Jurnal Strategi
Volume 3 Nomor 2 November 2021

Analisis dan Prediksi Pertempuran Game Of Thrones Menggunakan Algoritma Random Forest dan Logistic Regression

Stefanus Hermawan[#1], Setia Budi*2

[#]*Program Studi Teknik Informatika, Universitas Kristen Maranatha*
Jl. Surya Sumantri No.65 Bandung 40164
[1]1772023@maranatha.ac.id
[2]setia.budi@it.maranatha.edu

Abstract — **Game of Thrones is a popular television series with an IMDB rating of 9.3/10 adapted from a fantasy fiction novel written by George R R Martin that aired from 2011 to 2019 with a total of 8 seasons and 73 episodes. The story is about nine kingdoms that fight for the throne throughout the land of Westeros and defend the kingdom from the threat of enemies who reappear after thousands of years. The Game of Thrones dataset available on the Kaggle website is a dataset collected from various sources which then attracted the interest of a number of fans to carry out various analyses. The purpose of this Final Project is to explore and explain the steps involved in data processing using techniques such as data augmentation to oversampling the data, dividing data using stratified cross validation techniques, then train the model using the Random Forest and Logistic Regression algorithms with a dataset that has been oversampled and the original dataset as a comparison to get the best accuracy.**

Keywords— **Game of Thrones, Logistic Regression, Machine Learning, Random Forest.**

I. PENDAHULUAN

Game of Thrones merupakan serial televisi adaptasi dari novel fiksi fantasi ditulis oleh George R. R. Martin yang ditayangkan sejak tahun 2011 hingga 2019 dengan total delapan musim dan 73 episode. Berkisah tentang sembilan kerajaan yang berjuang untuk menguasai takhta di seluruh tanah Westeros serta mempertahankan kerajaan dari ancaman musuh yang muncul kembali setelah ribuan tahun.

Serial televisi Game of Thrones merupakan salah satu serial televisi yang sukses menarik banyak fans dengan alur cerita peperangan dan politik kerajaan yang ekstensif. Hingga tulisan ini dibuat terdapat hampir 1,8 juta review positif dan mendapatkan rating 9,3/10 di website IMDb[1].

Dengan banyaknya jumlah fans Game of Thrones dari seluruh dunia, terbentuklah berbagai website yang berguna untuk saling berbagi informasi seputar pembaharuan cerita dan produksi serial televisi hingga informasi lengkap profil karakter dan peristiwa-peristiwa yang terjadi dalam buku dan serial televisi.

Dataset Game of Thrones yang terdapat pada website Kaggle[2] merupakan dataset hasil pengumpulan dari berbagai sumber yang kemudian menarik minat sejumlah fans untuk melakukan berbagai analisis. Terdapat pula publikasi berupa artikel yang memuat hasil analisis dan diunggah ke dalam website The Wall Street Journal[3]

Penelitian ini berfokus pada analisis terhadap dataset pertempuran Game of Thrones untuk melakukan eksplorasi dan membangun model prediksi kemenangan pertempuran.

II. KAJIAN TEORI

A. Machine Learning

Machine learning merupakan salah satu cabang dari kecerdasan buatan yang memiliki kemampuan untuk mempelajari suatu tugas berdasarkan data yang diberikan. Menurut Mitchell (1997), *machine learning* adalah program komputer yang belajar dari pengalaman (*experience*) E terhadap suatu tugas (*task*) T yang memiliki ukuran kinerja (*performance*) P [1].

Terdapat 3 kategori dalam machine learning, yaitu:

1. *Supervised Learning*, merupakan kategori *machine learning* yang mempelajari data yang sudah diberikan label maupun suatu nilai dan melakukan prediksi pada data yang belum memiliki label atau nilai.

[1] https://www.imdb.com/
[2] https://www.kaggle.com/
[3] https://www.wsj.com/articles/fans-geek-out-over-game-of-thrones-data-1499877067

2. Jenis tugas pada *supervised learning* terbagi menjadi klasifikasi yang dapat menyelesaikan tugas untuk melakukan pengelompokan suatu data berdasarkan label dan regresi yang menentukan nilai pada suatu data dengan cara mencari korelasi antara variabel di dalam data yang diberikan.
3. *Unsupervised Learning*, merupakan teknik pengelompokan suatu data berdasarkan variabel di dalam dataset yang tidak diberikan label secara spesifik.
4. *Reinforcement Learning*, merupakan teknik pembelajaran mesin yang didasarkan dengan trial dan error dalam berinteraksi dengan suatu lingkungan atau skenario tertentu yang kemudian diberikan umpan balik kepada setiap aksi yang dilakukan.

B. Random Forest Classifier

Random Forest Classifier adalah metode ansambel yang melatih beberapa pohon keputusan secara paralel dengan *bootstrap* dan agregasi yang secara kolektif disebut sebagai *bagging*. *Bootstrap* menunjukkan bahwa beberapa pohon keputusan individu dilatih secara paralel pada berbagai subset dari set data pelatihan menggunakan subset berbeda dari fitur yang tersedia. *Bootstrap* memastikan bahwa setiap pohon keputusan di model ini unik, sehingga mengurangi variasi keseluruhan. Untuk keputusan akhir, *Random Forest Classifier* menggabungkan berbagai pohon keputusan. model seringkali mengungguli sebagian besar metode klasifikasi lainnya dalam hal akurasi tanpa masalah *overfitting* [2].

Pada penelitian ini, peneliti menggunakan model *Random Forest Classifier* yang tersedia di *library* scikit-learn untuk melakukan prediksi kemenangan dalam pertempuran *Game of Thrones*.

C. Logistic Regression

Logistic Regression adalah analisis regresi yang tepat untuk dilakukan ketika variabel dependen bersifat biner. Seperti semua analisis regresi, *Logistic Regression* adalah analisis prediktif. *Logistic Regression* digunakan untuk mendeskripsikan data dan menjelaskan hubungan antara satu variabel biner dependen dan satu atau lebih variabel independen nominal, ordinal, interval atau tingkat rasio [3].

Pada penelitian ini, *Logistic Regression* yang tersedia pada *library* scikit-learn digunakan sebagai model kedua sebagai model pembanding terhadap algoritma *Random Forest* untuk melakukan prediksi pertempuran *Game of Thrones*.

D. Decision Tree Regressor

Decision Tree adalah metode *supervised learning* yang digunakan untuk tugas klasifikasi dan regresi. Tujuannya adalah untuk membuat model yang memprediksi nilai variabel target dengan mempelajari aturan keputusan sederhana yang disimpulkan dari fitur data. Aturan keputusan umumnya dalam bentuk pernyataan *if-then-else*. Semakin dalam pohonnya, semakin kompleks aturannya dan modelnya sesuai [4].

Pada penelitian ini, *Decision Tree Regressor* yang tersedia pada *library* scikit-learn digunakan sebagai estimator untuk melakukan imputasi pada dataset untuk mengisi nilai-nilai yang hilang.

E. Synthetic Minority Oversampling Technique (SMOTE)

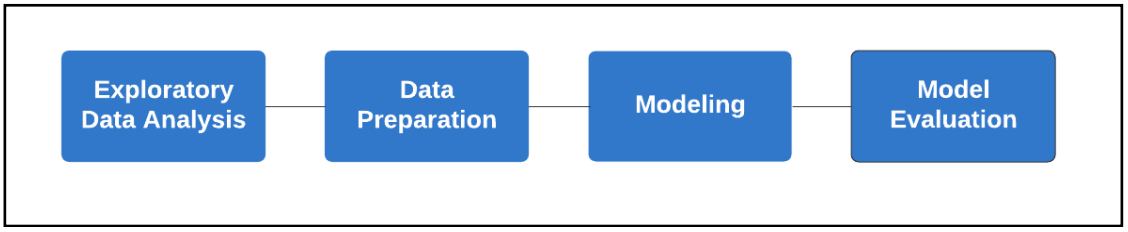
Metode SMOTE merupakan solusi dari masalah ketidakseimbangan data. Ketika metode *oversampling* memiliki berbagai prinsip secara acak, metode SMOTE meningkatkan jumlah kelas sekunder ke jumlah yang sama dengan kelas utama dengan menghasilkan data buatan. Data buatan atau sintetis dibuat berdasarkan *k-nearest neighbors* [5].

Pada penelitian ini, peneliti menggunakan fungsi SMOTE yang tersedia di *library* imbalanced-learn untuk mengatasi ketidakseimbangan data yang dimiliki oleh dataset pertempuran *Game of Thrones*.

III. METODE PENELITIAN

A. Metodologi

Garis besar dari tahapan penelitian ini terbagi menjadi empat proses terurut yang diilustrasikan pada Gambar 1.



Gambar 1 Tahapan Analisis dan Pemodelan

B. *Exploratory Data Analysis (EDA)*

Tahapan eksplorasi yang akan dilakukan meliputi kegiatan sebagai berikut:

1. Mengenal nama – nama kolom dan data yang terdapat dalam dataset.
2. Melakukan pengecekan jumlah baris dan kolom.
3. Mengenal komposisi tipe data pada dataset.
4. Melakukan penghitungan persentase nilai yang hilang pada setiap kolom.
5. Melakukan pengecekan jumlah kelas data untuk melihat keseimbangan dataset.
6. Mencari temuan – temuan berupa data yang kurang relevan dan bisa dilakukan koreksi maupun ekstraksi fitur/kolom baru.

C. *Data Preparation*

Tabel dibawah ini merupakan daftar proses yang dapat dilakukan untuk *data preparation* sebelum dilakukan pemodelan.

TABEL I
DATA PREPARATION YANG DILAKUKAN PADA DATASET

No	Nama kolom	Proses yang dilakukan	Keterangan
1	`attacker_commander` dan `defender_commander`	Ekstrasi fitur dan penghapusan kolom	Data dipisahkan dengan koma dan kurang relevan jika digunakan secara langsung untuk melatih model
2	`attacker_2` s/d `attacker_4`	Ekstrasi fitur dan penghapusan kolom	Kolom memiliki nilai null diatas 60% sehingga tidak relevan jika digunakan dalam pelatihan model
3	`defender _2` s/d `defender _4`	Ekstrasi fitur dan penghapusan kolom	Kolom memiliki nilai null diatas 60% sehingga tidak relevan jika digunakan dalam pelatihan model
4	`battle_number, name, year`	Penghapusan kolom	Kurang relevan dan tidak memiliki informasi yang kuat terkait pertempuran sehingga kurang signifikan untuk digunakan pada pelatihan model
5	Semua kolom dengan tipe data *object*	*Encoding*	Mengubah data teks atau kategorik menjadi nilai *integer* menggunakan `LabelEncoder`
6	Semua kolom dengan nilai *null*	Pengisian nilai *null*	Pengisian nilai *null* menggunakan fungsi `IterativeImputer` dan model *Decision Tree Regressor*
7	Semua kolom	Penskalaan data	Mengubah data menjadi rentang nilai yang lebih terukur
8	Semua kolom	*Oversampling*	Membuat data sintetik tambahan guna menambah jumlah data serta sehingga menghasilkan kelas data yang seimbang menggunakan fungsi `SMOTE`
9	Semua kolom	Pembagian dataset	Membagi dataset menjadi 2 bagian yaitu 75% *train set* untuk pelatihan awal model dan 25% *test set* untuk evaluasi terhadap model

D. *Modeling*

Tahapan pemodelan akan dibagi menjadi dua tahapan, yaitu *modeling* pada dataset asli yang memiliki ketidakseimbangan pada kelas data dan *modeling* menggunakan dataset yang telah dilakukan augmentasi berupa penyeimbangan kelas pada data guna membandingkan perbandingan skor antara model maupun perbandingan antara dataset.

Evaluasi awal pada pemodelan yang menggunakan *train set* ini akan dilakukan dengan fungsi `cross_val_score` untuk melakukan perhitungan skor akurasi, skor F1, serta skor ROC AUC dan fungsi `StratifiedKFold` untuk mengacak sampel dataset dan memastikan data pada setiap bagiannya merepresentasikan keseluruhan data yang akan digunakan dalam pemodelan.

Setelah pemodelan dilakukan, *feature importances* pada model *Random Forest* dan koefisien setiap fitur pada model *Logistic Regression* akan ditampilkan guna dilakukan peninjauan terhadap fitur-fitur penting yang terdapat pada dataset.

E. Model Evaluation

Setelah mendapatkan evaluasi awal berupa skor pada model yang dilatih pada *train set*, tahapan model evaluation akan dilakukan menggunakan data *test set* dengan menghitung skor evaluasi menggunakan fungsi `classification_report` yang menampilkan skor akurasi, skor F1, skor *precision* dan skor *recall* yang kemudian akan disertakan dalam subbab *Model Evaluation.*

IV. HASIL DAN DISKUSI

A. Exploratory Data Analysis (EDA)

Hasil EDA menunjukkan bahwa dataset terdiri dari 28 baris data dan 25 kolom fitur, tipe data yang terdapat pada dataset antara lain tipe *object* yang merupakan tipe data umum yang menunjukkan data teks, tipe data *integer* serta tipe data *float* serta distribusi kelas pada dataset memiliki ketimpangan antar kelas target.

B. Data Preparation

Pada data preparation dilakukan pengisian nilai *null* pada baris 37 kolom `battle_type` dan kolom `battle_outcome`, penghapusan kolom dengan nilai *null* lebih besar dari 60%, kemudian dilakukan ekstraksi fitur berupa penghitungan jumlah komandan perang.

Selanjutnya dilakukan transformasi data menggunakan `LabelEncoder`, pengisian nilai *null* menggunakan fungsi `IterativeImputer` dengan model `DecisionTreeRegressor`, *data scaling* menggunakan fungsi `StandardScaler`, dilakukan augmentasi data menggunakan fungsi SMOTE dengan metode *oversampling* pada data dengan target kelas minoritas sehingga menciptakan data sintetis baru untuk menyeimbangkan target kelas pada dataset dan kemudian dilakukan pembagian dataset hasil penskalaan dan dataset hasil augmentasi menjadi *train set* sebanyak 75% dan *test set* sebanyak 25% menggunakan fungsi `train_test_split` yang selanjutnya kedua jenis dataset tersebut akan digunakan dalam pemodelan.

C. Modeling

Hasil pelatihan model yang terdapat pada Tabel 2 menunjukkan bahwa pada dataset yang tidak dilakukan *oversampling* skor akurasi dan skor ROC AUC pada model *Random Forest* lebih tinggi dibandingkan model Logistic Regression, tetapi kedua model memiliki skor F1 yang sama.

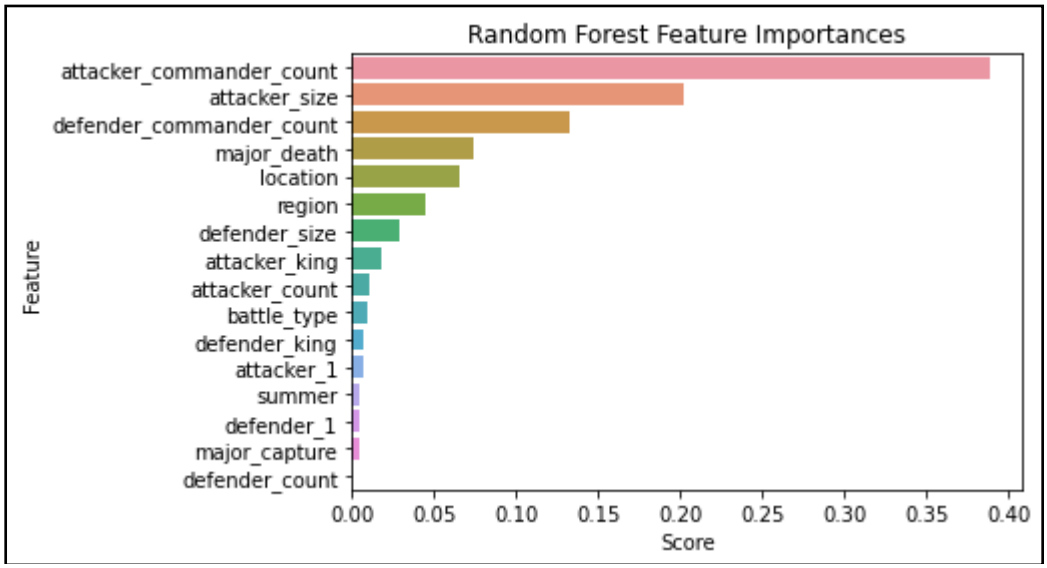
Hasil pelatihan model dengan dataset yang telah dilakukan *oversampling* menunjukkan hasil yang sebaliknya. Semua skor pada model *Logistic Regression* meningkat secara signifikan sedangkan pada model *Random Forest* tidak terjadi peningkatan pada skor akurasi dan skor F1.

TABEL 2

REKAPITULASI SKOR HASIL MODEL

Model	Akurasi	F1	ROC AUC
Random Forest	0.960000	0.959596	0.937500
Logistic Regression	0.926667	0.959596	0.812500
Random Forest + Oversampling	0.960000	0.959596	1.000000
Logistic Regression + Oversampling	0.980000	0.981818	1.000000

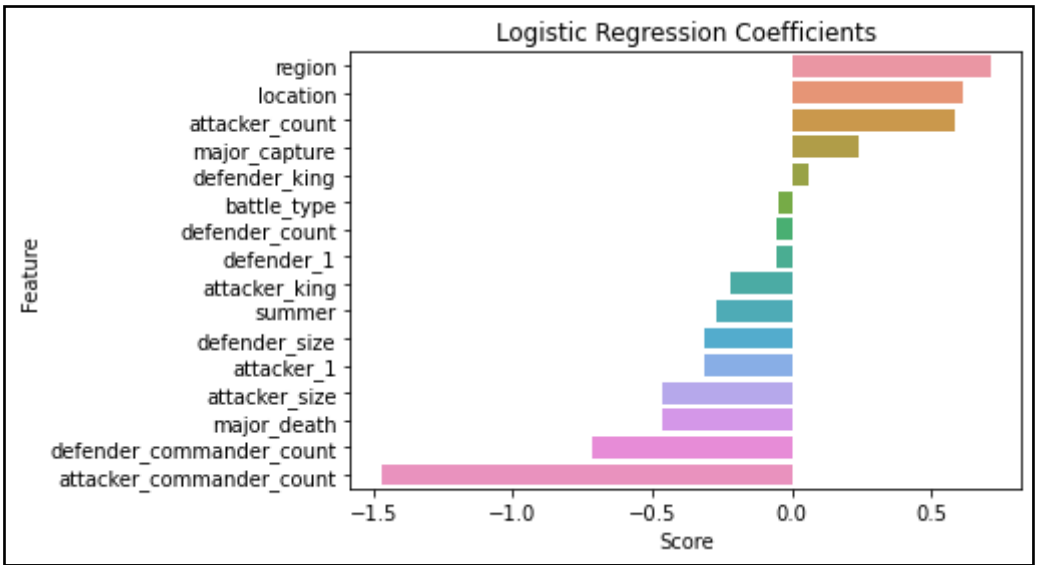
Pada Gambar 2 merupakan hasil dari perhitungan feature importances dari model Random Forest, terlihat bawa tiga fitur yang paling mempengaruhi kemenangan peperangan yaitu `attacker_commander_count` (0.389246) diikuti dengan `attacker_size`(0.202693) dan `defender_commander_count` (0.132379).



Gambar 2 Grafik Feature Importances Random Forest

Pada Gambar 3 menunjukkan nilai koefisien dari setiap fitur yang terdapat pada model *Logistic Regression*. Semakin besar nilai koefisien pada sebuah fitur maka semakin besar pula pengaruh terhadap model terlepas dari nilai koefisien tersebut memiliki nilai positif maupun negatif.

Terlihat bahwa tiga fitur yang paling mempengaruhi kemenangan peperangan adalah `attacker_commander_count` (-1.472369), `defender_commander_count` (-0.719559) dan `region` (0.712785).



Gambar 3 Grafik Nilai Koefisien Logistic Regression

D. Model Evaluation

Hasil evaluasi pada test set menggunakan model Random Forest dan Logistic Regression menunjukkan skor 100% pada matrik evaluasi skor akurasi, skor F1, skor precision dan skor recall seperti yang ditunjukkan pada Gambar 4 serta.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	1.00	1.00	1.00	9
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

Gambar 4 Classification Report Test Set

V. KESIMPULAN

Berdasarkan hasil eksplorasi serta pemodelan yang telah dilakukan, dapat disimpulkan bahwa model *Logistic Regression* dan model *Random Forest* cocok digunakan dalam dataset berukuran kecil seperti dataset peperangan *Game of Thrones*.

Penerapan ekstraksi fitur/kolom pada dataset yang tepat mampu menghasilkan fitur yang memiliki pengaruh besar terhadap model dan penggunaan metode oversampling pada dataset dengan target kelas tidak seimbang mampu meningkatkan akurasi secara signifikan.

DAFTAR PUSTAKA

[1] Mitchell, T., 1997. Machine learning. New York: MacGraw-Hill.
[2] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[3] Statistics Solutions. 2021. What is Logistic Regression? - Statistics Solutions. [online] Available: https://www.statisticssolutions.com/what-is-logistic-regression/ [Diakses 2 Maret 2021.
[4] HackerEarth. 2021. Decision Tree Tutorials & Notes | Machine Learning | HackerEarth. [online] Available: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/[Diakses 2 Maret 2021].
[5] Santoso, Noviyanti & Wibowo, Wahyu & Himawati, H.. (2019). Integration of synthetic minority oversampling technique for imbalanced class. Indonesian Journal of Electrical Engineering and Computer Science. 13. 102-108. 10.11591/ijeecs.v13.i1.pp102-108.