KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 6, No. 2, April (2025), pp. 488-497

Prediksi Jumlah Pasien Medical Check Up Berdasarkan Time Series Forecasting Menggunakan Algoritma XGBoost

Mohammad Aldinugroho Abdullah
Universitas Nasional, Jakarta, Indonesia
E-mail: mohammadaldinugrohoabdullah@civitas.unas.ac.id

Abstract

Medical Check Up, also known as MCU, is utilized by the public to assess their health condition. However, there is often an issue with inadequate availability of medical equipment. To address this problem, this research aims to forecast the number of patients seeking medical check-ups based on patient arrival data from 2020 to 2022. During this period, the data experienced outliers due to the COVID- 19 pandemic. The forecasting approach employed in this study involves machine learning using the XGBoost algo rithm. The outliers are handled, and an additional feature of average patient visitation movement is incorporated. The model achieved an accuracy of MAE 8.607, RMSE 10.583, and MSE 111,999.

Keywords: *time-series forecasting, health, data mining, XGBoost, MSE, RMS.*

1. Pendahuluan

Kesehatan menjadi hal yang sangat penting di Indonesia karena merupakan salah satu Hak Asasi Manusia. Pada UU No. 36 Tahun 2009 tentang kesehatan semua manusia memiliki hak yang sama dalam mendaptkan akses kesehatan dengan pelayanan yang aman, bermutu, dan terjangkau. Maka dari itu setiap orang memiliki aksesbilita dan pelayanan yang baik dan berkualitas tanpa melakukan diskriminatif [1]. Pelayanan kesehatan harus dilakukan dengan efesien dan efektif dalam memusatkan perhatian pada keinginan dan kebutuhan seorang pasien. Setiap pasien harus mendapatkan 5 aspek dalam mendapatkan pelayan *tangible, empathy, assurance, responsiveness* dan *reliability.* Pasien dapat dikatakan puas ketika keingnannya terpenuhi dan kebutuhan pasien terpenuhi [2]. Peningkatan pelayanan kesehatan merupakan tanggung jawab dari seluruh penyedia layanan kesehatan baik puskemas, klinik atau rumah sakit. Untuk mewujudkan kesehatan masyarkat yang optimal, pelayanan yang maksimal akan memudahkan masyarakat mendapatkan akses kesehatan yang lebih cepat. Berbanding lurus dengan tujuan pelayanan kesehatan yang memiliki visi pelayanan yang cepat, tepat, akurat dan mengutamakan kesehatan pasien [3]

Bidang teknologi sudah merambah pada seluruh bidang pada kehidupan manusia, baik dalam kehidupan sehari-hari maupun dalam bidang kesehatan. Dalam bidang kesehatan sendiri sudah banyak sebagai penunjang pelayanan kesehatan. Perkembangan ini didukung dengan 750.000 artikel kesehatan yang dipublikasi tiap tahun. Peningkatan kulaitas pelayanan merupakan salah satu kemajuan untuk menentukan daya saing [4]. Perkembangan teknologi membantu pengolahan data mentah menjadi suatu pemahaman yang dapat membantu untuk pengambilan keputusan yang tepat. Hasil pengolahan data juga dapat digunakan untuk menentukan strategi di masa depan atau sebagai pencegahan sebelum hal itu terjadi. Analisis data terbagi menjadi 2 kategori deksriptif dan prediktif. Deskriptif untuk mendapatkan informasi pola yang terjadi pada data. Prediktif untuk menganalisis kejadian yang akan terjadi pada masa mendatang berdasarkan data yang sudah ada [5].

Peramalan meruapakan tahap yang penting pada pengambilan keputusan dalam manajemen suatu organisasi. Kegiatan ini akan sangat meningkatkan agar suatu

ISSN: 2720-992X

organisasi dapat menghindari ketergantungan peluang pada suatu kejadian. Hasil analisis ini akan membantu efektifitas kinerja medical check-up dan mengoptimalkannya. Peningkatan pasien atau penurunan pasien akan membantu pada stock obat yang harus disediakan, banyak SDM yang bertugas, fasilitas kesehatan dan performa tenaga medis pada pasien. Kasus pada analisis deret waktu memiliki banyak algoritma sesuai dengan kebutuhan dan kecocokan dataset yang digunakan. Pendekatan dengan metode konvensional seperti ARIMA, SARIMAX dan Autoregressive. Pendekatan dengan metode machine learning seperti XGBoost Regressor dan Prophet. Pendekatan machine learning biasa digunakan jika dataset memiliki jumlah yang besar [6].

Berdasarkan penelitian yang dilakukan oleh [7] yang melakukan peramalan kedatangan jumlah pasien pada saat pandemi 2019 dengan memanfaatkan metode time series forecasting. Penelitian ini melakukan studi kasus pada peramalan jumlah kedatangan pasien unit gawat darurat. Algoritma yang digunakan dalam penelitian ini adalah SARIMAX/SARIMA, Facebook Prophet, Holt Winters, dan LSTM. Dengan nilai evaluasi matriks Mean Absolute Error SARIMA 22.28, Holt Winters 23.07, LSTM 23.92, dan Facebook Prophet 51.15. Penelitian yang dilakukan [8] yang membandingkan akurasi ARIMA dan XGBoost dalam melakukan time series forecasting. Tujuan penelitian ini untuk mendapatkan model dengan evaluasi matriks terkecil pada Mean Absolute Error untuk meningkatkan akurasi pada forecasting. Studi kasus yang digunakan penyakit brucellosis dengan hasil matriks XGBoost 189.332 dan ARIMA dengan 338,867. Evaluasi itu membuktikan XGBoost lebih baik dalam melakukan modeling. XGBoost dapat dilakukan dalam bentuk analisis klasifikasi dan regressi. XGBoost memiliki kelebih an untuk melakukan prediksi data yang bersifat nonlinier. XGBoost diminati dalam penelitian berkat kecepatan yang cepat dan efek klasifikasi yang baik. Kelemahan XGBoost jika banyak outlier pada data sehingga model akan mengalami penurunan akurasi [9] XGBoost memiliki regulasi L2 yang diperkenalkan dalam loss function sehingga dapat mencegah overfitting pada model dan secara efektif dapat mengurangi kompleksitas model. Dimana algoritma ini dapat secara efektif membantu dalam kasus imbalance data dan meningkatkan akurasi model. Sehingga XGBoost akan menghubungkan setiap attribute yang ada [10]. Moving Average digunakan dalam permodelan ARIMA untuk mengetahui fluktuasi pada data. Sulit untuk ARIMA jika hanya menggunakan autoregressive dalam memprediksi fluktuasi yang terjadi [11]. Moving Average sering digunakan dalam beberapa kasus seperti finansial, ekonomi, dan apapun selama memiliki data berbentuk time series. Moving Average digunakan untuk melihat pattern dan tren yang terjadi pada dataset. Sehingga Moving Average dapat digunakan untuk membantu prediksi berdasarkan data histori. Penulis melakukan analisis kedatangan pasien untuk meningkatkan pelayanan kesehatan yang diberikan. Berdasarkan informasi yang didapatkan adanya penurunan pasien yang akan melakukan medical check-up.

Analisis ini juga untuk menentukan strategi yang tepat agar pasien tetap ingin berkunjung untuk menggunakan fasilitas kesehatan yang diberikan. Data yang digunakan dalam penelitian ini adalah kedatangan pasien sepanjang 2020 sampai 2022. Data tersebut berbentuk tabular dan termasuk raw data. Pengolahan data yang baik dapat membentuk pengetahuan baru dari suatu kejadian [12]. Pada penelitian ini penulis melakukan prediksi kedatangan pasien dengan pendekatan *machine learning*. Mempertimbangkan bahwa data yang digunakan sudah lebih dari cukup untuk melakukan *time series forecasting* dengan model *machine learning* dan algoritma XGBoost. Model ini akan digunakan untuk melakukan prediksi sebanyak 30 hari. Pada prosesnya penulismenambahkan *attribute Moving Average*. Penambahan attribute ini berharapmeningkatkan akurasi yang model. Pada

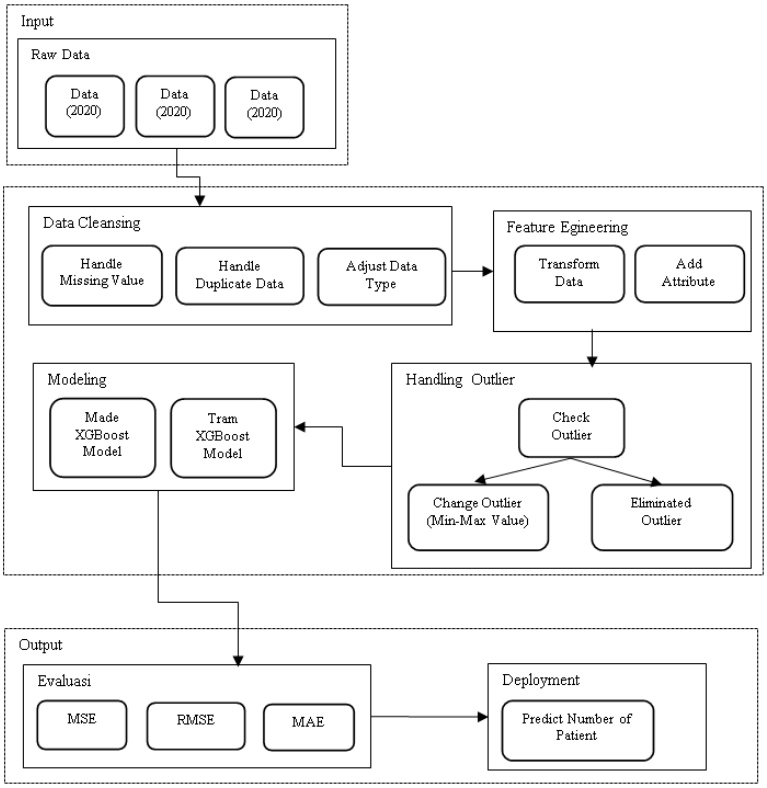
prosesnya penulis juga melakukan *Exploratory Data Analysis* (EDA) untuk mengetahui jumlah yang kedatanganpasien berdasarkan data 2020 sampai 2022.

2. Metodologi Penelitian

Metode penelitian yang digunakan penulis untuk menyelesaikan penelitian ini *Agile Metodology*. Tahapan pertama merupakan perencanaan dan melakukan desain sistem. Tahapan kedua melakukan pengembangan sistem sesuai dengan desain sistem. Tahapan ketiga sudah melakukan testing, *deployment*, dan uji hipotesis.

2.1. Perencanaan dan Desain Sistem

Pada tahap ini penulis melakukan analisis tentang apa yang akan dilakukan untuk membantu penulis mencapai tujuan dari penelitian. Proses input mulai dari pengumpulan data. Tahap kedua merupakan proses atau pengembangan sistem agar dapat mencapai tujuan penelitian. Tahapan ketiga merupakan hasil penelitian dan output yang akan digunakan dalam pengujian hipotesis penelitian.



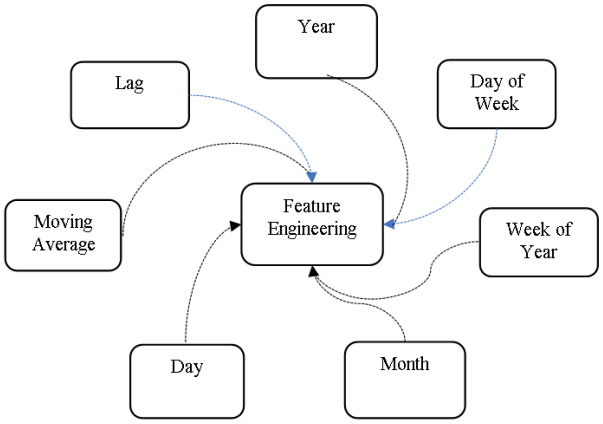
Gambar 1. Perancangan desain penelitian

2.2. Data Cleansing

Pada tahap ini penulis melakukan pembersihan data pada dataset. Pembersihan disini seperti menanggulangi missing value, data duplikasi dan memanipulasi tipe data sesuai strukturnya. Tahap ini digunakan agar data terhindar dari inkonsistensi dan dapat menurunkan error correction pada model yang akan dibuat.

2.3. *Feature Engineering*

Pada tahap ini penulis melakukan transformasi data karena data yang digunakan adalah kunjungan harian. Data tersebut harus dilakukan *feature engineering* dengan cara melakukan *grouping* berdasarkan jumlah pasien dari tanggal transaksi. Tahap ini akan menghasilkan jumlah pasien harian.



Gambar 2. *Feature* yang digunakan dalam penelitian

2.4. *Handling Outlier*

Penelitian menaggulangi outlier dengan cara mengeliminasi nilai tersebut. Pertimbangan ini karena algoritma XGBoost sangat sensitif pada nilai outlier. Penulis juga akan membandingkan nilai evaluasi model pada data yang outlier-nya dieliminasi dengan yang tidak.

2.5. Modeling

Pada tahap penelitian ini penulis melakukan modeling pada *machine learning* menggunakan dataset yang sudah siap digunakan. Tahap ini penulis menggunakan algoritma XGBoost untuk melakukan *time series forecasting*. Penulis juga akan melakukan *hyper parameter tunning* untuk mendapatkan hasil evaluasi yang maskimal.

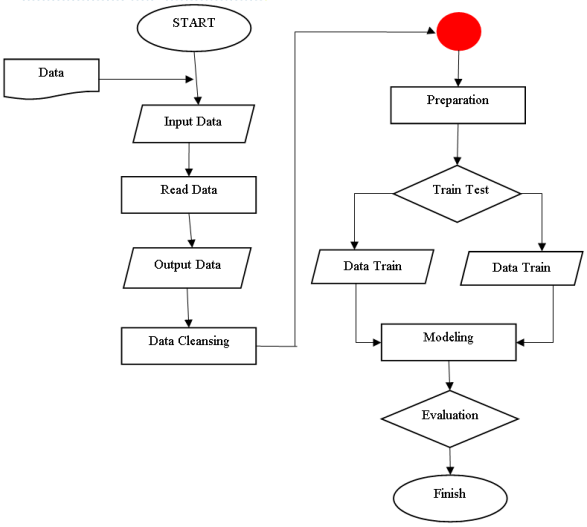
3. Hasil dan Pembahasan

3.1. Evaluasi

Proses evaluasi ini menggunakan MAE, MSE dan RMSE untuk mengukur bagaimana keakuratan model yang dibuat. Hasil evaluasi juga untuk menentukan apakah model underfitting, overfitting atau sudah termasuk model yang baik. Model ini akan dilakukan *feature importance* untuk melihat atribute mana yang paling berpengaruh dalam pembuatan model.

Metode pemilihan yang digunakan dalam penelitian ini adalah populasi. Metode ini dipilih karena pendekatan menggunakan machine learning akan lebih baik jika data yang digunakan terbilang banyak. Data akan dibagi menjadi data training dan testing

Pengumpulan data yang digunakan dalam penelitian ini merupakan data internal. Data internal dari kunjungan pasien yang akan melakukan medical check-up setiap hari. Data ini akan digunakan untuk menguji hipotesis yang telah penulis buat dan mencapai tujuan penelitian.



Gambar 3. Desain machine learning

a) Memulai program dan memasukan data pada program.
b) Melakukan input dengan data yang dimasukan.
c) Menggunakan pandas untuk membaca data.
d) Menghasilkan dataframe yang akan digunakan.
e) Membersihkan data.
f) Menyiapkan data yang akan digunakan untuk modeling.
g) Membagi data menjadi train dan test.
h) Modeling algoritma XGBoost untuk melakukan forecasting.
i) Evaluasi model yang sudah dibuat.

3.2. Data Preparation

Hasil pengujian dari penelitian yang dibuat penulis terbagi menjadi delapan model. Delapan model ini memiliki perbedaan pada *data preparation* atau *data feature*. Data yang digunakan penulis sebanyak 672 data dengan 2 kolom. Pada tabel 1 dapat diketahui bentuk dataset. Data tersebut terdiri dari tanggal dan jumlah kunjungan pasien per hari. Data yang digunakan dimulai dari 2 Januari 2020 sampai 8 Desember 2022

Tabel 1. Dataset yang digunakan

Date	Arrival Patient
2020-08-02	15
2020-08-03	20
2020-08-06	60
2020-08-07	24
...	...
2022-12-05	57
2022-12-06	39
2022-12-07	43
2022-12-08	41

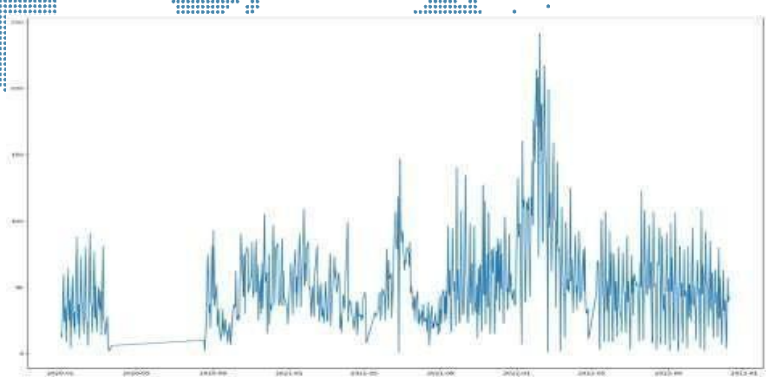
3.3. Pengecekan Tipe Data

Hasil pengecekan tipe data dataset yang digunakan dapat dilihat pada tabel 4.2

Tabel 2. Tipe data pada dataset

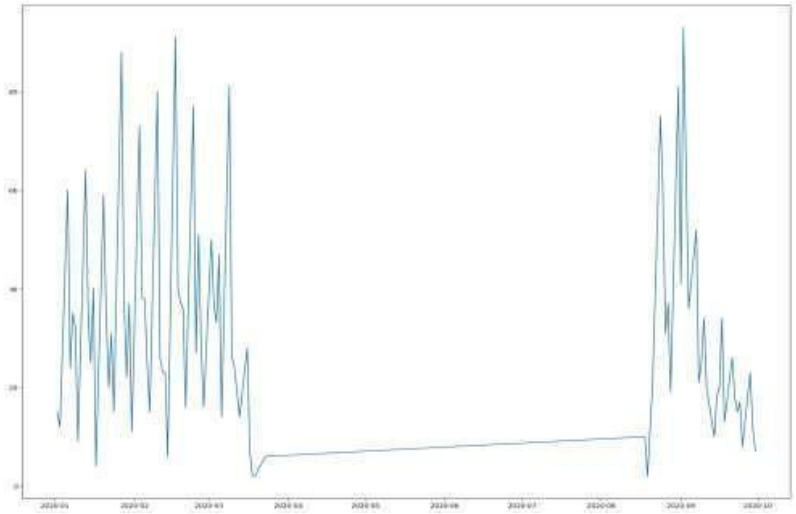
Nama Kolom	Tipe Data
Date	*datetime*
Count	integer

Berdasarkan Tabel 2 data yang akan digunakan sudah sesuai dengan kebutuhan untuk pengolahan data dan time series forecasting. Kolom date untuk menentukan baris waktu dan kolom count merujuk pada kunjungan pasien pada waktu tersebut.



Gambar 4. Grafik kunjungan pasien harian

Gambar 4 merupakan grafik kedatangan harian pasien. Diketahui garis lurus yang terdapat pada grafik tersebut dikarenakan ada penutupan akibat adanya pandemi covid 19. Sehingga data memiliki kekosongan rentang.

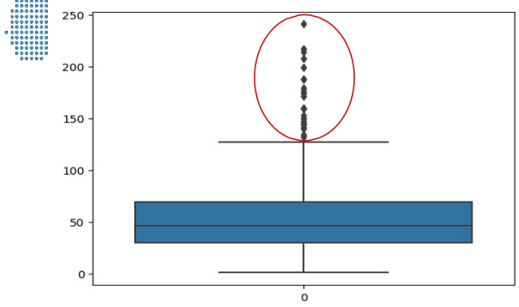


Gambar 5. Terdapat kekosongan yang terjadi pada data akibat pandemi

Pada Gambar 5 terdapat kekosongan data yang karena adanya pandemi covid-19. Data tersebut kosong dari tanggal 23-03-2020 sampai 18-08- 2020 tanggal. Dengan pertimbangan konsistensi data penulis tidak menggunakan data tersebut. Dataset yang digunakan penulis setelah melakukan analisis deret waktu tanggal sehingga didapatkan banyak data yang digunakan menjad 613 dari keseluruhan data.

3.4. Outlier

Pada Gambar 6 terdapat nilai outlier pada dataset yang digunakan. Pada hasil visualisasi boxplot outlier hanya berada pada sisi upper. Ini menandakan terjadi kunjungan pasien yang meningkat sang signifikan pada suatu waktu.



Gambar 6. Boxplot outlier

3.5. Korelai Fitur

Pada bagian ini penulis melakukan uji korelasi antar feature pada dataset. Penulis melakukan pengujian dengan metode pearson dan spearman. Fitur-fitur yang digunakan pada pengujian ini lag_1, lag_7, lag_14, MA_21, MA_50, dan count. Berikut hasil pengujian pearson pada dataset.

Gambar 7. Korelasi Pearson

Berdasarkan Gambar 7 mendapatkan korelasi positif tertinggi untuk count *feature* pada *moving average* 21 dengan skor 0,52. Sedangkan korelasi positif terendah pada lag_14 dengan skor 0,26. Teknik korelasi pearson mengasumsikan data berdistribusi normal dan memiliki hubungan linier. Selanjutnya melakukan pengujian menggunakan spearman. Pengujian korelasi menggunakan spearman yang memiliki ketahanan terhadap nilai outlier. Digunakan untuk mengetahui dua variabel memiliki hubungan linier atau tidak. Berikut hasil pengujian korelasi menggunakan spearman.

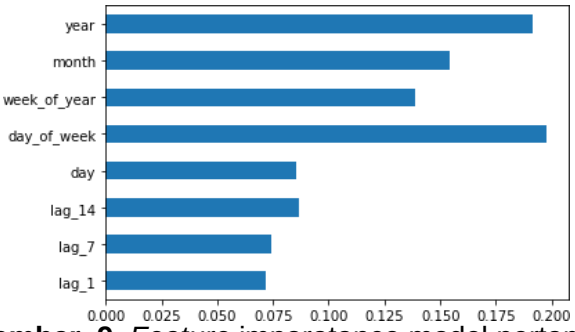
Gambar 8. Korelasi Spearman

Dalam pengujian ini fitur yang memiliki korelasi tertinggi terhadap feature count adalah moving average 21 dengan skor 0,36. Fitur terendah ditempati oleh lag_14 dengan

skor 0,16. Rata-rata korelasi dalam spearman memiliki korelasi yang rendah terhadap count feature.

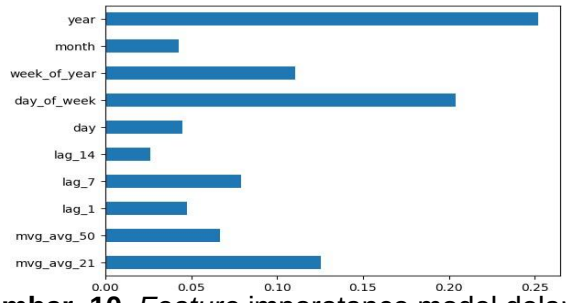
3.6. Modeling

Proses *modelling* penelitian ini akan dibagi kedalam delapan model. Model *machine learning* yang dibangun akan menggunakan XGBoost regressor dengan parameter terbaik dari setiap modelnya. Penulis melakukan hyper parameter tunning menggunakan metode *gridsearch*. Setelah mendapatkan model penulis melakukan *feature importance* untuk mengetahui fitur mana yang berpengaruh dalam permodelan.



Gambar 9. *Feature* imporatance model pertama

Gambar 9 terlihat grafik fitur yang paling berpengaruh adalah day of week. Day of week memiliki efek yang besar pada model untuk melakukan prediksi. Sedangkan Lag Feature tertinggi adalah lag_14 fitur daripada fitur lag lainnya. Dalam model satu fitur lag tidak terlalu memiliki efek pada model.



Gambar 10. *Feature* imporatance model delapan

Feature yang paling berpengaruh pada model delapan adalah year dan posisi kedua adalah day_of_week. Pada model delapan MA 21 menempati posisi ketiga untuk fitur yang paling berpengaruh pada model. Semetara lag_14 menjadi fitur yang paling rendah dari semua fitur yang digunakan.

Setelah melakukan train model dan mendapatkan model machine learning selanjutnya akan digunakan untuk melakukan prediksi (testing). Dalam penelitian ini penulis menguji delapan model menggunakan MAE, RMSE, dan MSE.

4. Kesimpulan

Setelah melakukan penelitian dengan sesuai lingkup penelitian terdiri dari batasan masalah, rumusan masalah, dan tujuan masalah. Penulis juga akan menjabarkan hasil yang didapatkan dari pengujian yang dilakukan untuk melakukan uji hipotesis yang telah dibuat. Penelitian ini bertujuan memprediksi kunjungan pasien medical checkup dengan menggunakan algoritma XGBoost. Saat melakukan pengecekan dataset terjadi kekosongan pada data. Kekosongan data terjadi akibat

penutupan pandemi COVID-19. Penulis memotong data kosong dan sebelumnya sehingga terjadi pengurangan data. Penulis juga menemukan nilai outlier pada dataset dengan bantuan visualisasi boxplot. Berdasarkan penelitian terdapat lonjakan pasien yang melakukan medical checkup sehingga menjadikannya outlier pada dataset. Sehingga penulis melakukan penanggulangan nilai outlier dengan dua metode menghilangkan dan memanipulasinya. Setelah menanggulangi nilai outlier penulis melakukan feature engineering pada dataset bertujuan agar model lebih presisi dalam melakukan prediksi. Feature yang ditambahkan pada dataset salah satunya moving average. Moving average yang menggambarkan pergerakan harian pasien menjadi suatu nilai informasi yang penting untuk membangun model machine learning. Penulis menambahkan dua moving average 21 dan 50. Tahap selanjutnya penulis melakukan uji korelasi pada feature. Penulis menggunakan dua metode untuk melakukan uji korelasi pearson dan spearman. Uji pearson berasumsi bahwa dataset memiliki distribusi normal. Hasilnya terdapat korelasi positif antara moving average 21 dan kunjungan pasien. Selanjutnya melakukan pengujian menggunakan spearman hasilnya terdapat korelasi positif antara moving average 21 dan kunjungan pasien. Setelah melakukan uji korelasi, penulis membagi dataset untuk melakukan modeling machine learning menggunakan algoritma XGBoost. Penulis menghasilkan 4 dataset berdasarkan data preparation yang berbeda-beda. Setelah itu penulis melakukan modeling dan menghasilkan delapan model. Pada tahap ini penulis melakukan feature importance untuk mengetahui fitur mana yang paling berpengaruh pada setiap model. Hasil dari kedelapan mode rata-rata menempatkan day of week sebagai feature yang paling berpengaruh. Setelah model machine learning penulis menguji dengan data validation dan setelahnya melakukan testing pada model. Dari kedelapan model tersebut empat model melakukan scaling pada feature dan empat model lainnya tidak menggunakan scaling. Model yang melakukan scaling memiliki rata-rata evaluasi MAE, RMSE, dan MSE yang tinggi. Nilai evaluasi terbaik ditempati oleh model 7 dengan nilai evaluasi testing MAE 8,607, RMSE. 10,583, dan MSE 111,999. Model tersebut melakukan manipulasi nilai outlier dengan mengubah nilai upper (outlier bagian atas/kanan) menjadi nilai kuartil 3 (Q3). Model 7 juga memiliki feature moving average untuk memberikan model atau memperluas pembelajaran model agar dapat melakukan prediksi. Pendekatan machine learning dengan nilai outlier baik dalam kasus prediksi kunjungan pasien dan menambahkan feature moving average meningkatkan akurasi model melakukan prediksi

Daftar Pustaka

[1] H. B. Sukendar, “Pemberdayaan Masyarakat Miskin Melalui Peningkatan Layanan Kesehatan oleh Rumah Sehat BAZNAS Yogyakarta di Desa Wukirsari,” *SANGKéP J. Kaji. Sos. Keagamaan*, vol. 1, no. 2, pp. 132–142, 2018.

[2] A. . Patattan, “Hubungan mutu pelayanan kesehatan dengan kepuasan pasien di Rumah Sakit Fatima Makale di era new normal,” *J. Keperawatan Florence Nightingale*, vol. 4, no. 1, pp. 14–19, 2021.

[3] S. Hade, A. Djalla, and A. D. . Rusman, “Management Information Systems in an Effort to Improve Health Services at Andi Makkasau Hospital Parepare,” *Ilm. Mns. Dan Kesehat.*, vol. 2, pp. 293–305, 2019.

[4] A. Yani, “Utilization of technology in the health of community health. PROMOTIF,” *J. Kesehat. Masy.*, vol. 8, no. 1, pp. 97–103, 2018.

[5] A. W. Maghfiroh, N. Ulinnuha, and A. Fanani, “Penerapan Fuzzy C-Means dalam Mengelompokkan Kabupaten/Kota Berdasarkan Fasilitas Pelayanan Kesehatan Di Jawa Timur,” *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun*, vol. 4, no. 1, pp. 8–14, 2019.

[6] C. Chandra and S. Budi, "Analisis Komparatif ARIMA dan Prophet dengan Studi Kasus Dataset Pendaftaran Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 278–287, 2020, doi: 10.28932/jutisi.v6i2.2676.

[7] E. E. Etu *et al.*, "A comparison of univariate and multivariate forecasting models predicting emergency department patient arrivals during the COVID-19 pandemic," *Healthc. MDPI.*, vol. 10, no. 6, p. 1120, 2022.

[8] M. Alim, G. H. Ye, P. Guan, D. S. Huang, B. Sen Zhou, and W. Wu, "Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: A time-series study," *BMJ Open*, vol. 10, no. 12, pp. 1–8, 2020, doi: 10.1136/bmjopen-2020-039676.

[9] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marqués, "Setting B2B digital marketing in artificial intelligence-based CRMs: A review and directions for future research," *Ind. Mark. Manag.*, vol. 98, pp. 161–78, 2021.

[10] M. Xue, L. Wu, Q. P. Zhang, J. X. Lu, X. Mao, and Y. Pan, "Research on load forecasting of charging station based on XGBoost and LSTM model," *J. Phys. Conf. Ser.*, vol. 1757, no. 1, p. 012145, 2021.

[11] D. Gunawan and W. Astika, "The Autoregressive Integrated Moving Average (ARIMA) Model for Predicting Jakarta Composite Index," *J. Inform. Ekon. Bisnis*, vol. 1, no. 6, 2022.

[12] C. X. Lv, S. Y. An, B. J. Qiao, and W. Wu, "Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model," *BMC Infect. Dis.*, vol. 21, pp. 1–13, 2021.