BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian serupa mengenai prediksi di bidang medis telah dilakukan ileh berbagai peneliti sebelumnya dengan pendekatan metode yang beragam. Studi-studi tersebut memberikan gambaran lebih mendetil mengenai efektivitas masing-masing metode yang digunakan peneliti. Untuk mendukung penyusunan model prediktif dalam penelitian ini, penulis merangkum beberapa penelitian terdahulu yang relevan dalam Tabel 2.1 berikut.

Tabel 2.1 Penelitian Terdahulu

No	Penulis	Tahun	Judul	Nama Jurnal	Akurasi	DOI
1	Arwin Datumaya Wahyudi Sumari Adhika Febrianto Yushintia Pramitarini [14]	2021	Sistem Prediksi Permintaan Darah Menggunakan Metode Regresi Linier	Jurnal Informatika Polinema, 7(2), 85–90	80,14%	doi.org/10.33795/jip.v7i2.495
2	Hafidza Nur ‘Ainika Rani, Tedy Rismawan, Cucu Suhery [15]	2023	Prediksi Jumlah Permintaan Darah menggunakan	Jurnal Komputer Dan Aplikasi, 10(03), 411	PRC gol. B 93,83%, alpha 0,2	doi.org/10.26418/coding.v10i03.56013

No	Penulis	Tahun	Judul	Nama Jurnal	Akurasi	DOI
			Metode Triple Exponential Smoothing (Studi Kasus: UTD PMI Kota Singkawang)		WB gol. AB 89,02%, alpha 0,1.	
3	Clement Twumasi, Juliet Twumasi [16]	2021	Machine Learning Algorithms for Forecasting and Backcasting Blood Demand Data with Missing Values and Outliers: A study of Tema General Hospital of Ghana	International Journal of Forecasting, 38(3), 1258–1277	MAE KNN: 12.55%	doi.org/10.1016/j.ijforecast.2021.10.008
4	Xinyi Ding, Xiao Zhang, Xiaofei Li, Jinlian Du [17]	2023	A Hybrid Neural Network-Based Model for Blood	Journal of Biomedical Informatics, 146,	R^2: 65.24% RMSE: 2.096 MAE: 1.760	doi.org/10.1016/j.jbi.2023.104488

No	Penulis	Tahun	Judul	Nama Jurnal	Akurasi	DOI
			Donation Forecasting	104488		
5	Mouigni Baraka Nafouanti, Junxia Li1, Edwin E. Nyakilla, Grant Charles Mwakipunda, Alvin Mulashani [11]	2023	A Novel Hybrid Random Forest Linear Model Approach for Forecasting Groundwater Fluoride Contamination	Environmental Science and Pollution Research, 30(17), 50661–50674	Hybrid RFLM: 95% Xgboost: 88% LightGBM: 88% RF: 85%	doi.org/10.1007/s11356-023-25886-w
6	Yajie Wang, Wei Zhang, Quan Rao, Yiming Ma, Xinyi Ding, Xiao Zhang, Xiaofei Li [18]	2024	Forecasting Demands of Blood Components Based on Prediction Models	Transfusion Clinique Et Biologique, 31(3), 141–148	SARIMAX R^2: 87.6% MAPE: 0.0037	doi.org/10.1016/j.tracli.2024.04.003
7	Emrehan Kutlug Sahin [12]	2020	Assessing the Predictive Capability	SN Applied Sciences, 2(7)	AUC Xgboost: 95.68%	doi.org/10.1007/s42452-020-

No	Penulis	Tahun	Judul	Nama Jurnal	Akurasi	DOI
			of Ensemble Tree Methods for Landslide Susceptibility Mapping Using Xgboost, Gradient Boosting Machine, and Random Forest		AUC RF: 93.77% AUC GBM: 92.15%	3060-1
8	Yannan Feng, Zhenhua Xu, Xiaolin Sun, Deqing Wang, Yang Yu [19]	2021	Machine Learning for Predicting Preoperative Red Blood Cell Demand	Transfusion Medicine, *31*(4), 262–270	LightGBM: AUC 0.908 (95% CI: 0.907–0.913)	doi.org/10.1111/tme.12794
9	Jie Song, Yuan Gao, Pengbin Yin, Yi Li, Yang Li, Jie Zhang, Qingqing Su, Xiaojie Fu, Hongying Pi [20]	2021	The Random Forest Model has The Best Accuracy Among The Four Pressure Ulcer Prediction	Risk Management and Healthcare Policy, Volume 14, 1175–1187	FI-Score: 99.88%	doi.org/10.2147/rmhp.s297838

No	Penulis	Tahun	Judul	Nama Jurnal	Akurasi	DOI
			Models using Machine Learning Algorithms			
10	Christian Kauten, Ashish Gupta, Xiao Qin, Glenn Richey [13]	2022	Predicting Blood Donors using Machine Learning Techniques	Information Systems Frontiers, 24(5), 1547–1562	RF: 99% GB: 98.9%	doi.org/10.1007/s10796-021-10149-1

Penelitian pertama yang berjudul "Sistem Prediksi Permintaan Darah Menggunakan Metode Regresi Linier" mengeksplorasi penggunaan regresi linier untuk memodelkan permintaan darah. Hasil evaluasi menunjukkan MAPE sebesar 80,14%, yang mengindikasikan keterbatasan model linier dalam menangkap kompleksitas pola dan variabel non-linear pada data permintaan darah [14]. Tingginya nilai MAPE menunjukkan bahwa regresi linier tidak mampu menangkap pola fluktuatif permintaan darah yang dipengaruhi oleh faktor musiman, hari libur, dan kondisi darurat medis. Model linier tidak dapat menangani interaksi kompleks antar variabel yang mempengaruhi permintaan darah di PMI. Penelitian ini terbatas pada pendekatan univariat dan tidak mempertimbangkan ensemble methods yang dapat menangani non-linearitas dan interaksi fitur yang kompleks dalam prediksi permintaan darah.

Penelitian yang berjudul "Prediksi Jumlah Permintaan Darah menggunakan Metode *Triple Exponential Smoothing* (Studi Kasus: UTD PMI Kota Singkawang)" menerapkan metode Triple Exponential Smoothing untuk meramalkan permintaan darah dengan hasil akurasi yang bervariasi berdasarkan golongan darah. Sebagai contoh, untuk PRC gol. B diperoleh akurasi sebesar 93,83% dengan *alpha* 0,2, sedangkan untuk WB gol. AB mencapai 89,02% dengan *alpha* 0,1 [15]. Variasi akurasi yang signifikan antar golongan darah menunjukkan bahwa setiap jenis darah memiliki karakteristik permintaan yang berbeda. Meskipun akurasi cukup baik, metode ini masih berbasis *time series* tradisional yang tidak dapat menangkap *feature importance* dan interaksi kompleks. Selain itu, metode smoothing tidak dapat mengidentifikasi faktor-faktor penting yang mempengaruhi permintaan darah seperti yang dapat dilakukan oleh Random Forest melalui *feature importance analysis.*

Dalam studi "Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: A study of Tema General Hospital of Ghana" yang diterbitkan di International Journal of Forecasting menginvestigasi berbagai algoritma *machine learning* untuk peramalan permintaan darah dalam kondisi data yang tidak sempurna. Hasilnya, model KNN

menunjukkan error sebesar 12,55%, menggambarkan tantangan dalam mengelola data nyata yang mengandung *missing values* dan *outlier* [16]. *Error rate* yang relatif rendah menunjukkan bahwa KNN dapat menangani data yang tidak sempurna dengan cukup baik. Namun, performa ini dapat menurun drastis pada dataset dengan dimensi tinggi atau pola yang kompleks. Serta, penelitian ini terbatas pada satu algoritma dan tidak mengeksplorasi potensi *ensemble methods* yang dapat memberikan *robustness* lebih baik terhadap data yang tidak sempurna.

Penelitian yang berjudul "A hybrid neural network-based model for blood donation forecasting" yang dipublikasikan di Journal of Biomedical Informatics mengembangkan model hybrid berbasis neural network untuk meramalkan jumlah donor darah. Dengan hasil evaluasi yang menunjukkan R^2 sebesar 65,24%, RMSE 2,096, dan MAE 1,760, penelitian ini menyoroti potensi integrasi berbagai teknik dalam meningkatkan akurasi prediksi pada domain donor darah [17]. Nilai R^2 yang moderat menunjukkan bahwa meskipun neural network dapat menangkap pola non-linier, kompleksitas model tidak selalu berkorelasi dengan peningkatan akurasi yang signifikan. Fokus penelitian ini hanya terletak pada donor darah, bukan permintaan darah aktual, menciptakan gap dalam pemahaman dinamika permintaan riil di lapangan. Selain itu, tidak ada perbandingan dengan metode *ensemble* yang lebih sederhana namun potensial lebih efektif.

Penelitian yang berjudul "Forecasting Demands of Blood Components Based on Prediction Models" yang diterbitkan di Transfusion Clinique Et Biologique menggunakan model SARIMAX untuk memprediksi permintaan komponen darah dengan capaian R^2 sebesar 87,6% dan MAPE yang sangat rendah (0,0037). Pendekatan time series ini efektif dalam menangkap tren jangka panjang dan pola musiman, namun keterbatasannya dalam menangani interaksi non-linier kompleks menekankan kebutuhan akan model prediksi yang lebih adaptif terhadap perubahan dinamis di lapangan, suatu aspek yang sangat penting untuk pengelolaan stok darah dalam sistem kesehatan [18].

Penelitian "A Novel Hybrid Random Forest Linear Model Approach for Forecasting Groundwater Fluoride Contamination" yang dipublikasikan di

Environmental Science and Pollution Research mengusulkan pendekatan hybrid yang menggabungkan model Random Forest dengan model linear, menghasilkan akurasi mencapai 95%, lebih tinggi dibandingkan dengan algoritma individual seperti XGBoost, LightGBM, atau Random Forest yang masing-masing menghasilkan akurasi antara 85–88% [11]. Kontribusi penelitian menegaskan bahwa kombinasi Random Forest dengan algoritma lain dapat mengkompensasi kelemahan individual dan menghasilkan prediksi yang lebih akurat. Namun, penelitian ini dilakukan dalam domain *environmental*, bukan *healthcare*, sehingga belum ada bukti empiris penerapan hybrid Random Forest dalam prediksi permintaan darah di PMI Indonesia.

Penelitian "Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using Xgboost, Gradient Boosting Machine, and Random Forest" yang diterbitkan di SN Applied Sciences membandingkan tiga metode *ensemble tree* dalam konteks pemetaan kerentanan longsor. Hasil penelitian menunjukkan AUC tertinggi sebesar 95,68% untuk XGBoost, sedangkan Random Forest dan Gradient Boosting Machine masing-masing mencapai AUC 93,77% dan 92,15% [12]. Perbedaan performa yang relatif kecil menunjukkan pentingnya evaluasi komprehensif untuk menentukan algoritma terbaik. Akan tetapi, penelitian ini dilakukan untuk *landslide prediction*, sehingga belum ada perbandingan langsung ketiga *ensemble methods* tersebut untuk prediksi permintaan darah di PMI Indonesia.

Melalui penelitian "Machine Learning for Predicting Perioperative Red Blood Cell Demand" yang diterbitkan di Transfusion Medicine menerapkan algoritma LightGBM untuk memprediksi kebutuhan transfusi sel darah merah pada pasien pra-operatif. Dengan AUC sebesar 0,908 (95% CI: 0,907–0,913), studi ini memberikan contoh aplikasi *machine learning* yang sukses dalam konteks klinis, dan menawarkan kontribusi penting sebagai benchmark prediksi dalam persiapan transfusi darah. Penelitian ini menekankan pentingnya penerapan algoritma prediktif untuk mengoptimalkan persiapan darah, meskipun cakupan prediksinya terbatas pada kebutuhan pra-operatif [19].

Dalam artikel “The Random Forest Model has The Best Accuracy Among the Four Pressure Ulcer Prediction Models using Machine Learning Algorithms” yang dipublikasikan di Risk Management and Healthcare Policy melaporkan bahwa model Random Forest mencapai FI-Score sebesar 99,88% dalam prediksi tekanan ulkus [20]. Akurasi Random Forest yang sangat tinggi dalam domain medis menegaskan potensi algoritma ini untuk aplikasi healthcare, termasuk prediksi permintaan darah. Random Forest terbukti unggul dibanding algoritma lain dalam menangani data medis yang kompleks. Meskipun Random Forest terbukti *excellent* dalam domain medis, peneliti belum secara khusus membandingkannya dengan Gradient Boosting dan pendekatan hybrid untuk prediksi permintaan darah.

Melalui studi “Predicting Blood Donors using Machine Learning Techniques” yang diterbitkan di Information Systems Frontiers menunjukkan bahwa model prediksi yang dibangun menggunakan algoritma Random Forest dan Gradient Boosting memperoleh akurasi masing-masing sebesar 99% dan 98,9% [13]. Kedua algoritma menunjukkan performa yang hampir sempurna dalam memprediksi donor darah dengan Random Forest sedikit unggul. Hasil ini sangat relevan karena menunjukkan bahwa kedua algoritma sangat cocok untuk domain darah dan transfusi. Namun, terdapat gap pada penelitian ini, yakni penelitian yg berfokus pada prediksi perilaku donor, bukan prediksi permintaan darah aktual yang dibutuhkan instansi kesehatan untuk pengelolaan stok.

Secara keseluruhan, penelitian-penelitian terdahulu telah mengeksplorasi berbagai metode, mulai dari regresi linier, peramalan *time series*, KNN, hingga beberapa model *ensemble* dalam konteks prediksi kebutuhan darah, donor, dan risiko klinis lainnya. Namun, terdapat *gap* dalam pemahaman mengenai performa masing-masing metode, terutama dalam menangani pola permintaan darah yang kompleks dan non-linier. Beberapa studi menunjukkan bahwa pendekatan statistik tradisional seperti regresi linier memiliki keterbatasan dalam menangkap pola non-linier. Sementara metode *time series* seperti Triple Exponential Smoothing dan SARIMAX efektif untuk tren jangka panjang, tetapi kurang adaptif terhadap perubahan mendadak. Selain itu, pendekatan berbasis *instance* seperti KNN

menunjukkan performa yang stabil tetapi kurang efisien untuk dataset berskala besar dan dengan interaksi fitur yang kompleks.

Di sisi lain, model *ensemble* seperti Random Forest dan Gradient Boosting telah terbukti memiliki akurasi tinggi dalam berbagai aplikasi prediksi. Namun, belum ada penelitian yang secara spesifik membandingkan kedua algoritma ini dalam konteks prediksi permintaan darah untuk pengelolaan stok UTD. Oleh karena itu, *novelty* dari penelitian ini terletak pada perbandingan langsung antara Random Forest, Gradient Boosting, dan hybrid Random Forest-Gradient Boosting dalam memprediksi jumlah permintaan darah. Dengan menganalisis performa kedua algoritma berdasarkan metrik evaluasi seperti MAE, MSE, dan R^2, penelitian ini bertujuan untuk mengidentifikasi metode yang lebih efektif dalam menangani dinamika permintaan darah di UTD PMI Kabupaten Tangerang. Pendekatan ini tidak hanya mengisi *gap* terkait efektivitas model *ensemble* dalam domain ini tetapi juga memberikan wawasan praktis bagi pengelola stok darah untuk mengambil keputusan strategis yang lebih akurat dan berbasis data.

2.2 Tinjauan Teori

2.2.1 Donor Darah

Donor darah merupakan aktivitas sukarela yang bertujuan untuk memenuhi kebutuhan darah di masyarakat [21]. Darah yang disumbangkan oleh pendonor menjadi sumber utama persediaan darah di UTD PMI. Ada dua jenis donor darah yang umum dilakukan, yaitu donor darah sukarela dan donor darah pengganti. Donor darah sukarela adalah bentuk ideal karena dilakukan tanpa paksaan dan lebih terjamin kontinuitasnya, sementara donor pengganti biasanya terjadi atas permintaan keluarga pasien. Penting untuk dicatat bahwa donor sukarela cenderung lebih aman daripada donor pengganti dalam hal risiko penularan infeksi yang dapat ditularkan melalui saluran transfusi [3]. Faktor yang memengaruhi jumlah donor darah meliputi kesadaran masyarakat, kampanye sosial, serta kemudahan akses ke lokasi donor. Kampanye digital

melalui media sosial dan platform online terbukti efektif meningkatkan jumlah donor, terutama di kalangan generasi muda.

Dalam konteks PMI Kabupaten Tangerang, melakukan prediksi jumlah permintaan darah juga dapat digunakan untuk mengidentifikasi kebutuhan donor darah secara spesifik. Hasil prediksi ini dapat memberikan rekomendasi terkait jumlah pengumpulan donor yang paling optimal berdasarkan data prediksi kebutuhan darah. Hal ini membantu meningkatkan efisiensi kampanye donor darah dan meminimalkan risiko kekurangan pasokan darah di wilayah tersebut. Dengan melibatkan algoritma prediksi kebutuhan darah, PMI dapat memastikan bahwa donor darah tidak hanya terkonsentrasi pada periode tertentu, tetapi juga merata sepanjang tahun. Beberapa penelitian terdahulu terkait prediksi donor darah telah dilakukan, seperti penggunaan hybrid neural network untuk forecasting jumlah donor darah dan penerapan machine learning techniques untuk memprediksi perilaku donor darah [13], [15], [17].

2.2.2 Stok Darah

Stok darah merujuk pada ketersediaan darah dan komponennya di fasilitas kesehatan seperti rumah sakit atau palang merah. Komponen darah yang dikelola meliputi sel darah merah, plasma, trombosit, dan cryoprecipitate. Komponen darah memiliki masa simpan yang berbeda. Sel darah merah, misalnya, dapat disimpan hingga 42 hari pada suhu 2-6°C, sedangkan trombosit hanya bertahan 5-7 hari pada suhu ruangan dengan agitasi kontinu [22]. Untuk memastikan kualitas, "*cold chain management*" menjadi sangat penting dalam mengontrol suhu selama pengumpulan, penyimpanan, hingga distribusi darah. Keberadaan stok darah yang cukup sangat penting karena darah merupakan komoditas medis yang tidak dapat diproduksi secara sintetis. Pengelolaan stok darah yang baik bertujuan untuk menjamin ketersediaan dan meminimalkan risiko kekurangan darah, terutama pada momen kritis seperti bencana alam, kecelakaan massal, atau meningkatnya permintaan akibat penyakit musiman.

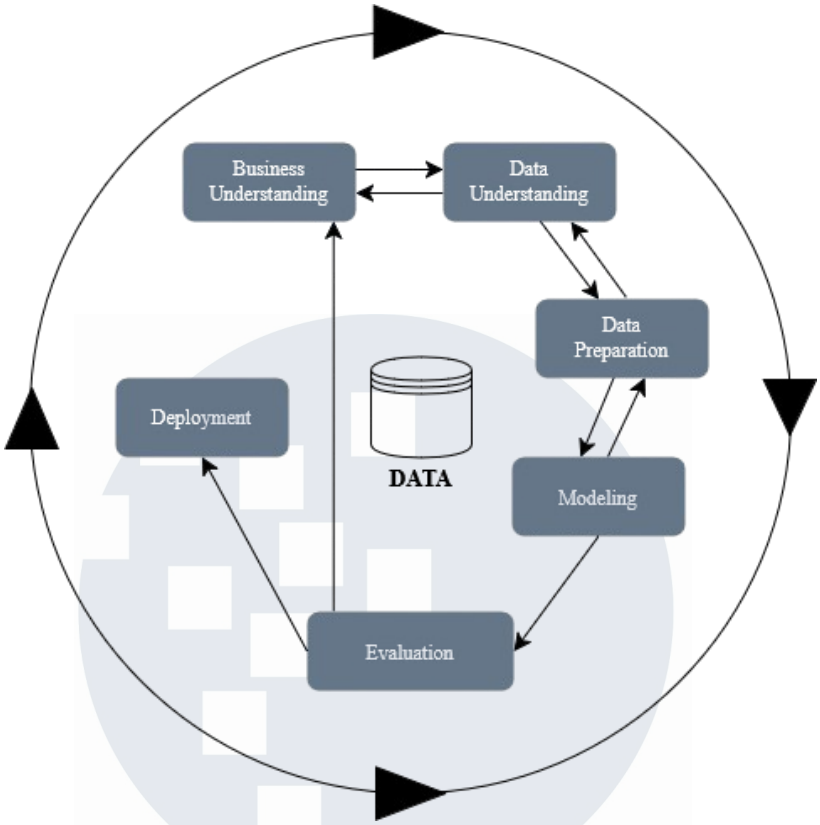
2.2.3 UTD PMI

Unit Transfusi Darah (UTD) merupakan bagian dari Palang Merah Indonesia (PMI) yang bertanggung jawab atas pengelolaan darah, mulai dari pengumpulan, pengolahan, penyimpanan, hingga distribusi kepada fasilitas kesehatan [15]. UTD PMI memiliki peran strategis dalam memastikan ketersediaan darah, khususnya di daerah seperti Kabupaten Tangerang yang memiliki populasi tinggi dan tingkat kebutuhan darah yang signifikan. Dalam operasionalnya, UTD PMI mengacu pada standar nasional dan internasional yang mengatur seluruh proses transfusi darah. Sebagai bagian dari sistem pelayanan kesehatan, UTD PMI juga bekerja sama dengan rumah sakit dan klinik untuk memonitor kebutuhan darah secara *real-time*.

2.3 Framework dan Algoritma

2.3.1 CRISP-DM

CRISP-DM adalah *framework* yang berlaku di berbagai industri dan bermanfaat terutama untuk proyek *data mining*. Gambar 2.1 menunjukkan langkah-langkah dalam CRISP-DM, yang mencakup enam tahapan berulang. Proses ini dimulai dengan *Business Understanding*. Setelah itu, langkah-langkah berikutnya adalah *Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment*. CRISP-DM dianggap sebagai metodologi *data mining* yang paling lengkap selama siklus pengembangannya untuk memenuhi kebutuhan proyek [23]. Gambar 2.1 menunjukkan bagaimana CRISP-DM digunakan dalam penelitian. Panah yang menghubungkan langkah-langkah tersebut menunjukkan adanya ketergantungan di antara langkah-langkah tersebut. Namun, urutan langkah-langkah tersebut dapat berubah-ubah tergantung pada hasil dari setiap langkah. Siklus data mining diwakili oleh cincin luar yang menutupi setiap fase. Tujuan utama model CRISP-DM adalah untuk mengoptimalkan proyek data mining dengan cara yang lebih ekonomis, dapat diandalkan, dapat diulang, mudah dikelola, dan cepat [24].



Gambar 2.1 *Framework CRISP-DM*

Fase pertama dari model CRISP-DM adalah *business understanding*, yang bertujuan untuk mendapatkan pemahaman tentang tujuan dan kebutuhan proyek dari sudut pandang bisnis. Kemudian, Informasi yang dikumpulkan kemudian dianalisis untuk menghasilkan pernyataan masalah *data mining* dan rencana proyek untuk mencapai tujuan tersebut. Kemudian, dilanjutkan dengan tahap *data understanding*, yang dimulai dengan pengumpulan data dan dilanjutkan dengan eksplorasi data, evaluasi data, dan pemahaman korelasi antar atribut. *Data Preparation* adalah langkah penting berikutnya setelah fase kedua selesai. Fase ini mencakup semua tindakan yang diperlukan untuk menyusun dataset akhir dari data mentah awal. Di tahap ini dilakukan beberapa tugas seperti pemilihan atribut, *pre-processing data*, konstruksi atribut baru, dan transformasi data untuk pemodelan. *Modeling* adalah langkah selanjutnya dalam proses CRISP-DM. Di fase ini ada berbagai teknik pemodelan yang dapat digunakan. Pada penelitian ini digunakan metode Random Forest dan Gradient Boosting. Tahap berikutnya adalah *evaluation*, di mana satu atau

lebih model dibangun untuk menghasilkan analisis data yang akurat. Untuk memastikan bahwa model mencapai tujuan bisnis, dilakukan evaluasi menyeluruh dan peninjauan langkah-langkah yang diambil untuk membangun model sebelum dilanjutkan ke implementasi akhir model. Tujuan utamanya adalah untuk mengidentifikasi masalah yang mungkin terlewatkan. *Deployment* adalah tahap terakhir dari proyek. Secara umum, informasi yang diperoleh harus diorganisir dan disajikan agar dapat dipahami dan digunakan oleh pengguna.

2.3.2 Random Forest

Random Forest adalah algoritma *machine learning* berbasis *ensemble* yang menggabungkan sejumlah besar pohon keputusan (*decision trees*) untuk menghasilkan prediksi yang lebih stabil dan akurat. Algoritma ini bekerja dengan membangun banyak model pohon keputusan independen, di mana setiap pohon dilatih menggunakan subset data yang diambil secara acak dengan metode *bootstrap*. Untuk prediksi akhir, Random Forest melakukan rata-rata hasil prediksi semua pohon dalam kasus regresi, atau mengambil hasil *voting* mayoritas dalam kasus klasifikasi [25]. Pendekatan ini membuat Random Forest sangat efektif dalam meminimalkan *overfitting* karena model tidak bergantung pada satu pohon tertentu, melainkan mengandalkan kontribusi kolektif dari seluruh pohon. Selain itu, algoritma ini juga fleksibel dan mampu menangani berbagai jenis data, baik regresi maupun klasifikasi. Adapun rumus utama untuk model *Random Forest* adalah sebagai berikut:

Jika terdapat M pohon keputusan $T_1, T_2, \dots, T_M$, maka prediksi akhir $\hat{y}$ untuk regresi adalah:

$$\hat{y} = \frac{1}{M} \sum_{i=1}^{M} T_i(x)$$

Rumus 2.1 Rumus Random Forest Regresi

Untuk klasifikasi, prediksi adalah hasil voting:

$$\hat{y} = mode\big(T_1(x), T_2(x), \dots, T_M(x)\big)$$

Rumus 2.2 Rumus Random Forest Klasifikasi

Keunggulan utama Random Forest adalah kemampuannya dalam menangani data yang memiliki *noise* dan fitur yang tidak relevan. Model ini juga dapat memberikan estimasi pentingnya setiap fitur, sehingga mempermudah interpretasi data. Namun, meskipun sangat andal, Random Forest memiliki keterbatasan dalam efisiensinya saat digunakan pada dataset yang sangat besar dengan banyak fitur, karena memerlukan banyak pohon yang dapat meningkatkan kompleksitas komputasi. Dalam konteks prediksi stok darah, Random Forest digunakan untuk memanfaatkan data historis yang sedikit kompleks, misalnya pola musiman.

2.3.3 Gradient Boosting

Gradient Boosting adalah algoritma *machine learning* berbasis *ensemble* yang membangun model secara bertahap dengan cara mengoptimalkan fungsi *loss* melalui *gradient descent*. Pendekatan ini dirancang untuk memperbaiki kelemahan model sebelumnya secara iteratif, sehingga setiap model baru berkontribusi untuk mengurangi error residual yang tersisa [26]. Gradient Boosting memulai prosesnya dengan membangun model awal yang sederhana, seringkali berupa pohon keputusan tunggal, dan kemudian secara bertahap menambahkan model baru untuk memperbaiki kesalahan prediksi dari iterasi sebelumnya [27], [28]. Setiap model baru dilatih untuk memperkirakan gradien dari fungsi *loss* terhadap error, dan kontribusi dari setiap model baru ini disesuaikan oleh faktor *learning rate* untuk memastikan proses pembelajaran tidak terlalu agresif. Adapun rumus utama algoritma ini adalah sebagai berikut:

Gradient Boosting bekerja iteratif untuk meminimalkan fungsi *loss L.* Pada iterasi ke-*t:*

1. Model prediksi awal:

$$F_0(x) = arg$$

Rumus 2.3 Rumus Inisialisasi Model Awal Gradient Boosting

2. Model ditingkatkan dengan menambahkan pohon baru:

$$F_t(x) = F_{t-1}(x) + \eta \, . h_t(x)$$

Rumus 2.4 Rumus Pembaruan Model Gradient Boosting

Dengan η adalah *learning rate,* dan $h_t(x)$ adalah pohon keputusan yang dibangun berdasarkan gradien fungsi *loss* [27].

Algoritma ini dikenal sangat akurat dan mampu menangkap pola yang kompleks dalam data. Fleksibilitas algoritma ini memungkinkan penggunaannya dalam berbagai skenario, termasuk untuk memprediksi stok darah. Gradient Boosting cocok digunakan dalam prediksi stok darah karena mampu menganalisis pola yang kompleks, seperti variasi permintaan akibat tren musiman atau faktor eksternal lainnya, dan memberikan estimasi yang sangat akurat untuk kebutuhan masa depan.

2.3.4 Hybrid Random Forest - Gradient Boosting

Pada *machine learning* terdapat istilah pendekatan hybrid. Pendekatan hybrid adalah metode yang menggabungkan dua atau lebih algoritma untuk memanfaatkan keunggulan masing-masing algoritma sekaligus mengatasi kelemahan yang ada jika algoritma digunakan secara terpisah. Dalam konteks hybrid untuk prediksi, pendekatan ini sering digunakan untuk meningkatkan akurasi, efisiensi, dan stabilitas model prediksi dengan mengintegrasikan karakteristik terbaik dari beberapa metode. Pada prinsipnya, pendekatan hybrid beroperasi melalui sinergi antara algoritma. Biasanya, salah satu algoritma digunakan untuk membangun fondasi atau memberikan prediksi awal, sementara algoritma lain berfungsi menyempurnakan hasilnya dengan memperbaiki kelemahan atau error yang masih ada. Hasil akhir dari pendekatan hybrid sering kali lebih unggul dibandingkan penggunaan algoritma tunggal karena menggabungkan dua sudut pandang analitis yang saling melengkapi.

Hybrid Random Forest Gradient Boosting (HRFGB) adalah kombinasi dari algoritma Random Forest dan Gradient Boosting yang memanfaatkan keunggulan masing-masing metode untuk menciptakan model prediksi yang lebih akurat dan stabil. Pendekatan ini dirancang untuk mengatasi kelemahan individu dari kedua algoritma tersebut, seperti risiko *overfitting* pada Gradient Boosting atau kebutuhan komputasi tinggi pada Random Forest, dengan menggabungkan keunggulan utama dari keduanya. Dalam HRFGB, Random Forest digunakan sebagai basis untuk menangani kompleksitas data dan mengurangi pengaruh fitur yang tidak relevan, sementara Gradient Boosting digunakan untuk meningkatkan akurasi prediksi dengan menyempurnakan model melalui optimasi iteratif.

Secara teknis, algoritma ini bekerja dengan menghasilkan prediksi awal menggunakan Random Forest, kemudian mengidentifikasi error residual yang masih ada, dan menyempurnakannya dengan menambahkan model Gradient Boosting. Kombinasi ini menghasilkan model akhir yang lebih akurat karena prediksi Random Forest memberikan stabilitas, sementara Gradient Boosting memperbaiki kelemahan residual secara iteratif. Dalam kasus regresi, hasil akhir dapat diperoleh dengan menggabungkan prediksi kedua algoritma menggunakan rata-rata tertimbang, di mana bobot diberikan berdasarkan tingkat kontribusi masing-masing algoritma terhadap hasil akhir.

2.3.5 Hyperparameter Tuning

Hyperparameter tuning adalah proses mengatur parameter model yang ditentukan sebelum proses pelatihan dimulai, berbeda dari parameter yang dipelajari secara otomatis selama *training*. Contohnya pada Random Forest adalah n_estimators dan max_depth, sementara pada Gradient Boosting meliputi learning_rate dan parameter regularisasi. *Tuning* yang tepat sangat penting agar model tidak mengalami *overfitting* maupun *underfitting* [29], terutama dalam konteks prediksi permintaan darah yang kompleks dan sensitif terhadap akurasi. Salah satu metode *tuning* yang digunakan dalam penelitian ini adalah RandomizedSearchCV, yaitu teknik pencarian acak pada ruang

hyperparameter yang ditentukan. Metode ini lebih efisien dibanding GridSearch karena tidak menguji semua kombinasi, melainkan memilih subset secara acak, lalu mengevaluasinya menggunakan *cross-validation* untuk mendapatkan konfigurasi terbaik [30]. Efisiensi ini sangat berguna saat ruang pencarian luas dan waktu komputasi terbatas.

Pada Random Forest, *tuning* mencakup parameter seperti n_estimators, max_depth, min_samples_split, dan max_features, yang memengaruhi keseimbangan bias-variance. Pada Gradient Boosting, tuning difokuskan pada learning_rate, n_estimators, max_depth, subsample, dan parameter regularisasi untuk menghindari overfitting. Dengan menggunakan RandomizedSearchCV, penelitian ini dapat mengoptimalkan performa masing-masing model secara adil dan efisien, sehingga hasil perbandingan lebih akurat dalam menentukan pendekatan terbaik untuk prediksi permintaan darah di PMI Kabupaten Tangerang.

2.4 Tools Penelitian

2.4.1 Microsoft Excel

Microsoft Excel merupakan salah satu program *spreadsheet* yang paling populer dan penting untuk analisis data, terutama dalam hal pengorganisasian, pengolahan, dan visualisasi data [32]. Selain itu, Microsoft Excel memungkinkan pengguna melakukan berbagai operasi perhitungan, mengelola data, dan membuat lembar kerja struktural. Di penelitian ini, Microsoft Excel digunakan untuk mengelompokkan data permintaan dan persediaan darah berdasarkan tahun nya. Data yang awalnya terpisah berdasarkan bulan dikelompokkan berdasarkan tahun nya. Selain itu Microsoft Excel digunakan untuk menyimpan dataset dalam bentuk format CSV (*Comma-Separated Values*) selama masa *pre-processing.*

2.4.2 Jupyter Notebook

Jupyter Notebook adalah tools pemrograman open source berbasis web yang digunakan untuk pengembangan interaktif dan penyajian proyek data

science. Setiap notebook terdiri dari sejumlah sel yang bisa berupa sel code, sel markdown, dan sel raw. Ketika sebuah sel code dieksekusi, semua output grafis atau teks (hasil numerik, gambar, atau tabel) akan disajikan dalam dokumen tepat di bawah sel tersebut. Jupyter Notebook dapat digunakan untuk berbagai bahasa pemrograman, termasuk Python, R, Julia, JavaScript, C, dan lainnya [33]. Jupyter Notebook menawarkan sejumlah keunggulan yang sangat menguntungkan penggunanya. Pertama, kemudahannya digunakan membuatnya sesuai untuk pengguna pemula maupun berpengalaman. Dengan *interface* yang sederhana, pengguna dapat dengan mudah membuat dokumen, menulis kode, dan menjalankannya. Kedua, fleksibilitas Jupyter Notebook menjadikannya alat yang sangat serbaguna. Pengguna dapat membuat program, menganalisis data, menciptakan visualisasi, dan membangun model pembelajaran mesin [33]. Keunggulan lainnya adalah interaktivitas. Pengguna dapat menjalankan kode dan melihat hasilnya secara langsung, memudahkan dalam *debugging* dan pengembangan. Hal itu memungkinkan pengguna untuk dengan cepat mengevaluasi kinerja *code* mereka dan membuat perubahan jika diperlukan.

2.4.3 Python

Python adalah bahasa pemrograman tingkat tinggi yang fleksibel, mudah dipahami, dinamis, dan dapat diterapkan pada berbagai domain aplikasi. Python diyakini sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaks kode yang sangat jelas, dan dilengkapi dengan fungsionalitas *library* yang besar serta komprehensif [34]. Python dapat digunakan untuk mengembangkan berbagai jenis aplikasi, seperti aplikasi web, aplikasi *desktop*, dan aplikasi seluler. Selain itu, bahasa pemrograman Python dapat dijalankan di berbagai platform sistem operasi, seperti Windows, Mac OS, Linux, dan sebagainya. Selain memiliki kemampuan membaca kode yang tinggi, sehingga *code* mudah dipahami, Python juga memiliki library yang sangat banyak dan luas. Hal ini membuat Python menjadi bahasa pemrograman yang ideal untuk pemula dan profesional. Meskipun Python dikenal sebagai bahasa tingkat

tinggi, Python termasuk salah satu bahasa pemrograman yang mudah dipelajari karena sintaks yang jelas, dikombinasikan dengan penggunaan modul-modul siap pakai dan struktur data tingkat tinggi yang efisien [34].

2.4.4 Streamlit

Streamlit adalah *framework open-source* yang digunakan untuk membuat aplikasi web *data science* dengan cepat dan efisien [35]. Platform ini dirancang untuk memudahkan para pengembang, *data scientist*, dan analis dalam membuat aplikasi interaktif tanpa perlu menguasai pengetahuan mendalam tentang pengembangan web. Dengan menggunakan Python, Streamlit memungkinkan pengguna untuk membuat antarmuka yang dinamis hanya dengan beberapa baris kode.

Streamlit menawarkan solusi praktis untuk membangun aplikasi berbasis data tanpa memerlukan HTML, CSS, atau JavaScript. Hal ini memberikan keuntungan besar dalam hal waktu dan sumber daya, terutama bagi data scientist yang lebih fokus pada analisis data dibandingkan pengembangan web. Selain itu, Streamlit juga sangat berguna dalam proyek-proyek penelitian yang melibatkan visualisasi data dan *machine learning*. Dengan integrasi yang mulus ke berbagai pustaka Python seperti Pandas, Matplotlib, Plotly, dan TensorFlow, Streamlit memungkinkan visualisasi yang lebih jelas dan interaktif dari hasil model pembelajaran mesin, mempercepat proses komunikasi hasil penelitian [35].

Keunggulan Streamlit juga terletak pada kemudahan penggunaannya. Satu-satunya yang dibutuhkan untuk memulai adalah pemahaman dasar Python, yang mana hal ini menjadikan Streamlit dapat diakses oleh banyak kalangan profesional, terutama mereka yang baru memasuki dunia pengembangan aplikasi berbasis data. Dengan antarmuka yang intuitif, Streamlit menyederhanakan proses pengembangan aplikasi menjadi lebih efisien tanpa mengurangi fungsionalitas.