DOKUMENTASI UJIAN PRAKTIKUM

“PREDIKSI RATING APLIKASI APP STORE MENGGUNAKAN ALGORITMA RANDOM FOREST”

Disusun Oleh :

Inas Najah Zhahirah (193105901160-6)

FAKULTAS VOKASI

UNIVERSITAS AIRLANGGA

2025

1. Dataset

App_Nam	Primary_G	Content_F	Size_Bytes	Required_	Released	Updated	Version	Price	Currency	Free	Developer	Developer	Average_U	Reviews	Current_V	Current_Version_Reviews
A+ Paper (	Education	4+	21993472	8	2017-09-2	2018-12-2	1.1.2	0	USD	TRUE	1.38E+09	HKBU ARC	0	0	0	0
A-Books	Book	4+	13135872	10	2015-08-3	2019-07-2	1.3	0	USD	TRUE	1.03E+09	Roman Dr	5	1	5	1
A-books	Book	4+	21943296	9	2021-04-1	2021-05-3	1.3.1	0	USD	TRUE	1.46E+09	Terp AS	0	0	0	0
A-F Book I	Book	4+	81851392	8	2012-02-1	2019-10-2	1.2	2.99	USD	FALSE	4.4E+08	i-editeur.c	0	0	0	0
A-Z Synon	Reference	4+	64692224	9	2020-12-1	2020-12-1	1.0.1	0	USD	TRUE	6.57E+08	Ngov chih	0	0	0	0
A. P. J. Ab	Book	4+	18073600	7.1	2016-09-1	2016-11-1	1.4	0	USD	TRUE	1.15E+09	Shera Maj	0	0	0	0
A.P. Telan	Book	4+	1.05E+08	8	2021-07-0	2021-07-1	1.3	0	USD	TRUE	1.23E+09	Bhagatjit S	0	0	0	0
A19BestPr	Book	4+	15737856	12	2020-10-1	2020-10-1	2.1	0	USD	TRUE	1.54E+09	Gregor An	0	0	0	0
A2 Directc	Book	4+	68765696	11	2020-10-2	2021-01-1	1.0.2	0	USD	TRUE	1.18E+09	SurfEdge (	0	0	0	0
A4 News	News	4+	13825024	13	2019-05-2	2020-09-2	2	0	USD	TRUE	5.87E+08	Pich Prath	0	0	0	0
AA Audio	Book	17+	26133504	8	2017-04-1	2017-08-2	3.6.1	0	USD	TRUE	1.25E+09	Kepler47 S	4.78132	1285	4.78132	1285
AA Big Boo	Book	17+	63112192	9	2015-05-1	2021-09-1	2.2.16	0	USD	TRUE	1.47E+09	Sobriety S	4.78902	1839	4.78902	1839
AA Big Boo	Lifestyle	4+	3095552	9	2012-04-0	2017-04-1	4	1.99	USD	FALSE	3.55E+08	Rob Laltre	4.67354	242	4.67354	242
AA Big Boo	Book	17+	2094080	8	2015-12-1	2018-10-1	1.4.2	0.99	USD	FALSE	2.96E+08	Dean Huff	3.09524	21	3.09524	21
AA Big Boo	Book	17+	45278208	11	2015-08-2	2019-05-2	1.2	4.99	USD	FALSE	9.75E+08	Big Book A	3.88333	60	3.88333	60
AA Big Boo	Book	17+	80515072	10	2015-03-2	2018-10-0	1.6	4.99	USD	FALSE	9.75E+08	Big Book A	4.13253	83	4.13253	83
AA Big Boo	Book	17+	26713088	8	2013-12-2	2017-10-2	3.6.1	0	USD	TRUE	1.25E+09	Kepler47 S	4.75047	533	4.75047	533
AA Daily R	Reference	17+	48664576	9	2020-10-2	2021-10-0	3	0	USD	TRUE	1.5E+09	Steve Bilo	4.06667	15	4.06667	15
Aa Gym C	Book	4+	61285376	8	2012-03-0	2017-03-0	1.9	0	USD	TRUE	4.22E+08	MAHONI (	3	1	3	1
AA Speake	Health & F	17+	28150784	8	2015-07-2	2017-09-1	3.6.1	0	USD	TRUE	1.25E+09	Kepler47 S	4.8041	3507	4.8041	3507
Aaabc Vui	Book	4+	47198208	7	2014-05-0	2014-05-0	1	0	USD	TRUE	3.81E+08	Gentouch	5	1	5	1
AAAÂ³ Litt	Games	4+	13932544	7.1	2014-12-0	2016-03-1	6	0	USD	TRUE	4.19E+08	Jochen He	5	1	5	1
B & O Boo	Book	4+	81585152	8	2020-02-2	2020-02-2	1	0	USD	TRUE	3.7E+08	B and O Tr	0	0	0	0

Dalam penelitian ini, aplikasi diklasifikasikan sebagai *rating rendah* jika memiliki rata-rata rating pengguna sebesar 0 hingga 3.5, dan sebagai *rating tinggi* jika memiliki rata-rata rating pengguna sebesar 4.0 hingga 5.0. Klasifikasi dilakukan menggunakan algoritma Random Forest, dengan prediksi berbasis fitur numerik aplikasi seperti jumlah ulasan, ukuran aplikasi, dan harga.

2. Eksplorasi Data

```
=== Info Dataset ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6000 entries, 0 to 5999
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   App_Name                6000 non-null   object
 1   Primary_Genre           6000 non-null   object
 2   Content_Rating          6000 non-null   object
 3   Size_Bytes              6000 non-null   int64
 4   Required_IOS_Version    6000 non-null   object
 5   Released                6000 non-null   object
 6   Updated                 6000 non-null   object
 7   Version                 6000 non-null   object
 8   Price                   6000 non-null   float64
 9   Currency                6000 non-null   object
 10  Free                    6000 non-null   bool
 11  DeveloperId             6000 non-null   int64
 12  Developer               6000 non-null   object
 13  Average_User_Rating     6000 non-null   float64
 14  Reviews                 6000 non-null   int64
 15  Current_Version_Score   6000 non-null   float64
 16  Current_Version_Reviews 6000 non-null   int64
dtypes: bool(1), float64(3), int64(4), object(9)
```

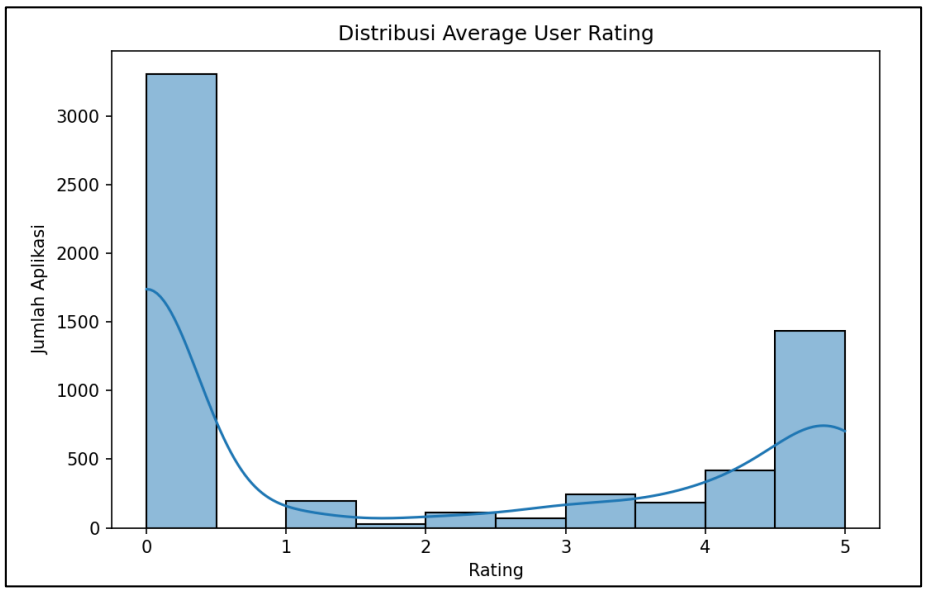
```
=== Missing Values ===
App_Name                  0
Primary_Genre             0
Content_Rating            0
Size_Bytes                0
Required_IOS_Version      0
Released                  0
Updated                   0
Version                   0
Price                     0
Currency                  0
Free                      0
DeveloperId               0
Developer                 0
Average_User_Rating       0
Reviews                   0
Current_Version_Score     0
Current_Version_Reviews   0
```

- Jumlah Data: 6.000 baris dan 17 kolom.
- Tipe Data: Terdiri dari data numerik (int, float), kategorik (object, bool), dan waktu (Released, Updated).
- Kolom Target: Average_User_Rating adalah nilai yang ingin kamu prediksi.
- Tidak Ada Missing Value: Semua kolom lengkap (0 nilai kosong).

```
=== Statistik Deskriptif ===
         Size_Bytes         Price  ...  Current_Version_Score  Current_Version_Reviews
count  6.000000e+03   6000.000000  ...            6000.000000              6000.000000
mean   8.164082e+07      1.028102  ...               1.815322               219.767000
std    1.442551e+08     10.971326  ...               2.166482              7608.897974
min    2.938880e+05      0.000000  ...               0.000000                 0.000000
25%    2.232550e+07      0.000000  ...               0.000000                 0.000000
50%    4.306022e+07      0.000000  ...               0.000000                 0.000000
75%    8.480077e+07      0.000000  ...               4.400000                 3.000000
max    2.804845e+09    499.990000  ...               5.000000            552210.000000
```

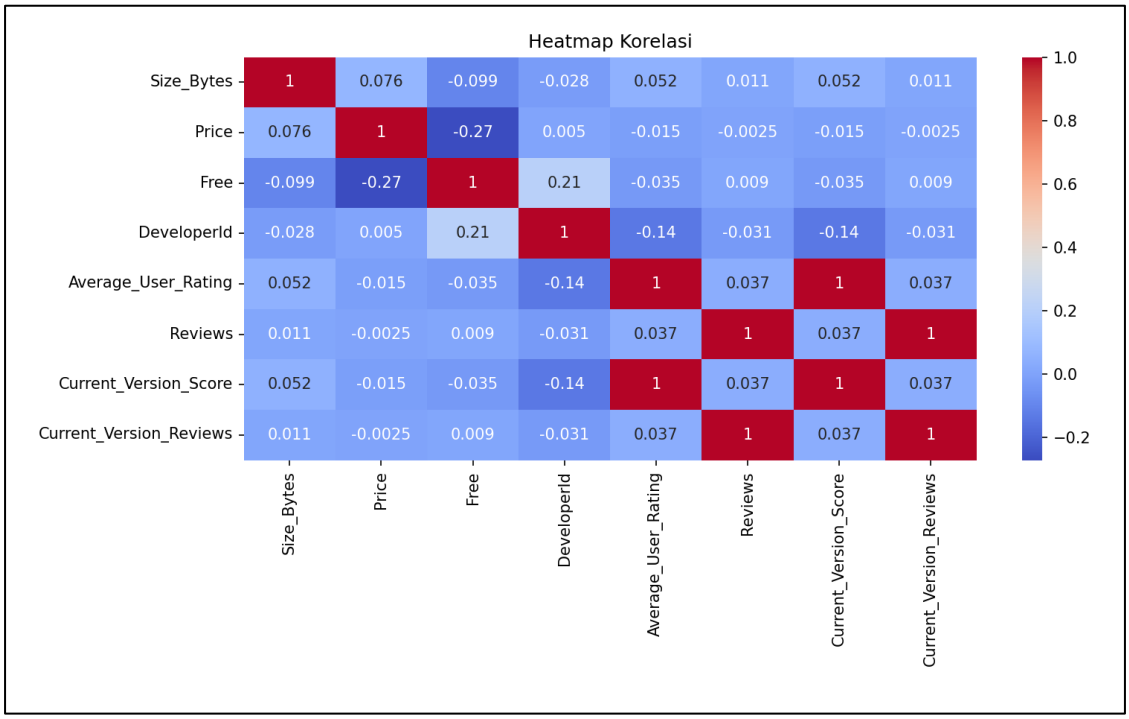
Statistik Deskriptif (Kolom Numerik)

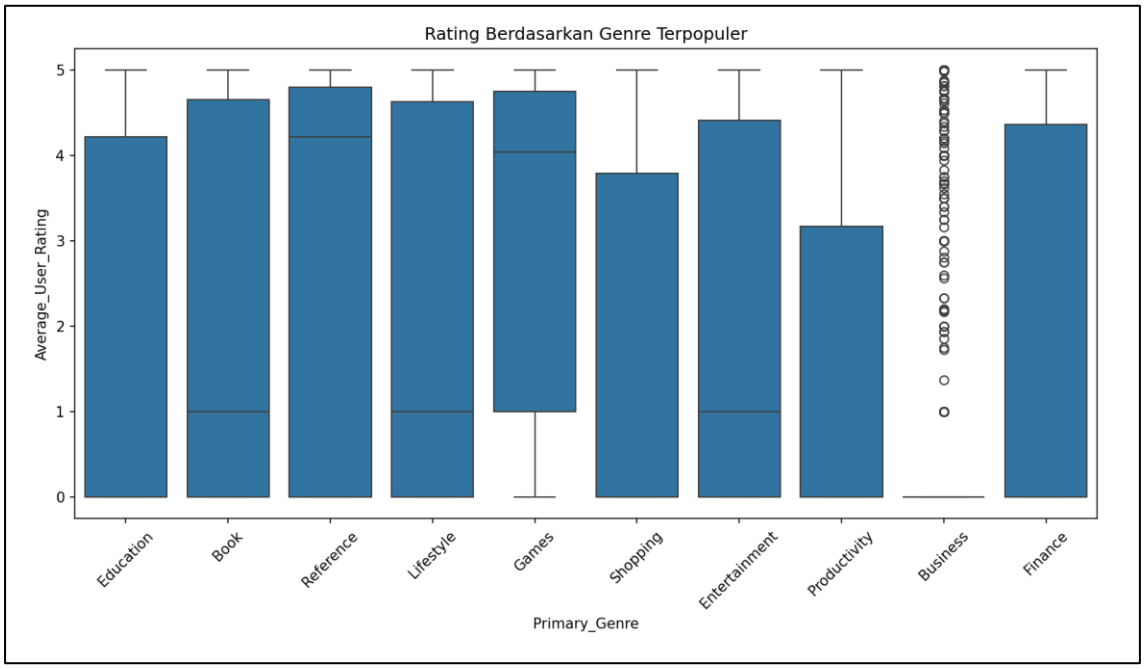
- Size_Bytes (Ukuran aplikasi):
    - Rata-rata: ~81 MB
    - Maksimum: ~2,8 GB
- Price:
    - 75% aplikasi adalah gratis (harga Rp 0)
    - Harga maksimum: $499.99 (kemungkinan aplikasi premium tertentu)
- Current_Version_Score & Reviews:
    - Banyak aplikasi belum mendapat review → nilai median dan mean mendekati 0
- Outlier:
    - Reviews, Current_Version_Reviews dan Price menunjukkan deviasi standar tinggi → ada outlier (bisa dibersihkan atau di-transform)



Distribusi Rating Pengguna

- Sebagian besar aplikasi memiliki rating 0 (mungkin karena belum direview).
- Puncak distribusi lain berada di rating 5, menunjukkan aplikasi populer atau berkualitas baik.
- Distribusi tidak normal dan skew ke kiri → rating 0 perlu dianalisis apakah valid atau perlu diabaikan.





Korelasi Fitur Numerik

- Korelasi dengan target (Average_User_Rating) rendah di semua fitur, artinya tidak ada fitur numerik yang sangat dominan terhadap rating.
- Free dan Price memiliki korelasi negatif (logis, karena aplikasi gratis berarti Price = 0).
- Korelasi tinggi antar Current_Version_Score, Current_Version_Reviews, dan Reviews, menunjukkan kemungkinan redundansi fitur.

3. Prapemrosesan data

Pada tahap pra-pemrosesan, dilakukan serangkaian langkah penting untuk menyiapkan data agar dapat digunakan dalam proses pemodelan machine learning dengan lebih optimal. Berikut langkah-langkah yang dilakukan:

1. Menghapus Data Rating Kosong (0): Baris data dengan Average_User_Rating = 0 dihapus karena dianggap tidak valid (kemungkinan aplikasi belum memiliki ulasan pengguna sama sekali). Hal ini penting untuk menjaga akurasi model prediksi.
2. Menghapus Kolom yang Tidak Relevan: Kolom seperti App_Name, Developer, Released, Updated, Version, Currency, dan Required_IOS_Version dihapus karena tidak memberikan kontribusi langsung terhadap prediksi rating atau bersifat unik/bernilai tinggi kardinalitas sehingga tidak efektif untuk model.
3. Memisahkan Fitur dan Target: Kolom Average_User_Rating dipisahkan sebagai target (label), sedangkan sisanya digunakan sebagai fitur (features).
4. Identifikasi Jenis Fitur:
    - Fitur numerik diidentifikasi berdasarkan tipe data int64, float64, dan bool.
    - Fitur kategorik diidentifikasi berdasarkan tipe data object.
5. Penanganan Missing Value (Data Hilang):
    - Untuk fitur numerik, strategi imputasi median digunakan agar tidak terpengaruh outlier.
    - Untuk fitur kategorik, imputasi nilai yang paling sering (most frequent) digunakan agar tetap representatif terhadap data dominan.

6. Transformasi Fitur Kategorik:
    - Dilakukan One-Hot Encoding (OHE) pada fitur kategorik agar dapat diproses oleh model regresi, karena model tidak dapat bekerja langsung dengan data kategorik.
7. Membuat Dataframe Akhir:
    - Fitur hasil transformasi digabungkan dan dibuat menjadi dataframe baru.
    - Target Average_User_Rating dikembalikan ke dataframe hasil pra-pemrosesan.
8. Menyimpan Hasil:
    - Dataset hasil pra-pemrosesan disimpan ke dalam file CSV dengan nama dataset_preprocessed.csv, siap digunakan untuk proses training dan evaluasi model regresi.

4. Data split

Pada tahap ini, dilakukan proses pemodelan dan evaluasi untuk memprediksi *Average User Rating* menggunakan beberapa algoritma regresi. Dataset yang telah melalui tahap pra-pemrosesan digunakan sebagai input.

1. Pemisahan Fitur dan Target

   Dataset dibagi menjadi dua bagian utama:
   - Fitur (X): seluruh kolom kecuali Average_User_Rating.
   - Target (y): nilai Average_User_Rating yang akan diprediksi.

   Langkah ini penting agar model dapat mempelajari hubungan antara fitur dan target secara optimal.

2. Validasi dengan K-Fold Cross Validation

   Untuk mengevaluasi performa model secara lebih objektif dan menghindari overfitting, digunakan metode K-Fold Cross Validation dengan:
   - 10 fold (n_splits=10): data dibagi menjadi 10 bagian,
   - Model dilatih pada 9 bagian dan diuji pada 1 bagian, diulang hingga semua bagian menjadi test set,
   - Shuffle=True: data diacak sebelum dibagi agar distribusinya merata.

Metode ini memberikan gambaran akurasi model secara umum, bukan hanya pada satu pembagian data saja.

3. Penggunaan Hyperparameter

Untuk mengoptimalkan performa dan mengurangi risiko overfitting, dilakukan penyesuaian hyperparameter pada beberapa model:

- Random Forest Regressor: menggunakan max_depth=10, min_samples_leaf=4, dan n_estimators=100 untuk membatasi kompleksitas pohon, menjaga generalisasi, dan meningkatkan stabilitas prediksi.
- Decision Tree Regressor: disesuaikan dengan max_depth=10 dan min_samples_leaf=4 untuk mencegah model terlalu dalam dan sensitif terhadap noise.
- XGBoost Regressor: menggunakan objective="reg:squarederror" dan parameter default, tetapi hasilnya menunjukkan kinerja lebih rendah dibanding model lain.

Penggunaan hyperparameter ini bertujuan agar model tidak hanya cocok pada data pelatihan tetapi juga dapat memberikan hasil prediksi yang akurat pada data baru (generalizable).

5. Pemodelan

Tiga model regresi diterapkan dan dievaluasi, yaitu:

- Random Forest Regressor: model ansambel berbasis pohon keputusan dengan teknik bagging.
- Decision Tree Regressor: model berbasis pohon tunggal yang memetakan fitur ke target.
- XGBoost Regressor: model boosting yang kuat dan efisien, cocok untuk menangani data kompleks.

Masing-masing model dilatih dan dievaluasi menggunakan metrik regresi:

- R^2 (R-squared): mengukur seberapa baik model menjelaskan variasi data.
- MAE (Mean Absolute Error): rata-rata selisih absolut antara prediksi dan nilai aktual.

- RMSE (Root Mean Squared Error): mengukur deviasi prediksi terhadap nilai aktual, lebih sensitif terhadap outlier.

6. Evaluasi

Hasil evaluasi dibandingkan untuk menentukan model terbaik. Model dengan R^2 tertinggi, serta MAE dan RMSE terendah dianggap memiliki performa terbaik dalam memprediksi *Average User Rating*.

```
=== Evaluasi K-Fold Cross Validation ===

Random Forest
R²   : 1.0000
MAE  : 0.0016
RMSE : 0.0052

Decision Tree
R²   : 0.9999
MAE  : 0.0025
RMSE : 0.0083

XGBoost
R²   : 0.9985
MAE  : 0.0268
RMSE : 0.0457
```

- Random Forest menjadi model terbaik karena:
  - $R^2 = 1.0000$ → artinya hampir seluruh variasi pada data target dijelaskan oleh model.
  - MAE dan RMSE juga yang paling kecil → prediksinya sangat dekat dengan nilai aktual.
- Decision Tree juga bagus, tetapi sedikit lebih overfit dibanding Random Forest (karena tidak menggabungkan banyak pohon).
- XGBoost memiliki performa paling rendah di sini, dengan MAE dan RMSE cukup tinggi, menandakan kemungkinan overfitting pada data tertentu atau sensitivitas terhadap outlier

7. Feature Importance

```
Top 10 Feature Importance - Random Forest:
                Feature   Importance
  Current_Version_Score 9.999757e-01
Current_Version_Reviews 1.102103e-05
                Reviews 8.074032e-06
             Size_Bytes 1.814149e-06
            DeveloperId 1.532996e-06
     Primary_Genre_Book 1.191781e-06
 Primary_Genre_Business 2.647760e-07
    Primary_Genre_Games 2.071701e-07
                  Price 7.830035e-08
Primary_Genre_Education 6.260909e-08
```

Berdasarkan hasil analisis feature importance menggunakan algoritma Random Forest, diketahui bahwa fitur Current_Version_Score memiliki pengaruh paling dominan terhadap prediksi Average_User_Rating. Hal ini menunjukkan bahwa skor pengguna pada versi terbaru aplikasi sangat menentukan penilaian keseluruhan aplikasi tersebut. Fitur-fitur lain seperti jumlah ulasan versi terbaru (Current_Version_Reviews), total ulasan (Reviews), dan ukuran aplikasi (Size_Bytes) juga berkontribusi meskipun dalam porsi yang jauh lebih kecil. Beberapa fitur kategorik seperti genre aplikasi (Primary_Genre_Book, Primary_Genre_Business, dan sebagainya) serta harga (Price) menunjukkan pengaruh yang rendah. Temuan ini mengindikasikan bahwa performa aplikasi pada versi terbaru menjadi perhatian utama pengguna dalam memberikan rating, sedangkan faktor-faktor lain hanya menjadi pelengkap dalam membentuk persepsi pengguna.

8. Kesimpulan

Model terbaik untuk regresi prediksi rating App Store pada dataset ini adalah:

Random Forest Regressor