BAB 2
LANDASAN TEORI

2.1 Justifikasi Solusi

Justifikasi solusi disusun berdasarkan kebutuhan penelitian dalam melakukan prediksi laba pada peternakan ayam berbasis data operasional dan keuangan menggunakan algoritma *Random Forest Regression*. Setiap metode dan teknologi yang digunakan dievaluasi berdasarkan kesesuaian terhadap karakteristik dataset, kompleksitas hubungan antar variabel, serta kebutuhan prediksi pada sistem peternakan ayam.

2.2 Peternakan Ayam

Peternakan ayam merupakan salah satu sektor agribisnis yang memiliki peran penting dalam memenuhi kebutuhan protein hewani. Terdapat dua jenis utama peternakan ayam, yaitu ayam broiler yang difokuskan pada produksi daging dan ayam layer yang difokuskan pada produksi telur.

Keberhasilan dalam peternakan ayam dipengaruhi oleh beberapa faktor utama, seperti kualitas pakan, manajemen kandang, kesehatan ayam, serta kondisi lingkungan. Pengelolaan yang tidak optimal dapat menyebabkan penurunan produktivitas dan peningkatan tingkat mortalitas.

Keberhasilan produksi ayam broiler pada satu periode pemeliharaan umumnya diukur melalui beberapa indikator performa, di antaranya *Feed Conversion Ratio* (FCR) dan Indeks Performa (IP). FCR menggambarkan efisiensi konversi pakan menjadi bobot tubuh, dihitung dari total pakan yang dikonsumsi (kg) dibagi dengan total bobot ayam yang dihasilkan (kg). IP merupakan indikator komposit yang menggabungkan tingkat kelangsungan hidup (*livability*), bobot rata-rata panen, FCR, dan umur panen ke dalam satu nilai, sebagaimana dirumuskan berikut:

$$IP = \frac{Livability\ (\%) \times Bobot\ Rata{-}rata\ (kg) \times 100}{Umur\ Panen\ (hari) \times FCR} \tag{2.1}$$

Semakin tinggi nilai IP, semakin baik performa pemeliharaan yang dicapai, karena mengindikasikan tingkat kematian yang rendah, efisiensi pakan yang baik, serta pertumbuhan bobot yang optimal dalam waktu pemeliharaan yang relatif

singkat [7].

2.3 Prediksi Laba

Laba merupakan selisih antara pendapatan dan biaya yang dikeluarkan dalam suatu kegiatan usaha. Dalam peternakan ayam, laba dipengaruhi oleh berbagai faktor, seperti biaya pakan, biaya operasional, harga jual ayam, serta tingkat produksi.

Berdasarkan penelitian terdahulu, sejumlah faktor produksi terbukti secara signifikan mempengaruhi keuntungan usaha tani ternak ayam broiler, di antaranya harga bibit (DOC), *Feed Conversion Ratio* (FCR), biaya tenaga kerja, biaya pemanas, harga jual, dan tingkat mortalitas [8]. Faktor-faktor tersebut secara individu memberikan kontribusi terhadap variasi keuntungan yang diperoleh peternak pada satu periode produksi, sehingga relevan untuk dipertimbangkan sebagai variabel dalam pemodelan prediksi laba.

Prediksi laba menjadi penting untuk membantu pengambilan keputusan bisnis. Dengan menggunakan pendekatan machine learning, hubungan kompleks antar variabel dapat dianalisis sehingga menghasilkan estimasi laba yang lebih akurat.

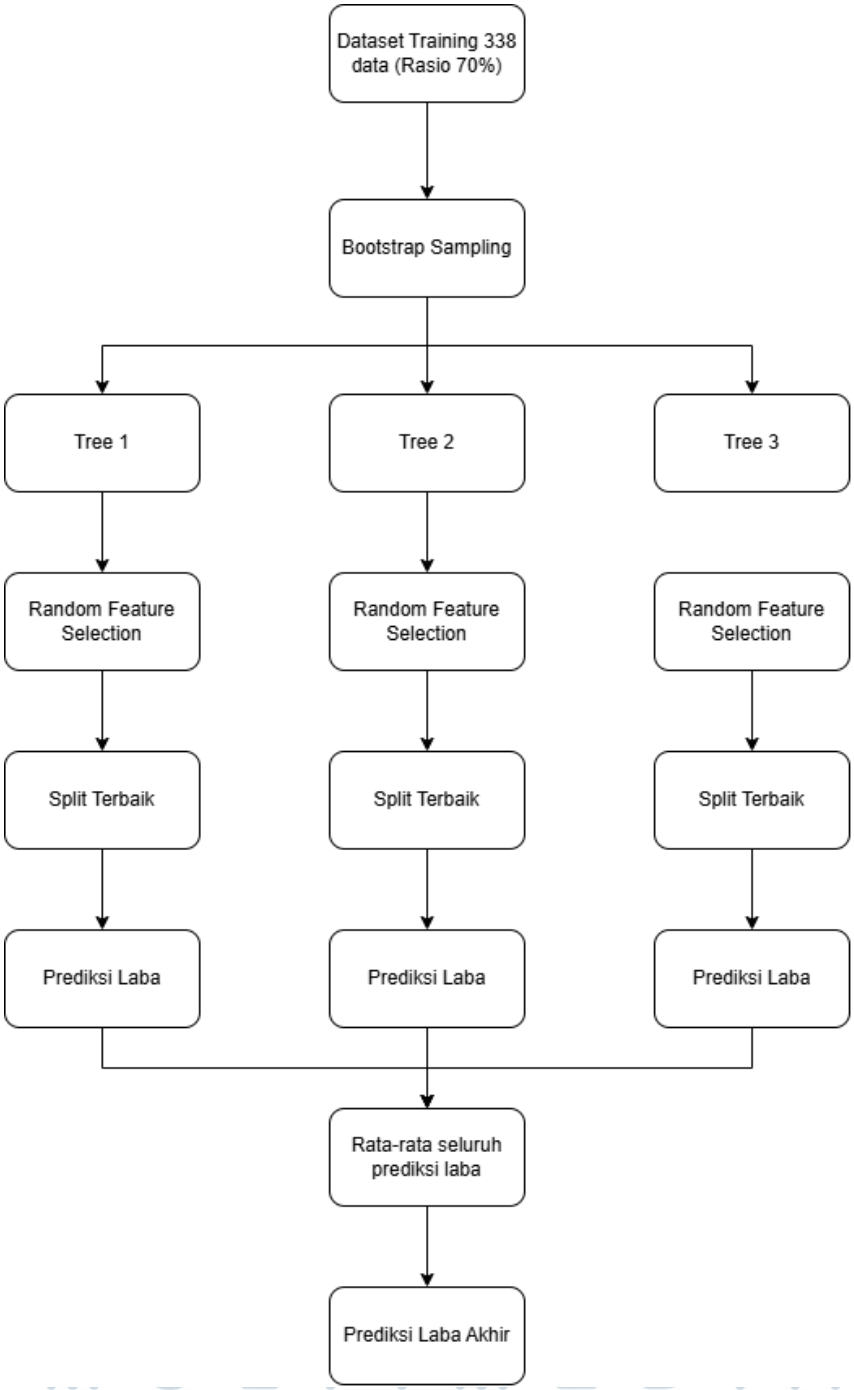
2.4 Algoritma *Random Forest Regression*

Random Forest Regression merupakan algoritma *ensemble* yang membangun banyak pohon keputusan dan menggabungkan hasil prediksinya sehingga menghasilkan model yang lebih stabil dibandingkan *decision tree* tunggal [9].

Penelitian ini menggunakan algoritma *Random Forest Regression* sebagai metode utama dalam melakukan prediksi laba. *Random Forest Regression* dipilih karena mampu menangani hubungan non-linear antar variabel serta memiliki tingkat akurasi yang baik pada data numerik dan tabular[5].

Secara umum, *Random Forest Regression* dibangun melalui beberapa tahapan utama, yaitu pembentukan sampel menggunakan teknik *bootstrap sampling*, pembangunan sejumlah *decision tree* secara independen, pemilihan subset fitur secara acak pada setiap proses pemisahan (*random feature selection*), serta penggabungan hasil prediksi seluruh pohon untuk menghasilkan prediksi akhir. Arsitektur kerja *Random Forest Regression* tersebut dapat dilihat pada

Gambar 2.1 [5, 10].



Gambar 2.1. Arsitektur *Random Forest Regression*

Random Forest Regression bekerja dengan membangun sejumlah *decision tree* secara independen menggunakan teknik *bootstrap aggregating* (bagging), yaitu setiap pohon dilatih pada sampel data yang diambil secara acak dengan

pengembalian (*resampling with replacement*) dari data training. Selain itu, pada setiap pemisahan (*split*) node, hanya subset acak dari fitur yang dipertimbangkan, sehingga menurunkan korelasi antar pohon dan meningkatkan keragaman model [5]. Untuk kasus regresi, prediksi akhir diperoleh dari rata-rata hasil prediksi seluruh pohon dalam *ensemble*, dirumuskan sebagai:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{2.2}$$

dengan B adalah jumlah pohon dalam forest dan $T_b(x)$ adalah hasil prediksi dari pohon ke-b terhadap input x [5].

Selain itu, *Random Forest Regression* memiliki kemampuan dalam mengurangi risiko *overfitting* melalui metode *ensemble learning* berbasis *decision tree* [10].

2.5 Preprocessing Data

Preprocessing data merupakan tahapan awal dalam proses *machine learning* yang bertujuan untuk meningkatkan kualitas data sebelum digunakan pada proses pelatihan model. Tahapan ini dilakukan untuk mengatasi permasalahan seperti data yang tidak lengkap, inkonsistensi format, nilai ekstrem, maupun atribut yang kurang relevan sehingga dataset menjadi lebih representatif dan sesuai untuk proses pemodelan. Kualitas data yang baik akan membantu algoritma dalam mempelajari pola hubungan antar variabel secara lebih efektif sehingga dapat meningkatkan performa prediksi [10, 11].

Tahapan *preprocessing data* umumnya meliputi beberapa proses sebagai berikut.

1. *Data Cleaning*

   Data cleaning merupakan proses untuk meningkatkan kualitas data dengan mengidentifikasi dan menangani *missing value*, data duplikat, inkonsistensi format, serta nilai yang tidak valid. Tahapan ini bertujuan agar data yang digunakan lebih konsisten dan dapat diproses oleh algoritma *machine learning*. Proses *data cleaning* juga dapat mencakup penanganan *outlier* maupun transformasi format data sehingga menghasilkan dataset yang lebih reliabel [12, 13].

2. *Feature Engineering*

Feature engineering merupakan proses membentuk atribut baru dari atribut yang telah tersedia dengan tujuan menghasilkan representasi data yang lebih informatif. Penambahan fitur yang relevan dapat membantu model menangkap hubungan yang sebelumnya tidak terlihat pada data asli sehingga berpotensi meningkatkan performa prediksi [11].

3. *Feature Selection*

   Feature selection merupakan teknik untuk memilih atribut yang paling relevan terhadap variabel target sehingga hanya informasi yang penting digunakan dalam proses pelatihan model. Pengurangan atribut yang kurang relevan dapat meningkatkan efisiensi komputasi, mengurangi kompleksitas model, serta membantu meningkatkan kemampuan generalisasi model terhadap data baru [9].

4. *Normalization*

   Normalization merupakan teknik untuk menyamakan skala antar fitur sehingga algoritma yang bergantung pada perhitungan jarak atau proses optimasi berbasis gradien dapat bekerja secara optimal. Namun, proses ini tidak selalu diperlukan pada seluruh algoritma. Algoritma berbasis *decision tree*, termasuk *Random Forest Regression*, melakukan pemisahan data berdasarkan nilai ambang (*threshold*) setiap atribut, bukan berdasarkan jarak antar data. Oleh karena itu, perbedaan skala antar fitur tidak memberikan pengaruh yang signifikan terhadap proses pembentukan pohon keputusan maupun hasil prediksi [5, 13].

5. *Encoding Data*

   Encoding merupakan proses mengubah data kategorikal menjadi representasi numerik agar dapat diproses oleh algoritma *machine learning*. Salah satu teknik yang paling umum digunakan adalah *One-Hot Encoding*, yaitu metode yang merepresentasikan setiap kategori ke dalam variabel biner bernilai 0 atau 1 tanpa membentuk hubungan ordinal antar kategori. Teknik ini banyak digunakan pada atribut nominal karena mampu mempertahankan informasi kategori tanpa memberikan urutan yang tidak semestinya [5, 13].

Secara umum, tahapan *preprocessing data* bertujuan menghasilkan dataset yang memiliki kualitas lebih baik sehingga proses pelatihan model dapat dilakukan secara lebih efektif. Penerapan tahapan *preprocessing* yang tepat berkontribusi

terhadap peningkatan akurasi, stabilitas, dan kemampuan generalisasi model *machine learning* [10, 11].

2.6 Data Leakage

Data leakage merupakan kondisi ketika informasi yang seharusnya tidak tersedia pada saat prediksi dilakukan, secara tidak sengaja ikut digunakan sebagai fitur input pada proses pelatihan model. Kondisi ini menyebabkan performa model terlihat sangat tinggi pada tahap evaluasi, namun tidak dapat direplikasi pada kondisi nyata karena model sesungguhnya tidak mempelajari hubungan yang valid, melainkan memanfaatkan informasi yang bocor dari variabel target [6].

Leakage juga dapat terjadi secara tidak langsung melalui variabel turunan, misalnya rasio yang dibentuk dari komponen pendapatan dan dibagi dengan variabel lain yang sudah menjadi fitur input, sehingga model dapat menyusun ulang nilai target melalui operasi aritmetika sederhana [6]. Oleh karena itu, setiap variabel turunan dalam penelitian ini diperiksa secara matematis sebelum digunakan sebagai fitur, untuk memastikan tidak terdapat hubungan langsung dengan variabel *laba*.

2.7 Split Data Training dan Testing

Split data training dan testing merupakan salah satu tahapan penting dalam proses pembangunan model *machine learning*. Tahapan ini bertujuan untuk membagi dataset menjadi dua bagian, yaitu data *training* yang digunakan untuk melatih model dan data *testing* yang digunakan untuk mengevaluasi performa model terhadap data yang belum pernah digunakan selama proses pelatihan. Dengan demikian, hasil evaluasi dapat menggambarkan kemampuan model dalam melakukan generalisasi terhadap data baru [10].

Proses pembagian data umumnya dilakukan secara acak (*random sampling*) untuk mengurangi potensi bias akibat urutan data. Pada penelitian yang menggunakan data tabular, pembagian data banyak dilakukan menggunakan fungsi *train_test_split* yang tersedia pada pustaka *Scikit-Learn*. Pendekatan ini bertujuan untuk memastikan bahwa data pelatihan dan data pengujian berasal dari distribusi yang sama sehingga hasil evaluasi model dapat dipercaya dan digunakan sebagai dasar dalam menilai performa algoritma [13].

Proporsi pembagian data training dan testing dapat di sesuaikan dengan karakteristik dataset dan tujuan penelitian , yaitu 60:40, 70:30, dan 80:20.

Ketiga skenario tersebut dibandingkan untuk mengetahui rasio pembagian data yang menghasilkan performa terbaik pada algoritma *Random Forest Regression*. Pendekatan ini dilakukan karena tidak terdapat satu rasio pembagian data yang selalu optimal untuk seluruh karakteristik dataset, sehingga diperlukan evaluasi pada beberapa skenario pembagian data [11].

2.8 Hyperparameter Tuning

Hyperparameter tuning dilakukan untuk memperoleh kombinasi parameter model yang optimal, seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), serta jumlah fitur yang dipertimbangkan pada setiap *split* (*max_features*). Penelitian ini menggunakan dua pendekatan, yaitu *Grid Search Cross-Validation* yang melakukan pencarian secara menyeluruh terhadap seluruh kombinasi parameter pada ruang pencarian yang telah ditentukan, dan *Randomized Search Cross-Validation* yang melakukan pencarian secara acak pada ruang parameter yang lebih luas dengan jumlah iterasi terbatas. Pendekatan pencarian acak terbukti secara empiris dapat menemukan kombinasi parameter yang setara atau lebih baik dibandingkan grid search, dengan waktu komputasi yang lebih efisien [14].

A GridSearchCV

GridSearchCV merupakan metode pencarian *hyperparameter* secara *exhaustive search*, yaitu dengan mencoba seluruh kombinasi nilai parameter yang telah ditentukan sebelumnya. Setiap kombinasi parameter dievaluasi menggunakan teknik *cross-validation*, kemudian kombinasi yang menghasilkan nilai evaluasi terbaik dipilih sebagai model akhir. Kelebihan *GridSearchCV* adalah mampu menemukan kombinasi parameter terbaik pada ruang pencarian yang telah ditentukan, namun waktu komputasi akan meningkat secara signifikan seiring bertambahnya jumlah kombinasi parameter yang diuji [13].

B RandomizedSearchCV

RandomizedSearchCV merupakan metode optimasi *hyperparameter* yang melakukan pencarian berdasarkan sejumlah kombinasi parameter yang dipilih secara acak dari distribusi parameter yang telah ditentukan. Berbeda dengan *GridSearchCV* yang mengevaluasi seluruh kombinasi parameter,

RandomizedSearchCV hanya menguji sejumlah kombinasi tertentu sesuai jumlah iterasi yang ditentukan. Pendekatan ini mampu mengurangi waktu komputasi secara signifikan, terutama ketika ruang pencarian *hyperparameter* sangat besar, namun tetap menghasilkan performa model yang kompetitif [14, 13].

2.9 Evaluasi Model

Evaluasi model dilakukan untuk mengetahui tingkat akurasi prediksi laba yang dihasilkan oleh algoritma *Random Forest Regression*. Pada penelitian ini digunakan tiga metrik evaluasi, yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error*(RMSE), dan *Coefficient of Determination*(R^2). Ketiga metrik tersebut dipilih karena mampu memberikan gambaran mengenai besar kesalahan prediksi serta kemampuan model dalam menjelaskan variasi data aktual.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) merupakan metrik yang mengukur rata-rata selisih absolut antara nilai aktual dengan nilai hasil prediksi. Nilai MAE yang semakin kecil menunjukkan bahwa hasil prediksi model semakin mendekati nilai aktual. Metrik ini mudah diinterpretasikan karena memiliki satuan yang sama dengan variabel target [5]. MAE dirumuskan sebagai berikut:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.3}$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) mengukur akar dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Berbeda dengan MAE, RMSE memberikan penalti yang lebih besar terhadap kesalahan prediksi yang besar sehingga lebih sensitif terhadap keberadaan *outlier*. Nilai RMSE yang lebih kecil menunjukkan performa model yang lebih baik [13]. Persamaan RMSE ditunjukkan sebagai berikut:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.4}$$

Coefficient of Determination (R^2)

Coefficient of Determination(R^2) merupakan metrik yang digunakan untuk mengukur kemampuan model dalam menjelaskan variasi data aktual. Nilai R^2 berada pada rentang 0 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan bahwa model semakin mampu menjelaskan hubungan antara variabel input dan target. Sebaliknya, nilai yang mendekati 0 menunjukkan bahwa kemampuan model dalam menjelaskan variasi data masih rendah [13]. Persamaan R^2 ditunjukkan sebagai berikut:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.5}$$

dengan:

- y_i = nilai laba aktual,
- $\hat{y}_i$ = nilai laba hasil prediksi,
- $\bar{y}$ = rata-rata nilai laba aktual,
- n = jumlah data pengujian.

Pada penelitian ini, MAE dan RMSE digunakan untuk mengukur besarnya kesalahan prediksi model, sedangkan R^2 digunakan sebagai metrik utama untuk mengevaluasi kemampuan model dalam menjelaskan variasi laba berdasarkan variabel operasional yang digunakan. Model yang memiliki nilai MAE dan RMSE lebih kecil serta nilai R^2 yang lebih tinggi dianggap memiliki performa prediksi yang lebih baik.

Tabel 2.1. Penelitian Terdahulu

No	Judul Penelitian	Penulis	Hasil Penelitian	Kebaruan
1	Deep Learning in Agriculture: A Survey	Kamilaris dan Prenafeta-Boldú [15]	Penelitian ini membahas penerapan deep learning di berbagai bidang pertanian (deteksi penyakit tanaman, klasifikasi gambar) secara umum.	Penelitian ini tidak membahas prediksi laba, tidak menggunakan data keuangan per siklus produksi, dan tidak menerapkan validasi *data leakage* pada variabel finansial.
2	Big Data Analytics in Animal Agriculture: A Review	Morota et al. [2]	Penelitian ini mengulas pemanfaatan big data untuk prediksi genetik dan efisiensi produksi ternak secara umum.	Penelitian ini tidak membangun model prediksi laba, tidak menggunakan data operasional dan keuangan per batch panen, dan tidak membahas feature engineering berbasis indikator performa industri peternakan.
3	Smart Poultry Management: Smart Sensors, Big Data, and the Internet of Things	Jake a Still. [4]	Penelitian yang relevan di bidang manajemen peternakan unggas berbasis data umumnya berfokus pada prediksi produksi (bobot, mortalitas, konsumsi pakan) menggunakan data sensor	Penelitian ini belum melalukan prediksi laba berbasis data keuangan historis. Integrasi data operasional dan keuangan per siklus panen sebagai satu model prediktif belum dibahas.