SKRIPSI

ANALISIS KESUKSESAN FILM DENGAN DATA MINING

Teuku Hashrul

NPM: 2016730067

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

UNDERGRADUATE THESIS

MOVIE PROFIT ANALYSIS USING DATA MINING

Teuku Hashrul

NPM: 2016730067

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2020

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS KESUKSESAN FILM DENGAN DATA MINING

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 19 Juni 2020

Teuku Hashrul
NPM: 2016730067

LEMBAR PENGESAHAN

ANALISIS KESUKSESAN FILM DENGAN DATA MINING

Teuku Hashrul

NPM: 2016730067

Bandung, 19 Juni 2020

Menyetujui,

Pembimbing

Kristopher David Harjono, M.T.

Ketua Tim Penguji	Anggota Tim Penguji
Dr.rer.nat. Cecilia Esti Nugraheni	**Luciana Abednego, M.T.**

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS KESUKSESAN FILM DENGAN DATA MINING

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 19 Juni 2020

Teuku Hashrul
NPM: 2016730067

ABSTRAK

Film merupakan media komunikasi yang bersifat audio visual untuk menyampaikan suatu pesan atau cerita kepada penontonnya dan dijadikan sebagai media hiburan. Film yang dibuat ada karena kumpulan orang dibalik layar. Perusahaan film berlomba-lomba untuk membuat film yang memperoleh keuntungan maksimum. Terdapat banyak kemungkinan faktor yang dapat dijadikan film dapat memperoleh keuntungan maksimum. Penelitian ini adalah analisis kesuksesan film dengan *data mining* untuk memperoleh faktor-faktor yang dapat memprediksi kesuksesan sebuah film. Penelitian ini merupakan eksperimen untuk membandingkan beberapa metode *machine learning* seperti regresi dalam memprediksi kesuksesan sebuah film. Penelitian ini menggunakan bahasa pemrograman Python dan memanfaatkan beberapa *library* untuk melakukan *data mining*.

Data mining adalah proses menemukan suatu pola dari kumpulan data yang besar. Dengan *data mining*, manusia dapat menemukan sebuah informasi / pemahaman baru dari data. Kumpulan proses *data mining* adalah *Data cleaning* untuk menghilangkan *noise* dan data yang tidak konsisten. *Data integration* adalah proses menggabungkan data dari beberapa sumber. *Data transformation* adalah mengubah bentuk data menjadi lebih mudah dan relevan untuk kebutuhan analisis. *Data Selection* adalah proses memilih data yang relevan untuk kebutuhan analisis. *Data Mining* adalah tahap untuk menggunakan metode *Machine Learning* untuk menemukan pola dari sebuah data. *Pattern Evaluation* adalah tahap untuk memeriksa pola yang dihasilkan apakah menghasilkan kebenaran.

Penelitian ini menghasilkan informasi berupa hasil visualisasi data dari *dataset* film yang dianalisis. Perangkat lunak membaca *dataset* yang berupa data film dari tahun 2006 sampai 2016 lalu melakukan sekumpulan proses *data mining* seperti membersihkan data dengan menghilangkan *noise*, melakukan pengumpulan data tambahan media sosial seperti Youtube, melakukan integrasi data dengan menggabungkan *dataset* dengan data tambahan, melakukan pemilihan fitur, melakukan prediksi keuntungan menggunakan fitur yang sudah dipilih sebelumnya dan melakukan evaluasi terhadap model prediksi yang dibuat. Selain itu, perangkat lunak melakukan *clustering* untuk mengelompokkan data film pada *dataset* berdasarkan aktor dan *genre*.

Berdasarkan penelitian dan hasil evaluasi data yang dilakukan, dapat disimpulkan bahwa faktor yang dapat berpengaruh dalam kesuksesan film adalah jumlah penonton yang menyukai film tersebut (*votes*), besar biaya yang dikeluarkan untuk membuat film (*budget*) dan jumlah penonton *trailer* film pada situs Youtube. Selain itu, dapat disimpulkan bahwa selera penonton (*votes*) lebih berpengaruh dalam memperoleh kesuksesan film dibanding dengan selera kritikus (*review*).

Kata-kata kunci: *Data Mining* , *Machine Learning*, Regresi,*Clustering*, *Dataset*

ABSTRACT

Movie is an audio visual media to communicate and deliver some story to its viewer. People tend to watch movie as an entertainment purpose. There are movie makers in movie production house who wants to make their movies gain maximum profit. There are many possible factor that affects movie to gain maximum profit. This research is about movie profit analysis using data mining to acquire insights about movie industry and factor that will be used as a predictor to predict profit of a movie. There are several programs that will be made to collect, clean, select, analyze and evaluate movie data. This research also will do experiment to compare Machine Learning techniques such as regression to predict profit of a movie. This research will use Python as a programming language and several libraries to help develop Data Mining script.

Data Mining is a process of discovering interesting patterns and knowledge from large amounts of data. Data mining can help human to obtain new knowledge and information from data. There are several process in data mining such as data cleaning to remove noise data. Data integration is a process to combine data from multiple sources. Data transformation is a process to transform data to become more readable and more relevant for analysis. Data selection is a process to choose feature that correlates and more relevant to the model. Data mining is a process to use Machine Learning methods to find interesting patterns from a data. Pattern evaluation is a process to validate the Machine Learning model.

This research will produce interesting insights and data visualization from the choosen dataset. the program will read the dataset of a popular movie from 2006 to 2016. This research also will implementing data mining process such as remove noise from dataset, collect additional social media data features such as Youtube, integrate the additional data and the original dataset, select the most relevant data feature, do some predictive analysis to predict movie profit and evaluate the model. This dataset also will be clustered by actor and genre that similar.

According to the research and the data analysis that has been done, it could be concluded that the most important factor to increase the possibility to gain maximum profit from a movie are how many viewer that liked the movie (votes), how much the production cost (budget) and how many movie trailer views gained in Youtube. This research also conclude that how many viewer that liked the movie (votes) is more important factor that acquiring a good grade from professional reviewer.

Keywords: Data Mining, Machine Learning, Regression, Clustering, Dataset

Skripsi ini saya persembahkan kepada ibunda dan almarhum ayahanda . . .

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan yang Maha Esa karena atas berkat dan rahmat-Nya penulis berhasil menyelesaikan penyusunan skripsi ini yang berjudul "Analisis Kesuksesan Film Menggunakan Dengan Data Mining". Penulis menyadari bahwa penyusunan skripsi ini tidak akan berhasil tanpa dukungan doa dari berbagai pihak, oleh karena itu penulis ingin mengucapkan terima kasih kepada:

- Orang tua penulis yang telah bekerja keras mendoakan, mendukung dan memenuhi kebutuhan penulis selama proses penyusunan skripsi.
- Kedua kakak penulis yang selalu mendoakan dan mendukung penulis.
- Paman dan Tante yang selalu mendorong penulis untuk berkembang dan belajar.
- Bapak Kristopher David Harjono, M.T.yang telah memberikan bimbingan dan arahan selama proses penyusunan skripsi.
- Sahabat semasa kuliah Giovanni, Timothy, Rashif, Naofal, Alif dan Shafira
- Sahabat semasa SMA Evander, Dewi, Reky, Cory, Annissa, Gintar dan Agung.
- Teman-teman Himpunan Mahasiswa Program Studi Teknik Informatika (HMPSTIF) sebagai tempat pengembangan diri.

Penulis berharap semoga skripsi ini dapat berguna bagi segenap pihak yang berkepentingan. Akhir kata, penulis memohon maaf apabila terdapat kekurangan dalam hasil penyusunan skripsi ini.

Bandung, Juni 2020

Penulis

DAFTAR ISI

Kata Pengantar xv

Daftar Isi xvii

Daftar Gambar xxi

Daftar Tabel xxv

1 Pendahuluan 1
- 1.1 Latar Belakang . . . 1
- 1.2 Rumusan Masalah . . . 1
- 1.3 Tujuan . . . 2
- 1.4 Batasan Masalah . . . 2
- 1.5 Metodologi . . . 2
- 1.6 Sistematika Pembahasan . . . 3

2 Landasan Teori 5
- 2.1 Penelitian Kaggle . . . 5
- 2.2 Measuring the Central Tendency [1] . . . 6
  - 2.2.1 Mean . . . 7
  - 2.2.2 Median . . . 7
  - 2.2.3 Modus . . . 7
- 2.3 Data Mining [1] . . . 7
  - 2.3.1 Data Cleaning . . . 9
  - 2.3.2 Data Integration . . . 9
  - 2.3.3 Data Reduction . . . 9
  - 2.3.4 Data Transformation . . . 10
  - 2.3.5 Data Selection . . . 10
- 2.4 Machine Learning [2] . . . 11
  - 2.4.1 Regression [3] . . . 11
  - 2.4.2 Classification . . . 13
  - 2.4.3 Clustering . . . 17
- 2.5 Data Visualization [1] . . . 19
  - 2.5.1 Boxplot . . . 20
  - 2.5.2 Histogram . . . 21
  - 2.5.3 Scatter Plot . . . 21
  - 2.5.4 Pie Chart . . . 22
- 2.6 One Hot Encoding [2] . . . 23
- 2.7 Web Scraping [4] . . . 23
- 2.8 Istilah Dalam Bisnis Film . . . 24

3 Analisis 25
- 3.1 Analisis Masalah . . . 25

3.2 Tahap Pengerjaan Penelitian . . . 26
3.2.1 Analisis Data Utama . . . 26
3.2.2 Analisis Data Tambahan IMDB . . . 26
3.2.3 Analisis Data Sosial Media . . . 26
3.3 Analisis Penerapan Data Cleaning . . . 27
3.4 Analisis Penerapan Data Transformation . . . 27
3.4.1 Attribute Construction . . . 28
3.4.2 Normalization . . . 28
3.5 Analisis Penerapan Data Integration . . . 29
3.6 Analisis Penerapan Data Selection . . . 30
3.7 Analisis Penerapan Linear Regression . . . 31
3.8 Analisis Penerapan Polynomial Regression . . . 33
3.9 Analisis Penerapan Evaluasi Regresi . . . 36
3.10 Analisis Penerapan K-NN . . . 37
3.11 Analisis Penerapan Evaluasi Klasifikasi . . . 38
3.12 Analisis Penerapan K-Means . . . 39
3.13 Analisis Penerapan Agglomerative . . . 40
3.14 Analisis Penerapan Evaluasi Clustering . . . 42
3.15 Analisis Visualisasi Data . . . 43
3.16 Analisis Web Scrapping . . . 45

4 Implementasi **47**
4.1 Lingkungan Perangkat Keras . . . 47
4.2 Lingkungan Perangkat Lunak . . . 47
4.3 Deskripsi Dataset . . . 47
4.4 Proses Analisis Data Utama . . . 49
4.4.1 Data Cleaning . . . 50
4.4.2 Analisis Data Utama Menggunakan Visualisasi . . . 51
4.4.3 Data Selection . . . 61
4.4.4 Percobaan Prediksi Fitur Data Utama . . . 64
4.5 Proses Analisis Data Tambahan . . . 69
4.5.1 Data Collection . . . 70
4.5.2 Data Integration . . . 74
4.5.3 Data Transformation . . . 74
4.5.4 Analisis Data Tambahan Menggunakan Visualisasi . . . 76
4.5.5 Data Selection . . . 81
4.5.6 Percobaan Prediksi Fitur Data Tambahan . . . 87
4.6 Proses Analisis Data Sosial Media . . . 88
4.6.1 Data Collection . . . 89
4.6.2 Data Integration . . . 95
4.6.3 Data Transformation . . . 96
4.6.4 Analisis Data Media Sosial Menggunakan Visualisasi . . . 97
4.6.5 Data Selection . . . 100
4.6.6 Percobaan Prediksi Fitur Data Sosial Media . . . 104
4.7 Analisis Clustering . . . 110
4.7.1 Clustering Berdasarkan Aktor . . . 110
4.7.2 Clustering Berdasarkan Genre . . . 117
4.8 Prediksi Berdasarkan Cluster . . . 122

5 KESIMPULAN DAN SARAN **127**
5.1 Kesimpulan . . . 127
5.2 Saran . . . 128

Daftar Referensi 129

A Kode Program 131

B Hasil Eksperimen 167

DAFTAR GAMBAR

2.1 Proses Data Mining . . . 8
2.2 Proses Klasifikasi . . . 14
2.3 Ilustrasi Decision Tree . . . 15
2.4 Boxplot dataset iris . . . 20
2.5 Histogram dataset iris . . . 21
2.6 Scatter plot dengan korelasi positif (a) dan negatif (b) . . . 22
2.7 Scatter plot tanpa korelasi . . . 22
2.8 Pie chart dataset iris . . . 23

3.1 Contoh dataset yang memiliki data kotor . . . 27
3.2 Kode analisis data cleaning . . . 27
3.3 Kode contoh Attribute Construction . . . 28
3.4 Kode contoh *Normalization* . . . 29
3.5 Kode contoh *integration* . . . 30
3.6 Kode contoh *pearson correlation* . . . 31
3.7 Kode contoh *linear regression* . . . 33
3.8 Kode contoh *polynomial regression* . . . 35
3.9 Perbandingan kurva linear dan polinom . . . 35
3.10 Kode contoh *evaluasi regresi* . . . 37
3.11 Kode contoh *K-NN* . . . 38
3.12 Kode contoh evaluasi klasifikasi . . . 39
3.13 Kode contoh *clustering K-Means* . . . 40
3.14 Dendogram example . . . 42
3.15 Kode contoh *agglomerative clustering* . . . 42
3.16 Kode contoh evaluasi clustering . . . 43
3.17 Contoh Analisis Histogram . . . 44
3.18 Contoh Analisis Boxplot . . . 44
3.19 Kode contoh visualisasi . . . 45
3.20 Flowchart Octoparse . . . 45
3.21 Tampilan utama Octoparse . . . 46
3.22 Tampilan scrapping pada Octoparse . . . 46

4.1 Ilustrasi sebuah baris film pada dataset . . . 48
4.2 Barchart deteksi jumlah baris yang kolomnya terdapat null . . . 50
4.3 Wordcloud kolom genre pada dataset . . . 51
4.4 Piechart persebaran film berdasarkan genre . . . 52
4.5 Top 10 Kombinasi genre dengan pendapatan kotor (Revenue) Terbaik . . . 53
4.6 Hubungan korelasi selera penonton dengan *review* . . . 54
4.7 Distribusi nilai kolom metascore . . . 54
4.8 Distribusi nilai kolom rating . . . 55
4.9 Perbandingan Distribusi nilai kolom review . . . 55
4.10 Histogram kolom metascore dan rating . . . 56
4.11 Distribusi nilai kolom runtime . . . 56

4.12 Distribusi kelas nilai runtime . . . 57
4.13 Hubungan korelasi runtime dan votes . . . 57
4.14 Distribusi nilai kolom votes . . . 58
4.15 Histogram distribusi kolom votes . . . 58
4.16 Distribusi nilai kolom year . . . 59
4.17 Histogram distribusi kolom year . . . 60
4.18 Line plot Jumlah Revenue dari tahun ke tahun . . . 60
4.19 Histogram Revenue Tahun 2011 . . . 61
4.20 Analisis Korelasi antara situs review dengan revenue . . . 62
4.21 Analisis Korelasi runtime year dan revenue . . . 62
4.22 Scatter plot votes dan revenue . . . 63
4.23 Pengujian perbandingan korelasi tiap fitur dengan revenue menggunakan pearson correlation . . . 63
4.24 Plot linear regression prediksi revenue dengan votes . . . 65
4.25 Distribusi nilai *squared error* . . . 65
4.26 Plot polynomilal regression prediksi revenue dengan votes . . . 66
4.27 Distribusi nilai squared error prediksi revenue menggunakan *polynomial regression* 67
4.28 Distribusi nilai squared error prediksi revenue dan perbandigan tiap algoritma regresi yang digunakan . . . 67
4.29 Fitur advance search pada situs IMDB . . . 70
4.30 Pengaturan octoparse pada hasil pencarian advanced search . . . 71
4.31 Detail page sebuah film pada situs IMDB . . . 71
4.32 Flowchart cara kerja scraping budget IMDB menggunakan Octoparse . . . 72
4.33 Flowchart cara kerja scraping budget IMDB menggunakan library IMDBpy . . . 73
4.34 Distribusi nilai kolom budget . . . 76
4.35 Distribusi nilai kolom profit . . . 76
4.36 Perbandingan film yang rugi dan untung pada dataset . . . 77
4.37 Distribusi nilai kolom ROI . . . 77
4.38 Histogram kolom Budget . . . 78
4.39 Histogram kolom Profit . . . 78
4.40 Histogram kolom ROI . . . 79
4.41 Line plot akumulasi profit dan budget semua film tiap tahun . . . 79
4.42 Top 10 Kombinasi genre dengan votes terbaik . . . 80
4.43 10 kombinasi genre dengan perbandingan budget dan revenue tertinggi (profit) . . . 81
4.44 Analisis korelasi kolom Budget dan Revenue . . . 82
4.45 Pengujian perbandingan korelasi tiap fitur dengan revenue menggunakan pearson correlation ke-2 . . . 83
4.46 Analisis korelasi situs review dan profit . . . 83
4.47 Analisis korelasi runtime votes dan profit . . . 84
4.48 Analisis korelasi budget dan profit . . . 84
4.49 Pengujian perbandingan korelasi tiap fitur dengan profit menggunakan pearson correlation ke-2 . . . 85
4.50 Analisis korelasi situs review dan ROI . . . 85
4.51 Analisis korelasi runtime votes dan roi . . . 86
4.52 Analisis korelasi budget dan profit . . . 86
4.53 Pengujian perbandingan korelasi tiap fitur dengan ROI menggunakan pearson correlation ke-2 . . . 87
4.54 Contoh JSON hasil request search youtube menggunakan API . . . 90
4.55 Contoh JSON hasil request seach youtube menggunakan API . . . 91
4.56 Contoh JSON Error Youtube Quota Exceeded . . . 91
4.57 Tampilan octoparse untuk scraping youtube view . . . 92

4.58 Contoh fitur Hashtag Instagram . . . 93
4.59 Boxplot view count . . . 97
4.60 Hashtag Instagram title boxplot . . . 98
4.61 Hashtag Instagram actor boxplot . . . 99
4.62 Analisis tren *youtube view trailer* . . . 100
4.63 Analisis tren *Instagram Hashtag* . . . 100
4.64 Pearson correlation fitur prediktor terhadap *revenue* . . . 101
4.65 Pearson correlation fitur prediktor terhadap *revenue* . . . 102
4.66 Pearson correlation fitur prediktor terhadap *profit* . . . 103
4.67 Pearson correlation fitur prediktor terhadap *profit* . . . 104
4.68 Distribusi nilai *squared errors* $((ypred - ytest)^2)$ *Linear regression* pada prediksi *revenue* menggunakan fitur *votes,budget* dan *viewCount* . . . 105
4.69 Distribusi nilai *squared errors* $((ypred - ytest)^2)$ *Polynomial regression* pada prediksi *revenue* menggunakan fitur *votes,budget* dan *viewCount* . . . 106
4.70 Distribusi nilai *squared errors* $((ypred - ytest)^2)$ *Linear regression* pada prediksi *profit* menggunakan fitur *votes,budget* dan *viewCount* . . . 108
4.71 Distribusi nilai *squared errors* $((ypred - ytest)^2)$ *Polynomial regression* pada prediksi *profit* menggunakan fitur *votes,budget* dan *viewCount* . . . 109
4.72 Kualitas Clustering Aktor berdasarkan jumlah menggunakan *Silhouette Score* . . . 111
4.73 Distribusi nilai *intracluster* pada beberapa percobaan clustering actor . . . 112
4.74 Distribusi jumlah anggota tiap *cluster* aktor . . . 113
4.75 Pengujian sebuah cluster merepresentasikan satu aktor dominan . . . 114
4.76 Wordcloud distribusi aktor cluster 13 . . . 115
4.77 Wordcloud distribusi aktor cluster 28 . . . 115
4.78 Wordcloud distribusi aktor cluster 161 . . . 115
4.79 Pengujian aktor dominan memiliki *genre* favorit pada tiap *cluster* . . . 116
4.80 Pengujian *genre* favorit pada tiap *cluster* juga berkontribusi tinggi untuk keuntungan 117
4.81 Kualitas Clustering Genre berdasarkan jumlah menggunakan *Silhouette Score* . . . 118
4.82 Distribusi nilai *intracluster* pada beberapa percobaan clustering *genre* . . . 119
4.83 Perbandingan jumlah *cluster* yang memiliki aktor utama dan tidak . . . 121
4.84 Perbandingan jumlah *cluster* yang aktor utama berkontribusi berdasarkan *revenue* pada kombinasi *genre* . . . 122
4.85 Segmented Cluster Flowchart . . . 123
4.86 Distribusi R2 hasil prediksi revenue berdasarkan *cluster genre* . . . 124
4.87 Distribusi R2 hasil prediksi revenue berdasarkan *cluster* aktor . . . 124

B.1 20 Kombinasi genre dengan akumulasi *revenue* tertinggi . . . 167
B.2 20 Kombinasi genre yang paling banyak dibuat . . . 168
B.3 Distribusi revenue semua kombinasi genre . . . 169
B.4 Distribusi votes semua kombinasi votes . . . 170
B.5 Distribusi Jumlah anggota hasil cluster aktor . . . 171
B.6 Jumlah perbandingan kelompok aktor yang genre favoritnya berkontribusi tertinggi pada *profit* . . . 172
B.7 Jumlah perbandingan kelompok aktor yang genre favoritnya berkontribusi tertinggi pada *ROI* . . . 172

DAFTAR TABEL

3.1 Tabel contoh yang perlu diubah . . . 28
3.2 Tabel dengan kolom baru setelah *Data Transformation* . . . 28
3.3 Tabel Dataset yang ingin dinormalisasi . . . 29
3.4 Tabel dataset setelah dinormalisasi . . . 29
3.5 Tabel kumpulan buku . . . 29
3.6 Tabel pengarang . . . 30
3.7 Tabel buku dan pengarang sudah tergabung . . . 30
3.8 Tabel data performa belajar siswa . . . 30
3.9 Tabel hasil perhitungan pearson . . . 31
3.10 Tabel dataset regresi . . . 32
3.11 Perhitungan komponen konstanta dan koefisien . . . 32
3.12 Tabel perbandingan volume penjualan prediksi dan volume penjualan asli . . . 33
3.13 Tabel dataset polinomial . . . 33
3.14 perhitungan komponen matriks hubungan polinom orde 2 . . . 34
3.15 Perbandingan prediksi volume penjualan (Y) polynomial regression . . . 35
3.16 Tabel perhitungan R2 berdasarkan contoh prediksi Linear Regression . . . 36
3.17 Tabel dataset k-nearest neighbor . . . 37
3.18 Tabel perhitungan euclidean distance dengan data baru . . . 37
3.19 Tabel 3 tetangga terdekat berdasarkan perhitungan euclidean distance . . . 38
3.20 Tabel dataset evaluasi klasifikasi . . . 38
3.21 Tabel dataset k-Means . . . 39
3.22 Tabel perhitungan k-Means iterasi ke-1 . . . 39
3.23 Tabel perubahan centroid iterasi 2 . . . 40
3.24 Tabel perhitungan k-Means iterasi ke-2 . . . 40
3.25 Tabel dataset perhitungan agglomerative . . . 40
3.26 Tabel perhitungan euclidean distance iterasi pertama . . . 41
3.27 Tabel perubahan nilai cluster iterasi 1 . . . 41
3.28 Tabel perhitungan euclidean distance iterasi kedua . . . 41
3.29 Tabel perubahan nilai cluster iterasi 2 . . . 41
3.30 Tabel dataset contoh yang akan di evaluasi . . . 42
3.31 Tabel dataset analisis visualisasi . . . 43

4.1 Deskripsi dataset . . . 49
4.2 5 Contoh Data pada dataset . . . 49
4.3 Tabel sebelum *split* . . . 51
4.4 Tabel setelah *split* . . . 51
4.5 Film *outlier* berdasarkan *votes* . . . 59
4.6 Tabel Percobaan Regresi Linear untuk prediksi revenue menggunakan votes . . . 64
4.7 Tabel Percobaan Regresi Polinom untuk prediksi revenue menggunakan votes . . . 66
4.8 Tabel perbandingan skor akurasi r2 menggunakan regresi . . . 67
4.9 Sampel contoh data hasil scraping budget IMDB menggunakan Octoparse . . . 72
4.10 Tabel 10 sampel hasil data scrapping menggunakan IMDBpy . . . 74

4.11 Tabel sampel data integrasi antara dataset dengan budget 74
4.12 Tabel sampel beberapa konversi budget pada dataset 75
4.13 Tabel 5 sampel pembuatan kolom profit pada dataset 75
4.14 Tabel 5 sampel pembuatan kolom roi pada dataset 75
4.15 Tabel Skor R2 Revenue 87
4.16 Tabel sampel percobaan prediksi Revenue menggunakan regresi fitur votes dan budget 88
4.17 Tabel skor R2 prediksi profit 88
4.18 Tabel sampel percobaan prediksi profit menggunakan budget dan votes dan perbandingannya dengan profit asli 88
4.19 10 sampel url yang digenerate 92
4.20 10 sampel hasil *scrapping youtube view* menggunakan *octoparse* 93
4.21 10 Sampel URL Instagram Hashtag dengan Menggunakan Judul Film 94
4.22 10 Sampel hasil *scraping* data hashtag instagram menggunakan *octoparse* 94
4.23 10 Sampel hasil pengumpulan data *hashtag* aktor utama setiap film pada dataset . 95
4.24 10 Sampel hasil *scraping* data *hashtag* aktor dengan *Octoparse* 95
4.25 Sampel data integrasi antara *dataset* dengan data sosial media 96
4.26 Operasi dan contoh data *transformation* pada data media sosial 96
4.27 5 Sampel *dataset* setelah konversi *data transformation* 96
4.28 5 Film *outlier* dengan *View Count* Terbanyak 98
4.29 5 Film *outlier* berdasarkan Hashtag judul terbanyak 98
4.30 5 Film *outlier* berdasarkan *Actor Hashtag* terbanyak 99
4.31 Tabel percobaan *Linear Regression* untuk prediksi *revenue* menggunakan fitur *votes,budget* dan *viewCount* 105
4.32 Hasil nilai R2 *linear regression* pada prediksi revenue menggunakan data sosial media 105
4.33 Tabel percobaan *Polynomial Regression* untuk prediksi *revenue* menggunakan fitur *votes,budget* dan *viewCount* 106
4.34 Hasil nilai R2 *polynomial regression* pada prediksi revenue menggunakan data sosial media 107
4.35 Tabel percobaan *Linear Regression* untuk prediksi *profit* menggunakan fitur *votes,budget* dan *viewCount* 107
4.36 Hasil nilai R2 *linear regression* pada prediksi *profit* menggunakan data sosial media 108
4.37 Tabel percobaan *Polynomial Regression* untuk prediksi *profit* menggunakan fitur *votes,budget* dan *viewCount* 109
4.38 Contoh Aktor pada *Dataset* 110
4.39 Contoh Hasil One Hot Encoding Aktor 110
4.40 Perbandingan waktu *clustering* menggunakan KMeans dan Agglomeratives 112
4.41 Contoh Anggota *Cluster* 13,28 dan 161 hasil *cluster Actor* 113
4.42 Contoh jumlah kontribusi aktor utama pada *cluster* 13 , 28 dan 161 117
4.43 Contoh Genre pada *Dataset* 118
4.44 Contoh Hasil One Hot Encoding Genre 118
4.45 Contoh Anggota *Cluster* 13,28 dan 62 hasil *cluster Genre* 120
4.46 Hasil evaluasi prediksi Revenue menggunakan *Votes* dengan R2 Clustering Genre . 123
4.47 Hasil evaluasi prediksi Revenue menggunakan *Votes* dengan R2 Clustering Actor . 124

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Film merupakan media komunikasi yang bersifat audio visual untuk menyampaikan suatu pesan kepada penontonnya. Keberadaan film membuat masyarakat menjadikan film sebagai media hiburan. Beragam cara dapat dilakukan untuk menikmati sebuah film yaitu datang ke bioskop, membeli kaset DVD dan *streaming* menggunakan aplikasi *desktop* dan *smartphone*.

Film yang dibuat ada karena ada kumpulan orang di balik layar yang bekerja untuk membuatnya. Terdapat beragam perusahaan produksi film yang berlomba-lomba untuk membuat film yang dapat memperoleh keuntungan maksimum. Dengan menciptakan film yang sesuai dengan keinginan penontonnya,maka peluang keuntungan yang diperoleh pun akan semakin meningkat dan dapat menutup *budget* yang digunakan sebelumnya untuk biaya produksi.

Berdasarkan penelitian analisis data film yang ada sebelumnya pada *kaggle*, data film memiliki beberapa atribut umum yaitu judul,*genre*, *rating*, keuntungan, *budget*, nilai *review* dari situs internet dan aktor yang terlibat dan lama tayang. Data film yang ada digunakan untuk membantu menganalisis sifat dari film. Penelitian yang dilakukan adalah analisis prediksi nilai IMDB *score*, yaitu situs yang berisi data film.

Genre adalah sebutan untuk membedakan berbagai jenis film. Sebuah film memiliki satu atau beragam *genre*. Terdapat banyak film yang menggunakan kombinasi dari beberapa *genre*. *Genre* yang ada berupa *action, adventure, animation ,drama ,comedy ,horror ,romance* dan lain-lain.

Terdapat banyak kemungkinan faktor yang dapat dijadikan sebuah film dapat memperoleh keuntungan maksimum. Faktor kesuksesan film berupa faktor *rating* dari situs *review* film, nama aktor yang terlibat, nama sutradara yang terlibat dan jumlah *budget* yang dikeluarkan. Salah satu faktor dan kombinasi beberapa faktor dapat memengaruhi kesuksesan film.

Berdasarkan uraian diatas, akan dilakukan sebuah penelitian mengenai data film. Penelitian ini adalah analisis kesuksesan film dengan *data mining* untuk memperoleh faktor-faktor yang ada dapat memprediksi kesuksesan sebuah film. Dari faktor yang diperoleh, maka akan diprediksi *revenue*/pendapatan sebuah film berdasarkan data film yang sudah ada sebelumnya.

Pada penelitian ini dibuat sekumpulan perangkat lunak yang digunakan untuk mengumpulkan, membersihkan data, analisis ,pembuatan model, evaluasi kerja model dan visualisasi data. Perangkat lunak yang dibuat akan membantu menganalisis data film yang digunakan. Pembuatan perangkat lunak akan menggunakan bahasa pemrograman *Python* dan memanfaatkan beberapa *library* dari *Python.Pandas* digunakan untuk integrasi data. *Sci-kit learn* digunakan untuk implementasi regresi, teknik *clustering* dan *classification* untuk memprediksi keuntungan sebuah film. Penelitian ini akan melakukan eksperimen untuk membandingkan beberapa metode *machine learning* dalam memprediksi kesuksesan sebuah film.

1.2 Rumusan Masalah

Berkaitan dengan identifikasi masalah yang ada pada deskripsi di atas, masalah-masalah yang ada dapat dirumuskan sebagai berikut.

- Apa saja faktor yang dapat digunakan untuk menentukan kesuksesan sebuah film ?
- Bagaimana langkah dalam melakukan analisis kesuksesan film dengan *data mining* ?
- Bagaimana hasil pengujian pada penelitian ini ?

1.3 Tujuan

Tujuan yang ingin dicapai dari penelitian ini adalah:

- Mengeksplorasi data yang dikumpulkan
- Membuat perangkat lunak yang dapat menggunakan metode *data mining* untuk melakukan analisis faktor-faktor yang dapat berpengaruh pada kesuksesan film
- Menguji metode-metode yang digunakan pada penelitian ini

1.4 Batasan Masalah

Pelaksanaan penelitian ini permasalahannya dibatasi pada:

1. Dataset yang digunakan pada penelitian berasal dari situs penyedia dataset seperti *Kaggle*
2. Data yang digunakan merupakan data IMDB dari tahun 2006 sampai 2016
3. Penelitian ini akan membuat mengimplementasikan tahapan *data mining* dengan memanfaatkan *library* dari *Python*
4. Penelitian ini tidak membuat antarmuka perangkat lunak dalam memrediksi kesuksesan film sehingga penjabaran dalam penelitian ini menggunakan visualisasi data

1.5 Metodologi

Langkah-langkah yang akan dilakukan dalam melakukan penelitian ini, yaitu:

1. Melakukan studi literatur dengan mencari jurnal, *paper* mengenai penelitian sejenis dari berbagai sumber untuk membantu penulis dalam menulis
2. Melakukan studi literatur langkah-langkah teknik *data mining* untuk memahami konsep
3. Melakukan studi literatur mengenai metode-metode *machine learning* yaitu regresi, *clustering* dan *classification* yang relevan
4. Melakukan studi literatur mengenai teori dan implementasi visualisasi data seperti *histogram,scatter plot, box plot* untuk membantu mengetahui sifat data yang dikumpulkan menggunakan *matplotlib*
5. Melakukan penelitian sejenis mengenai industri perfilman untuk mengetahui relevansi antar faktor yang ada
6. Mempelajari bahasa pemrograman *Python* dan beberapa *library* dari *Python* seperti *Pandas*,*Sci-Kit learn* dan *matplotlib*
7. Mencari sumber data yang relevan untuk melakukan pengumpulan data dari situs *review* film dan media sosial

8. Melakukan integrasi data dari sumber yang digunakan
9. Melakukan eksplorasi untuk menemukan sifat data
10. Melakukan analisis data secara statistik dengan teknik visualisasi data
11. Menerapkan metode-metode *machine learning* regresi, *clustering* dan *classification*
12. Melakukan pengujian dan eksperimen
13. Menulis dokumen skripsi

1.6 Sistematika Pembahasan

Sistematika penulisan ini berguna untuk memberikan gambaran secara umum mengenai penelitian yang akan dibuat. Berikut ini adalah uraian dari sistematika pembahasan :

- Bab 1. Pendahuluan, membahas tentang latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, metode penelitian dan sistematika pembahasan mengenai skripsi
- Bab 2. Landasan Teori, membahas pengertian *data mining*, langkah-langkah *data mining*, algoritma *Machine Learning*, visualisasi data dan *web scraping*
- Bab 3. Analisis, membahas tentang bagaimana implementasi teori yang dijelaskan pada bab sebelumnya. Pada bab ini juga menjelaskan implementasi menggunakan *library Python*.
- Bab 4. Implementasi, membahas tentang deskripsi *dataset*, hasil analisis data, pengujian prediksi dan interpretasi pola menarik dari visualisasi.
- Bab 5. Kesimpulan dan Saran membahas kesimpulan yang diperoleh setelah melakukan analisis data serta saran yang dapat diberikan untuk pengembangan lebih lanjut tentang analisis data film