjukkan keadaan tersembunyi tertentu.
- π_i: Probabilitas untuk memulai pada keadaan tersembunyi i.

Probabilitas transisi antara keadaan tersembunyi, model GHMM yang dibuat dengan komponen n dan dilatih pada data tutup (*close*). Selama pelatihan, model akan mempelajari parameter yang mencakup matriks transisi antar state (a_{ij}):

$$P(q_t = j \mid q_{t-1} = i) = a_{ij} \tag{2.2}$$

Keterangan:

- P: Probabilitas terjadinya sebuah peristiwa.

- q_t: Keadaan tersembunyi pada langkah waktu t.
- i, j: Indeks yang menunjukkan keadaan tersembunyi tertentu.
- a_{ij}: Probabilitas transisi dari keadaan tersembunyi i ke keadaan tersembunyi j.

Probabilitas emisi (kemungkinan dari pengamatan) untuk observasi o pada waktu t ketika model berada di state i, perhitungan ini dilakukan untuk menghitung hasil *log likelihood* dengan menggunakan fungsi *model.score*:

$$P(O_t = o \mid q_t = i) = \mathcal{N}(o \mid \mu_i, \Sigma_i) \tag{2.3}$$

Keterangan:

- P: Probabilitas terjadinya suatu peristiwa.
- O_t: Keluaran yang diamati pada langkah waktu t.
- o: Nilai spesifik dari keluaran yang diamati pada langkah waktu t.
- q_t: Keadaan tersembunyi pada langkah waktu t.
- i: Indeks yang menunjukkan keadaan tersembunyi tertentu.
- μ_i: Rata-rata dari distribusi Gaussian untuk keadaan tersembunyi i.
- Σ_i: Matriks kovariansi dari distribusi Gaussian untuk keadaan tersembunyi i.
- $\mathcal{N}$: Fungsi kepadatan probabilitas dari distribusi Gaussian.

Probabilitas dari sebuah urutan pengamatan menunjukkan probabilitas total dari observasi $O_1, O_2, ..., O_T$ dalam model HMM:

$$P(O_1, O_2, ..., O_T) = \sum_{q_1, q_2, ..., q_T} P(O_1, O_2, ..., O_T, q_1, q_2, ..., q_T) \tag{2.4}$$

Keterangan:

- P: Probabilitas terjadinya suatu peristiwa.
- $O_1, O_2, ..., O_T$: Urutan keluaran yang diamati.
- $q_1, q_2, ..., q_t$: Urutan keadaan tersembunyi.

- $\sum$: Penjumlahan dari semua kemungkinan urutan keadaan tersembunyi.

Perhatikan bahwa variabel t dan T mewakili langkah waktu, sedangkan variabel i dan j mewakili kondisi tersembunyi tertentu. Variabel o, O_t, μ_i, dan Σ_i merepresentasikan keluaran yang diamati dan parameter terkait. Variabel π_i dan a_{ij} masing-masing merepresentasikan probabilitas keadaan awal dan probabilitas transisi.

Evaluasi kinerja model akan menggunakan data dari *explained variance score* (EVS), *root mean squared error* (RMSE), *mean square error* (MSE), *mean absolute error* (MAE), dan *R-Squared Score* (Skor R2). *Root mean square error* (RMSE) telah digunakan sebagai metrik statistik standar untuk mengukur kinerja model dalam penelitian meteorologi, kualitas udara, dan iklim. *mean absolute error* (MAE) adalah ukuran lain yang berguna yang banyak digunakan dalam evaluasi model [18]. EVS digunakan untuk mengukur sejauh mana model regresi dapat menjelaskan variasi dalam data aktual dan Skor R2 digunakan untuk mengukur seberapa baik model regresi cocok dengan data [19]. Pemilihan metrik ini dipilih untuk menentukan variabel terbaik dengan model dan seberapa jauh model regresi dapat mengukur akurasi data.

MAE digunakan untuk mengetahui seberapa jauh prediksi model secara rata-rata. MAE juga biasanya digunakan ketika ada outlier atau nilai ekstrem dalam data, karena kurang sensitif terhadap *outlier* dibandingkan metrik lain seperti RMSE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{2.5}$$

Keterangan:

- n adalah jumlah sampel
- y_i adalah nilai aktual dari sampel ke-i
- $\hat{y}_i$ adalah nilai prediksi dari sampel ke-i
- $|\cdot|$ merupakan fungsi nilai absolut

RMSE digunakan ketika kesalahan dalam prediksi harus dihukum lebih berat karena besar, mirip dengan MAE tetapi memberikan bobot yang lebih tinggi pada kesalahan yang besar.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\, y_i - \hat{y}_i \,)^2} \tag{2.6}$$

Keterangan:

- n adalah jumlah sampel
- $\hat{y}_i$ adalah nilai prediksi dari sampel ke-i
- $(\cdot)^2$ mewakili operasi penguadratan
- $\sqrt{\cdot}$ mewakili fungsi akar kuadrat

MSE digunakan ketika ingin memberikan kesalahan yang lebih besar dengan lebih berat dan bersedia bekerja dengan metrik yang tidak dapat ditafsirkan secara langsung dengan cara yang sama seperti variabel target asli.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\, y_i - \hat{y}_i \,)^2 \tag{2.7}$$

Keterangan:

- n adalah jumlah sampel
- y_i adalah nilai aktual dari sampel ke-i
- $\hat{y}_i$ adalah nilai prediksi dari sampel ke-i
- $(\cdot)^2$ mewakili operasi penguadratan

Nilai R2 biasanya digunakan ketika ingin mengetahui seberapa besar variasi dalam variabel target yang dijelaskan oleh model. Kisarannya dari 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan kinerja yang lebih baik.

$$\text{R2 Score} = 1 - \frac{\text{SSres}}{\text{SStot}} \tag{2.8}$$

Keterangan:

- SS_{res} adalah jumlah kuadrat residual (yaitu, jumlah kuadrat perbedaan antara nilai aktual dan nilai prediksi)
- SS_{tot} adalah jumlah kuadrat total (yaitu, jumlah perbedaan kuadrat antara nilai aktual dan rata-rata nilai aktual)

EVS biasanya digunakan ketika ingin mengetahui seberapa besar variasi dalam variabel target yang dijelaskan oleh model, mirip dengan Skor R2. Kisarannya dari 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan kinerja yang lebih baik.

$$\text{EVS} = 1 - \frac{\text{Var}(y_i - \hat{y}_i)}{\text{Var}(y_i)} \tag{2.9}$$

Keterangan:

- Var adalah fungsi varians
- y_i adalah nilai aktual dari sampel ke-i
- $\hat{y}_i$ adalah nilai prediksi dari sampel ke-i

2.2 Akaike Information Criterion (AIC)

Metode *Akaike Information Criterion* (AIC) adalah teknik analisis yang digunakan untuk menghasilkan model faktor produksi terbaik. Metode ini menggunakan estimasi *maximum likelihood* sebagai perhitungan yang tepat. [20]. *Akaike Information Criterion* (AIC) adalah cara yang ampuh untuk membedakan antar model. AIC mempertimbangkan kesesuaian dan jumlah parameter dalam model, tetapi belum banyak digunakan untuk data imitansi. AIC dapat digunakan untuk membandingkan skema pembobotan yang berbeda dan juga model yang berbeda [21]. Persamaan dalam metode *Akaike Information Criterion* (AIC) sebagai berikut.

$$AIC = 2 \times k - 2 \times ln(L) \tag{2.10}$$

Keterangan:

- k adalah jumlah parameter dalam model
- L adalah nilai maksimum fungsi likelihood untuk model
- n adalah ukuran sampel

2.3 Bayesian Information Criterion (BIC)

Dalam statistik, *Bayesian information criterion* (BIC) adalah kriteria untuk model pemilihan model di antara sekumpulan model yang terbatas. Hal ini

didasarkan, sebagian, pada fungsi likelihood, dan terkait erat dengan *Akaike kriteria informasi* (AIC). Saat menyesuaikan model, dimungkinkan untuk meningkatkan kemungkinan dengan menambahkan parameter, tetapi hal ini dapat mengakibatkan over fitting. BIC menyelesaikan masalah ini dengan memperkenalkan istilah penalti untuk jumlah parameter dalam model [22]. Kriteria BIC didasarkan pada teori Bayesian dan bertujuan untuk memaksimalkan probabilitas posterior dari suatu model yang diberikan data. Hal ini memungkinkan untuk meningkatkan kemungkinan fitting model dengan menambahkan parameter, namun hal ini dapat mengakibatkan overfitting. Dengan memasukkan istilah penalti untuk jumlah parameter dalam model, BIC dapat mengatasi masalah ini [15]. Persamaan dalam metode *Bayesian Information Criterion* (BIC) sebagai berikut. *Overfitting* terjadi ketika sebuah algoritma mengurangi kesalahan melalui penghafalan pelatihan. *Underfitting* terjadi ketika sebuah algoritma tidak memiliki kapasitas model yang memadai atau pelatihan yang cukup untuk mempelajari hubungan yang sebenarnya, baik melalui hafalan atau tidak [23].

$$BIC = k \times ln(n) - 2 \times ln(L) \tag{2.11}$$

Keterangan:

- k adalah jumlah parameter dalam model
- L adalah nilai maksimum fungsi *likelihood* untuk model
- n adalah ukuran sampel