# BAB II

# LANDASAN TEORI

## 2.1 Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai landasan dalam memahami perkembangan metode analisis data dan *Machine Learning* yang berkaitan dengan prediksi klaim asuransi. Kajian terhadap penelitian sebelumnya difokuskan pada penggunaan berbagai metode statistik dan algoritma *Machine Learning*, seperti *Linear Regression, Random Forest, XGBoost*, serta metode lainnya dalam menganalisis dan memodelkan data klaim asuransi.

Beberapa penelitian menunjukkan bahwa data klaim asuransi memiliki karakteristik yang kompleks, melibatkan banyak variabel seperti nilai pertanggungan, jenis produk, profil tertanggung, serta faktor risiko lainnya. Oleh karena itu, pendekatan berbasis data mining dan *Machine Learning* menjadi solusi yang banyak digunakan untuk mengidentifikasi pola hubungan antar variabel serta meningkatkan akurasi prediksi nilai klaim.

Selain itu, penelitian terdahulu juga memberikan gambaran mengenai pentingnya tahapan *preprocessing data*, pemilihan fitur, serta evaluasi model dalam menghasilkan model prediksi yang optimal. Tantangan seperti data tidak seimbang, *missing values*, serta hubungan *non-linear* antar variabel menjadi faktor penting yang mempengaruhi performa model.

Dengan demikian, penelitian terdahulu tidak hanya berfungsi sebagai referensi teoritis, tetapi juga sebagai dasar dalam menentukan metode yang digunakan dalam penelitian ini, yaitu *Linear Regression dan Random Forest*, serta dalam merancang alur analisis data yang sistematis.

Berikut adalah beberapa penelitian terdahulu yang relevan,

Tabel 2.1 Penelitian Terdahulu

| No. | Judul | Tahun | Penulis | Metode | Penjelasan |
|---|---|---|---|---|---|
| **1.** | Development of ML Models to Predict Health Insurance Claim Costs | 2026 | Yeni Mahwati dkk | LR, RF, XGBoost | Hasil menunjukkan XGBoost memiliki performa terbaik, diikuti Random Forest, sedangkan Linear Regression memiliki akurasi lebih rendah. Disimpulkan bahwa metode Machine Learning lebih unggul dalam |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | prediksi nilai klaim.[12] |
| **2.** | Pengaruh Rasio Beban Klaim & Hasil Investasi | 2026 | Anggun Fitria Novianti dkk | Regresi Panel | Ditemukan bahwa rasio klaim tidak berpengaruh signifikan, sedangkan hasil investasi berpengaruh positif. Disimpulkan bahwa faktor keuangan tertentu lebih dominan dibanding klaim.[13] |
| **3.** | Comparison LR, RF, Logistic Regression | 2023 | Surya Wijaya, Fauziah | LR, RF, Logistic | Random Forest menunjukkan performa terbaik di semua skenario. Disimpulkan bahwa metode berbasis ensemble lebih unggul untuk prediksi dibanding regresi klasik.[14] |
| **4.** | Car Insurance Compensation ML | 2024 | Yichen Han | RF, NN | Hasil menunjukkan model *Machine Learning* mampu memprediksi klaim dengan akurasi tinggi. Disimpulkan bahwa RF dan NN efektif untuk analisis klaim.[15] |
| **5.** | Responsible AI Insurance Claims | 2024 | Ashrafe Alam dkk | RF, LR, XGBoost | XGBoost dan Random Forest memiliki akurasi tertinggi. Disimpulkan bahwa variabel seperti BMI dan kebiasaan hidup sangat berpengaruh terhadap klaim.[16] |
| **6.** | *Random Forest* Prediksi Pendapatan Asuransi | 2025 | Wilsen Mokodaser dkk | Random Forest | Random Forest memiliki performa tinggi ($R^2$ 0.85). Disimpulkan bahwa metode ini efektif |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | dalam menangkap pola data kompleks.[17] |
| **7.** | Forecasting Health Insurance Claims | 2025 | Iyad Alkrunz dkk | RF, MARS | Random Forest menunjukkan hasil terbaik setelah preprocessing data. Disimpulkan bahwa pengolahan data meningkatkan akurasi model.[18] |
| **8.** | Perbandingan RF dan Poisson | 2025 | Luthfir Adrell dkk | RF, Poisson | Random Forest lebih unggul dalam akurasi, sedangkan Poisson lebih interpretatif. Disimpulkan bahwa pemilihan metode tergantung tujuan analisis.[19] |
| **9.** | Model Lognormal Premi Asuransi | 2025 | Putri Isnaini Baiti dkk | GLM, Lognormal | Model lognormal terbaik berdasarkan AIC/BIC. Disimpulkan bahwa distribusi data sangat mempengaruhi hasil model.[20] |
| **10.** | Regresi Poisson Klaim Asuransi Jiwa | 2025 | Lathifatul Aulia dkk | Poisson | Regresi Poisson mampu memodelkan hubungan signifikan antara polis dan klaim. Disimpulkan bahwa model ini efektif untuk data count.[21] |
| **11.** | Generalised Linear Models Actuarial | 2025 | Haberman & Renshaw | GLM | GLM memiliki fleksibilitas tinggi untuk berbagai jenis data klaim. Disimpulkan bahwa GLM menjadi dasar banyak model prediksi asuransi.[22] |
| **12.** | Pemodelan Klasifikasi Penyakit Jantung | 2026 | Eugene Kaparang, Marcellus | Random Forest | Penelitian ini bertujuan untuk meningkatkan |

| | | | | | |
|---|---|---|---|---|---|
| | Menggunakan Logistic Regression dan Random Forest dengan Teknik Borderline SMOTE dan K-Means SMOTE | | | | performa model klasifikasi penyakit jantung menggunakan algoritma Logistic Regression dan Random Forest dengan menerapkan teknik penanganan ketidakseimbangan data, yaitu Borderline- SMOTE dan K-Means SMOTE. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-score, dan Area Under the Curve (AUC) [14][23] |

Berdasarkan sebelas penelitian terdahulu yang telah dikaji pada Tabel 2.1, dapat diketahui bahwa penelitian mengenai prediksi klaim asuransi telah berkembang dengan memanfaatkan berbagai pendekatan, mulai dari metode statistik klasik hingga metode *Machine Learning*. Beberapa penelitian menggunakan metode statistik seperti *Generalized Linear Model* (*GLM*), regresi *Poisson*, dan *Linear Regression* untuk menganalisis hubungan antar variabel klaim asuransi[15][16][17]. Metode tersebut memiliki keunggulan dari sisi interpretasi karena mampu menjelaskan pengaruh masing-masing variabel terhadap target yang diprediksi. Namun, metode statistik memiliki keterbatasan dalam menangani hubungan *non-linear* dan interaksi kompleks antar variabel yang sering ditemukan pada data asuransi.

Di sisi lain, berbagai penelitian menunjukkan bahwa metode *Machine Learning* seperti *Random Forest*, *XGBoost*, *Gradient Boosting*, dan *Artificial Neural Network* mampu menghasilkan performa prediksi yang lebih baik dibandingkan metode statistik tradisional[18][19][20][21] Hal tersebut disebabkan oleh kemampuan algoritma *Machine Learning* dalam mempelajari pola data yang kompleks, menangani hubungan *non-linear*, serta memanfaatkan informasi dari banyak variabel secara simultan. Selain itu, beberapa penelitian juga menunjukkan bahwa kualitas data, proses *preprocessing*, pemilihan fitur, dan transformasi data menjadi faktor penting yang memengaruhi performa model prediksi[20][22][23].

Jika dikaitkan dengan kebutuhan bisnis PT Jasaraharja Putera, hasil penelitian terdahulu menunjukkan bahwa pendekatan *Machine Learning* memiliki potensi untuk membantu perusahaan dalam meningkatkan efisiensi proses analisis klaim dan pengelolaan risiko. Perusahaan memiliki data historis polis dan klaim yang mengandung berbagai karakteristik seperti nilai pertanggungan, premi, durasi polis, jumlah klaim, saluran distribusi (*channel*), serta jenis produk asuransi. Informasi tersebut dapat dimanfaatkan untuk memprediksi nilai klaim sekaligus membantu perusahaan memahami karakteristik polis yang paling berpengaruh terhadap besarnya klaim yang diajukan oleh nasabah.

Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada peningkatan akurasi prediksi atau perbandingan performa algoritma menggunakan dataset publik maupun data asuransi secara umum [16][18][19][20][22][23]. Penelitian sebelumnya juga lebih banyak menitikberatkan pada hasil prediksi akhir tanpa membahas secara mendalam pengaruh karakteristik polis terhadap nilai klaim. Selain itu, masih terbatas penelitian yang menggunakan data historis aktual perusahaan asuransi kendaraan di Indonesia serta mengombinasikan pendekatan statistik yang mudah diinterpretasikan dengan metode *Machine Learning* berbasis *ensemble learning* dalam satu penelitian.

Berdasarkan kondisi tersebut, terdapat *research gap* berupa masih terbatasnya penelitian yang tidak hanya berfokus pada akurasi prediksi, tetapi juga menganalisis hubungan karakteristik polis terhadap nilai klaim sebagai dasar pengambilan keputusan bisnis perusahaan asuransi. Selain itu, masih sedikit penelitian yang memanfaatkan data historis aktual perusahaan asuransi kendaraan di Indonesia untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap nilai klaim.

Sebagai upaya untuk mengisi *research gap* tersebut, penelitian ini memiliki beberapa unsur kebaruan (*novelty*), yaitu penggunaan data historis PT Jasaraharja Putera sebagai studi kasus perusahaan asuransi kendaraan di Indonesia, analisis pengaruh karakteristik polis terhadap nilai klaim melalui pendekatan *feature importance*, perbandingan metode *Linear Regression* dan *Random Forest* untuk melihat perbedaan pendekatan statistik dan *Machine Learning*, serta implementasi model prediksi ke dalam aplikasi berbasis web menggunakan *Streamlit* untuk mendukung proses analisis risiko dan evaluasi klaim perusahaan.

Meskipun beberapa penelitian menunjukkan bahwa algoritma seperti *XGBoost* dan *LightGBM* mampu menghasilkan performa prediksi yang sangat tinggi, penelitian ini memilih *Linear Regression* dan *Random Forest* dengan beberapa pertimbangan. *Linear Regression* digunakan karena memiliki tingkat interpretabilitas yang tinggi sehingga dapat menjelaskan hubungan antara karakteristik polis dan nilai klaim secara lebih

mudah dipahami oleh pihak bisnis. Sementara itu, *Random Forest* dipilih karena mampu menangani hubungan *non-linear*, lebih tahan terhadap *overfitting*, serta menyediakan informasi *feature importance* yang dapat digunakan untuk mengidentifikasi faktor-faktor polis yang paling berpengaruh terhadap nilai klaim. Selain itu, kedua metode tersebut relatif lebih sederhana untuk diimplementasikan dan dievaluasi dibandingkan algoritma *boosting* yang memerlukan optimisasi parameter yang lebih kompleks.

## 2.2 Teori yang berkaitan

### 2.2.1 Asuransi

Asuransi merupakan suatu mekanisme pengalihan risiko dari pihak tertanggung kepada pihak penanggung melalui pembayaran premi. Dalam operasionalnya, perusahaan asuransi seperti PT. Jasaraharja Putera bertanggung jawab dalam memberikan perlindungan finansial terhadap risiko kerugian yang mungkin terjadi di masa mendatang.

Klaim asuransi merupakan realisasi dari perlindungan tersebut, di mana tertanggung mengajukan permintaan pembayaran atas kerugian yang dialami. Besarnya nilai klaim dipengaruhi oleh berbagai faktor yang berasal dari karakteristik polis, seperti nilai pertanggungan, jenis produk, serta durasi pertanggungan. Oleh karena itu, analisis terhadap data klaim menjadi sangat penting untuk memahami pola risiko serta meningkatkan efektivitas pengelolaan portofolio asuransi.

Dalam penelitian ini, klaim asuransi digunakan sebagai variabel dependen yang akan dianalisis berdasarkan karakteristik polis yang dimiliki oleh PT. Jasaraharja Putera.

### 2.2.2 Data Mining

Data mining merupakan proses penggalian informasi dari data dalam jumlah besar untuk menemukan pola atau hubungan yang tidak terlihat secara langsung. Dalam konteks penelitian ini, data mining digunakan untuk mengolah data polis dan klaim yang dimiliki oleh PT. Jasaraharja Putera.

Proses data mining meliputi beberapa tahapan penting seperti data preprocessing, transformasi data, pemodelan, serta evaluasi hasil. Tahapan ini sangat relevan dengan penelitian ini karena data yang digunakan berasal dari sistem operasional perusahaan yang memiliki potensi masalah seperti missing values, duplikasi data, serta perbedaan format.

Dengan menggunakan pendekatan data mining, penelitian ini bertujuan untuk menemukan pola hubungan antara karakteristik polis

dan nilai klaim yang dapat digunakan sebagai dasar dalam pengambilan keputusan.

### 2.2.3 Karakteristik Polis

Karakteristik polis merupakan sekumpulan atribut yang menggambarkan kondisi pertanggungan, profil risiko, serta informasi administratif yang melekat pada suatu polis asuransi. Karakteristik tersebut menjadi dasar bagi perusahaan asuransi dalam melakukan evaluasi risiko, menentukan besaran premi, menetapkan cakupan perlindungan, serta mendukung proses *underwriting* sebelum polis diterbitkan.

Dalam penelitian ini, karakteristik polis digunakan sebagai variabel prediktor untuk mempelajari hubungan antara informasi polis dengan nilai klaim yang dibayarkan perusahaan. Karakteristik polis yang digunakan meliputi nilai pertanggungan (TSI_OC), premi asuransi (Premi_IDR), jumlah klaim historis (Jumlah_Claim), durasi polis (Durasi_Polis), jenis lini bisnis asuransi (COB_desc), dan saluran pemasaran polis (Channel).

Variabel TSI_OC merepresentasikan besarnya nilai pertanggungan yang diberikan perusahaan kepada tertanggung. Variabel Premi_IDR menunjukkan besaran premi yang dibayarkan nasabah sebagai kompensasi atas perlindungan yang diterima. Variabel Jumlah_Claim menggambarkan frekuensi klaim historis yang pernah terjadi pada polis yang bersangkutan sehingga dapat digunakan sebagai indikator risiko. Variabel Durasi_Polis menunjukkan lamanya masa pertanggungan, sedangkan COB_desc dan Channel digunakan untuk menggambarkan karakteristik bisnis dan mekanisme distribusi polis.

Pemilihan variabel tersebut dilakukan berdasarkan ketersediaan data historis perusahaan serta relevansinya terhadap proses pengelolaan risiko asuransi. Karakteristik polis tidak digunakan untuk memprediksi nilai klaim secara langsung, melainkan digunakan sebagai sumber informasi yang memungkinkan model *Machine Learning* mempelajari pola hubungan antara karakteristik polis dengan nilai klaim historis yang pernah terjadi. Dengan demikian, model dapat menghasilkan estimasi nilai klaim untuk polis baru berdasarkan pola yang telah dipelajari dari data sebelumnya.

### 2.2.4 Klaim Asuransi

Klaim asuransi merupakan permintaan pembayaran yang diajukan oleh tertanggung kepada perusahaan asuransi atas kerugian yang dialami akibat suatu peristiwa yang dijamin dalam polis asuransi. Proses klaim biasanya melibatkan verifikasi terhadap dokumen dan bukti

kejadian untuk memastikan bahwa peristiwa tersebut termasuk dalam cakupan perlindungan polis.

Nilai klaim yang dibayarkan oleh perusahaan asuransi dapat dipengaruhi oleh berbagai faktor, termasuk karakteristik polis, tingkat risiko yang diasuransikan, serta kondisi kejadian yang menimbulkan kerugian. Oleh karena itu, perusahaan asuransi perlu melakukan implementasi terhadap data klaim untuk memahami pola kerugian yang terjadi.

Dalam beberapa tahun terakhir, penggunaan metode implementasi data dan *Machine Learning* dalam prediksi nilai klaim semakin berkembang. Pendekatan ini memungkinkan perusahaan asuransi untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap besarnya nilai klaim serta meningkatkan akurasi dalam pengelolaan risiko.

### 2.2.5 *Machine Learning*

*Machine Learning* merupakan metode analisis data yang memungkinkan sistem untuk belajar dari data historis dan menghasilkan prediksi atau keputusan tanpa pemrograman eksplisit. Dalam penelitian ini, *Machine Learning* digunakan untuk membangun model prediksi nilai klaim berdasarkan data polis dan klaim PT. Jasaraharja Putera.

Keunggulan utama *Machine Learning* adalah kemampuannya dalam menangani data dengan jumlah besar dan hubungan yang kompleks. Hal ini sangat penting dalam analisis data asuransi, di mana hubungan antara variabel tidak selalu bersifat linear.

Penggunaan *Machine Learning* dalam penelitian ini didasarkan pada hasil penelitian terdahulu yang menunjukkan bahwa metode ini memiliki performa yang lebih baik dibandingkan metode statistik konvensional dalam memprediksi nilai klaim.

### 2.2.6 *Linear Regression*

*Linear Regression* merupakan metode statistik yang digunakan untuk memodelkan hubungan antara variabel independen dan dependen secara linear. Dalam penelitian ini, metode ini digunakan untuk mengidentifikasi pengaruh karakteristik polis terhadap nilai klaim.

Keunggulan utama *Linear Regression* adalah kemampuannya dalam memberikan interpretasi yang jelas terhadap hubungan antar variabel. Hal ini penting dalam penelitian ini karena selain melakukan prediksi, juga ingin mengetahui variabel mana yang memiliki pengaruh signifikan terhadap nilai klaim.

Namun, metode ini memiliki keterbatasan dalam menangani hubungan non-linear, sehingga perlu dikombinasikan dengan metode lain seperti *Random Forest*.

### 2.2.7 *Random Forest*

*Random Forest* merupakan algoritma *Machine Learning* berbasis ensemble yang mampu menangani data dengan kompleksitas tinggi. Metode ini bekerja dengan membangun banyak decision tree dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih akurat.

Dalam penelitian ini, *Random Forest* digunakan untuk melengkapi *Linear Regression* dalam menangkap hubungan non-linear antara karakteristik polis dan nilai klaim. Berdasarkan penelitian terdahulu, *Random Forest* terbukti memiliki performa yang tinggi dalam berbagai kasus prediksi klaim asuransi.

Penggunaan *Random Forest* dalam penelitian ini sangat relevan dengan kondisi data PT. Jasaraharja Putera yang memiliki banyak variabel dan potensi hubungan yang kompleks.

### *2.2.8 Feature Importance*

*Feature Importance* merupakan metode yang digunakan untuk mengukur tingkat kontribusi setiap variabel independen terhadap hasil prediksi yang dihasilkan oleh model *Machine Learning*. Pada algoritma *Random Forest*, *feature importance* dihitung berdasarkan seberapa besar suatu fitur berkontribusi dalam mengurangi *impurity* atau kesalahan pada proses pembentukan pohon keputusan.

Nilai *feature importance* digunakan untuk mengetahui variabel mana yang memiliki pengaruh paling besar terhadap target prediksi. Semakin tinggi nilai *feature importance* suatu variabel, maka semakin besar kontribusi variabel tersebut dalam membantu model menghasilkan prediksi yang akurat. Informasi ini dapat digunakan untuk memahami pola data, melakukan seleksi fitur, serta memberikan insight yang lebih mudah dipahami oleh pengguna bisnis.

Dalam penelitian ini, *feature importance* digunakan untuk mengidentifikasi karakteristik polis yang paling berpengaruh terhadap nilai klaim asuransi pada PT Jasaraharja Putera. Hasil analisis tersebut diharapkan dapat membantu perusahaan dalam memahami faktor risiko yang berkontribusi terhadap besarnya nilai klaim serta mendukung proses pengambilan keputusan berbasis data.

### *2.2.9 Hyperparameter Tuning*

*Hyperparameter Tuning* merupakan proses optimisasi parameter model yang dilakukan sebelum proses pelatihan model untuk memperoleh performa prediksi yang optimal. Berbeda dengan parameter model yang dipelajari secara otomatis selama proses *training*, *hyperparameter* harus ditentukan terlebih dahulu oleh peneliti sebelum model dijalankan.

Pada algoritma *Random Forest*, beberapa *hyperparameter* yang umum digunakan antara lain jumlah pohon keputusan (*n_estimators*), kedalaman maksimum pohon (*max_depth*), jumlah minimum sampel pada setiap node (*min_samples_split*), serta jumlah minimum sampel pada setiap daun (*min_samples_leaf*). Pemilihan kombinasi *hyperparameter* yang tepat dapat meningkatkan kemampuan model dalam mempelajari pola data sekaligus mengurangi risiko *overfitting*.

Dalam penelitian ini, proses *hyperparameter tuning* dilakukan untuk memperoleh konfigurasi *Random Forest* yang menghasilkan performa prediksi terbaik berdasarkan metrik evaluasi yang digunakan. Dengan adanya optimisasi *hyperparameter*, model diharapkan mampu menghasilkan prediksi nilai klaim yang lebih akurat dan stabil.

### 2.2.10 *Cross Validation*

*Cross Validation* merupakan teknik evaluasi model yang digunakan untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Teknik ini dilakukan dengan membagi dataset menjadi beberapa subset atau fold, kemudian model dilatih dan diuji secara bergantian menggunakan kombinasi subset tersebut.

Salah satu metode yang paling umum digunakan adalah *K-Fold Cross Validation*. Pada metode ini, dataset dibagi menjadi K bagian dengan ukuran yang relatif sama. Selanjutnya, satu bagian digunakan sebagai data pengujian dan bagian lainnya digunakan sebagai data pelatihan. Proses tersebut diulang hingga seluruh subset pernah digunakan sebagai data pengujian.

Penggunaan *Cross Validation* dapat mengurangi bias evaluasi yang mungkin terjadi akibat pembagian data secara acak serta memberikan gambaran yang lebih representatif mengenai performa model. Dalam penelitian ini, *Cross Validation* digunakan untuk mengevaluasi konsistensi performa model *Linear Regression* dan *Random Forest* sehingga hasil evaluasi yang diperoleh menjadi lebih reliabel.

### 2.2.11 Evaluasi Model

Evaluasi model merupakan tahap untuk mengukur performa model dalam memprediksi nilai klaim. Dalam penelitian ini, digunakan metrik evaluasi seperti *MAE, MSE, RMSE, dan* $R^2$.

Pemilihan metrik ini didasarkan pada penelitian terdahulu yang menunjukkan bahwa metrik tersebut efektif dalam mengukur akurasi model regresi. Evaluasi model juga digunakan untuk membandingkan performa antara *Linear Regression* dan *Random Forest* dalam konteks data PT. Jasaraharja Putera.

Hasil evaluasi ini akan menjadi dasar dalam menentukan model terbaik yang dapat digunakan oleh perusahaan dalam mendukung pengambilan keputusan berbasis data.

Dalam konteks prediksi nilai klaim asuransi, interpretasi metrik evaluasi tidak hanya berfokus pada aspek statistik, tetapi juga dampaknya terhadap proses bisnis perusahaan. Nilai *MAE* dan *RMSE* digunakan untuk mengukur besarnya rata-rata kesalahan prediksi yang dihasilkan model terhadap nilai klaim aktual. Semakin kecil nilai *MAE* dan *RMSE*, maka semakin kecil pula potensi kesalahan estimasi yang dapat memengaruhi perencanaan cadangan klaim, pengelolaan risiko, serta pengambilan keputusan perusahaan.

Sementara itu, nilai $R^2$ digunakan untuk menunjukkan kemampuan model dalam menjelaskan variasi data nilai klaim. Semakin mendekati nilai 1, maka semakin besar proporsi variasi nilai klaim yang dapat dijelaskan oleh model. Dalam industri asuransi, kemampuan model untuk menjelaskan variasi nilai klaim menjadi penting karena dapat membantu perusahaan memahami hubungan antara karakteristik polis dengan risiko klaim yang mungkin terjadi di masa mendatang.

## 2.3 Framework/Algoritma yang digunakan

### 2.3.1 *Linear Regression*

*Linear Regression* merupakan salah satu metode implementasi statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dengan satu variabel dependen[17][18]. Metode ini bertujuan untuk mengetahui bagaimana perubahan pada variabel independen dapat mempengaruhi variabel dependen melalui hubungan *linear.*

Dalam model *Linear Regression*, hubungan antara variable dinyatakan dalam bentuk persamaan matematis sebagai berikut:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \epsilon$$

di mana:

Y = variabel dependen (nilai klaim asuransi)

$X_1, X_2, \ldots, Xn$ = variabel independen (karakteristik polis)

$\beta_0$ = konstanta (intercept)

$\beta_1, \beta_2, \ldots, \beta n$ = koefisien regresi

$\varepsilon$ = error atau residual

Koefisien regresi dalam model ini menunjukkan seberapa besar pengaruh masing-masing variabel independen terhadap variabel dependen. Dengan demikian, model ini tidak hanya dapat digunakan untuk melakukan prediksi, tetapi juga untuk memahami hubungan antar variabel dalam suatu dataset[20][23].

Dalam konteks industri asuransi, *Linear Regression* sering digunakan untuk mengimplementasi hubungan antara faktor-faktor risiko dengan nilai klaim yang terjadi[15][18]. Variabel seperti nilai pertanggungan, premi, masa pertanggungan, serta karakteristik tertanggung dapat diimplementasi untuk mengetahui sejauh mana pengaruhnya terhadap besarnya nilai klaim.

Metode *Linear Regression* dipilih dalam penelitian ini karena memiliki beberapa keunggulan, antara lain:

1. **Mudah diinterpretasikan**
   Koefisien regresi dapat memberikan gambaran langsung mengenai pengaruh masing-masing karakteristik polis terhadap nilai klaim.
2. **Cocok untuk implementasi hubungan antar variabel**
   Model ini sangat efektif digunakan ketika tujuan penelitian adalah untuk mengimplementasi pengaruh variabel independen terhadap variabel dependen.
3. **Sering digunakan dalam penelitian ekonomi dan keuangan**
   Banyak penelitian di bidang keuangan dan asuransi menggunakan regresi linear untuk mengimplementasi faktor-faktor yang mempengaruhi risiko dan kerugian.

Dalam penelitian ini, *Linear Regression* digunakan sebagai model dasar untuk mengimplementasi pengaruh karakteristik polis terhadap nilai klaim asuransi. Hasil dari model ini akan memberikan gambaran mengenai hubungan linear antara variabel-variabel yang diimplementasi[17][23].

### 2.3.2 *Random Forest*

*Random Forest* merupakan salah satu algoritma *Machine Learning* yang termasuk dalam metode ensemble learning[17][19]. Algoritma ini bekerja dengan membangun sejumlah besar *Decision Tree* dan kemudian menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan prediksi akhir yang lebih stabil dan akurat.

Konsep utama dari *Random Forest* adalah menggabungkan beberapa model *Decision Tree* yang dilatih menggunakan subset data yang berbeda. Proses ini dilakukan dengan metode bootstrap sampling, di mana setiap pohon keputusan dilatih menggunakan sampel data yang diambil secara acak dari dataset asli[20].

Setiap *Decision Tree* dalam *Random Forest* akan melakukan proses pemisahan data berdasarkan fitur tertentu hingga menghasilkan prediksi. Hasil prediksi dari seluruh pohon kemudian digabungkan dengan cara menghitung rata-rata nilai prediksi untuk menghasilkan output akhir pada model regresi.

Keunggulan utama dari *Random Forest* antara lain:

1. **Mampu menangani hubungan non-linear** Berbeda dengan regresi linear yang hanya memodelkan hubungan linear, *Random Forest* mampu menangkap hubungan yang lebih kompleks antara variabel.
2. **Lebih tahan terhadap overfitting** Dengan menggunakan banyak *Decision Tree*, model ini dapat menghasilkan prediksi yang lebih stabil dibandingkan satu pohon keputusan tunggal.
3. **Mampu menangani dataset dengan banyak variabel** Algoritma ini dapat bekerja dengan baik meskipun terdapat banyak variabel independen dalam dataset.
4. **Dapat mengukur pentingnya fitur (*feature importance*)** *Random Forest* dapat menunjukkan variabel mana yang paling berpengaruh dalam proses prediksi.

Keunggulan-keunggulan tersebut menyebabkan *Random Forest* banyak digunakan dalam penelitian prediksi klaim asuransi karena mampu menghasilkan performa prediksi yang tinggi pada data dengan karakteristik kompleks dan *non-linear*[15][15][21].

Dalam penelitian ini, *Random Forest* digunakan sebagai model pembanding terhadap *Linear Regression*. Penggunaan algoritma ini bertujuan untuk mengetahui apakah pendekatan berbasis *Machine Learning* mampu menghasilkan prediksi nilai klaim yang lebih akurat dibandingkan metode statistik tradisional[15][17].

Selain itu, karakteristik polis dalam data asuransi sering kali memiliki hubungan yang kompleks dan tidak selalu linear. Oleh karena itu, penggunaan *Random Forest* diharapkan dapat menangkap pola hubungan yang lebih kompleks antara karakteristik polis dan nilai klaim asuransi[15][20].

### 2.3.3 Kombinasi Framework dan Algoritma

Penelitian ini menggunakan dua pendekatan model yaitu *Linear Regression* dan *Random Forest* untuk mengimplementasi pengaruh karakteristik polis terhadap nilai klaim asuransi[15][17]. Penggunaan dua model ini bertujuan untuk melakukan perbandingan performa model dalam memprediksi nilai klaim berdasarkan variabel-variabel yang tersedia dalam dataset.

*Linear Regression* digunakan sebagai model dasar karena memiliki interpretasi yang jelas terhadap hubungan antara variabel independen dan variabel dependen. Model ini dapat memberikan gambaran mengenai seberapa besar pengaruh masing-masing karakteristik polis terhadap nilai klaim yang terjadi.

Sementara itu, *Random Forest* digunakan untuk menangkap pola hubungan yang lebih kompleks yang mungkin tidak dapat dijelaskan secara optimal oleh model regresi linear. Algoritma ini mampu memodelkan hubungan non-linear serta mengurangi risiko overfitting melalui pendekatan ensemble learning.

Dengan menggunakan kedua model tersebut, penelitian ini dapat melakukan evaluasi terhadap performa model menggunakan beberapa metrik evaluasi seperti *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, dan *Root Mean Squared Error (RMSE*)[10][11]. Hasil evaluasi ini akan menunjukkan model mana yang memiliki tingkat akurasi prediksi yang lebih baik dalam mengimplementasi nilai klaim asuransi.

Pendekatan ini diharapkan dapat menghasilkan model prediksi yang lebih optimal serta memberikan pemahaman yang lebih mendalam mengenai pengaruh karakteristik polis terhadap nilai klaim asuransi pada PT Jasaraharja Putera[15][18][20].

### 2.3.4 Alasan Pemilihan Algoritma

Pada penelitian ini dipilih dua algoritma, yaitu *Linear Regression* dan *Random Forest*, sebagai metode utama untuk memprediksi nilai klaim asuransi berdasarkan karakteristik polis. Pemilihan kedua algoritma tersebut dilakukan dengan mempertimbangkan karakteristik data penelitian, tujuan penelitian, serta hasil penelitian terdahulu yang menunjukkan bahwa kedua metode memiliki keunggulan yang saling melengkapi dalam permasalahan regresi.

*Linear Regression* dipilih sebagai model dasar (*baseline model*) karena merupakan salah satu metode regresi yang paling umum digunakan dalam analisis statistik maupun *machine learning*. Algoritma ini memiliki kemampuan untuk memodelkan hubungan

linear antara variabel independen dan variabel dependen sehingga hasil model lebih mudah diinterpretasikan. Selain menghasilkan prediksi, *Linear Regression* juga memungkinkan peneliti melakukan analisis terhadap pengaruh masing-masing variabel independen melalui nilai koefisien regresi, uji signifikansi statistik, serta pengujian asumsi regresi linear. Oleh karena itu, penggunaan *Linear Regression* pada penelitian ini tidak hanya bertujuan memperoleh hasil prediksi, tetapi juga untuk mengetahui apakah karakteristik polis memiliki hubungan yang signifikan terhadap nilai klaim asuransi.

Selain memiliki tingkat interpretabilitas yang tinggi, *Linear Regression* juga memiliki proses komputasi yang relatif sederhana sehingga dapat digunakan sebagai acuan awal dalam mengevaluasi kemampuan prediksi dataset. Apabila performa *Linear Regression* masih rendah, maka dapat disimpulkan bahwa hubungan antara variabel independen dan variabel dependen tidak sepenuhnya bersifat linear sehingga diperlukan algoritma yang mampu menangkap hubungan yang lebih kompleks.

Sebagai pembanding, penelitian ini menggunakan algoritma *Random Forest*. *Random Forest* merupakan metode *ensemble learning* berbasis kumpulan decision tree yang dibangun menggunakan teknik *bootstrap sampling* dan *random feature selection*. Berbeda dengan *Linear Regression* yang mengasumsikan hubungan linear antarvariabel, *Random Forest* mampu mempelajari hubungan non-linear, interaksi antarvariabel, serta pola-pola kompleks yang sering ditemukan pada data dunia nyata, termasuk data klaim asuransi. Selain menghasilkan performa prediksi yang tinggi, *Random Forest* juga menyediakan informasi mengenai tingkat kepentingan setiap variabel (*feature importance*), sehingga dapat digunakan untuk mengidentifikasi karakteristik polis yang paling berpengaruh terhadap nilai klaim.

Pemilihan *Random Forest* juga didasarkan pada karakteristik dataset penelitian. Data klaim asuransi memiliki variasi nilai yang cukup tinggi, distribusi yang tidak selalu normal, serta kemungkinan terdapat hubungan *non-linear* antara nilai pertanggungan, premi, jumlah klaim, durasi polis, dan nilai klaim yang dibayarkan. Kondisi tersebut menyebabkan algoritma berbasis *decision tree* lebih mampu menangkap pola hubungan dibandingkan metode regresi *linear* konvensional. Selain itu, *Random Forest* relatif lebih tahan terhadap *noise*, *outlier*, serta risiko *overfitting* dibandingkan *decision tree* tunggal karena proses pembelajaran dilakukan menggunakan banyak pohon keputusan yang kemudian digabungkan melalui mekanisme agregasi.

Berdasarkan kajian penelitian terdahulu, algoritma berbasis *boosting* seperti *XGBoost* memang banyak digunakan pada penelitian prediksi klaim asuransi karena memiliki kemampuan menghasilkan akurasi yang tinggi. *XGBoost* bekerja dengan membangun model secara bertahap (*sequential boosting*) sehingga setiap pohon baru berfungsi memperbaiki kesalahan prediksi pada pohon sebelumnya. Pendekatan tersebut terbukti efektif dalam menangani hubungan data yang kompleks dan sering menghasilkan performa yang lebih baik dibandingkan algoritma regresi konvensional.

Meskipun demikian, penelitian ini tidak menggunakan *XGBoost* sebagai algoritma utama karena tujuan penelitian tidak hanya berfokus pada pencapaian akurasi prediksi tertinggi, tetapi juga pada analisis hubungan karakteristik polis terhadap nilai klaim serta interpretasi hasil model. *Linear Regression* dipilih untuk memberikan dasar analisis statistik yang mudah dipahami, sedangkan *Random Forest* dipilih sebagai model machine learning yang mampu menangkap hubungan *non-linear* sekaligus memberikan informasi mengenai tingkat kepentingan setiap variabel. Dengan kombinasi kedua algoritma tersebut, penelitian dapat membandingkan pendekatan statistik klasik dengan pendekatan machine learning modern secara lebih komprehensif.

Selain itu, penggunaan dua algoritma dengan karakteristik yang berbeda memungkinkan proses evaluasi dilakukan secara lebih objektif. *Linear Regression* digunakan sebagai model dasar (*baseline*) untuk mengetahui kemampuan prediksi apabila hubungan antarvariabel diasumsikan *linear*. Selanjutnya, *Random Forest* digunakan untuk mengevaluasi apakah pendekatan *non-linear* mampu meningkatkan performa prediksi terhadap dataset penelitian. Perbandingan kedua model dilakukan menggunakan metrik evaluasi yang sama, yaitu *Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* dan *Coefficient of Determination* ($R^2$), sehingga hasil evaluasi dapat dibandingkan secara adil dan konsisten.

Dengan demikian, pemilihan *Linear Regression* dan *Random Forest* pada penelitian ini tidak hanya didasarkan pada popularitas algoritma, tetapi juga mempertimbangkan karakteristik data, tujuan penelitian, tingkat interpretabilitas model, kemampuan menangkap pola hubungan data, serta hasil penelitian terdahulu yang relevan. Pendekatan tersebut diharapkan mampu menghasilkan model prediksi yang akurat sekaligus memberikan informasi yang dapat mendukung proses pengambilan keputusan pada pengelolaan klaim asuransi.

Tabel 2.2 Perbandingan Algoritma

| Algoritma | Kelebihan | Kekurangan |
|---|---|---|
| Linear Regression | Mudah diinterpretasi | Hubungan linear |
| Random Forest | Akurat dan non-linear | Lebih kompleks |
| XGBoost | Akurasi tinggi | Parameter kompleks |
| ANN | Menangkap pola rumit | Sulit dijelaskan |

## 2.4 Tools/software yang digunakan

Dalam penelitian ini, beberapa perangkat lunak digunakan untuk mendukung implementasi data dan pemodelan prediktif. Alat-alat tersebut meliputi Python (melalui Jupyter Notebook), Streamlit, dan Google Colab. Berikut ini adalah deskripsi dan fungsi dari masing-masing alat yang digunakan:

### 2.4.1 Phyton dengan Jupyter Notebook

*Python* merupakan salah satu bahasa pemrograman yang paling populer dan banyak digunakan dalam bidang *data science, Machine Learning, artificial intelligence*, serta pengembangan perangkat lunak. *Python* memiliki sintaks yang sederhana dan didukung oleh berbagai pustaka seperti *Pandas, NumPy, Scikit-learn*, dan *Matplotlib* yang membantu proses analisis data dan *Machine Learning*.[24]

Sementara itu, Jupyter Notebook merupakan aplikasi berbasis web yang digunakan untuk menulis dan menjalankan kode *Python* secara interaktif dalam bentuk notebook. *Jupyter Notebook* memungkinkan pengguna menggabungkan kode program, visualisasi data, serta dokumentasi dalam satu dokumen yang terintegrasi sehingga sangat sesuai digunakan dalam penelitian berbasis *data science* dan *Machine Learning*.

Dalam penelitian ini, *Python* dan *Jupyter Notebook* digunakan sebagai tools utama dalam proses *preprocessing data, exploratory data analysis (EDA)*, pembangunan model *Linear Regression* dan *Random Forest*, evaluasi model, serta visualisasi hasil prediksi nilai klaim asuransi.

Keunggulan Jupyter Notebook:

a. Interaktif: Memungkinkan eksekusi kode secara langsung, sehingga pengguna dapat melihat hasil secara instan, baik berupa teks, grafik, atau visualisasi data.
b. Dokumentasi dan Visualisasi: Jupyter memungkinkan penulisan dokumentasi dalam bentuk teks (markdown), memudahkan pengguna untuk memberikan penjelasan terkait implementasi atau eksperimen yang sedang dilakukan. Selain itu, Jupyter juga mendukung visualisasi

data menggunakan berbagai pustaka, seperti Matplotlib, Seaborn, atau Plotly.

c. Reproduksi dan Kolaborasi: Jupyter Notebook dapat disimpan dan dibagikan dalam format .ipynb, memungkinkan kolaborasi antar tim atau pembagian hasil eksperimen dengan mudah.
d. Kompatibilitas: Dapat digunakan untuk berbagai macam bahasa pemrograman selain Python, seperti R, Julia, dan Scala. Namun, Python adalah bahasa yang paling umum digunakan dalam Jupyter Notebook.
e. Integrasi dengan Pustaka Data Science: Jupyter Notebook memudahkan integrasi dengan pustaka- pustaka penting di bidang data science, seperti Pandas untuk manipulasi data, Scikit-learn untuk pembelajaran mesin, serta TensorFlow dan Keras untuk deep learning

### 2.4.2 Google Collab

Google Colab (Collaboratory) adalah platform berbasis cloud yang memungkinkan pengguna untuk menulis dan menjalankan kode Python secara interaktif dalam notebook yang terhubung langsung ke Google Drive. Platform ini dikembangkan oleh Google dan menyediakan lingkungan pemrograman Python yang memungkinkan eksekusi kode tanpa perlu mengatur instalasi perangkat keras atau perangkat lunak secara lokal.[25]

Google Colab memiliki banyak kemiripan dengan Jupyter Notebook, tetapi dengan keuntungan tambahan karena semuanya berjalan di cloud dan terintegrasi dengan layanan Google lainnya.

Ini memudahkan kolaborasi, berbagi notebook, dan akses ke sumber daya komputasi yang lebih kuat.

Keunggulan Google Colab:

a. Akses ke Komputasi Gratis:
   Google Colab menawarkan akses gratis ke GPU dan TPU (Tensor Processing Unit), yang sangat berguna dalam pelatihan model deep learning yang membutuhkan sumber daya komputasi besar. Ini memungkinkan pengguna untuk menjalankan model deep learning dan eksperimen yang lebih kompleks tanpa memerlukan perangkat keras mahal.
b. Berbasis Cloud:
   Karena berbasis cloud, semua proyek yang dikerjakan di Google Colab dapat diakses dari mana saja dan kapan saja selama ada koneksi internet. Kolaborasi menjadi lebih mudah karena notebook dapat dibagikan kepada orang lain

melalui Google Drive, yang memungkinkan beberapa orang untuk bekerja bersama dalam satu dokumen secara simultan.

c. Integrasi dengan Google Drive:
   Google Colab terintegrasi langsung dengan Google Drive, yang memungkinkan pengguna untuk menyimpan, mengakses, dan berbagi notebook mereka secara mudah. Pengguna dapat menyimpan hasil pekerjaan mereka di Google Drive, yang dapat diakses dari perangkat manapun dan diunduh ke format .ipynb (Jupyter Notebook).
d. Dukungan untuk Pustaka Data Science:
   Google Colab mendukung banyak pustaka Python untuk data science dan pembelajaran mesin, seperti Pandas, NumPy, Matplotlib, Scikit-learn, TensorFlow, Keras, dan lainnya. Pustaka-pustaka ini sudah terinstal secara default, sehingga pengguna tidak perlu menginstalnya secara manual.
e. Kemudahan Berbagi dan Kolaborasi:
   Seperti dokumen Google lainnya, notebook di Google Colab dapat dengan mudah dibagikan dengan rekan kerja, dosen, atau tim. Pengguna dapat mengatur izin akses (misalnya, hanya tampilan atau dapat mengedit). Ini memudahkan kolaborasi pada proyek data science atau pembelajaran mesin, di mana banyak orang dapat memberikan kontribusi secara bersamaan.
f. No Setup Required:
   Google Colab memungkinkan pengguna untuk langsung menulis dan menjalankan kode tanpa memerlukan konfigurasi atau pengaturan lingkungan pengembangan lokal. Ini sangat memudahkan pemula atau pengguna yang tidak ingin repot mengatur Python di komputer mereka.
g. Penggunaan GPU/TPU untuk Deep Learning:
   Untuk tugas yang melibatkan deep learning, Google Colab memungkinkan penggunaan GPU dan TPU secara gratis, yang mempercepat proses pelatihan model-model besar. Ini memberi keuntungan besar bagi para peneliti atau pengembang yang bekerja dengan dataset besar atau model neural network yang kompleks.

### 2.4.3 Streamlit

Streamlit merupakan framework open-source berbasis Python yang digunakan untuk membangun aplikasi web interaktif secara cepat, khususnya pada bidang data science, machine learning, dan visualisasi data. Streamlit memungkinkan pengguna mengubah script Python menjadi aplikasi berbasis web tanpa memerlukan pengembangan frontend yang kompleks seperti HTML, CSS, maupun JavaScript. Framework ini banyak digunakan dalam pengembangan prototype machine learning, dashboard analitik, serta deployment model prediksi

karena memiliki integrasi yang sederhana dengan berbagai library Python seperti Pandas, NumPy, Scikit-learn, dan Matplotlib [26].

Pada penelitian ini, Streamlit digunakan sebagai media deployment model Random Forest untuk melakukan simulasi prediksi nilai klaim asuransi berdasarkan karakteristik polis. Dengan menggunakan Streamlit, model machine learning yang telah dibangun dapat diimplementasikan ke dalam aplikasi berbasis web sehingga pengguna dapat melakukan prediksi secara interaktif melalui antarmuka sederhana.

Keunggulan Streamlit:

a. Kemudahan Implementasi
Streamlit memungkinkan pengembang membuat aplikasi berbasis web hanya menggunakan bahasa pemrograman Python tanpa memerlukan pengembangan frontend yang kompleks. Hal ini membuat proses deployment model machine learning menjadi lebih cepat dan mudah untuk diimplementasikan.

b. Integrasi dengan Library Machine Learning
Streamlit memiliki kompatibilitas yang baik dengan berbagai library machine learning seperti Scikit-learn, TensorFlow, Pandas, dan NumPy sehingga memudahkan proses implementasi model prediksi berbasis Python.

c. Tampilan Interaktif
Streamlit menyediakan berbagai komponen interaktif seperti input box, slider, button, dan grafik visual yang memungkinkan pengguna melakukan simulasi prediksi secara langsung melalui aplikasi berbasis web.

d. Deployment Cepat dan Ringan
Aplikasi Streamlit dapat dijalankan secara lokal maupun dihosting secara online dengan proses deployment yang relatif sederhana. Hal ini membuat Streamlit banyak digunakan dalam pengembangan prototype machine learning dan dashboard analitik.

e. Mendukung Visualisasi Data
Streamlit mendukung integrasi dengan berbagai library visualisasi data seperti Matplotlib dan Plotly sehingga hasil analisis dan prediksi dapat ditampilkan secara lebih informatif dan mudah dipahami.

f. Mendukung Implementasi Machine Learning
Streamlit mampu menghubungkan model machine learning dengan antarmuka pengguna secara langsung sehingga hasil prediksi dapat digunakan secara interaktif dan real-time pada aplikasi berbasis web.