ANALISA CLUSTER APLIKASI PADA APP STORE DENGAN MENGGUNAKAN METODE K-MEANS

Sofian Wira Hadi[1], Muhammad Fahmi Julianto[2], Syaifur Rahmatullah[3], Windu Gata[4]
STMIK Nusa Mandiri[1,2,3,4],
14002361@nusamandiri.ac.id[1], 14002389@nusamandiri.ac.id[2] , syaifur.syl@nusamandiri.ac.id[3]
windu@nusamandiri.ac.id[4]

Abstrak- Bagi para pengguna iphone, salah satu tempat untuk mengunduh ratusan ribu aplikasi android adalah *App Store*. Aplikasi-aplikasi iOS di bagi menjadi ketegori-ketegori yang unik. Di dalam aplikasi iOS ini terdapat aplikasi-aplikasi yang berbayar dan gratis. Dengan kategori tersebut pengguna bisa dengan mudah mencari aplikasi yang dibutuhkannya. Pada penelitian ini kami menggunakan metode K-Means untuk melihat ciri-ciri dari atribut yang ada. Dataset *App Store* diambil dari website resmi kaggle. Tujuan dari penelitian ini adalah untuk menganalisa hasil cluster dari *K-Means.* Hasil dari penelitian adalah adanya sebuah cluster yang memiliki ciri-ciri aplikasi yang ideal, yaitu nilai user rating tinggi, harga yang cukup lumayan dan memiliki ukuran aplikasi yang rendah.
Kata Kunci : *Clustering*, K-Means, App Store, Kaggle

Abstract *- For iphone users, one place to download hundreds of thousands of android applications is the App Store. IOS applications are divided into unique categories. In this iOS application, there are paid and free applications. With these categories users can easily find the application they need. In this study we use the K-Means method to see the characteristics of the existing attributes. The App Store dataset is taken from the official Kaggle website. The purpose of this study is to analyze the results of clusters from K-Means. The results of the study are the existence of a cluster that has the characteristics of an ideal application, namely a high user rating value, a pretty decent price and has a low application size*
Keywords: Clustering, K-Means, App Store, Kaggle

I. PENDAHULUAN

Bagi para pengguna iphone, salah satu tempat untuk mengunduh ratusan ribu aplikasi android adalah *App Store*. *App Store* adalah pasar *platform* distribusi aplikasi untuk iOS yang dikembangkan dan dikelola *AppleInc*. Layanan ini memungkinkan pengguna menjelajah dan mengunduh aplikasi yang dikembangkan dengan *Apple* iOS SDK aplikasi dapat diunduh langsung ke sebuah perangkat iOS atau komputer pribadi (*Macintosh* atau PC) Melalui *iTunes*.

Apple Store dan *google play* diluncurkan pada tahun 2008, dan sejak itu keduanya telah mengakumulasi lebih dari 1 juta yang dapat diunduh dan aplikasi yang dapat diunduh dan aplikasi dapat di nilai. Google mengumumkan bahwa ada 1,4 miliar perangkat android yang di aktifkan pada bulan September 2015 (Effendi & M Jorgi, 2018).

Pada *AppStore,* Aplikasi-aplikasi iOS di bagi menjadi ketegori-ketegori yang unik. Di dalam aplikasi iOS ini terdapat aplikasi-aplikasi yang berbayar dan gratis. Dengan kategori tersebut pengguna bisa dengan mudah mencari aplikasi yang dibutuhkannya.

Toko aplikasi seluler juga sangat menguntungkan. Set toko aplikasi seluler diproyeksikan bernilai 25 miliar USD pada tahun 2015(Martin et al., 2017). Keberhasilan aplikasi dilihat dari tingkat kebutuhan konsumen dalam mengadopsi perangkat *smartphone. Smartphone* ada sebelum peluncuran toko-toko ini, tetapi tidak sampai 2008 bahwa pengguna benar-benar mengeksploitas kekuatan komputasi ekstra mereka dan menghasilkan *fleksibilitas* melalui aplikasi yang dapat diunduh. Jumlah unduhan aplikasi seluler diseluruh dunia pada tahun 2017 adalah 178.1 Miliar. Terdapat 3,8 juta aplikasi seluler di *Google Play* store yang merupakan toko aplikasi terbesar dan ada 2 juta aplikasi seluler di *Apple App Store* yang merupakan toko terbesar pada quartal pertama 2018(Bozanta & Co, 2018). Namun tidak semua aplikasi yang diunggah akan mendapatkan unduhan dan rating yang tinggi , *Developer* harus memperhitungkan factor-faktor yang dapat mempengaruhi rating dan jumlah unduhan tersebut.

Pada penelitian sebelumnya memahami perbedaan masalah aplikasi pada platform Seperti *Google Play, App Store* dan windows(Bozanta & Co, 2018). Selanjutnya penelitian yang membahas tentang masalah Aplikasi Lintas-Platform dari ulasan Pengguna (Man et al., 2016).Penelitian lain juga menganalisa sentiment analis pada *App Store* dengan cara melihat ulasan(Sangani & Ananthanarayanan, 2013). Pada penelitian lain terdapat menggunakan data *Google Play.* Untuk menganalisa *cluster* aplikasi pada *Google Play Store* dengan menggunakan metode K-Mean (Effendi & M Jorgi, 2018). Metode K-Means berguna untuk melihat ciri-ciri dari aplikasi

berdasarkan atribut yang ada dan ini adalah salah satu *tekhnik* pada *data mining*.

Data Mining merupakan proses yang menggunakan teknik *statistic,* perhitungan, kecerdasan buatan dan Machine Learning untuk mengekstraksi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar (Febrianti et al., 2016). Dalam data mining terdapat sebuah metode yang digunakan untuk mengklaster data, yaitu K-Means.

K-Means merupakan algoritma *clustering* yang berulang-ulang. Algoritma K-Means dimulai dengan pemilihan secara acak, K disini merupakan banyaknya *cluster* yang ingin di bentuk(Putra & Wadisman, 2018). yang nantinya nilai-nilai K secara *random* untuk sementara nilai nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid,* atau *means*.

Pada penilitan ini penulis menggunakan data pada *App Store* dengan metode K-Means. Metode K-Means digunakan untuk mengelompokan data sesuai klaster yang dibuat. Metode klaster memiliki kelebihan yaitu, mudah untuk diimplemntasikan dan mampu mengelompokan data yang besar dan waktu komputasinya yang cepat dan efisien.

II. METODOLOGI PENELITIAN

1. Data Mining

Penelitian menggunakan tekhnik Data Mining yang memiliki arti yaitu *knowledge discovery* ataupun *pattern recognition* merupakan suatu istilah yang digunakan untuk mendapatkan pengetahuan yang tersembunyi dari kumpulan data yang berukuran sangat besar, tujuan utama data mining adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki (Setiawan, 2016). Proses *data mining* dalam menemukan hubungan yang berarti, pola dan tren dengan memeriksa data berukuran besar dalam suatu penyimpanan dengan menggunakan tekhnologi pengenalan pola, misalnya statistic dan matematika (Abdurrahman, 2016).

2. Data Set

Pada penelitian ini, dataset yang dipakai adalah kumpulan data tentang aplikasi pada *Apple App Store*, Dataset ini diambil pada website kagle. Dataset ini memiliki 11 ribu aplikasi pada App Store dan 16 kolom atribut. Sebelum data di pakai, data sudah dirapihkan dengan membuang missing value. Setelah itu data dinormalisasikan dengan metode *MaxMinScaling.*

Atribut yang di ambil untuk model K-Means ada sebanyak 3 atribut yaitu : *Size Byte*, *Price,* dan *User Rating*. Melihat dari ukuran aplikasi, harga, dan rating pengguna

3. Clustering dengan K-Means

Perhitungan jumlah Cluster dilakukan dengan perhitungan jarak data dengan centroid. Perhitungan ini dilakukan dengan menghitung jarak dari setiap nilai pada data dengan jarak centroidnya. Apabila data berdekatan dengan centroidnya., data tersebut akan dimasukan dengan centroid, data tersebut akan dimasukan ke cluster dengan nilai centroid tersebut.

4. Algoritma K-Means

K-Means *Cluster Analysis* merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam *cluster* yang lain (Sandi et al., 2018).

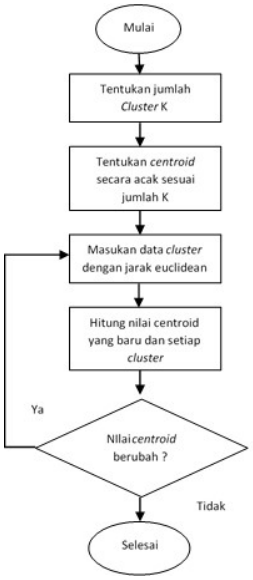
Data *clustering* menggunakan metode *K*-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut (Yudi Agusta, 2007):

a. Tentukan jumlah *cluster*
b. Alokasikan data ke dalam *cluster* secara random
c. Hitung centroid/rata-rata nilai data yang ada di masing-masing cluster
d. Alokasikan masing-masing data ke centroid/rata-rata terdekat
e. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan.

Perhitungan K-Means dimulai dengan menentukan jumlah cluster yaitu K *Cluster.* Selanjutnya, tentukan *centroid* secara acak sesuai dengan jumlah *cluster.* Setelah itu, masukan setiap data ke *centroid* yang terdekat engan jarak *Euclidean.* Maka akan terbentuk *cluster* sesuai jarak data dengan *centroid. Cluster* yang terbentuk akan dihitung kembali nilai Centroidnya. Setelah itu, data akan dimasukan lagi ke cluster *centroid* yang terdekat. Ulangi langkahilangkah diatas sampai nilai centroidnya tidak berubah lagi dan stabil (Effendi & M Jorgi, 2018).

5. RapidMiner

Rapid miner adalah sebuah aplikasi atau perangkat lunak untuk mengolah sebuah dataset, rapid miner juga merupakan aplikasi yang bersifat open source yang banyak diminati karna mudah untuk dipelajari dan rapidminer juga memiliki lisensi AGPL *(GNU Affero General Public License)* yang mampuh untuk mengolah data mining yang dikembangkan oleh Ralf Klinkenrberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund (Setiawan, 2016).



Gambar 1. Flowchart K-Means

6. CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining)* merupakan suatu metodologi data mining yang dikembangkan dan disusun oleh satu konsorsium sebuah perusahaan yang didirikan oleh komisi eropa pada tahun 1996 dan metodologi CRISP-DM sudah ditetapakn sebagai standar dalam pengelolahan data mining. Menurut larose, data mining mempunyai enam fase yaitu (Setiawan, 2016):

a. Fase Pemahaman Bisnis *(Business Understanding Phase)*
   Pada fase pertama ini adalah bertujuan untuk memahami sebuah bisnis yang sedang dijalankan atau yang akan dijalankan, kemudian setelah memahami apa yang terjadi selanjutnya menterjemahkan pengetahuan yang sudah didapatkan ke dalam pendefenisian masalah yang terjadi dalam data mining.
b. Fase Pemahaman Data *(Data Understanding Phase)*
   Fase ini terjadi dan dimulai dengan pengumpulan data yang selanjutnya akan dilanjutkan dengan proses untuk mendapatkan pemahaman dan pengetahuan yang dalam tentang data, mengindentifikasi masalah kualitas data, atau untuk mengetahui adanya bagian yang menarik dari sebuah data untuk dijadikan hipotesa untuk informasi yang bersifat terbatas.
c. Fase Pengolahan Data (*Data Preparation Phase*)
   Fase ini dimana semua kegiatan melakukan untuk membangun sebuah dataset akhir dalam artian data yang ingin diproses pada tahap pemodelan dari data yang belum diolah atau mentah. Fase ini dapat dilakukan berkali kali. Pada tahap ini juga mencakup pemilihan table, *record*, dan atribut atribut apa saja yang mau digunakan, termasuk proses pembersihan data yang ganda dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan.
d. Fase Pemodelan (Modeling *Phase)*
   Dalam tahapan ini akan dilakukan pemilihan dan penerapan berbagai model dan tekhnik dari beberapa parameternya akan meneysuaiakan untuk mendapatkan sebuah nilai yang optimal.
e. Fase Evaluasi (*Evaluation Phase*)
   Pada tahapan ini, model yang dilakukan pada tahapan sebelumnya sudah terbentuk dan sangat diharapkan memiliki kualitas yang baik jika dilihat dari sudut pandang analisa data. Pada tahapan ini dilakukan proses evaluasi terhadap model yang digunakan dilihat dari kefektifan dan kualitas model sebelum di implementasikan apakah model yang digunakan dapat mencapai tujuan yang ditetapkan seperti fase awal.
f. Fase Peneybaran (*Deployment Phase*)
   Pada tahap ini merupakan proses untuk mempresentasikn dari informasi yang sudah didapatkan atau data yang sudah diolah dalam bentuk khusus sehingga dapat digunakan oleh si pengguna. Tahap Deployment dapat berupa laporan sederhana atau mengimpelemntasikan proses data mining yang berulang dalam perusahaan, tahap deployment melibatkan konsumen karena hal yang sangat penting bagi konsumen untuk memahami apa yang terjadi dan memahami tindakan apa saja yang harus dilakukan untuk menggunakan model yang sudah dibuat.

III. HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Data yang digunakan berasal dari website kaggle, Dataset ini memiliki 11 ribu aplikasi pada App Store dan 16 kolom atribut. Atribut yang di ambil untuk model K-Means ada sebanyak 3 atribut yaitu : *Size Byte, Price,* dan *User Rating.* Melihat dari ukuran aplikasi, harga, dan rating pengguna, yang bertujuan adanya sebuah *cluster* yang memiliki ciri-ciri aplikasi yang ideal dengan atribut yang dipilih.

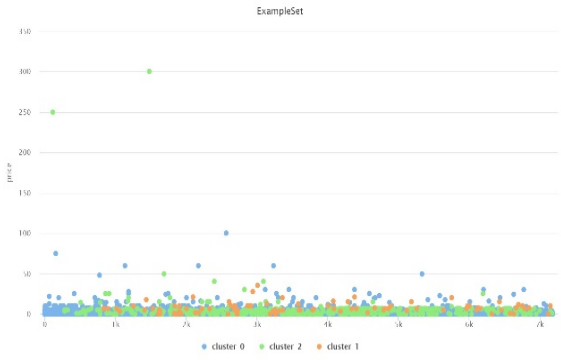
2. Praproses Data

Pada tahap ini data yang telah dikumpulkan akan masuk kedalam filter untuk melakukan proses *cleaning* pada data yang menjadi focus. Proses *cleaning* mencakup beberapa seperti membuang duplikasi data, memeriksa data yang inkonsistensi, serta memperbaiki beberapa kesalahan pada data dan selanjutnya memilih atribut yang dipilih untuk melakukan tahap pengolahan data selanjutnya.

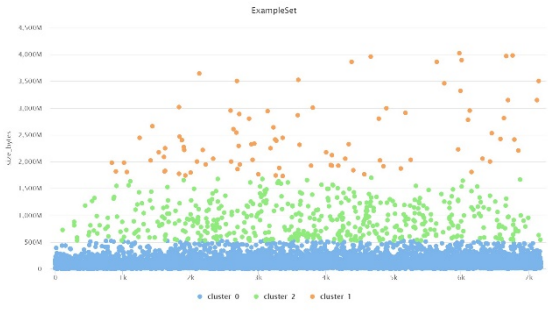
3. Proses Klastering dengan Algoritma K-Means

Pada penelitian ini, pengelompokan data dilakukan dengan metode K-Means, Selanjutnya, analisa *cluster* dilakukan dengan menggunakan *cluster Model.* Perhitungan K-Means terlebih dahulu dimulai dengan menentukan jumlah cluster yaitu K cluster. Selanjutanya menentukan centroid secara acak sesuai dengan jumlah cluster yang kiti pilih atau yang kita ingin gunakan di dalam penelitian. Setelah dipilih selanjutnya, masukan setiap dat ke dalam centroid yang paling terdekat dengan menggunakan jarak Euclidean. Maka nantinya akan terbentuk sebuah cluster sesuai dengan jarak data dengan *centroid* yang sudah di sesuaikan.

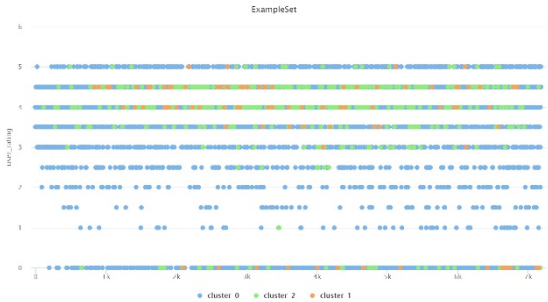
Sebelum melakukan pemodelan menggunakan algoritma K-Means. Diharuskan untuk menentukan nilai K atau berapa banyak nilai yang akan di *cluster*. Penelitian ini menggunakan nilai K sebanyak 3 maka akan ada 3 *cluster* yang terbagi. Hasil cluster dalam bentuk *scatter* plot bisa dilihat pada gambar berikut.



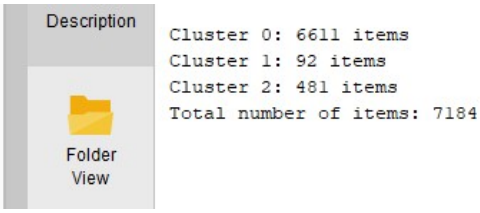
Gambar 2. *Cluster* pada Atribut Price



Gambar 3. Cluster pada Atribut *Size_Bytes*



Gambar 4. Cluster pada Atribut *User Rating*



Gambar 5. *Cluster* Model

Pada gambar diatas terdapat salah satu *cluster* yang jumlah datanya lebih besar dari pada *cluster* yang lainnya. Untuk lebih memahami isi dari setiap atribut dan cluster, kami melihat nilai rata-rata setiap atribut pada setiap *cluster,* hasil nilai rata-rata setiap atribut bisa dilihat pada table 1.

Tabel1. Hasil rata-rata setiap atribut

Atribut	*Cluster 0*	*Cluster 1*	*Cluster 2*
Size_bytes	114mb	2,4gb	932mb
Price($)	1.391	7.492	5.251
User rating (1-5)	3,5	3,5	3,8

Pada cluster 0 memiliki nilai rata-rata *Size_bytes, Price*, dan User rating yang terendah dari *cluster* kedua. Hal ini berarti bahwa cluster 0 dikelompokan berdasarkan nilai yang terkecil dari Cluster 1.

Pada cluster 1 nilai atribut *Size_bytes* dan *price* memiliki nilai yang terbesar dari setiap *cluster* kecuali pada atribut user rating yang memiliki ukuran rendah dari *cluster* 2. Namun nilainya tetap diatas rata-rata *cluster* 0, Hal ini menunjukan bahwa nilai dari cluster 2 dikelompokan berdasarkan atribut *Size byte* dan *price* yang terbesar.

Pada *cluster* 2, rata-rata dari Atribut User rating memiliki nilai yang terbesar dari setiap cluster. Pada table 1, terdapat cluster 3 dengan ciri-ciri aplikasi yang tergolong ideal. Ciri-ciri ini adalah dengan rating tertinggi, harga, dan memiliki ukuran aplikasi yang kecil.

IV. KESIMPULAN

Hasil dari penelitian ini menunjukan bahwa setiap *cluster* yang telah dibagi terdapat perbedaan pada nilai rata-rata antara setiap cluster-nya.

Pada *cluster* 0, pengelompokan dilakukan berdasarkan aplikasi yang kurang bagus karena memiliki ukuran aplikasi sangat rendah, harga dan rata-rata user rating sangat rendah.

Pada *cluster* 1, terdapat ciri-ciri aplikasi yang cukup bagus karena rata-rata ukuran aplikasi memiliki nilai tinggi, dan tentu dengan harga yang tinggi namun memiliki user rating cukup bagus.

Pada *cluster* 2, terdapat ciri-ciri aplikasi yang ideal, yaitu nilai *user_rating* tinggi, harga yang cukup lumayan dan memiliki ukuran aplikasi yang rendah.
Dengan hasil penelitian ini kami mengharapkan dapat digunakan pada penelitian-penelitian selanjutnya. Dataset didapatkan dari website kaggle, dan pembagian data menjadi 3 cluster diharapkan mampu membantu dalam menemukan ciri-ciri dan perbedaan pada tiap atribut pada setiap cluster yang ada. Serta membantu kepada developer dalam menganalisa sebuah aplikasi yang terdapat di *App Store.*

V. REFERENSI

Abdurrahman, G. (2016). Clustering Data Ujian Tengah Semester (UTS) Data Mining Menggunakan Algoritma K-Means. *Jurnal Sistem Dan Teknologi Informasi Indonesia*, *2*(1), 71–79. https://doi.org/10.32528/justindo.v1i2.566

Bozanta, A., & Co, M. (2018). *K-Means vs . Fuzzy C-Means : A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark*.

Effendi, J., & M Jorgi, R. (2018). *Analisis Cluster Aplikasi pada Google play Store dengan Menggunakan Metode K-Mean*. *4*(1), 978–979.

Febrianti, F., Hafiyusholeh, M., & Asyhar, A. H. (2016). Perbandingan Pengklusteran Data Iris Menggunakan Metode K-Means Dan Fuzzy C-Means. *Jurnal Matematika "MANTIK,"* *2*(1), 7. https://doi.org/10.15642/mantik.2016.2.1.7-13

Man, Y., Gao, C., Lyu, M. R., & Jiang, J. (2016). Experience Report: Understanding Cross-Platform App Issues from User Reviews. *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, 138–149. https://doi.org/10.1109/ISSRE.2016.27

Martin, W., Sarro, F., Jia, Y., Zhang, Y., & Harman, M. (2017). A survey of app store analysis for software engineering. *IEEE Transactions on Software Engineering*, *43*(9), 817–847. https://doi.org/10.1109/TSE.2016.2630689

Putra, R. R., & Wadisman, C. (2018). *Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritam K-Means Alghorithm*. *1*, 72–77.

Sandi, T. A. A., Raharjo, M., Putra, J. L., & Ridwan, R. (2018). Clustering Kesetiaan Pelanggan Dengan Model Rfm (Recency, Frequency, Monetary) Dan K-Means. *Jurnal Pilar Nusa Mandiri*, *14*(2), 239. https://doi.org/10.33480/pilar.v14i2.950

Sangani, C., & Ananthanarayanan, S. (2013). Sentiment Analysis of App Store Reviews. *Technical Report, Stanford University.*, 1–5. http://cs229.stanford.edu/proj2013/CS229-ProjectReport-ChiragSangani-SentimentAnalysisOfAppStoreReviews.pdf

Setiawan, R. (2016). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru (Studi Kasus : Politeknik Lp3i Jakarta). *J. Lentera Ict*, *3*(1), 76–92.

Yudi Agusta. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem Dan Informatika*, *3*(Februari)