PENDEKATAN ALGORITMA *COST SENSITIVE DECISION TREE* PADA KLASIFIKASI FILM BERDASARKAN PEROLEHAN KOMPILASI DARI *INTERNET MOVIE DATABASE* (IMDB)

SKRIPSI

untuk memenuhi persyaratan
dalam menyelesaikan program sarjana Strata-1 Statistika

Oleh
ALDY NUR PRATAMA
NIM. 2011017210021

PROGRAM STUDI S-1 STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JANUARI 2024

SKRIPSI

PENDEKATAN ALGORITMA *COST SENSITIVE DECISION TREE* PADA KLASIFIKASI FILM BERDASARKAN PEROLEHAN KOMPILASI DARI *INTERNET MOVIE DATABASE* (IMDB)

Oleh
Aldy Nur Pratama
NIM. 2011017210021

Telah dipertahankan pada hari Rabu, tanggal 03-01-2024 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

Pembimbing I

Yuana Sukmawaty, S.Si., M.Si.
NIP. 198810152015042002

Penguji I

Oni Soesanto, S.Si., M.Si.
NIP. 197301262005011003

Pembimbing II

Selvi Annisa, S.Si., M.Si.
NIP. 199212262022032016

Penguji II

Sigit Dwi Prabowo, S.Mat., M.Stat.
NIP. -

Banjarbaru, 03 Januari 2024
Mengetahui,
Koordinator Program Studi Statistika
FMIPA ULM

Prof. Dewi Anggraini, S.Si., M.AppSci., Ph.D.
NIP. 198303282005012001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 03 Januari 2024

Aldy Nur Pratama
NIM. 2011017210021

ABSTRAK

Pendekatan Algoritma *Cost Sensitive Decision Tree* pada Klasifikasi Film berdasarkan Perolehan Kompilasi dari *Internet Movie Database* (Imdb)
(Oleh: Aldy Nur Pratama; Pembimbing: Yuana Sukmawaty, S.Si., M.Si. dan Selvi Annisa, S.Si., M.Si., 2023; 71 halaman)

Internet Movie Database atau IMDb adalah situs yang berkaitan dengan film dan produksinya yang berisikan informasi suatu film. Preferensi film dapat mengalami profit yang tinggi bisa dilihat pada genre film, pemasaran, karya adaptasi, sutradara, pemain, latar, dan rumah produksi. Preferensi masyarakat ini harus menjadi perhatian sebelum dilakukannya produksi film, agar dapat meminimalisir adanya ketidakseimbangan pada profit suatu film khususnya pada film yang memiliki profit terendah. Data tidak seimbang terjadi ketika suatu kelas mengalami perbedaan yang jauh dibandingkan kelas lainnya. Pada kasus ini, data diperoleh dari *website Kaggle* berdasarkan 100 film teratas yang populer setiap tahunnya dari tahun 2003 sampai 2022 pada IMDb. Hal ini menyebabkan terjadinya ketidakseimbangan data dikarenakan perolehan data yang didasari oleh 100 film teratas. Penanganan data tidak seimbang harus dilakukan agar meningkatkan keseimbangan pada data. Hal ini dibantu dengan hasil pengukuran sensitivitas, spesifisitas, dan *G-Mean*. Proses klasifikasi dilakukan menggunakan metode *Decision Tree* dengan bantuan penanganan data yang tidak seimbang menggunakan *Cost Sensitive Learning*. Penerapan ini dapat disebut sebagai *Cost Sensitive Decision Tree*. Mengacu pada rasio pembagian 85% data *training* dan 15% data *testing*, menghasilkan bahwa metode *Cost Sensitive Decision Tree* dinilai baik dalam penanganan ketidakseimbangan data. Adapun berdasarkan perolehan dari nilai *G-Mean*, untuk *Cost Sensitive Decision Tree* memperoleh sebesar 66.36% yang menunjukkan hasil empat kali lebih banyak daripada *Decision Tree* dengan total sebesar 16.75%.

Kata kunci: *Decision Tree*, *Cost Sensitive Learning*, *Internet Movie Database*

ABSTRACT

Cost Sensitive Decision Tree Algorithm Approach to Movie Classification based on Compiled Acquisition from Internet Movie Database (Imdb) (By: Aldy Nur Pratama; Advisors: Yuana Sukmawaty, S.Si., M.Si. and Selvi Annisa, S.Si., M.Si., 2023; 71 pages)

Internet Movie Database or IMDb is a site related to movies and their production that contains information about a movie. Movie preferences can experience high profits can be seen in movie genres, marketing, adaptation works, directors, players, settings, and production houses. This public preference must be a concern before film production is carried out, in order to minimize the imbalance in the profit of a film, especially in films that have the lowest profit. Unbalanced data occurs when a class experiences a large difference compared to other classes. In this case, the data is obtained from the Kaggle website based on the top 100 movies that are popular every year from 2003 to 2022 on IMDb. This causes data imbalance due to data acquisition based on the top 100 movies. Handling unbalanced data must be done in order to improve the balance in the data. This is assisted by the results of sensitivity, specificity, and G-Mean measurements. The classification process is performed using the Decision Tree method with the help of handling unbalanced data using Cost Sensitive Learning. This application can be referred to as Cost Sensitive Decision Tree. Referring to the division ratio of 85% training data and 15% testing data, it results in that the Cost Sensitive Decision Tree method is considered good in handling data imbalance. As for the acquisition of the G-Mean value, the Cost Sensitive Decision Tree obtained 66.36% which showed four times more results than the Decision Tree with a total of 16.75%.

Keywords: Decision Tree, Cost Sensitive Learning, Internet Movie Database

PRAKATA

Puji dan syukur dipanjatkan atas kehadirat Allah SWT yang telah memberikan rahmat dan karunianya sehingga dapat menyelesaikan penulisan Tugas Akhir dengan judul "Pendekatan Algoritma *Cost Sensitive Decision Tree* pada Klasifikasi Film berdasarkan Perolehan Kompilasi dari *Internet Movie Database* (Imdb)". Penyusunan ini bertujuan untuk memenuhi persyaratan dalam rangka menyelesaikan program sarjana strata-1 Statistika di Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat. Penyusunan ini juga tidak terlepas dari bantuan dan dukungan oleh berbagai pihak terkait. Oleh karena itu, disampaikan ucapan rasa terima kasih yang sebesar-besarnya kepada pihak terkait yang terlampir sebagai berikut.

1. Orang tua, adik, dan keluarga besar yang senantiasa mendukung serta mendoakan untuk penyelesaian Tugas Akhir;
2. Ibu Yuana Sukmawaty, S.Si., M.Si. dan Ibu Selvi Annisa, S.Si., M.Si selaku dosen pembimbing yang telah bersedia meluangkan waktu, tenaga, dan pikiran dalam pelaksanaan penelitian serta penyelesaian Tugas Akhir;
3. Bapak Oni Soesanto, S.Si., M.Si. dan Bapak Sigit Dwi Prabowo, S.Mat., M.Stat. selaku dosen penguji yang telah memberikan masukan, saran, dan bantuan dalam perbaikan penulisan Tugas Akhir;
4. Koordinator Program Studi Statistika FMIPA ULM beserta seluruh dosen dan staf yang telah memberikan ilmu, motivasi, nasihat, kelancaran, dan mendukung selama masa perkuliahan;
5. Teman-teman Statistika FMIPA ULM Angkatan 2020 khususnya anak-anak Gaada Yang Tahu yang selalu ada, mendukung, memotivasi satu sama lain dalam penyelesaian Tugas Akhir;
6. Terakhir adalah diri saya sendiri yang sudah mampu berjuang untuk menghilangkan rasa malas dan sering tunda-menunda, sehingga bisa menyelesaikan Tugas Akhir dengan tepat waktu.

Penulisan Tugas Akhir ini tentunya masih jauh dari kata sempurna, untuk itu sangat diharapkan saran, masukkan, dan tambahan yang dapat membangun untuk penyempurnaan Tugas Akhir ini. Meskipun demikian, diharapkan penulisan ini dapat bermanfaat bagi semua yang tertarik dengan penelitian ini.

Banjarbaru, 18 Desember 2023

Aldy Nur Pratama

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

H	: *entropy*
S	: himpunan kasus
A	: variabel
n	: jumlah partisi variabel A
$\|S_\alpha\|$	: jumlah kasus pada partisi ke-i
$\|S\|$	: jumlah kasus dalam S
p_i	: proporsi dari S_i terhadap S
$P(j\|x)$	: probabilitas kelas j berdasarkan kondisi pada x
$P(j\|x, M_i)$	: probabilitas kelas j berdasarkan kondisi pada x dari model *Decision Tree*
$\arg min_i$	: argumen minimum dari kelas i
$C(i, j)$	: *cost* dari kelas i dan j
TP	: *true positive*
TN	: *true negative*
FP	: *false positive*
FN	: *false negative*
DT	: *decision tree*
CSDT	: *cost sensitive decision tree*