

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Pertumbuhan pasar aplikasi *mobile* yang signifikan memiliki dampak besar pada teknologi digital dengan jumlah aplikasi yang tersedia di *Google Play store* hingga maret 2019 sekitar 2,6 juta dan akan terus bertambah seiring berjalannya waktu (Appbrain, 2019). Pengembang dan pengguna aplikasi merupakan kunci utama pada dampak pasar terhadap pengembangan aplikasi (Hengshu Zhu *et al.*, 2014). Dalam melakukan pengembangan aplikasi para developer perlu memprediksi aplikasi di *market* dengan akurat, karena hasil prediksi yang akurat sangat penting dalam menentukan pengembangan aplikasi di *Google Play* (Shen, Lu and Hu, 2017). Tahun 2017, Hartmann-Boyce *et al* melakukan tinjauan terhadap aplikasi *Google Play Store* untuk mengeksplorasi apa yang disukai dan tidak disukai pengguna pada aplikasi penurunan berat badan dan pelacakan berat badan. Hasil penelitian Hartmann-Boyce *et al* menunjukkan bahwa rating dari pengguna mempengaruhi kesuksesan sebuah aplikasi (Hartmann-Boyce *et al.*, 2017). Rating aplikasi juga mempengaruhi system rekomendasi populer aplikasi pada *market google play* dengan kriteria menggunakan parameter kategori, jumlah install, penilaian, ulasan (Zhu *et al.*, 2014).

Untuk melakukan prediksi rating aplikasi terdapat beberapa metode yang digunakan seperti pada tahun 2017, chen *et al* melakukan komparasi antara metode *Logistic Model Tree (LMT)*, *Random Forest (RF)* dan *Decision Tree (CART)* untuk melakukan prediksi pada kerawanan tanah longsor. Hasil penelitian menunjukkan bahwa hasil perbandingan ketiga metode tersebut menghasilkan model *random forest* memiliki prediksi yang terbaik dibandingkan dengan model LMT atau CART dengan nilai *Area Under Curve (UAC)* sebesar 0.837 dan dengan nilai *predictive accuracy* sebesar 0.772. pada penelitian lainnya yang melakukan perbandingan antara Random Forest dengan K-Nearest Neighbors pada dataset HAR (human activity recognition, hasil dari perbandingan

tersebut didapatkan hasil akurasi terbaik dengan menggunakan metode Random forest dengan nilai 93.13%(Bindu, BhanuJyothi and Suryanarayana, 2017). pada penelitian lainnya yang dimana dilakukan perbandingan antara SVM dikombinasikan dengan classifiers lain seperti BayesNet, AdaBoost, Logistic, IBK, J48, Random Forest, JRip, OneR dan SimpleCart, hasil dari penelitian tersebut mendapatkan SVM dikombinasikan dengan Random Forest mendapatkan hasil yang baik dengan nilai 97,50% dibandingkan dengan hanya menggunakan SVM dengan nilai 91,81% (Chand *et al.*, 2016)

Untuk dapat memprediksi *rating* sebuah aplikasi dengan baik maka dibutuhkan adanya data yang lengkap. Namun, permasalahan yang sering muncul pada suatu data adalah adanya ketidaklengkapan data pada suatu variabel atau sering disebut dengan *missing value*. Untuk mengatasi *missing value* pada *dataset* yang digunakan adalah dengan mengisi *missing value* dengan nilai yang mungkin berdasarkan informasi yang tersedia pada data atau biasanya disebut dengan imputasi (Aljuaid and Sasi, 2017). Tahun 2019, Jadhav, Pramod dan Ramanathan melakukan penelitian tentang perbandingan metode *mean imputation*, *median imputation*, *KNN imputation*, *predictive mean matching*, *Bayesian linear regression*, *linier regression* dan *random sample* untuk mengatasi *missing value* pada dataset numerik (Jadhav, Pramod and Ramanathan, 2019). Ketujuh metode tersebut diuji dengan beberapa dataset seperti *wine*, *concrete*, *liver patient* dan *seed* untuk melakukan proses imputasi data. Hasil dari pengujian yang dilakukan menunjukkan bahwa metode KNN imputation memiliki nilai NRMSE yang paling rendah yaitu 0.087871 dibandingkan dengan metode lainnya. pada penelitian lainnya yaitu dengan melakukan komparasi antara KNN, *litewise deletion* dan *Mean Imputation* untuk mengatasi *multiple imputation* pada dataset. Hasil dari perbandingan ketiga metode tersebut KNN merupakan metode imputasi yang mendapatkan hasil akurasi yang tinggi dengan nilai 74.5% (Minakshi, Vohra and Gimpy, 2014)

Berdasarkan uraian di atas maka dalam penelitian ini akan dilakukan kajian optimalisasi pada data *google-play-store-apps* dengan menggunakan algoritma KNN *Imputation* untuk mengatasi *missing value* dan memprediksi

rating aplikasi pada *Google Play* menggunakan metode *Random Forest* pada atribut aplikasi untuk meningkatkan akurasi dalam melakukan prediksi dengan judul “**KNN Imputation Missing Value untuk Prediksi Rating Aplikasi Pada Google Play Menggunakan Metode Random Forest**”.

## 1.2 Masalah Penelitian

Berdasarkan masalah pada latar belakang di atas, maka masalah dalam penelitian ini dapat diuraikan menjadi dua bagian yaitu identifikasi masalah dan rumusan masalah.

### 1.2.1 Identifikasi Masalah

Adapun yang menjadi identifikasi masalah pada penelitian ini adalah adanya *missing value* pada *dataset* yang dapat mempengaruhi hasil prediksi rating.

### 1.2.2 Rumusan Masalah

Berdasarkan identifikasi masalah yang ada, maka rumusan masalah pada penelitian ini adalah bagaimana menghilangkan *missing value* untuk meningkatkan prediksi rating aplikasi berdasarkan layanan dari *google play*.

## 1.3. Tujuan dan Manfaat Penelitian

Tujuan dari penelitian ini adalah untuk meningkatkan akurasi prediksi *rating* aplikasi di *play store* dan melakukan imputasi *missing value* pada *dataset* dengan penerapan Algoritma *k-nearest neighbor* dan memprediksi hasil dari imputasi dengan menggunakan *random forest*.

Manfaat dari penelitian ini antara lain:

1. Hasil penelitian ini dapat dijadikan referensi untuk penelitian lebih lanjut.
2. Membantu para pengembang aplikasi untuk menentukan aplikasi apa yang akan dibuat.

## 1.4. Batasan masalah

Adapun yang menjadi batasan masalah pada penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan adalah hasil dari *web scrapped Google Play store* yang diambil dari *kaggle*. *Dataset* tersedia secara umum dan dapat diakses pada link <https://www.kaggle.com/lava18/google-play-store-apps>
2. Atribut yang digunakan adalah *App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Ver, Android Ver*.
3. Arsitektur *machine learning* yang digunakan adalah dengan *random forest*

### 1.5. Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

#### 1. Studi Literatur

Pada tahap ini dilakukan pendalaman terhadap konsep *missing value* dalam KNN *Imputation* dan metode *Random Forest* untuk proses prediksi rating aplikasi di *Google Play* dan bagaimana cara menyelesaikan masalah tersebut serta mempelajari beberapa penelitian terdahulu.

#### 2. Analisis masalah

Pada tahap ini dilakukan analisis berdasarkan hasil studi literatur untuk mengidentifikasi masalah yang harus diselesaikan, data yang dibutuhkan, dan menentukan metode yang diusulkan untuk menyelesaikan masalah.

#### 3. Perancangan model

Pada tahap ini dilakukan perancangan model dengan membuat *flowchart* yang menggambarkan proses penerapan KNN-*Imputation* untuk mengatasi *missing value* dan *Random Forest* dalam melakukan prediksi rating pada aplikasi.

#### 4. Pengujian

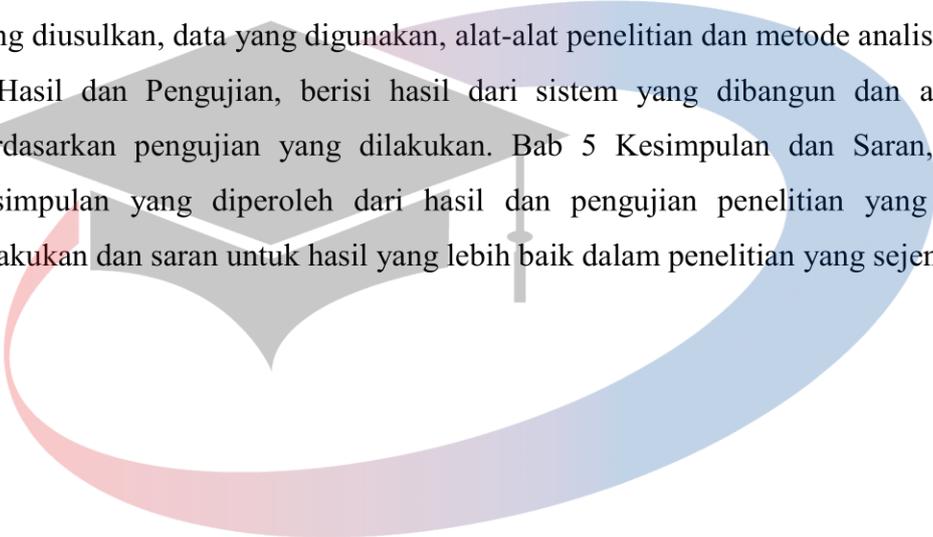
- a. Melakukan pengujian untuk mengetahui nilai akurasi dan presisi yang lebih baik dengan menentukan nilai  $k$  terbaik pada proses imputasi *missing value*.
- b. Melakukan pengujian untuk mengetahui nilai prediksi terbaik dengan mengubah jumlah pohon dari random forest

#### 5. Menarik kesimpulan dari hasil pengujian

#### 6. Menyusun laporan Tesis

## 1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini terdiri dari 5 bab, dimana secara garis besar masing-masing bab membahas hal – hal berikut ini. Bab 1 Pendahuluan, berisi penjelasan umum, masalah dan solusi yang akan dilakukan penelitian. Bab 2 berisi studi literatur dan tinjauan singkat terkait masalah dan metode yang berhubungan dengan penelitian yang akan dilakukan. Bab 3 Metodologi Penelitian, berisi identifikasi masalah, langkah-langkah dari metode yang diusulkan, data yang digunakan, alat-alat penelitian dan metode analisis. Bab 4 Hasil dan Pengujian, berisi hasil dari sistem yang dibangun dan analisis berdasarkan pengujian yang dilakukan. Bab 5 Kesimpulan dan Saran, berisi kesimpulan yang diperoleh dari hasil dan pengujian penelitian yang sudah dilakukan dan saran untuk hasil yang lebih baik dalam penelitian yang sejenis.



UNIVERSITAS  
MIKROSKIL