Analisis dan Prediksi Data Penjualan Menggunakan Machine Learning dengan Pendekatan Ilmu Data

Ferdy Riza [1*]
[1]Universitas Muhamadiyah Sumatera Utara, Fakultas Teknik dan Ilmu Komputer, Sistem Informasi

[1*] ferdydna89@gmail.com

Received: 29 Desember 2021 Accepted: 9 Januari 2022 Published: 12 Januari 2022

*** Ferdy Riza**

Keywords: Analisis Data Penjualan, CRIS-DM, Prediksi Penjualan, XGBoost, LightGBM

Abstrak : Pendekatan *Data Science* (ilmu data) memberi peluang besar untuk menggunakan data history dan mengubahnya menjadi wawasan yang berguna untuk membangun model prediksi penjualan masa depan. akan tetapi, model prediksi membutuhkan analisis data tertentu untuk menghasilkan model yang kuat, termasuk jumlah pelanggan, jumlah pelanggan yang hilang, tingkat penjualan rata-rata serta kecenderungan musiman. Makalah ini menganalisis data penjualan menggunakan kerangka kerja ilmu data dengan desain sesuai prinsip CRIS-DM yang terdiri dari tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan. Pemodelan digunakan algoritma *Machine Learning* untuk memprediksi penjualan di masa depan yang hasil kinerjanya dievaluasi dengan RMSE, MEA dan R^2. Berdasarkan hasil pengujian algoritma XGBoost dan LightGBM menghasilkan nilai R^2 mencapai 60% dengan tingkat kesalahan deteksi terendah dibandingkan algoritma lainnya..

PENDAHULUAN

Dalam beberapa tahun terakhir ini, analisis bisnis dengan menggunakan *data science* (Ilmu data) memiliki peran penting dalam menentukan kebijakan strategi bisnis khususnya meningkatkan pendapatan. Setiap industri dapat mengambil manfaat dari keputusan berbasis data, terstruktur dengan baik. Industri penjualan secara aktif menerapkan solusi ilmu data untuk keuntungannya [1]. Inovasi yang dibawa oleh Ilmu Data ke penjualan sebagian besar meningkatkan pengalaman pelanggan dan sebagai hasilnya meningkatkan penjualan. KPI (*Key Performance Indicators*) penjualan dan ROI (*Return on Investment*) dapat ditingkatkan dengan upaya yang lebih sedikit juga. Tentu saja, untuk mencapai tujuan ini banyak pengumpulan data, pemrosesan, pembersihan akan dibutuhkan.

Pendekatan ilmu data memberi peluang besar untuk menggunakan data history dan mengubahnya menjadi wawasan yang berguna untuk meningkatkan pendapatan. Prediksi penjualan masa depan membawa kelegaan besar bagi perusahaan yang bekerja dengan penjualan. Pengelolaan manajemen stok yang baik dapat mengoptimalkan kerugian atau kelebihan stok [2], karena jika terlalu banyak stok item dari satu produk dapat berisiko tidak memiliki ruang penyimpanan untuk item lainya atau terpaksa menjual dengan harga diskon [3]. Jika sebaliknya, penjualan akan menurun apabila jumlah barang yang terlalu sedikit dan tentu akan berdampak pada pelayanan maupun pendapatan[4]. Prediksi penjualan masa depan memungkinkan untuk menghindari masalah ini dan membuat keputusan yang lebih baik.

Model prediksi membutuhkan data tertentu termasuk jumlah pelanggan yang diperoleh, jumlah pelanggan yang hilang, tingkat penjualan rata-rata serta kecenderungan musiman [5]. Selain itu, asumsi penjualan keadaan yang berubah yang dapat mempengaruhi penjualan secara signifikan harus ditentukan sebelumnya [6]. Algoritme perkiraan penjualan mencari pola dalam data ini. Pola yang terdeteksi selanjutnya digunakan untuk menilai kecenderungan umum dari transaksi dalam alur untuk membangun prediksi

dengan tingkat akurasi yang tinggi [7]. Sangat penting memami data yang akan memberikan informasi berharga dalam membangun algoritma perkiraan. Berbagai pendekatan telah banyak diusulkan oleh peneliti, seperti *Adaptive Neuro Fuzzy Inference System* [6], *Recurrent Neural Networks* [8], *Fuzzy Inferences System* [9], *Deep Learning* [10], *Prophet* [11], *Generative Adversarial Networks* [12]. Akan tetapi, hasil evaluasi kinerja algoritma *machine learning* memberikan hasil yang perlu di pertimbangkan untuk membangun model peramalan yang kuat, selain itu algoritma ini juga dapat diterapkan untuk peramalan kuantitatif maupun kualitatif [13].

Dalam makalah ini fokus menganalisis data penjualan dengan menggunakan kerangka kerja ilmu data untuk memahami dan menemukan wawasan yang berharga dari kumpulan data transaksi penjualan sehingga menghasilkan fitur yang optimal diterapkan pada algoritma perkiraan penjualan. Selanjutnya, hasil analisis data penjualan ini akan diterapkan pada algoritma *machine learning* untuk menemukan algoritma yang paling akurat memprediksi penjualan di masa depan.

TINJAUAN LITERATUR

Peramalan bisnis adalah proses memprediksi perkembangan bisnis di masa depan berdasarkan analisis tren pada data masa lalu dan sekarang. Perusahaan melakukan prakiraan bisnis untuk menentukan tujuan, target, dan rencana proyek mereka untuk setiap periode baru, baik perencanaan triwulanan, tahunan, atau bahkan 2–5 tahun. Peramalan bisnis mengacu pada alat dan teknik yang digunakan untuk memprediksi perkembangan dalam bisnis, seperti penjualan, pengeluaran, dan keuntungan [14]. Tujuan dari peramalan bisnis untuk mengembangkan strategi yang lebih baik berdasarkan prediksi informasi. Data masa lalu dikumpulkan dan dianalisis melalui model kuantitatif atau kualitatif sehingga pola dapat diidentifikasi dan dapat mengarahkan perencanaan permintaan, operasi keuangan, produksi masa depan, dan operasi pemasaran. Ketika dilakukan dengan benar, peramalan menambah keunggulan kompetitif dan dapat menjadi perbedaan antara perusahaan yang sukses dan tidak berhasil [13].

Peramalan bisnis yang baik memungkinkan organisasi untuk mendapatkan wawasan unik dan eksklusif tentang kemungkinan peristiwa di masa depan, memanfaatkan sumber daya mereka, mengatur tim produk OKR, dan menjadi pemimpin pasar. Peran bisnis yang paling signifikan adalah memperkirakan penjualan di masa depan, sehingga prediksi masa lalu harus akurat untuk pengembangan dan peningkatan perusahaan dengan melakukan prakiraan bisnis yang cermat dan terperinci untuk menjamin pengambilan keputusan yang baik berdasarkan data dan logika, bukan emosi atau firasat. Peramalan dan perencanaan bisnis dapat dilakukan dengan dua pendekatan utama, yaitu peramalan kuantitatif dan kualitatif [15].

Peramalan kuantitatif adalah metode peramalan bisnis jangka panjang yang hanya berkaitan dengan data terukur seperti statistik dan data historis. Kinerja masa lalu digunakan untuk mengidentifikasi tren atau tingkat perubahan. Jenis peramalan bisnis ini sangat berguna untuk peramalan jangka panjang dalam bisnis [14]. Peramalan kuantitatif diterapkan ketika ada data masa lalu yang akurat yang tersedia untuk menganalisis pola dan memprediksi kemungkinan kejadian di masa depan dalam bisnis atau industri. Peramalan kuantitatif mengekstrak tren dari data yang ada untuk menentukan hasil yang lebih mungkin. Ini menghubungkan dan menganalisis variabel yang berbeda untuk menetapkan sebab dan akibat antara peristiwa, elemen, dan hasil. Contoh data yang digunakan dalam peramalan kuantitatif adalah angka penjualan masa lalu. Model kuantitatif bekerja dengan data, angka, dan rumus, sedikit campur tangan manusia dalam analisis ini [13].

Peramalan bisnis kualitatif adalah prediksi dan proyeksi berdasarkan pendapat para ahli dan pelanggan. Metode ini paling baik bila ada data masa lalu yang tidak cukup untuk dianalisis untuk mencapai perkiraan kuantitatif. Peramalan kualitatif bergantung pada pakar industri atau "pakar pasar" untuk membuat prediksi jangka pendek. Teknik-teknik ini sangat berguna dalam meramalkan pasar yang data historisnya tidak cukup untuk membuat kesimpulan yang relevan secara statistik. Dalam kasus ini, pakar industri dan peramal mengumpulkan data yang tersedia untuk membuat prediksi kualitatif. Model kualitatif paling berhasil dengan proyeksi jangka pendek [16].

Sangat penting untuk di perhatikan bahwa perubahan signifikan dalam perusahaan, seperti fokus produk

baru, pesaing baru atau strategi bersaing, atau perubahan persyaratan kepatuhan mengurangi hubungan antara tren masa lalu dan masa depan. Ini membuat pemilihan metode peramalan yang tepat menjadi lebih penting [17].

BAHAN DAN METODE

Penelitian ini menggunakan pendekatan ilmu data dengan desain sesuai prinsip data mining yang disebut CRISP-DM [18]. Metodologi Ini terdiri dari enam langkah: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Penerapan.

1. Pemahaman Bisnis

Memahami pernyataan masalah adalah langkah pertama dan terpenting untuk membantu dalam memberikan intuisi tentang apa yang akan dilakukan sebelumnya. Berdasarkan informasi data sumber dataset, kumpulan data ini bertujuan membangun model prediktif untuk mencari tahu setiap toko, faktor-faktor kunci yang dapat meningkatkan penjualan dan perubahan apa yang dapat dilakukan pada karakteristik produk atau toko. Dalam masalah ini akan ditentukan empat level hipotesis yaitu Tingkat Toko, Tingkat Produk, Tingkat Pelanggan, dan Tingkat Makro.

2. Pemahaman Data

Data yang dikumpulkan dan diakses untuk penelitian dalam penelitian ini adalah dataset Big Mart Sales Data. Dataset ini merupakan kumpulan data penjualan dari 10 toko yang terletak di lokasi berbeda dengan masing-masing toko memiliki 1559 produk berbeda sesuai pengumpulan data tahun 2013 dengan atribut tertentu dari setiap produk dan toko telah ditentukan.

3. Persiapan Data

Data mentah dapat berisi berbagai jenis pola yang mendasarinya. Hal Ini merupakan bagian penting dalam penelitian ini sebelum menerapkannya pada algoritma, karena dapat berisi nilai null, nilai yang terlalu tinggi, atau berbagai jenis ambiguitas yang juga memerlukan pra-pemrosesan data. Oleh karena itu, dataset harus diselidiki sebanyak mungkin.

4. Pemodelan Data

Pengembangan model prediktif, terdiri dari *Lasso Regressor, Linear Regression, Random Forest* (RF), LightGBM (LG) dan XGBoost.

5. Evaluasi

Pada bagian ini akan eilakukan evaluasi kinerja keseluruhan model terdiri dari MAE (*Mean Absolute Error*), RMSE (*Root Mean Square Error*) dan *R-Square*

6. Penerapan

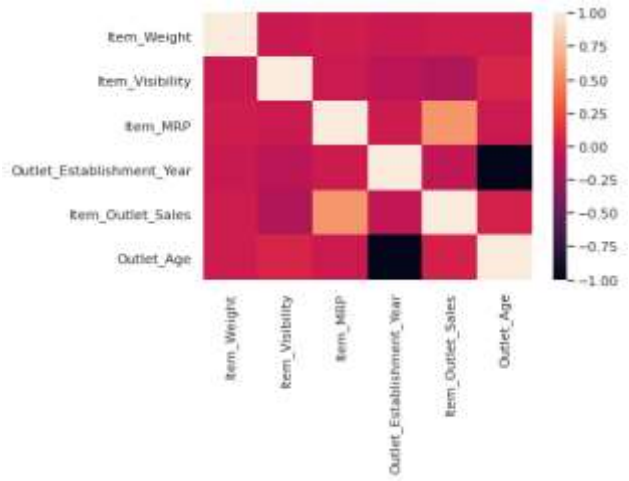
Keseluruhan aplikasi dibangun pada lingkungan kerja Jupyter Notebook dengan python 3 dan beberapa pustakan yang digunakan untuk manipulasi dan visualisasi data seperti numpy, pandas, sklearn, seaborn, matplotlib dan lainnya.

HASIL PENELITIAN

Pada bagian ini diuraikan hasil pengamatan dan akurasi pemodelan yang diusulkan. Pustaka numpy, pandas, matplotlib, seaborn digunakan untuk pemrosesan data termasuk dalam perhitungan ilmiah, visualisasi dan analisis data. Semua kode sumber menggunakan bahasa pemrograman python versi 3 pada lingkungan kerja jupyter notebook yang terbukti bekerja dengan baik karena keunggulannya dalam 'pemrograman literate', di mana kode mudah diterapkan diselingi dalam blok kode.

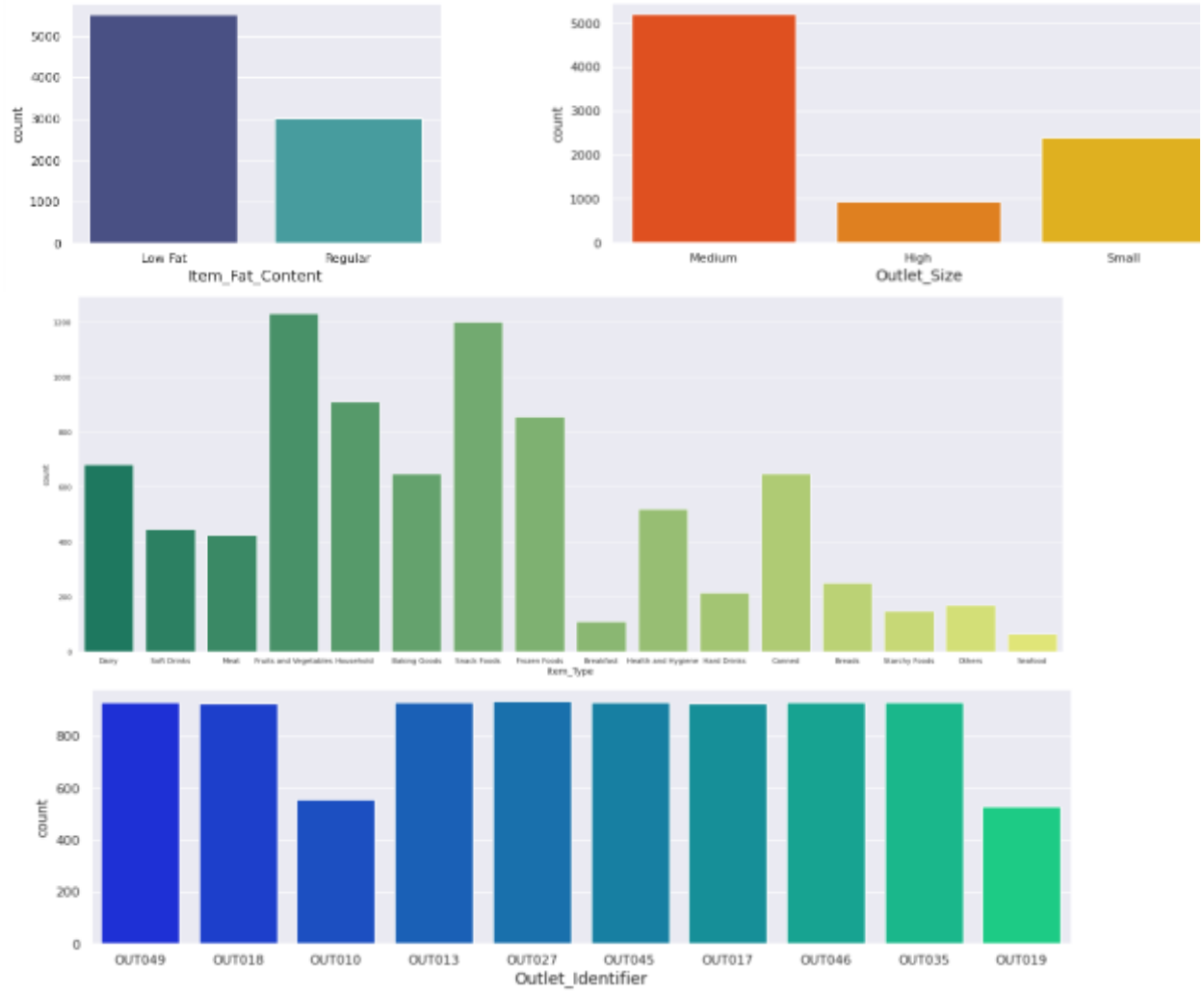
Analisis Data

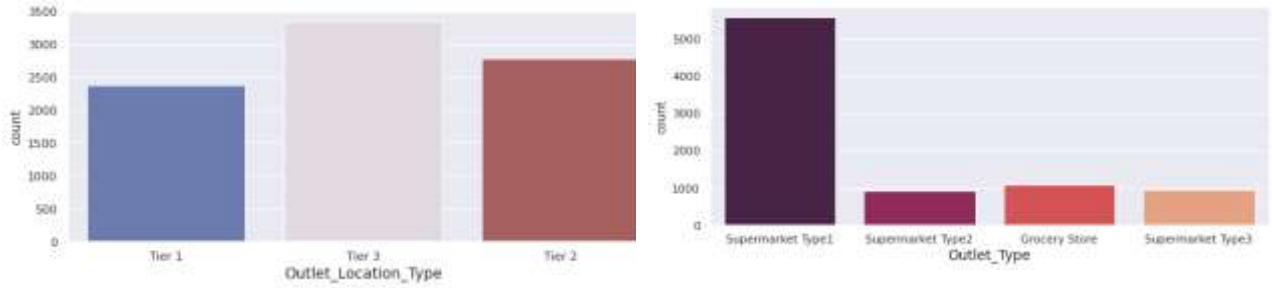
Sebelum menerapkan model, pertama dilakukan analisis korelasi antar variabel dengan memeriksa korelasi antara variabel dependen dan variabel target. Tahapan ini bertujuan untuk menemukan fitur yang akan digunakan pada model. Hasil korelasi dapat dilihat pada gambar 1, dimana Item_MRP memiliki korelasi paling positif dan Item_Visibility memiliki korelasi terendah dengan variabel target.



Gambar 1 Korelasi Variabel

Pada gambar 1 terlihat bahwa hasilnya sangat berbeda dengan hipotesis awal, variabel-variabel ini diperkirakan akan berdampak tinggi pada peningkatan penjualan. Namun demikian, karena ini bukan perilaku yang diharapkan. Oleh karena itu, langkah selanjutnya dianalisis variabel kategori dan melihat variabel yang mengandung beberapa wawasan tentang hipotesis yang telah dibuat sebelumnya. Visualisasi distribusi variabel kategori dapat dilihat pada gambar 2.





Gambar 2 Analisis Variabel Kategori

Pada gambar 2 merupakan visualisasi variabel kategori dan dapat dilihat bahwa realisasi kolom kategoris Item_Fat_Content sebagian besar item yang dijual rendah dan Produk Low Fat tampaknya memiliki penjualan yang lebih tinggi daripada produk Reguler. Item_Type merupakan Jenis item yang sangat populer adalah vegetables dan snack foods. Outlet_Identifier merupakan barang yang terjual didistribusikan secara merata di antara outlet tidak termasuk OUT010 dan OUT019 yang jauh lebih rendah. Outlet_Size - Outlet Bigmart sebagian besar berukuran sedang dalam data dan Outlet_Location_Type merupakan jenis yang paling umum adalah Tier3. Outlet_type adalah Tipe Supermarket1. Sebagian besar toko Supermarket Type1 berukuran Tinggi dan tidak memiliki hasil terbaik. sedangkan Supermarket Type3 (OUT027) adalah toko ukuran Medium dan memiliki hasil terbaik.

Berdasarkan hasil analisis ini maka dapat ditarik kesimpulan bahwa terdapat perbedaan jenis barang menurut penjualan sangat kecil. Outlet 27 adalah yang paling menguntungkan dan ada perbedaan besar antara setiap penjualan outlet tertentu. Anehnya supermarket tipe 3 adalah yang paling menguntungkan dan bukan tipe 1. Ukuran outlet menengah dan tinggi cukup banyak bahkan dalam penjualan. Tingkat 2 dan 3 hampir menjadi penjualan tertinggi (2 sedikit lebih besar). Dari hasil ini, Kolom Outlet_Establishment_Year, Item_Identifier dan Outlet_Identifier tidak memiliki nilai signifikan sehingga akan dihapus. Semua variabel Ordinal Item_Fat_Content, Outlet_Size, Outlet_Location_Type akan dikodekan Label sebagai fitur. Kolom Outlet_Type dan Item_Type akan di encode nilai antara 0 dan n_classes-1 di mana n adalah jumlah label yang berbeda. Jika label berulang, itu memberikan nilai yang sama seperti yang ditetapkan sebelumnya.

Membangun Model

Pembelajaran yang diawasi menunjukkan hubungan antara dua kumpulan data. Data yang diamati X dan variabel eksternal y yang kita coba prediksi, biasanya disebut "target" atau "label". Semua fitur yang diawasi dalam scikit-learn menerapkan metode fit(X, y) agar sesuai dengan model dan metode prediksi(X), dengan pengamatan tak berlabel X, mengembalikan label prediksi y. Nilai target Item_Outlet_Sales ditetapkan ke y, variabel X sebagai fitur yang kita definisikan sebelumnya. Selanjutnya, data dibagi dengan rasio 80:20 yang artinya 80% data untuk pelatihan dan 20% data untuk pengujian. Proses pembagian data menggunakan fungsi train_test_split pada pustaka scikit-learn untuk membagi array data menjadi dua subset: untuk data pelatihan dan untuk data pengujian sehingga, tidak perlu membagi dataset secara manual. Hasil kelima model yang diusulkan dapat dilihat pada tabel 1

Tabel 1. Hasil akurasi model

Algoritma	MAE	RSME	R^2
Lasso Regressor	838.07	1128.81	0.5594
Linear Regression	838.19	1285.72	0.5593
Random Forest	1084.52	1125.61	0.3268
LightGBM	754.28	1075.89	0.6094
XGBoost	755.61	1084.49	0.6142

Pada tabel 1 merupakan hasil pengujian terhadap lima model yang digunakan yaitu *Lasso Regressor, Linear Regression, Random Forest* (RF), LightGBM (LG) dan XGBoost. Berdasarkan hasil tersebut model XGBoost menghasilkan nilai r2 tertinggi dengan nilai skor 0.61 (61%), kemudian LightGBM 0.60 (60%) dan Random Forest paling rendah sebesar 0.32 (32%) dengan nilai akurasi tingkat kesalahan evaluasi prediksi terbesar berdasarkan MAE 1084.52, namun berbeda dengan evaluasi kesalahan berdasarkan RSME algoritma Linear

Regression menghasikan tingkat kesalahan tertingi sebesar 1285.72 dan terendah adalah algoritma LightGBM dengan tingkat rata-rata kesalahan sebesar 1075.89.

Diskusi

Berdasarkan hasil analisis data dimana lokasi terbesar tidak menghasilkan penjualan tertinggi. Lokasi yang menghasilkan penjualan tertinggi adalah lokasi OUT027, yang pada gilirannya merupakan Supermarket Type3, memiliki ukurannya dicatat sebagai media dalam dataset. Dapat dikatakan bahwa kinerja outlet ini jauh lebih baik daripada lokasi outlet lainnya dengan ukuran apa pun yang disediakan dalam dataset yang dipertimbangkan. Median variabel target Item_Outlet_Sales dihitung menjadi 3364,95 untuk lokasi OUT027. Lokasi dengan skor median tertinggi kedua (OUT035) memiliki nilai median 2109,25 dan untuk Nilai kuadrat R yang disesuaikan lebih tinggi untuk model XGBoost daripada yang lain. Oleh karena itu, model yang digunakan lebih cocok dan menunjukkan akurasi. terakhir, nilai skor model XGBoost dan LightGBM dapat mencapai 60 %, jika dibangun dengan lebih banyak pertimbangan dan analisis hipotesis, dapat meningkatkan akurasi.

KESIMPULAN

Penelitian ini menyajikan analis data penjualan dan dasar-dasar *supervised machine learning* untuk tugas prediksi penjualan di pusat perbelanjaan Big Mart di lokasi yang berbeda. Berdasarkan hasil evalusi tingkat kesalahan deteksi berdasarkan MAE dan RMSE, algoritma XGBoost dan LightGBM manghasilkan tingkat kesalahan paling rendah dibandingkan algoritma lainnya. Sedangkan hasil evaluasi skor R^2 kedua algoritma ini mencapai nilai tertinggi 0.61 (61%) XGBoost dan LightGBM 0.60 (60%). Selain itu juga, hasil prediksi menunjukkan kegembiraan corr di antara atribut yang berbeda dipertimbangkan dan bagaimana lokasi tertentu dari ukuran menengah mencatat penjualan tertinggi, menunjukkan bahwa lokasi belanja lainnya harus mengikuti pola yang sama untuk peningkatan penjualan. Namun, dari hasil pengujian ini masih perlu mempertimbangkan beberapa paremeter instans untuk membuat prediksi penjualan ini lebih inovatif dan sukses seperti penerapan hyper parameter Gridsearch, Random Search, Bayesian optimization akan menjadi pertimbangan untuk meningkatkan akurasi pada algoritma XGBoost dan LightGBM di penelitian selanjutnya.

Supplementary Materials (optional)

Sumber dataset tersedia di https://www.kaggle.com/devashish0507/big-mart-sales-prediction

Konstribusi Penulis

Semua Penulis memiliki konstribusi yang sama dalam makalah ini Semua penulis telah membaca dan menyetujui versi manuskrip yang diterbitkan.

Konflik kepentingan

Para penulis menyatakan tidak ada konflik kepentingan.

REFERENCES

[1] N. Calixto and J. Ferreira, "Salespeople performance evaluation with predictive analytics in B2B," *Appl. Sci.*, vol. 10, no. 11, p. 4036, Jun. 2020, doi: 10.3390/app10114036.

[2] A. M. Husein, A. M. Simarmata, M. Harahap, S. Aisyah, and A. Dharma, "Implementation ANFIS Method for Prediction Needs Drug-based Population Diseases and Patient," in *2019 International Conference of Computer Science and Information Technology (ICoSNIKOM)*, 2019, pp. 1–5.

[3] J. Kang, H. J. Lee, S. H. Jeong, H. S. Lee, and K. J. Oh, "Developing a forecasting model for real estate auction prices using artificial intelligence," *Sustain.*, vol. 12, no. 7, p. 2899, Apr. 2020, doi:

10.3390/su12072899.

[4] M. Harahap, A. M. Husein, S. Aisyah, F. R. Lubis, and B. A. Wijaya, "Mining association rule based on the diseases population for recommendation of medicine need," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, p. 12017, 2018, doi: 10.1088/1742-6596/1007/1/012017.

[5] H. Alimohammadi, H. Rahmanifard, and N. Chen, "Multivariate time series modelling approach for production forecasting in unconventional resources," *Proc. - SPE Annu. Tech. Conf. Exhib.*, vol. 2020-Octob, pp. 1–13, 2020, doi: 10.2118/201571-ms.

[6] A. M. Husein and A. M. Simarmata, "Drug Demand Prediction Model Using Adaptive Neuro Fuzzy Inference System (ANFIS)," *SinkrOn*, vol. 4, no. 1, p. 136, 2019, doi: 10.33395/sinkron.v4i1.10238.

[7] A. M. Husein, M. Harahap, S. Aisyah, W. Purba, and A. Muhazir, "The implementation of two stages clustering (k-means clustering and adaptive neuro fuzzy inference system) for prediction of medicine need based on medical data," *J. Phys. Conf. Ser.*, vol. 978, no. 1, p. 12019, 2018, doi: 10.1088/1742-6596/978/1/012019.

[8] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent Neural Networks for Time Series Forecasting: Current status and future directions," *Int. J. Forecast.*, vol. 37, no. 1, pp. 388–427, 2021, doi: 10.1016/j.ijforecast.2020.06.008.

[9] C. P. P. Maibang, A. M. Husein, and others, "Prediksi Jumlah Produksi Palm Oil Menggunakan Fuzzy Inference System Mamdani," *J. Teknol. dan Ilmu Komput. Prima*, vol. 2, no. 2, p. 19, 2019, doi: 10.34012/jutikomp.v2i2.528.

[10] P. Hähnel, J. Mareček, J. Monteil, and F. O'Donncha, "Using deep learning to extend the range of air pollution monitoring and forecasting," *J. Comput. Phys.*, vol. 408, pp. 1–13, 2020, doi: 10.1016/j.jcp.2020.109278.

[11] A. M. Husein, J. P. Hutabarat, J. E. Sitorus, T. Giawa, and M. Harahap, "Predicting the Spread of the Corona Virus (COVID-19) in Indonesia: Approach Visual Data Analysis and Prophet Forecasting," *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, p. 151, 2020, doi: 10.29099/ijair.v5i1.192.

[12] A. M. Husein, M. Arsyal, S. Sinaga, and H. Syahputa, "Generative Adversarial Networks Time Series Models to Forecast Medicine Daily Sales in Hospital," *SinkrOn*, vol. 3, no. 2, p. 112, 2019, doi: 10.33395/sinkron.v3i2.10044.

[13] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *J. Big Data*, vol. 7, no. 1, pp. 1–22, Jul. 2020, doi: 10.1186/s40537-020-00329-2.

[14] W. Kratsch, J. Manderscheid, M. Röglinger, and J. Seyfried, "Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction," *Bus. Inf. Syst. Eng.*, vol. 63, no. 3, pp. 261–276, Apr. 2021, doi: 10.1007/s12599-020-00645-0.

[15] J. Huikku, T. Hyvönen, and J. Järvinen, "The role of a predictive analytics project initiator in the integration of financial and operational forecasts," *Balt. J. Manag.*, vol. 12, no. 4, pp. 427–446, Sep. 2017, doi: 10.1108/BJM-05-2017-0164.

[16] J. Babcock, *Mastering predictive analytics with Python: exploit the power of data in your business by building advanced predictive modeling applications with Python*. .

[17] I. Raeesi Vanani and S. Majidian, "Literature Review on Big Data Analytics Methods," in *Social Media and Machine Learning*, IntechOpen, 2020.

[18] R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000, [Online]. Available: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining.