

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Dalam kehidupan sehari-hari, manusia tidak akan terlepas dari data. Bahkan sejak lahir, manusia sudah menghasilkan data seperti nama, tanggal lahir, jam lahir, berat badan, dan sebagainya. Sagioglu & Sinanc [1] menyatakan bahwa jumlah data yang telah dihasilkan manusia pada tahun 2003 adalah sebesar 5 *exabit* ( $10^{18}$  bit) sedangkan pada tahun 2013, jumlah data ini dapat dihasilkan hanya dalam waktu dua hari. Fenomena data yang melimpah ini dikenal sebagai *big data*, yaitu sekumpulan data terstruktur maupun tidak terstruktur yang volumenya sangat besar. Pengolahan *big data* akan sangat sulit jika menggunakan metode yang tradisional, sehingga dapat digunakan aplikasi dari *artificial intelligence* yaitu *machine learning* (ML).

Wang & Alexander [2] menjelaskan bahwa ML dapat dimanfaatkan untuk memperoleh informasi dan wawasan yang akan digunakan untuk membuat keputusan tertentu. Algoritma ML dapat diklasifikasikan menjadi dua kategori, yaitu *supervised learning* dan *unsupervised learning*. Beberapa contoh metode yang menggunakan algoritma *supervised learning* adalah *Naïve Bayes* dan *Support Vector Machine* (SVM). Berbeda dengan algoritma *supervised learning* yang menggunakan data yang sudah memiliki label, *unsupervised learning* menggunakan data yang tidak memiliki label dan melakukan klasifikasi dengan membandingkan fitur dari data tersebut. Metode *Clustering* adalah salah satu contoh metode algoritma *unsupervised learning*.

Menurut Clark & Pregibon [3], *tree-based models* adalah metode *supervised learning* yang digunakan untuk menemukan struktur di data. Metode ini menggunakan beberapa pernyataan bersyarat untuk mempartisi data menjadi beberapa himpunan bagian. Beberapa contoh dari *tree-based models* adalah *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest*. Prasad *et al.* [4] menjelaskan perbedaan antara *Regression Tree* dengan teknik regresi klasik. Berbeda dengan teknik regresi klasik di mana hubungan antara variabel prediktor dan variabel responsnya sudah ditentukan seperti linear atau kuadratik, *Regression Tree* tidak mengasumsikan hubungan apa pun. *Output* dari metode tersebut berupa diagram pohon dengan cabang-cabang yang ditentukan dari aturan *splitting* dan sejumlah *terminal nodes* yang merupakan rata-rata dari variabel respons. ML

dapat dimanfaatkan di berbagai industri dan salah satunya adalah di industri penerbangan. Industri penerbangan di Indonesia menghasilkan data dengan volume yang sangat besar setiap harinya.

Fatimah [5] menjelaskan bahwa Indonesia adalah negara kepulauan yang mencakup lima pulau besar, ratusan pulau sedang, dan ribuan pulau kecil. Luas wilayah Indonesia yang sangat besar juga mengakibatkan diperlukannya sarana transportasi yang dapat memadai kebutuhan mobilisasi banyak orang dari satu tempat menuju ke tempat yang lainnya, sehingga tidak jarang orang yang memanfaatkan sarana transportasi udara untuk melakukan kegiatan bepergian. Oleh karena itu, perusahaan maskapai penerbangan tentunya menginginkan pendapatan yang maksimum sebagai jasa penyedia transportasi udara tersebut. Untuk mencapai target pendapatan tersebut, perusahaan menggunakan *revenue management* untuk menentukan harga tiket pesawat yang paling optimum. Menurut Li & Peng [6], dalam menentukan harga tiket penerbangan, perusahaan mempertimbangkan berbagai banyak hal seperti tingkat permintaan dan *inventory level*. Data ini tidak terbuka bagi umum dan sulit bagi calon penumpang untuk memprediksi harga tiket pesawat karena ketidakterseediaannya data tersebut.

Untuk mengatasi masalah ini, dapat dimanfaatkan data penerbangan yang terbuka untuk umum sebagai faktor untuk memprediksi harga pesawat. Data yang dapat digunakan untuk membangun model prediksi yaitu jam kedatangan, jam keberangkatan, jumlah transit, hari penerbangan, dan sebagainya. Algoritma *tree-based models* yaitu *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest* adalah metode yang dipilih untuk memprediksi harga tiket pesawat. Ketiga metode tersebut juga akan dibandingkan akurasi untuk mendapatkan model yang terbaik. Rute penerbangan domestik di Indonesia yaitu Jakarta-Surabaya dan Jakarta-Manado adalah salah satu rute dengan jumlah penerbangan yang cukup banyak yaitu 20-30 penerbangan setiap harinya. Kedua rute ini akan digunakan untuk membuat model prediksi.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah pada penelitian ini adalah sebagai berikut.

1. Bagaimana cara memprediksi harga tiket pesawat maskapai penerbangan di Indonesia pada rute Jakarta-Surabaya dan Jakarta-Manado menggunakan

metode *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest*?

2. Bagaimana perbandingan akurasi metode *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest* dalam memprediksi harga tiket pesawat maskapai penerbangan di Indonesia pada rute Jakarta-Surabaya dan Jakarta-Manado?

### 1.3 Tujuan Penelitian

Berawal dari rumusan masalah yang akan dipecahkan, penelitian ini memiliki dua tujuan utama yaitu sebagai berikut.

1. Memperoleh prediksi harga tiket pesawat maskapai penerbangan di Indonesia pada rute Jakarta-Surabaya dan Jakarta-Manado dengan metode *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest*.
2. Membandingkan keakuratan metode *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest* dalam memprediksi harga tiket pesawat maskapai penerbangan di Indonesia pada rute Jakarta-Surabaya dan Jakarta-Manado.

### 1.4 Batasan Masalah

Berikut adalah batasan masalah beserta asumsi yang dipakai untuk mencapai tujuan penelitian.

1. Data yang digunakan adalah seluruh data penerbangan dari 9 Juni 2022 hingga 31 Agustus 2022.
2. Data penerbangan dikumpulkan setiap hari dari situs web *tiket.com* menggunakan *browser extension* yaitu *Web Scraper*.
3. Rute penerbangan yang digunakan dalam penelitian adalah rute penerbangan domestik yaitu Jakarta-Surabaya (CGK-SUB) dan Jakarta-Manado (CGK-MDC).
4. Maskapai penerbangan yang digunakan dalam penelitian adalah maskapai penerbangan di Indonesia yaitu Batik Air, Citilink, Garuda Indonesia, Lion Air, Nam Air, dan Super Air Jet.
5. Kelas kabin yang digunakan dalam penelitian ini adalah ekonomi.

6. Zona waktu yang digunakan dalam penelitian adalah Waktu Indonesia Barat (WIB).
7. Metode *tree-based models* yang digunakan adalah *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest*.
8. Bahasa pemrograman yang digunakan adalah *R*.
9. *Package R* yang digunakan untuk metode *Regression Tree* adalah *rpart* [7].
10. *Package R* yang digunakan untuk metode *Bagging Regression Tree* adalah *caret* [8].
11. *Package R* yang digunakan untuk metode *Random Forest* adalah *caret* [8].

## **1.5 Manfaat Penelitian**

Pada penelitian ini, terdapat dua jenis manfaat yaitu manfaat teoritis dan manfaat praktis. Berikut dijelaskan beberapa manfaat yang bisa diambil dari penelitian ini.

### **1.5.1 Manfaat Teoritis**

Pada penelitian ini, manfaat teoritisnya adalah untuk menemukan metode *tree-based models* yang paling optimal untuk memprediksi harga tiket pesawat maskapai penerbangan di Indonesia pada rute Jakarta-Surabaya dan Jakarta-Manado.

### **1.5.2 Manfaat Praktis**

Pada penelitian ini, manfaat praktis yang dapat diambil adalah mengurangi ketidakpastian harga tiket pesawat dan dapat digunakan calon penumpang pesawat untuk menentukan jadwal penerbangan yang diinginkan dengan harga paling optimal.

## **1.6 Sistematika Penulisan**

Berikut adalah sistematika penulisan skripsi yang diterapkan untuk penulisan dari penelitian ini.

1. BAB I - Pendahuluan

Latar belakang, rumusan masalah, tujuan penelitian, batasan dan asumsi penelitian, manfaat penelitian, dan sistematika penulisan penelitian ini akan dijelaskan pada Bab I.

2. BAB II - Landasan Teori

Pada bab ini dijelaskan beberapa teori yang akan digunakan untuk mendukung penulisan penelitian. Teori yang dibahas antara lain *machine learning*, *tree-based models*, *Regression Tree*, *bias and variance trade-Off*, *overfitting and underfitting*, cara meningkatkan akurasi *Regression Tree*, *k-fold cross-validation* dan evaluasi model. Selain itu, tinjauan pustaka yang dimanfaatkan sebagai referensi bagi penelitian ini juga dimuat dalam bab ini.

3. BAB III - Metodologi

Pada Bab III dijelaskan proses atau langkah pengerjaan yang akan dilaksanakan secara rinci.

4. BAB IV - Analisis dan Pembahasan

Hasil yang diperoleh dari pengolahan data menggunakan metode *Regression Tree*, *Bagging Regression Tree*, dan *Random Forest*, beserta perbandingan akurasi setiap model dijelaskan pada bab IV. Hasil analisis tersebut berupa model yang terbaik untuk memprediksi harga tiket pesawat.

5. BAB V - Kesimpulan dan Saran

Pada bab ini dijelaskan penarikan kesimpulan yang didapatkan dari hasil penelitian, beserta saran-saran untuk pengembangan penelitian yang serupa kedepannya.