1 PENDAHULUAN

1.1 Latar Belakang

Perkembangan *Educational Data Mining* (EDM) dalam mengembangkan metode untuk melakukan analisis pada macam-macam tipe data. Lingkungan akademis berperan besar untuk melakukan trobosan baru dalam mencapai suatu intensi pada dataset yang sedang dihadapi. Hubungan terserbut terkosolidasi pada *Learning Analytics* (LA) yang mana akan berpengaruh pada analisis, penilaian, dan laporan pada suatu data. Sehingga dapat diketahui pemahaman dan pengertian pada data secara mendalam (Romero & Ventura, 2020).

Peramalan deret waktu atau biasa dikenal dengan *time series forecasting* untuk melakukan prediksi kejadian yang akan terjadi kedepannya berdasarkan analisis deret waktu (Castán-Lascorz et al., 2022). Algoritma baik dari konvensional, *machine learning,* dan *deep learning* dikembangkan untuk menjawab tantangan industri. Proses meliputi observasi pada data sebelumnya atau sekarang dan akan dilakukan *forecasting* pada masa yang akan datang (Karanikola et al., 2022). Peran ini juga dapat membantu pengelola web untuk mengetahui trafik yang akan terjadi berdasarkan kebiasan (*behaviour*) dari data sebelumnya.

Prediksi *web traffic* pada dasarnya dapat dimodelkan secara matematis dengan rangkaian waktu. Rangkaian waktu tersebut biasanya digunakan dalam variabel harian dan jumlah kunjungan visitor pada web tersebut. Menganalisis *web traffic* memberikan informasi tentang lalu lintas web seperti *trend* dan *seasonality* biasanya akan dilakukan dengan data yang sudah memiliki rentang waktu yang signifikan (Casado-Vara et al., 2021). Data yang banyak menghasilkan model dan hasil *forecast* yang akurat "semakin banyak data semakin baik model belajar dalam melakukan prediksi". Analisis dan prediksi ini akan memberikan manfaat pada *administrator* dan *user* untuk mendapatkan *win win solution*.

Definisi keuntungan ini mengacu pada tiga hal kecepatan, volume, dan okupansi pada *traffic* yang terjadi. Tiga hal ini yang akan membantu pengelola (*administrator*) untuk memberikan layanan dan meramajakan sistem (Ma et al., 2020). Kecepatan membantu user dalam menjangkau layanan yang diberikan. Volume *traffic* memberikan informasi pada *administrator* untuk mengetahui jumlah pengunjung. Okupansi mengacu pada jumlah *instance* yang digunakan *administrator* agar tidak terjadi *over occupation*. Dari penjabaran diatas memprediksi hasil kunjungan akan mebantu memberikan keuntungan baik sisi *user* atau *administrator*.

Pada dasarnya sebuah *service* berjalan pada suatu sistem (*hardware*) yang dilakukan secara prosedural sehingga mendapatkan output yang diinginkan. Banyaknya input yang diterima akan berbanding lurus pada proses yang dijalankan. Proses tersebut akan menggunakan banyak *resource* yang digunakan oleh sistem. Dalam komputasi awan (*cloud computing*) menjadi sangat populer karena fleksibiltasnya dalam memberikan sistem infrastuktur. Keunggulannya sistem ini dapat menyediakan *instance* dengan cepat sesuai dengan permintaan pengguna. Kelemahan sistem ini perlu waktu untuk membentuk *instance* jika diperlukan dan maksimal *instance* ditentukan oleh *user* tersebut (Ouhame et al., 2021).

Analisis teknikal (*technical analysis*) kerap dihubungkan dalam suatu kegiatan ekonomi. Analisis ini berlandaskan pergerakan suatu nilai (harga) pada suatu rentang waktu. *Exponential Moving Average* (EMA) atau *Single Exponential Smoothing* salah satu metode yang dapat digunakan dalam analisis teknikal. Indikator EMA memberikan bobot yang besar pada nilai terkini dan memberikan bobot yang kecil pada nilai yang sudah terlampau jauh (Lusindah & Sumirat, 2021).

EMA lebih cepat bereaksi pada perubahan nilai dibandingkan *Moving average* (MA) biasa. Melakukan kalkulasi lebih cepat untuk memuluskan fluktuasi harga, membuat tren pergerakan nilai sehingga menjadi jelas. Analisis ini merupakan keunggulan menggunakan EMA, kesesuaian yang terjadi pada pergerakan nilai pada identfikasi pola atau tren yang terjadi dan menghasilkan fluktuasi nilai. Mekanisme pembobotan memungkinkan EMA bereaksi lebih cepat pada informasi baru (Gruevski, 2021).

Pemilihan fitur (*feature* selection) merupakan *subset* teperting untuk mewakilkan kondisi data asli. Untuk melakukan *forecasting* setiap fitur yang memiliki potensial dievaluasi untuk membangun model atau persamaan matematis. Fitur dapat dikurangi maupun ditambahkan (*feature engineering*) ini berfungsi untuk mengeleminasi data yang tidak digunakan atau memberikan informasi tambahan berdasarkan apa yang sedang terjadi pada data. *Feature selection* terbagi menjadi tiga bagian *filtering* berdasarkan kalkulasi statistika, *wrapper* melakukan operasi pada fitur, dan *embedded* berdasarkan kontribusi fitur saat melakukan *model training* (Karasu et al., 2020).

Feature selection merupakan tahapan penting pada analisis data, *machine learning* (ML), dan *data mining*. Terlebih pada data yang memiliki dimensi yang sangat besar, tahapan ini dibutuhkan agar data tetap mempunyai data yang terintegritas. Untuk permodelan ML ini dapat mempengaruhi tingkat akurasi dari suatu model yang dibangun untuk melakukan prediksi (Bommert et al., 2020). Pada kasus *forecasting web traffic* fitur-fitur akan diseleksi apakah ada hubungan antara jumlah kunjungan "hari ini" dengan hari-hari sebelumnya. Analisis ini yang akan membantu membangun model ML dalam melakukan *forecasting*.

XGBoost merupakan algoritma ML yang biasa digunakan untuk melakukan model regressi dan klasifikasi. XGBoost ditemukan pertama kali oleh Chen dan Guestrin pada tahun 2016 yang dikembang dari arsitektur *Gradient Boost*. Algoritma ini sangat efektif karena gabungan dari beberapa pohon keputusan (*tree-based ensemble learning*) (Shahani et al., 2021). Cara kerja bagaimana setiap pohon memiliki bobot yang disempurnakan agar mencapai suatu prediksi yang sangat baik. Biasanya XGBoost digunakan untuk melakukan *supervised learning* seperti klasifikasi dan regressi, namun sekarang dikembangkan untuk melakukan *time series forecasting*.

Kemampuan *Gradient Boosting* (GB) memperbaiki model sebelumnya atau memberikan model baru berdasarkan residu yang ada, sehingga memiliki hasil prediksi dengan memaksimalkan semua fitur. Tidak seperti *Decision Tree* (DT) biasa, GB yang memulai pohon keputusan berdasarkan *leaf* karena algoritma ini mengizinkan pola yang kompleks dan hubungan setiap data agar mendapatkan hasil prediksi yang kuat (Liew et al., 2021). Ini memungkinkan untuk mengetahui melakukan prediksi berdasarkan *time series forecasting* yang harus mengetahu pola-pola yang rumit berdasarkan data sebelumnya dan fitur-fitur yang kompleks.

Berdasarkan penjabaran XGBoost dan GB dapat memberikan hasil akurasi dengan optimasi setiap *feature* yang ada. Dalam segi meningkatkan permodelan salah satunya menambahkan fitur untuk pembelajaran mesin (*feature engineering*) dan melakukan pemilihan pada setiap fitur (*feature selection*) yang ada untuk mendapatkan model yang kuat (*robust*) (Mudassir et al., 2020a). EMA memberikan tambahan informasi pada model ML berdasarkan perubahan data. Fokus EMA memberikan bobot lebih besar pada data baru sehingga memberikan *trend analysis* lebih jelas dan membantu dalam mengurangi dampak fluktasi acak dan *noise*.

Leela dan kawan-kawan melakukan prediksi pada web trafik menggunakan LSTM dan CNN dengan banyak data sebanyak 4800. Fitur-fitur data merupakan traffic interval jam dan banyak kunjungan user. Eksplorasi dengan melakukan visualisasi harian dan mingguan untuk medapatkan pola pada dataset. LSTM mendapatkan hasil lebih baik daripada CNN. Pada penelitian ini tidak terlalu jelas informasi pada saat eksplorasi data dan tidak ada fitur tambahan atau melakukan *feature engineering* untuk menangkap pola tersembunyi pada dataset (Dharani et al., 2023).

Litha dan Tasrif melakukan analisis perbandingan MA antara *Simple Moving Average* dan *Exponential Moving Average* (EMA) untuk memprediksi jumlah penderita covid-19. Pada dasarnya penelitian ini hanya memprediksi jumlah rata-rata penderita covid harian. Berdasarkan hasil yang didapatkan EMA memiliki hasil lebih baik dalam melakukan prediksi berdasarkan nilai *error* yang didapatkan. Penelitian memberikan informasi metode EMA dapat beradaptasi lebih cepat dengan data baru (Litha Sari, 2020a).

Zhenchang dan kawan-kawan melakukan *forecasting* pada penjualan mobil menggunakan XGBoost. Hasilnya penelitian yang dilakukan mendapatkan posisi pertama pada kompetisi "*Alibaba Cloud & the Yancheng Municipal Government*" karena memiliki akurasi terbaik dari peserta lainnya. Dengan menggunakan *sliding window* untuk menemukan hubungan antara histori data dengan data sekarang dapat meningkatkan akurasi prediksi. *Slinding window* merupakan teknik digunakan untuk menambah fitur berdasarkan data sebelumnya (Xia et al., 2020a).

Berdasarkan analisa yang dilakukan dengan acuan penelitian-penelitian sebelumnya, penulis akan melakukan analisa *time series forecasting* pada trafik web. Pentingnya penelitian ini dilakukan untuk mendapatkan manfaat pengetahun dalam analisis *time-series forecasting* dan menentukan kebutuhan sistem pada suatu server. Penelitian ini menggunakan algoritma XGBoost yang dipertimbangkan dari penelitian sebelumnya dengan analisa setiap fitur yang mendalam untuk menangkap informasi tersembunyi. Penambahan fitur EMA dan *sliding window* untuk memperkuat dan memperluas pembelajaran mesin. Fitur-fitur seperti *day, day of week, month,* dan *year* akan dibandingkan dengan fitur tambahan menggunakan metode *fieature importance*. Penulis menggunakan data tabular dengan deskripsi 2167 baris data dan 8 fitur untuk melakukan penelitian ini. Data yang dilakukan merupaka data observasi dari 14 September 2014 sampai 19 Agustus 2020. Pengolahan data yang baik dapat membentuk suatu pengetahuan yang baru (Lv et al., 2021) dan akan mendapatkan manfaat dari hal tersebut.

1.2 Identifikasi Masalah

Penulis mengidentifikasi masalah berdasarkan penjabaran latar belakang yang telah dibahas sebelumnya. Berikut permasalahan yang dapat penulis temukan:

1. Menggunakan Deep Learning untuk melakukan time series forecasting memiliki kekurangan pada sumberdaya komputasi yang besar untuk jumlah data yang terbilang sedikit.
2. Kurangnya analisis tambahan yang dilakukan pada saat *data preparation*.
3. Kurangnya analisis akhir untuk mengetahui fitur-fitur yang sangat berpengaruh pada sebuah model prediksi.
4. Belum memanfaatkan atau menambahkan fitur-fitur lain untuk memperkaya pembelajaran model dan menignkatkan hasil prediksi.

1.3 Rumusan Masalah

Berdasarkan identifikasi masalah-masalah yang ditemukan penulis di sub-bab sebelumnya, penulis telah merumuskan masalah-masalah sebagai pokok penelitian menjadi dua bagian antara lain:

1. Bagimana hasil akurasi pada model ML setelah ditambahkan fitur EMA dan *sliding window*?
2. Fitur mana yang paling berpengaruh membangun model untuk melakukan prediksi?

1.4 Tujuan

Tujuan penulis melakukan penelitian ini dibuat untuk mengetahui hal-hal sebagai berikut:

1. Mengetahui hasil akurasi ML setelah mendapatkan fitur tambahan EMA dan Sliding windows.
2. Mengetahui fitur mana yang memberikan dampak paling besar pada model yang dibangun.

1.5 Manfaat

Adapun manfaat yang diberikan penelitian ini baik dalam akademis dan para *stakeholder*:

1. Untuk para *stakeholder* penelitian ini akan memberikan prediksi kunjungan sehingga dapat memberikan estimasi biaya dalam membentuk sistem yang tepat, pemeliharaan sistem dan mengurangi kerugian akibat resiko yang ditimbulkan.
2. Untuk para akademisi, penelitian ini akan memberikan informasi tambahan cara pengolahan data berbasis *time series forecasting* menggunakan XGBoost dengan penambahan fitur EMA. Penilitian ini akan membantu para akademisi atau pembaca sebagai tambahan referensi untuk kemajuan pendidikan dan teknologi informasi.

1.6 Batasan Masalah

Berdasarkan rumusan masalah yang dijabarkan di atas, penulis memberikan batasan masalah agar mencapai tujuan dari penelitian ini. Penulis membatasi penelitian ini pada hasil akurasi ML menggunakan XGBoost dengan penambhan fitur EMA dan mencari fitur yang paling berpengaruh pada pembuatan model ML. Penulis akan membandingkan model *machine learning* dengan evaluasi akhir menggunakan *Mean Absolute Error, Root Mean Square Error,* dan *Mean Absolute Percentage Error*. Penulis juga akan menggunakan parameter dengan nilai konstan untuk setiap model yang dibangun agar tidak terjadi bias saat perbandingannya. Batasan ini diharapkan menjadi landasan jika terdapat kritik atau saran yang akan diberikan.

1.7 Sistematika Pembahasan

1. Bab 1. Pendahuluan

   Bab ini berisi tentang latar belakang, rumusan masalah, tujuan masalah, manfaat, batasan masalah, dan sistematika pembahasan

2. Bab 2. Landasan Teori

   Landasan teori berisi uraian dan pembahasan tentang teori, konsep, model, metode, atau sistem dari literatur ilmiah,yang berkaitan dengan tema, masalah, atau pertanyaan penelitian.

3. Bab 3. Metodologi

   Metodologi penelitian berisi tipe penelitian, strategi dan rancangan penelitian, subjek atau partisipan penelitian, lokasi penelitian, metode/teknik pengumpulan data, metode/teknik analisis data dan pembahasan hasilnya, peralatan pendukung, metode/teknik lainnya, dan metodologi terfokus.

4. Bab 4. Implementasi

   Deskripsi struktur dan setiap komponen utama Representasi data dalam model data dan basis data. Detil implementasi dari fungsi-fungsi utama yang menjadi fokus.

5. Bab 5. Analisa dan Evaluasi Hasil

   Ringkasan hasil pengujian perangkat lunak, termasuk data dan analisisnya (detilnya di Lampiran). Evaluasi hasil proyek secara keseluruhan.

6. Bab 6. Penutup

   Kesimpulan memuat secara singkat dan jelas tentang hasil penelitian yang diharapkan sesuai dengan tujuan penelitian.