PENGEMBANGAN ALGORITMA *UNSUPERVISED LEARNING TECHNIQUE* PADA *BIG DATA ANALYSIS* DI MEDIA SOSIAL SEBAGAI MEDIA PROMOSI *ONLINE* BAGI MASYARAKAT

Nurhayati [1], Busman [2], Rayi Pradono Iswara [3]

[1,3] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta
[2] STIE Gotong Royong Jakarta

ABSTRACT

Large data collection or known as big data can be analyzed with various techniques. One technique for processing big data is Unsupervised Technique. There are various kinds of algorithms that apply this technique. Each algorithm has its own ways and characteristics. This study focuses on developing an algorithm that implements an unsupervised learning technique, one of which is the K-Means algorithm by taking data samples to people who are doing creative and independent efforts. The Society utilized online and offline business in marketing. The researcher conducted an experimental test and simulation of the algorithm by producing output in the form of software applications as well as tables and graphs that were able to combine data obtained from social media and questionnaires fromline. The results of the analysis of data processing can be used as a DSS (Decision Support System) by the community in making their next production marketing development decisions.

Keywords: *Big Data, Machine Learning, Unsupervised Learning, K-Means DSS (Decision Support System)*

ABSTRAK

Kumpulan data yang besar atau dikenal dengan istilah *big data* dapat dianalisis dengan berbagai macam teknik. Salah satu teknik untuk mengolah *big data* adalah *Unsupervised Technique*. Ada berbagai macam algoritma yang menerapkan teknik ini. Setiap algoritma memiliki cara dan karakteristik masing-masing. Penelitian ini berfokus pada pengembangan algoritma yang menerapkan *unsupervised learning technique* salah satunya algoritma K-*Means* dengan mengambil sampel data pada masyarakat yang melakukan usaha kreatif dan mandiri. Masyarakat dalam yang memanfaatkan usaha *online* dan *offline* dalam pemasarannya. Peneliti melakukan uji eksperimen dan simulasi terhadap algoritma tersebut dengan menghasilkan output berupa aplikasi *software* serta tabel dan grafik yang mampu menggabungkan data yang didapat dari media social dan kuesioner secara *offline*. Hasil analisa pengolahan data tersebut dapat digunakan sebagai DSS (*Decision Support System*) oleh masyarakat dalam mengambil keputusan pengembangan pemasaran produksinya selanjutnya.

Kata Kunci: *Big Data, Machine Learning, Unsupervised Learning, K-Means DSS (Decision Support System)*

I. PENDAHULUAN

Saat ini pertumbuhan yang sangat pesat dari akumulasi data telah menciptakan kondisi kaya data tapi minim informasi. Informasi yang dibutuhkan tidak dapat diperoleh dengan mudah dikarenakan volume data yang sangat besar. Sehingga dibutuhkan suatu metode untuk mendapatkan pengetahuan yang tidak terlihat di dalam data namun potensial untuk digunakan yaitu metode *data mining*.

Big data merupakan istilah populer yang digunakan untuk menggambarkan pertumbuhan eksponensial dan ketersediaan data, baik terstruktur dan tidak terstruktur. *Big data* sudah menjadi hal yang penting bagi bisnis dan masyarakat, seperti halnya internet. Teknologi *big data* tersebut bisa digunakan melakukan analisa di media sosial. Salah satu data yang dapat diolah untuk suatu keperluan tertentu menggunakan *big data* adalah data Twitter. Media sosial seperti Twitter dan Facebook menyediakan layanan untuk berhubungan dengan teman-teman *online* yang meningkatkan efektivitas iklan Internet. Sehingga data dan informasi dari Twitter dan Facebook dapat kita gunakan untuk media promosi usaha secara *online* bagi masyarakat.

Teknologi yang dapat digunakan dalam *big data* ini salah satunya adalah *Machine Learning* (ML) yang merupakan salah satu varian dari sistem kecerdasan buatan yang memungkinkan komputer dapat belajar tanpa diprogram secara eksplisit. Secara umum, pekerjaan *Machine Learning* (ML) yang seringkali digunakan adalah untuk mengklasifikasikan satu permasalahan menjadi beberapa kelompok. Dalam kehidupan sehari-hari, obyek dapat diidentifikasi dengan mudah oleh manusia, namun belum tentu dapat dijelaskan secara spesifik. Di sinilah peran *Machine Learning* dalam mengenali, mengidentifikasi, ataupun memprediksi data tertentu dengan mempelajari data histori. Dengan *Machine Learning*, model dibuat baik secara langsung ataupun tidak, dengan mengekstrak pengetahuan dari pakar ataupun dari data yang bahkan belum diketahui hubungannya dengan cara mempelajarinya dengan algoritma tertentu.

Machine Learning mempunyai 2 tipe teknik yaitu *supervised learning* dan *unsupervised learning*. Mayoritas praktis dari *machine learning* menggunakan *supervised learning* [1]. *Supervised learning* adalah salah satu tipe algoritma *machine learning* yang menggunakan dataset yang dikenal (*training dataset*) untuk membuat prediksi. Penelitian ini menggunakan *Unsupervised Learning*. *Unsupervised learning* adalah salah satu tipe algoritma *machine learning* yang digunakan untuk menarik kesimpulan dari *datasets* yang terdiri dari input *data labeled response*. Metode *unsupervised learning* yang paling umum adalah analisa *cluster*, yang digunakan pada analisa data untuk mencari pola-pola tersembunyi atau pengelompokan dalam data ("*Machine learning technique for building predictive models from known input and response data*," n.d.).

Salah satu algoritma yang digunakan metode *unsupervised learning* adalah *K-Means* algoritma. Pada penelitian ini peneliti akan memanfaatkan algoritma *K-Means* ini. Algoritma *K-Means* adalah metode partisi yang terkenal untuk *clustering* [2]. *K-Means* merupakan salah satu metode data *cluster*ing non hierarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lainnya [3].

Berdasar uraian di atas maka peneliti mengangakat topik penelitian ini adalah "Pengembangan Algoritma *Unsupervised Learning Technique* pada *Big data Analysis* di media sosial sebagai media promosi usaha *online* bagi masyarakat". Penelitian diharapkan dapat menentukan faktor apa yang memengaruhi dampak promosi usaha *online* terhadap masyarakat atau konsumen. Penelitian memakai metode algoritma *K-Means* dan *big data analysis* dengan simulasi dan eksperimen logika bisnis *big data*. Hasil riset dan development ini sangat berguna untuk perkembangan pengetahuan di bidang IT terutama bidang *unsupervised machine learning* dan Penelitian ini akan menghasilkan output berupa *software* aplikasi akan dapat di gunakan sebagai *decision support system* (DSS) bagi masyarakat dalam mengelola dan mengembangkan usaha secara *online*.

II. TINJAUAN PUSTAKA

Berdasarkan studi literatur, yakni pertama adalah "*A Comparison of Unsupervised Learning Techniques for Encrypted Traffic Identification"* oleh Carlos Bacquet, Kubra Gunus, Dogukan Tizer, A. Nur Zincir-Heywood dan Malcolm I. Heywood. Penulis dalam jurnal mengatakan penggunaan *traffic* terenkripsi digabung dengan port non-standar membuat tugas identifikasi *traffic* menjadi lebih sulit. Penulis mengukur kemampuan 5 algoritma yakni: *Basic K-Means*, *Semi-supervised K-Means*, DBSCAN, EM dan MOGA untuk mengidentifikasi lalu lintas terenskripsi, terutama SSH pada *dataset*. Algoritma *K-Means* dan MOGA mendapatkan hasil terbaik dan juga meng-*cluster* data menjadi sangat kecil [4].

Pada studi literatur kedua yaitu "*Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning"* Shan Suthaharan. Penulis dalam jurnal ini mendiskusikan tentang tantangan sistem yang ada pada permasalahan *Big data* terkait prediksi penyusup. Prediksi pada kemungkinan serangan penyusup terjadi dalam jaringan membutuhkan pengumpulan data *traffic* secara terus menerus dan mempelajari karakteristi*K*-karakteristiknya. Pengumpulan data *traffic* secara terus menerus merupakan permasalahan dalam bentuk *Big data* yang disebabkan oleh properti *Big data volume*, *variety* dan *velocity*. Jurnal ini menyarankan pengintegrasian teknologi *Hadoop Distributed File Systems* dan *Cloud Technologies* dengan teknik *representation-learning* terbaru dan mendukung *vector machine* untuk memprediksi penyusup pada jaringan melalui klasifikasi strategi *Big data* [5].

Studi literatur ketiga adalah Simon Hudson, Li Huang, Martin S. Roth, dan Thomas J. Madden dalam jurnalnya berjudul "The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors" mengatakan perusahaan meningkatkan pemasarannya menggunakan social media saat ini. Melalui relasi anatar social media orang dapat melakukan pemasarannya saat ini. Metode pemasaran semacam itu mengalami peningkatan [6].

III. METODOLOGI

Big data dapat didefinisikan sebagai sekumpulan data yang ukurannya melampaui kemampuan dari *tool* perangkat lunak basis data untuk mengambil, menyimpan, mengatur dan menganalisa. Kumpulan data tersebut secara umum dihasilkan melalui internet, perangkat *mobile*, sensor jaringan, sistem *enterprise* dan organisasi [7]. *Big data* tidak hanya terfokus pada *volume*, v*elocity* dan *variety* juga termasuk pada fokus *Big data*. Hasil dari *big data* bisa terstruktur, tidak terstruktur dan semi terstruktur [8].

3.1 *Machine Learning*

Machine Learning adalah salah satu disiplin ilmu dari *Computer Science* yang mempelajari bagaimana membuat komputer atau mesin itu agar mempunyai suatu kecerdasan, komputer atau mesin harus dapat belajar. Dengan kata lain, *Machine Learning* adalah suatu bidang keilmuan yang berisi tentang pembelajaran komputer atau mesin untuk menjadi cerdas [9].

K-Means merupakan algoritma yang paling sering digunakan untuk keperluan *cluste*ring dokumen. Prinsip utama *K-Means* adalah menyusun x prototype atau pusat massa (*centroid*) dari sekumpulan data berdimensi n [10]. Sebelum diterapkan proses algoritma *K-Means*, akan dilakukan proses *preprocessing* terlebih dahulu terhadap data. Algoritma *K-Means* termasuk dalam *partitioning cluster*ing yang memisahkan data ke k daerah bagian yang terpisah. Algoritma *K-Means* sangat sering digunakan karena kemudahan dan kemampuannya untuk melakukan *cluster* data besar dan outlier dengan waktu yang sangat cepat.

Algoritma *K-Means* cukup efektif untuk diterapkan dalam proses pengelompokkan karakteristik terhadap objek penelitian. Menurut MacQueen [11], *K-Means* merupakan metode klasterisasi yang paling terkenal dan banyak digunakan dalam berbagai bidang karena bentuknya yang sangat sederhana, mudah diimplementasikan, memiliki kemampuan untuk mengklaster data yang cukup besar, mampu menangani data *outlier*, dan kompleksitas waktunya linear O (nKT) dengan n adalah jumlah dokumen, K adalah jumlah klaster, dan T adalah jumlah iterasi. *K-Means* merupakan metode pengklasteran secara *partitioning* yang memisahkan data ke dalam kelompok yang berbeda. Dengan *partitioning* secara iteratif,

K-Means mampu meminimalkan rata-rata jarak masing-masing data ke klasternya.

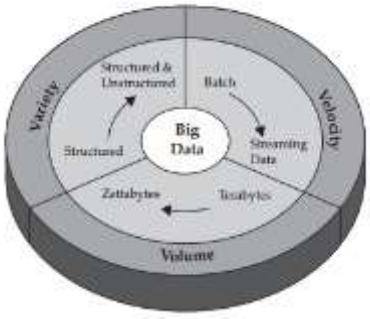
*K-Means cluster*ing merupakan sebuah metode dari *unsupervised learning* yang bertujuan untuk mempartisi peninjauan n ke kelompok K dimana tiap peninjauan dimiliki kelompok yang mempunyai nilai rata-rata terdekat [12].

Algoritma *K-Means* pada dasarnya bekerja dalam 2 proses yakni proses pendeteksian lokasi pusat *cluster* dan proses pencarian anggota dari tiap-tiap *cluster*. Proses *clustering* dimulai dengan mengidentifikasi data yang akan diklaster, C_{ij} (i=1,…,n; j=1,...,m) dengan n adalah jumlah data yang akan diklaster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap klaster ditetapkan secara sembarang (terserah peneliti), C_{kj} (k=1,...,k; j=1,...,m). Kemudian dihitung jarak antara setiap data ke masing-masing pusat klaster. Untuk melakukan penghitungan jarak data ke-I (xi) pada pusat klaster ke-k (ck), diberi nama (dik), dapat digunakan fungsi Euclidean. Suatu data akan menjadi anggota dari klaster ke-k apabila jarak data ke pusat klaster k tersebut bernilai paling kecil jika dibandingkan dengan jarak ke pusat klaster lainnya.

3.2 *Big Data*

Teknologi *Big data* adalah pengelolaan aset informasi dengan volume dan kecepatan yang tinggi serta kompleks yang membantu perusahaan mengelola data dengan biaya efektif dan mendorong inovasi pengolahan informasi untuk pengambilan keputusan dan peningkatan pengetahuan atau wawasan. *Big data* menjamin pemrosesan solusi data dengan varian baru maupun eksisting untuk memberikan manfaat nyata bagi bisnis [7].

Ada 3 karakteristik atau dimensi awal dalam *Big data* yaitu 3V: *Volume*, *Variety* dan *Velocity*. IBM menggambarkan karakteristik *big data* sebagai berikut:



Gambar 1. Karakteristik *big data* [7]

3.3 Metode Simulasi

Ada berbagai jenis *lifecycle* yang dapat digunakan untuk studi pada pemodelan dan simulasi. Langkah-langkah dalam metode simulasi adalah sebagai berikut [13]:

1. *Problem Formulation*
   Proses simulasi dimulai dengan masalah yang memerlukan pemecahan atau pemahaman. Sebagai contoh seperti kasus sebuah perusahaan kargo mencoba untuk mengembangkan strategi baru untuk truk pengiriman atau astronom mencoba untuk memahami bagaimana nebula terbentuk. Pada tahap ini, harus dipahami perilaku dari sebuah sistem, *organize* operasi sistem sebagai obyek dalam rangka percobaan. Kemudian perlu dianalisis berbagai alternatif solusi dengan menyelidiki hasil sebelumnya untuk masalah yang sama. Solusi yang paling diterima harus dipilih (menghilangkan tahap ini dapat menyebabkan pemilihan solusi yang salah). Jika masalah melibatkan analisis kinerja, ini adalah titik di mana bisa didefinisikan metrik kinerja (berdasarkan variabel output) dan fungsi tujuan (yaitu, kombinasi dari beberapa metrik).

2. *Conceptual Model*
   Langkah ini terdiri dari pengembangan deskripsi tingkat tinggi dari struktur dan perilaku atau *behavior* sebuah sistem dan mengidentifikasi semua benda dengan atribut dan *interface* mereka. Pada tahap ini harus ditentukan apa saja variabel *state*nya, bagaimana mereka berhubungan, dan mana yang penting untuk penelitian. Pada langkah ini, aspek-aspek kunci dari *requierements* dinyatakan. Selama definisi model konseptual, perlu diungkapkan fitur yang penting. Kemudian mendokumentasikan informasi-untuk non-fungsional misalnya, perubahan masa depan, perilaku *unintuitive*, dan hubungan sistem dengan lingkungan.

3. *Collection of Input/Output Data*
   Pada tahap ini, kita harus mempelajari sistem untuk memperoleh data input/output. Untuk melakukannya, harus diamati dan mengumpulkan atribut yang dipilih pada tahap sebelumnya. Isu penting lainnya selama fase ini adalah

pemilihan ukuran sampel yang valid secara statistik dan format data yang dapat diproses dengan komputer. Akhirnya, kita harus memutuskan mana atribut yang stokastik dan yang deterministik. Dalam beberapa kasus, tidak ada sumber data yang bisa dikumpulkan (misalnya, untuk sistem yang belum ada). Dalam kasus tersebut, kita perlu mencoba untuk mendapatkan set data dari sistem yang sama (jika tersedia). Pilihan lain adalah dengan menggunakan pendekatan stokastik untuk menyediakan data yang diperlukan melalui generasi nomor acak.

4. *Modelling Phase*
   Pada tahap pemodelan, kita harus membangun representasi rinci dari sistem berdasarkan model konseptual dan koleksi data yang dikumpulkan. Model ini dibangun dengan mendefinisikan objek, atribut, dan metode menggunakan paradigma yang dipilih. Pada titik ini, spesifikasi model dibuat, termasuk set persamaan yang mendefinisikan perilaku dan struktur. Setelah menyelesaikan definisi ini, kita harus berusaha untuk membangun struktur awal model (mungkin berkaitan variabel sistem dan metrik kinerja). Harus berhati-hati dalam menjelaskan setiap asumsi dan penyederhanaan dan juga dalam mengumpulkan atribut ke EF (*Entity Framework*) model.

5. *Simulation Phase*
   Selama tahap simulasi, kita harus memilih mekanisme untuk menerapkan model (dalam banyak kasus menggunakan komputer dan bahasa pemrograman yang memadai serta *tools* yang tepat), dan model simulasi yang dibangun. Selama langkah ini, mungkin diperlukan untuk menentukan algoritma dan menerjemahkannya ke dalam program komputer. Pada tahap ini, kita juga harus membangun model EF untuk proses simulasi.

6. *Verification, Validation, and Experimentation*
   Pada tahap-tahap sebelumnya, tiga model yang berbeda dibangun: model konseptual (spesifikasi), model sistem (desain), dan model simulasi (*executable program*). Kita perlu untuk memverifikasi dan memvalidasi model ini. Verifikasi terkait dengan konsistensi internal antara tiga model. Validasi difokuskan pada korespondensi antara model dan realitas: adalah hasil simulasi yang konsisten dengan sistem yang dianalisis. Sementara itu pada fase *experimentation*, kita harus mengevaluasi hasil dari simulator, menggunakan korelasi statistik untuk menentukan tingkat presisi untuk metrik kinerja. Fase ini dimulai dengan desain eksperimen, dengan menggunakan teknik yang berbeda. Beberapa teknik ini meliputi analisis sensitivitas, optimasi, dan seleksi (dibandingkan dengan sistem alternatif).

7. *Output Analysis Phase*
   Pada tahap analisis output, output simulasi dianalisis untuk memahami perilaku sistem. Output ini digunakan untuk memperoleh tanggapan tentang perilaku sistem yang asli. Pada tahap ini, alat visualisasi dapat digunakan untuk membantu proses tersebut. Tujuan dari visualisasi adalah untuk memberikan pemahaman yang lebih dalam tentang sistem yang sedang diselidiki dan membantu dalam mengeksplorasi set besar data numerik yang dihasilkan oleh simulasi.

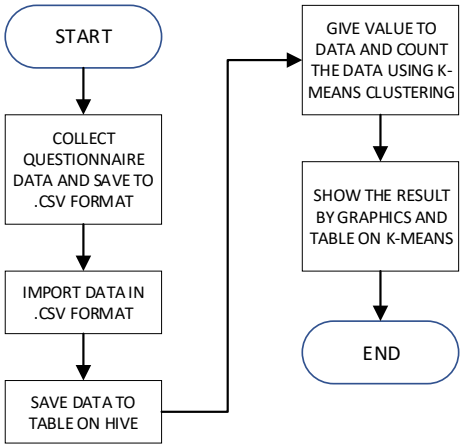
IV. HASIL DAN PEMBAHASAN

4.1 *Problem Formulation*

Formulasi masalah merupakan tahap awal dalam perancangan pada model metode simulasi. Formulasi masalah merupakan suatu kegiatan untuk memilih satu permasalahan yang dianggap paling penting untuk diselesaikan saat itu dari sekian banyak permasalahan. Pada penelitian ini, penulis memformulasikan sebuah masalah yaitu banyaknya aspek yang mempengaruhi dalam suatu metode pemasaran suatu usaha. Sehinga diperlukan sistem pengambil keputusan untuk dapat memantau jenis pemasaran tersebut masuk kedalam *cluster online* maupun *cluster offline*. Dalam kasus ini, solusi terbaik yaitu dengan memanfaatkan *Hadoop* dan *Machine Learning* yaitu algoritma *K-Means Cluster*ing agar dapat mengolah data yang besar dan melakukan kategorisasi dari data tersebut dengan tepat.

4.2 *Conceptual Model*

Pemodelan secara konsep menggambarkan konsep sistem secara keseluruhan (*overall solution*), mulai dari awal input, proses, sampai dengan output yang dihasilkan oleh sistem. *Conceptual model* ini dengan Hadoop dan Machine Learning yaitu *K-Means Clustering* untuk diimplementasikan pada sistem yang akan dibangun. Penggunaan Hadoop pada sistem digambarkan mulai dari melakukan koneksi sistem dengan Hadoop untuk penyimpanan data kuesioner. Kemudian melakukan koneksi sistem dengan Hive untuk membuat dan menyimpan tabel pada Hadoop. Setelah melakukan koneksi, pembuatan GUI dilakukan. Input pada program "AplikasiKMean" berupa data kuesioner dalam format file .CSV yang dapat diimpor ke dalam Aplikasi dan disimpan ke dalam tabel Hive ke dalam Hadoop. Konsep penggunaan *K-Means Cluster*ing yaitu untuk dapat digunakan pada sistem pada saat menganalisa data kuesioner, setelah itu data dianalisa dengan algoritma *K-Means*.



Gambar 2. *Flowchart* proses analisa data

Merujuk pada Gambar 2. *Flowchart* Proses Analisa Data, terdapat 4 tahapan proses untuk melakukan analisa data kuesioner. Proses pertama yaitu *Collect Questionaire Data and Save to .CSV format*. Proses tersebut merupakan proses pengumpulan data kuesioner *online* maupun *offline* dikumpulkan dan disimpan ke dalam format .CSV, selanjutnya proses kedua yaitu *Import Data with .CSV format*. Proses kedua ini yaitu dengan mengambil data kuesioner yang berformat .CSV untuk diimpor pada program "AplikasiKMean". Proses ketiga yaitu *Save the Data to tabel on Hive*. Pada proses ini data yang diimpor, disimpan ke dalam tabel pada Hive, dan data yang tersimpan akan tampil pada Aplikasi *K-Means* berupa tabel data. Proses terakhir *yaitu Give value to data and count the data with K-Means Clustering.* Pada proses ini program "AplikasiKMean" memberikan nilai pada data kuesioner, yaitu setiap jawaban pada data kuesioner diberikan nilai dan dihitung menggunakan algoritma *K-Means Cluster*ing. Hasil yang memiliki nilai terkecil terdekat pada *cluster* 1 maka akan masuk *cluster* 1 dan berlaku juga pada *cluster* 2. Terakhir, hasilnya ditampilkan di program "AplikasiKMean" berupa tabel dan grafik.

4.3 *Collection of Input/Output Data*

Dalam membangun sebuah sistem dan melakukan simulasi tentu dibutuhkan sebuah sumber data. Data diperlukan untuk kebutuhan dalam melakukan proses pemodelan (*modelling*). Dalam analisis pengujian akan mengukur seberapa tepat pemodelan yang dibuat sehingga dapat memproses sumber data yang diperoleh untuk menjadi output yang bermanfaat. Data bisa diperoleh melalui berbagai sumber tergantung sistem yang dibuat. Pada penlitian ini sumber data yang akan digunakan diambil dari data kuesioner *online* dan kuesioner *offline*. Data yang diperoleh selanjutnya diolah dan dianalisa sehingga dapat menjadi informasi yang ditampilkan dalam bentuk tabel data dan grafik. Informasi tersebut berupa nilai *cluster online* maupun *offline* dalam soal-soal yang dijawab responden pada kuesioner.

Pada penelitian ini, input data didapatkan dari data kuesioner yang akan diproses, terlihat pada tabel di bawah ini:

Tabel 1. Data input pada penelitian

No	Data Kuesioner	Tipe Data
1	Nama Usaha	*String*
2	Nama Pengusaha	*String*
3	Jenis_Kelamin	*String*
4	Usia	*String*
5	Badan_Hukum	*String*
6	Pendidikan_terakhir	*String*
7	Lama_Usaha	*String*
8	Badan_hukum	*String*
9	Alamat_usaha	*String*
10	Contact_person	*String*
11	Jenis_usaha	*String*
12	Jumlah_asset	String
13	Volume_produksi	String
14	Lokasi_usaha	String
15	Wilayah_pemasaran	*String*

No	Data Kuesioner	Tipe Data
16	Sektor_usah	*String*
17	Sistem_penjualan	*String*
18	Target_pasar	*String*
19	Tipe_produk	*String*
20	Metode_pembayaran	*String*
21	Q1	*String*
22	Q2	*String*
23	Q3	*String*
24	Q4	*String*
25	Q5	*String*
26	Q6	*String*
27	Q7	*String*
28	Q8	*String*
29	Q9	*String*
30	Q10	*String*
31	Q11	*String*
32	Q12	*String*
33	Q13	*String*
34	Q14	*String*
35	Q15	*String*

Selanjutnya data-data tersebut diproses dan dianalisa oleh AplikasiKMean dan disimpan ke dalam tabel Hive pada Hadoop. Terdapat beberapa tambahan data hasil analisa, namun hanya akan ditampilkan pada AplikasiKMean dan tidak semua data disimpan ke Hadoop. Berikut merupakan data yang disimpan:

Tabel 2. Data yang disimpan pada Hadoop

No	Data Kuesioner	Tipe Data
1	Id	Int
2	Nama Usaha	*String*
3	Nama Pengusaha	*String*
4	Jenis_Kelamin	*String*
5	Usia	*String*
6	Badan_Hukum	*String*
7	Pendidikan_terakhir	*String*
8	Lama_Usaha	*String*
9	Badan_hukum	*String*
10	Alamat_usaha	*String*
11	Contact_person	*String*
12	Jenis_usaha	*String*
13	Jumlah_asset	*String*
14	Volume_produksi	*String*
15	Lokasi_usaha	*String*
16	Wilayah_pemasaran	*String*
17	Sektor_usah	*String*
18	Sistem_penjualan	*String*
19	Target_pasar	*String*
20	Tipe_produk	*String*
21	Metode_pembayaran	*String*
22	Q1	*String*
23	Q2	*String*
24	Q3	*String*
25	Q4	*String*
26	Q5	*String*
27	Q6	*String*
28	Q7	*String*
29	Q8	*String*
30	Q9	*String*
31	Q10	*String*
32	Q11	*String*
33	Q12	*String*
34	Q13	*String*
35	Q14	*String*
36	Q15	*String*

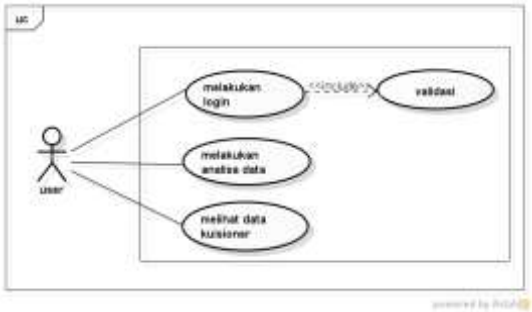
Semua data pada tabel 2 disimpan dalam bentuk format String, kecuali Id dengan format int. Data-data diatas disimpan dan hasil output analisa ditampilkan pada program "AplikasiKMean". Outputnya dalam bentuk tabel dan grafik yang berisi nilai-nilai yang dihasilkan dari data tersebut. Penjelasan mengenai hal ini dijelaskan lebih lanjut pada fase pemodelan.

4.4 *Modelling Phase*

Fase pemodelan adalah tahap dilakukannya pembuatan sebuah skenario pengujian yang dilakukan pada sistem sesuai dengan variabel yang sudah ditentukan. Skenario dilakukan dengan berpatok pada pembandingan hasil output simulasi dengan menjalankan sistem. Pada fase pemodelan ini dilakukan pemodelan diagram UML yang digunakan ada lima yaitu *usecase diagram, class diagram, object diagram, sequence diagram,* dan *activity diagram.* Kemudian dilakukan pemodelan konstruksi *K-Means* dan melakukan pengkodean pada program "AplikasiKMean".

1. Pemodelan Diagram UML

   Pemodelan *usecase diagram* pada program "AplikasiKMean" digambarkan sebagai berikut.



Gambar 3. *Use case diagram* pada AplikasiKMeans

Pemodelan diagram UML selanjutnya yaitu *class diagram*, selanjuttnya adalah *object diagram, sequence diagram,* dan *activity diagram.*

2. Konstruksi Algoritma *K-Means*

Pada fase pemodelan ini, dilakukan konstruksi Algoritma *K-Means* yang dihitung secara manual. Data manual didapat dari membandingkan jumlah total pemasaran *online* dan jumlah total pemasaran *offline* lalu diambil nilai terbesar. Jika jumlah total *online* > jumlah total *offline* maka masuk ke dalam *cluster online,* dan sebaliknya. Proses perhitungan manualnya yakni sebagai berikut.

1. Proses *Clustering* Algoritma *K-Means*:

Pada tahap ini akan dilakukan proses utama yaitu segmentasi data nilai yang diakses dari database yaitu sebuah metode *clustering* algoritma *K-Means*. Berikut ini merupakan diagram *flowchart* dari algoritma *K-Means* dengan asumsi bahwa parameter *Input* adalah jumlah data set sebanyak n data dan jumlah inisialisasi centroid K=2 sesuai dengan *cluster* yang diinginkan.

Konstruksi *K-Means* dapat dijelaskan beberapa langkah yang dilalui oleh *cluster*ing algoritma *K-Means* memuat bagian-bagian sebagai berikut ini:

1) N data: data set yang akan diolah sebanyak N data dimana N data tersebut terdiri dari atribut-atributnya N (Jumlah Nilai A, Jumlah Nilai B) yang berarti data N memiliki atribut sebanyak 2.
2) K *centroid*: Inisialisasi dari pusat *cluster* data adalah sebanyak K dimana pusat-pusat awal tersebut digunakan sebagai banyaknya kelas yang akan tercipta. *Centroid* didapatkan secara random dari N data set yang ada.
3) *Euclidian Distance*: merupakan jarak yang didapat dari perhitungan antara semua N data dengan K *centroid* dimana akan memperoleh tingkat kedekatan dengan kelas yang terdekat dengan populasi data tersebut. Jarak *euclidian* untuk menandai adanya persamaan antar tiap *cluster* dengan jarak minimum dan mempunyai persamaan yang lebih tinggi.

$$D_{ik} = \sqrt{\sum_{j}^{m} (Cij - Ckj)^2}$$

C_{ij} : Titik Data Pertama
C_{kj} : Titik Data Kedua
D_{ik} : *Euclidian distance* yaitu jarak antara data pada titik x dan titik y menggunakan kalkulasi matematika

4) Pengelompokkan data: setelah sejumlah populasi data tersebut menemukan kedekatan dengan salah satu *centroid* yang ada maka secara otomatis populasi data tersebut masuk kedalam kelas yang memiliki *centroid* yang bersangkutan.
5) *Update centroid* baru: tiap kelas yang telah tercipta tadi melakukan *update centroid* baru. Hal ini dilakukan dengan menghitung nilai rata-rata dari kelas masing-masing. Apabila belum memenuhi optimal hasil proses pengukuran *ecluidian distance* dilakukan kembali.
6) Batas iterasi: apabila dalam proses *clustering* belum optimal namun sudah memenuhi batas iterasi maksimum, maka proses dihentikan.

Berikut ini contoh dari fungsi algoritma *K-Means* yang penulis gunakan:

Dari total data yakni 650 koresponden diambil 10 koresponden sebagai contoh yang akan digunakan untuk konstruksi algoritma *K-Means* secara manual pada pemahaman tentang aliran besar dalam Islam. Percobaan dilakukan dengan

menggunakan parameter-parameter berikut:
Jumlah *cluster* : 2
Jumlah data :10
Jumlah atribut : 2

Berikut ini merupakan tabel yang digunakan untuk melakukan percobaan perhitungan manual.

Tabel 3. Daftar cara pemasaran produk

No	Nama	Jumlah Nilai *Online*	Jumlah Nilai *Offline*
1	Victor Afrizal	675	650
2	Hezmi Emimah	1275	200
3	Rizki	750	450
4	Viva	925	750
5	Luh Kesuma Wardhani	1175	475
6	Ahmad Fauzan	575	1050
7	Komarudin	450	900
8	Adelia Rusliani	1300	525
9	Afifah Fitria Lestari	1150	475
10	Gugun Hidayansyah	625	825

Iterasi ke-1

1. Penentuan pusat awal *cluster*
   Untuk penentuan awal di asumsikan:
   Diambil data ke- 2 lalu dibagi 2 sebagai pusat *Cluster* Ke-1: (637.5, 100).
   Diambil data ke- 5 sebagai pusat *Cluster* Ke-2 lalu dibagi 2: (587.5, 237.5).
2. Perhitungan jarak pusat *cluster*
   Untuk mengukur jarak antara data dengan pusat *cluster* digunakan *Euclidian distance*, kemudian akan didapatkan matrik jarak sebagai berikut:

Rumus *Euclidian distance:*

$$d_{ik} = \sqrt{\sum_{j}^{m} (Cij - Ckj)^2}$$

C_{ij}: Pusat *Cluster*
C_{kj}: Data

Sebagai contoh, perhitungan jarak dari data ke-1 terhadap pusat *cluster* adalah:

$$C1 = \sqrt{(637.5 - 675)^2 + (100 - 650)^2} = 551.28$$

$$C2 = \sqrt{(587.5 - 675)^2 + (237.5 - 650)^2} = 421.68$$

Dan seterusnya dilanjutkan untuk data ke 2, 3, … n
Kemudian akan didapatkan matrik jarak sebagai berikut:

D^1 =

1	2	3	4	5	6	7	8	9	10	
551.28	645.30	367.64	710.74	655.39	952.05	821.68	787.10	635.04	725.11	C1
421.68	688.52	267.51	613.65	633.69	812.60	676.62	768.32	610.58	588.70	C2

Setiap kolom pada matrik menunjukkan nilai jarak data terhadap pusat *cluster*. Baris pertama pada matrik menunjukkan nilai jarak data terhadap titik pusat *cluster* pertama, baris kedua pada matrik menunjukkan nilai jarak data terhadap titik pusat *cluster* kedua dan seterusnya.

3. Pengelompokkan data
   Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat. Berikut ini akan ditampilkan data matriks pengelompokkan group, nilai 1 berarti data tersebut berada dalam *group*.

G^1 =

1	2	3	4	5	6	7	8	9	10	
0	1	0	0	0	0	0	0	0	0	C1
1	0	1	1	1	1	1	1	1	1	C2

Keterangan:
Jika jarak data ke-1 (D1) dengan pusat *cluster* ke-1 (C1) atau pusat *cluster* ke-2 (C2) lebih dekat, maka G1 bernilai 1 dan termasuk grup atau kelompok *cluster* baru.
Jika jarak data ke-1 (D1) dengan pusat *cluster* ke-1 (C1) atau pusat *cluster* ke-2 (C2) lebih jauh, maka G1 bernilai 0 dan tidak termasuk grup atau kelompok *cluster* baru.

4. Penentuan pusat *cluster* baru
   Setelah diketahui anggota tiap-tiap *cluster* kemudian pusat *cluster* baru dihitung berdasarkan data anggota tiap-tiap *cluster* sesuai dengan rumus

pusat anggota *cluster*. Sehingga didapatkan perhitungan sebagai berikut: Karena C1 hanya memiliki 1 anggota maka perhitungan *cluster* baru menjadi:

C1 =
$\left(\frac{1275}{1}, \frac{200}{1}\right)$

C1 = (1275, 200)

Karena C2 mempunyai 9 anggota maka perhitungan *cluster* baru menjadi:

C2=
$\left(\frac{675+750+925+575+450+1300+1150+625}{9}, \frac{650+450+750+475+1050+900+525+475+825}{9}\right)$

C2 = (925;590.91)

G^2 =

1	2	3	4	5	6	7	8	9	10	
1	0	0	1	1	1	0	0	0	0	C1
0	1	1	0	0	0	1	1	1	1	C2

Iterasi ke-3

6. Karena G2! = G1 ulangi langkah ke 4 (empat) dan dilanjuti langkah ke 2

D^3 =

1	2	3	4	5	6	7	8	9	10	
596.64	224.39	476.03	446.91	75.26	906.08	912.26	130.05	93.75	724.60	C1
121.12	834.22	331.48	259.17	588.15	293.83	252.25	679.37	566.68	68.34	C2

Langkah selanjutnya sama dengan langkah pada nomor 3 jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat. Berikut ini akan ditampilkan data matriks pengelompokkan group, nilai 1 berarti data tersebut berada dalam group.

G^3 =

1	2	3	4	5	6	7	8	9	10	
1	0	0	1	1	1	0	0	0	0	C1
0	1	1	0	0	0	1	1	1	1	C2

Karena G2 = G1 memiliki anggota yang sama maka tidak perlu dilakukan iterasi/perulangan lagi. Hasil *cluster*ing telah mencapai stabil dan konvergen.

Iterasi ke-2

5. Ulangi langkah ke 2 (kedua) hingga posisi data tidak mengalami perubahan.

D^2 =

1	2	3	4	5	6	7	8	9	10	
750	0	581.49	651.92	292.62	1101.14	1081.95	325.96	302.08	901.73	C1
277.84	727.04	524.86	212.10	454.60	382.16	232.12	445.52	452.16	167.36	C2

Langkah selanjutnya sama dengan langkah pada nomor 3 jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat. Berikut ini akan ditampilkan data matriks pengelompokkan *group*, nilai 1 berarti data tersebut berada dalam *group*.

(kedua) hingga posisi data tidak mengalami perubahan.

4.5 *Simulation Phase*

Dalam fase simulasi dilakukan proses simulasi terhadap program "AplikasiKMean" yang terkoneksi dengan Hadoop dan menggunakan Algorita *K-Means*. Simulasi dijalankan dari tahap *login*, penyimpanan data, dan analisa data.

Adapun faktor–faktor dalam proses simulasi dapat dilihat pada tabel berikut.

Tabel 4. Variabel dan faktor dalam proses simulasi

Variabel / parameter simulasi	*Data Filtering*
Faktor 1	Data difilter berdasarkan jenis kelamin
Faktor 2	Data difilter berdasarkan usia
Faktor 3	Data difilter berdasarkan jenis usaha
Faktor 4	Data difilter berdasarkan jumlah asset
Faktor 5	Data difilter berdasarkan tingkat pendidikan
Faktor 6	Data difilter berdasarkan pemanfaatan Teknologi Informasi
Faktor 7	Data difilter berdasarkan wilayah pemasaran
Faktor 8	Data difilter berdasarkan tingkat persaingan

Variabel yang digunakan yaitu filter data pada hasil analisa program "AplikasiKMean". Filter data yang dimaksud adalah data diurutkan berdasarkan kondisi tertentu dan hasil analisa data juga berbeda pada setiap filter.

Proses simulasi pada *scenario* dimulai dengan menjalankan *Hadoop* dan *Hive* pada *Terminal Linux*. Kemudian membuka program "AplikasiKMean".

Dalam melakukan analisa data kuesioner, data di *import* terlebih dahulu dengan mengklik *button File* pada program "AplikasiKMean". dan akan muncul *window open file* untuk memilih file berformat .csv.

Data kuesioner yang dipilih nama *file* nya akan tampil pada *textBox* di program "AplikasiKMean". Selanjutnya klik *button Save* untuk menyimpan data kedalam tabel *Hive* pada *Hadoop*.

Selanjutnya klik *button* Mulai untuk melakukan analisa data kuesioner yang telah tersimpan pada *Hadoop*. Data yang dianalisa akan tampil pada tabel hasil analisa dan jumlah data yang masuk ke dalam *cluster* pemasaran *online* maupun *cluster* pemasaran *offline* ditampilkan dalam bentuk *text* di atas tabel hasil analisa.

Hasilnya koresponden cenderung masuk kedalam *cluster* suni. Dari 650 koresponden terdapat 474 koresponden masuk ke dalam *cluster online* dan 176 koresponden masuk ke dalam *cluster offline* dan memiliki tingkat akurasi sebesar 85%.

Pada proses simulasi berikut ini menggunakan parameter dan faktor-faktor yang telah dibahas sebelumnya. Faktor pertama yaitu data di filter berdasarkan jenis kelamin "Perempuan".

Pada hasil analisa di filter berdasarkan jenis kelamin yaitu Perempuan. Hasilnya koresponden perempuan lebih dominan masuk ke dalam *cluster online*. Dimana terdapat 240 koresponden yang masuk ke dalam *cluster online* dan 82 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 85%.

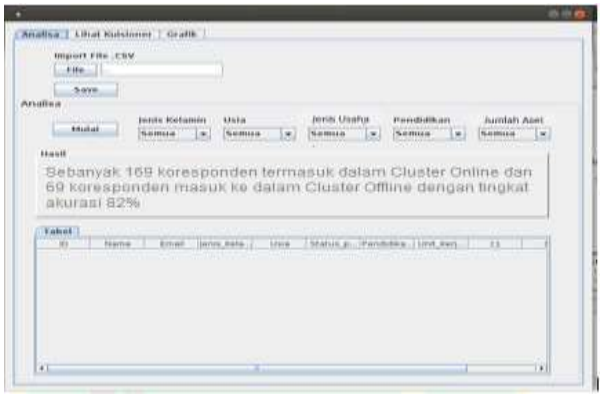
Faktor kedua yaitu simulasi hasil analisa berdasarkan usia dapat dilihat pada gambar berikut ini.

Gambar 4. *Screenshoot* tampilan hasil analisa difilter berdasarkan usia

Pada gambar di atas hasil analisa berdasarkan usia 15 sampai dengan 25 tahun. Hasilnya yakni koresponden *cluster online* lebih dominan dibandingkan *cluster offline*, yaitu sebesar 113 koresponden masuk dalam *cluster online* dan 35 koresponden masuk kedalam *cluster offline*. Dan memiliki tingkat akurasi sebesar 87%. Faktor ketiga yaitu simulasi hasil analisa berdasarkan jumlah *asset* dapat dilihat pada gambar berikut ini.

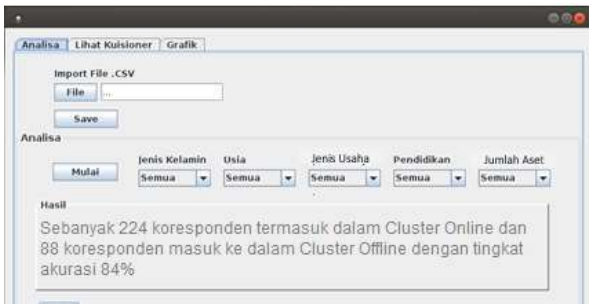
Gambar 5. *Screenshoot* tampilan hasil analisa difilter berdasarkan jumlah *asset*

Pada Gambar 5 terlihat hasil filter berdasarkan jumlah *asset* yang dipilih yaitu kurang dari 300 juta. Hasilnya koresponden lebih banyak masuk ke dalam *cluster online*. Dimana terdapat 273 koresponden yang masuk ke *cluster online*, sedangkan 49 koresponden masuk ke dalam *cluster offline* dan memiliki tingkat akurasi sebesar 96%. Faktor selanjutnya merupakan filter berdasarkan pendidikan dapat dilihat pada gambar berikut.



Gambar 6. *Screenshoot* tampilan hasil analisa difilter berdasarkan pendidikan

Pada Gambar 6 menampilkan hasil analisa berdasarkan pendidikan yaitu SMA/SMK/Aliyah/Pesantren Sederajat. Hasilnya yaitu lebih banyak koresponden yang masuk ke dalam *cluster online*. Pada filter berdasarkan pendidikan ini, sebanyak 169 koresponden masuk ke dalam *cluster online* dan 69 koresponden masuk ke dalam *cluster offline* dengan tingkat akurasinya sebesar 82%. Faktor selanjutnya merupakan filter berdasarkan wilayah pemasaran dapat dilihat pada gambar berikut.



Gambar 7. *Screenshoot* tampilan hasil analisa difilter berdasarkan wilayah pemasaran

Gambar 8. *Screenshoot* tampilan hasil analisa difilter berdasarkan tingkat persaingan

Faktor berikutnya yaitu filter data berdasarkan wilayah pemasaran. Hasilnya koresponden lebih dominan masuk ke *cluster online*, dimana sebanyak 224 koresponden masuk ke *cluster online* dan 88 koresponden masuk ke *cluster online* dengan akurasi sebesar 84%.

Faktor terakhir yaitu filter data berdasarkan tingkat persaingan. Hasilnya koresponden lebih dominan masuk ke *cluster online*, dimana sebanyak 124 koresponden masuk ke *cluster online* dan 30 koresponden masuk ke *cluster online* dengan akurasi sebesar 88%.

Setelah dilakukan simulasi berdasarkan faktor-faktor di atas, penulis melakukan pengujian untuk menghitung akurasi AplikasiKMean dengan rumus sebagai berikut:

$$\text{Akurasi} = \left(\frac{\text{Jumlah TRUE}}{\text{Jumlah TRUE} + \text{Jumlah FALSE}} \; x \; 100\%\right)$$

Dari rumus di atas, seluruh data dihitung dan dibandingkan hasil perhitungan manual dengan hasil AplikasiKMean untuk mendapatkan data tersebut bernilai *TRUE* atau *FALSE*. Hasil akurasinya berupa besaran persentase. Hasilnya dapat dilihat pada bab selanjutnya.

4.6 *Verification, Validation, and Experimentation (Verifikasi, Validasi, dan Eksperimen)*

Dalam tahap penerapan metode simulasi pada AplikasiKMean ini yaitu proses verifikasi, validasi, dan eksperimen. Verifikasi dan validasi dilakukan pada konseptual model, model sistem, dan model simulasi. Pada penulisan penelitian ini, peneliti memulai dengan mengimplementasikan *framework Hadoop* dan *Machine Learning* yaitu algoritma *K-Means* pada sebuah aplikasi berbasis Java.

Tahap verifikasi yang bertujuan untuk memastikan kinerja aplikasi berjalan dengan baik dan memberikan output yang sesuai dengan konsep yang diharapkan. Dengan adanya hubungan antara ketiga tahapan ini (yakni model konseptual, pemodelan sistem, dan model simulasi), maka AplikasiKMean sudah terverifikasi dengan baik sesuai dengan syarat yang ada.

Pada proses validasi dalam AplikasiKMean ini, variabel yang menentukan adalah *Data Filtering*. Proses melakukan validasi dengan membandingkan hasil perhitungan manual *K-Means* dengan hasil program "AplikasiKMean", sehingga dapat menentukan tingkat akurasi dari program "AplikasiKMean" ini.

Sementara pada *experimentation* dilakukan evaluasi hasil dari AplikasiKMean. Tahap ini dimulai dari desain eksperimen sesuai dengan penulis buat pada tahapan simulasi dengan teknik tertentu berdasarkan faktor penguji nilai parameter untuk melakukan analisa pada output hasil dari proses simulasi. Pada penelitian ini penulis membandingkan perbedaan jika data *filtering* yang ada pada program "AplikasiKMean" diubah. Penulis menggunakan factor-faktor data *filtering* berdasarkan jenis kelamin, usia, status pekerjaan, pendidikan terakhir, dan unit kerja. Proses analisa output akan dijelaskan pada poin selanjutnya.

4.7 *Output Analysis Phase (*Fase Analisa Output*)*

Pada fase analisa output dilakukan pengecekan output yang dihasilkan sesuai atau tidak dengan target yang diinginkan pada saat melakukan pemodelan konsep maupun pemodelan simulasi. Pada penelitian ini, hasil output akhir ditampilkan pada sebuah tabel dan grafik dari proses simulasi pada AplikasiKMean yang disimulasikan berdasarkan variabel yang digunakan yaitu data *filtering*.

Penulis melakukan pengecekan output dengan menghitung akurasi AplikasiKMean dengan rumus yang telah dijelaskan di sub bab sebelumnya. Hasil dari perhitungan akurasi program "AplikasiKMean" dapat dilihat pada Tabel 5.

Tabel 5. Akurasi AplikasiKMean

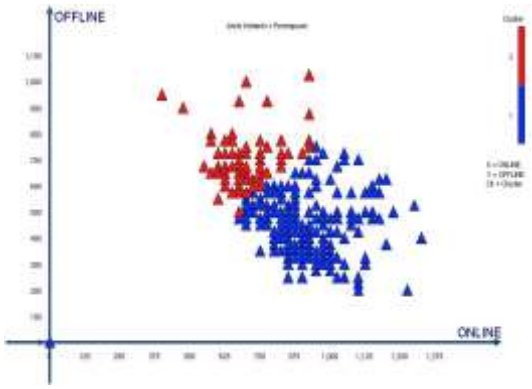
Data *Filtering*	Manual		AplikasiKMean		Akurasi		
	Online	*Offline*	*Online*	*Offline*	T	F	Persentase
Filter data berdasarkan jenis kelamin	288	34	240	82	274	48	85%
Filter data berdasarkan usia "15-25 tahun"	131	17	113	35	130	18	87%
Filter data berdasarkan Jumlah Asset <300 juta	283	39	273	49	312	10	96%
Filter data berdasarkan pendidikan "SMA/SMK/Aliyah/Pesantren Sederajat"	211	27	169	69	196	42	82%
Filter data berdasarkan jenis usaha "Usaha Mikro"	141	13	124	30	137	17	88%
Filter data berdasarkan pemanfaatan Teknologi Informasi "IT sebagai support"	341	25	293	73	318	48	87%
Filter data berdasarkan Wilayah Pemasaran "Lokal"	274	38	224	88	262	50	84%
Filter data berdasarkan jumlah tenaga kerja "1-25"	481	63	398	146	461	83	84%
Filter data berdasarkan tingkat persaingan "sedang"	358	44	286	116	330	72	82%

Dari tabel di atas perbandingan hasil perhitungan data secara manual dan program "AplikasiKMean". Dari koresponden yang masuk dalam *cluster* sunni maupun *cluster* syiah, dan akurasi yang bernilah true maupun false, didapatkan akurasi yang memiliki persentase yang rata – rata yaitu 82%. Sehingga dari hasil akurasi tersebut dapat menjamin bahwa hasil perhitungan algoritma KMeans pada program "AplikasiKMean" sesuai dengan target yang diinginkan pada saat melakukan pemodelan konsep maupun pemodelan simulasi.

Berikut merupakan hasil output simulasi data secara keseluruhan dapat dilihat pada tabel berikut.

Tabel 6. Hasil output simulasi tanpa filter

No	Nama	Nilai C1	Nilai C2	Hasil
1	RESP00001	742.04	781.02	*Online*
3	RESP00002	305.16	380.79	*Online*
4	RESP00003	403.11	482.83	*Online*
5	RESP00004	313.25	254.95	*Offline*
7	RESP00005	500.00	485.41	*Offline*
8	RESP00006	340.04	390.51	*Online*
9	RESP00007	201.56	250.00	*Online*
10	RESP00008	340.04	390.51	*Online*
11	RESP00009	583.63	585.23	*Online*
12	RESP000010	550.57	530.33	*Offline*



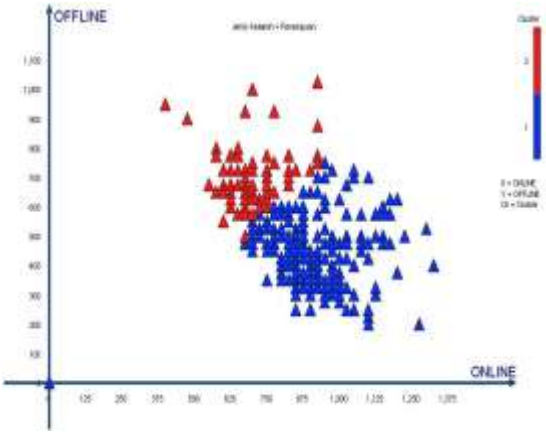
Gambar 9. Grafik hasil output simulasi

Gambar 9 menampilkan grafik hasil output simulasi tanpa filter data. Pada iterasi pertama sebagian besar masuk kedalam *cluster online*, yakni terlihat pada banyaknya segitiga berwarna biru yang mendekati pusat *cluster online* dibandingkan dengan *cluster offline*. Namun pada gambar di atas yang menunjukkan iterasi kedua terlihat adanya penambahan pada *cluster offline* yang terlihat dari banyaknya belah ketupat berwarna merah yang mendekati pusat *cluster offline*. Akan tetapi pada iterasi kedua *cluster online* (segitiga berwarna biru) masih terlihat lebih dominan. Iterasi ini dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai, sehingga dapat disimpulkan bahwa responden cenderung masuk ke *cluster online*. Dari 650 koresponden, sebanyak 474 koresponden masuk ke *cluster online* dan 176 koresponden masuk ke *cluster offline*.

Adapun hasil output simulasi berdasarkan data filtering yang dipilih yakni jenis kelamin perempuan. Penulis mendapatkan hasil outputnya dapat dilihat pada tabel berikut.

Tabel 7. Hasil output simulasi faktor 1 (berdasarkan jenis kelamin perempuan)

No	Nama	Nilai C1	Nilai C2	Hasil
1	RESP00001	742.04	781.02	*Online*
3	RESP00002	305.16	380.79	*Online*
4	RESP00003	403.11	482.83	*Online*
5	RESP00004	313.25	254.95	*Offline*
7	RESP00005	500.00	485.41	*Offline*
8	RESP00006	340.04	390.51	*Online*
9	RESP00007	201.56	250.00	*Online*
10	RESP00008	340.04	390.51	*Online*
11	RESP00009	583.63	585.23	*Online*
12	RESP000010	550.57	530.33	*Offline*



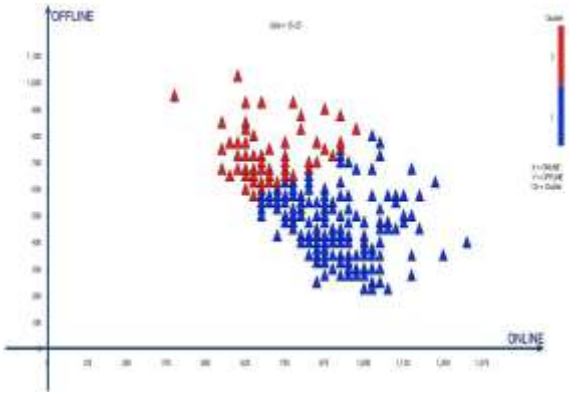
Gambar 10. Grafik hasil output simulasi faktor 1 (berdasarkan jenis kelamin perempuan)

Gambar 10 menampilkan grafik hasil output dari simulasi faktor pertama yaitu data difilter berdasarkan Jenis Kelamin Perempuan. Pada Pada iterasi pertama sebagian besar masuk kedalam *cluster online*, yakni terlihat pada banyaknya segitiga berwarna biru yang mendekati pusat *cluster online*. Namun pada

gambar di atas yang menunjukkan iterasi kedua terlihat adanya penambahan pada *cluster offline* yang terlihat dari banyaknya belah ketupat berwarna merah yang mendekati pusat *cluster offline*. Akan tetapi pada iterasi kedua *cluster online* (segitiga berwarna biru) masih terlihat lebih dominan. Iterasi ini dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai. Sehingga dapat disimpulkan bahwa responden yang berjenis kelamin perempuan cenderung masuk ke *cluster online*. Dari 322 koresponden, sebanyak 240 koresponden masuk ke *cluster online* dan 82 koresponden masuk ke *cluster offline*.

Tabel 8. Hasil output simulasi faktor 2 (berdasarkan usia 15-25 tahun)

No	Nama	Nilai C1	Nilai C2	Hasil
2	RESP00001	292.62	215.06	*Offline*
3	RESP00002	111.80	201.56	*Online*
4	RESP00003	90.14	158.11	*Online*
6	RESP00004	456.21	364.01	*Offline*
7	RESP00005	604.67	601.04	*Offline*
8	RESP00006	456.21	459.62	*Online*
10	RESP00007	375.00	425.73	*Online*
11	RESP00008	654.31	649.04	*Offline*

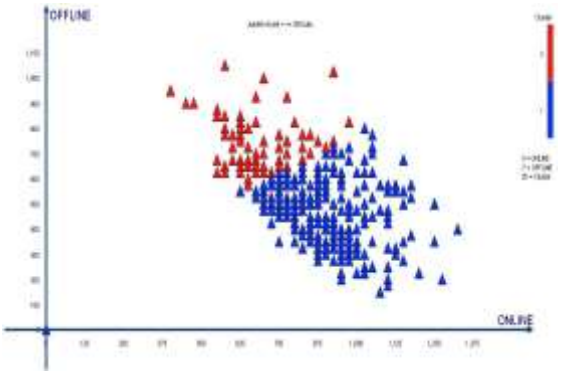


Gambar 11. Grafik hasil output simulasi faktor 2 (berdasarkan usia 15-25 tahun)

Gambar 11 menampilkan grafik hasil output dari simulasi faktor kedua yang difilter berdasarkan usia 15-25 tahun atau kurang. Pada iterasi pertama sebagian besar masuk kedalam *cluster online*, yakni terlihat pada banyaknya titik biru yang mendekati pusat *cluster online*. Namun pada gambar di atas yang menunjukkan iterasi kedua terlihat adanya penambahan pada *cluster offline* yang terlihat dari banyaknya belah ketupat berwarna merah yang mendekati pusat *cluster offline*. Akan tetapi pada iterasi kedua *cluster online* masih terlihat lebih dominan. Iterasi ini dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai. Sehingga dapat disimpulkan bahwa responden yang berumur 15-25 tahun atau kurang cenderung masuk ke *cluster online*. Dari 148 koresponden, sebanyak 113 koresponden masuk ke *cluster online* dan 35 koresponden masuk ke *cluster offline*.

Tabel 9. Hasil output simulasi faktor 3 (jumlah asset < 300 juta)

No	Nama	Nilai C1	Nilai C2	Hasil
2	RESP00001	55.90	50.00	*Offline*
3	RESP00002	742.04	781.02	*Online*
4	RESP00003	395.28	442.30	*Online*
5	RESP00004	111.80	25.00	*Offline*
6	RESP00005	456.21	364.01	*Offline*
7	RESP00006	360.56	301.04	*Offline*
8	RESP00007	285.04	340.04	*Online*
10	RESP00008	388.91	450.69	*Online*
11	RESP00009	70.71	55.90	*Offline*
12	RESP000010	25.00	127.48	*Online*



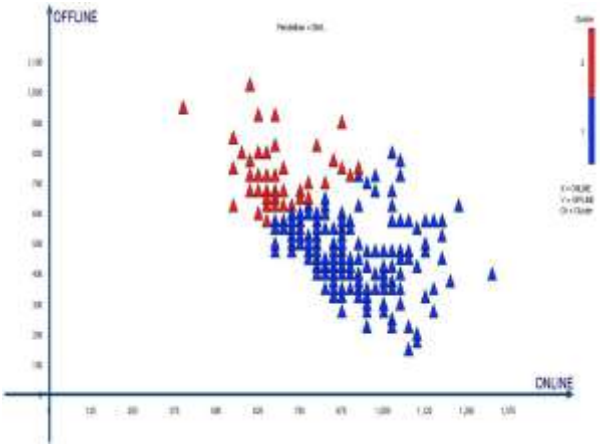
Gambar 12. Grafik hasil output simulasi faktor 3 (berdasarkan jumlah asset <300 juta)

Gambar 12 menampilkan grafik hasil output dari simulasi faktor ketiga yang difilter berdasarkan jumlah asset yaitu kurang dari 300 juta. Pada iterasi pertama sebagian besar masuk kedalam *cluster online*, terlihat pada banyaknya segitiga berwarna biru yang mendekati pusat *cluster online*. Namun pada iterasi kedua *cluster offline* bertambah banyak, terlihat dari belah ketupat berwarna merah yang mendekati pusat *cluster offline*. Iterasi hanya dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai. Sehingga dapat disimpulkan bahwa responden dengan jumlah asset kurang

dari 300 juta cenderung masuk ke *cluster online*, dimana 273 koresponden masuk ke *cluster online* dan 49 koresponden masuk ke *cluster offline*.

Tabel 10. Hasil output simulasi faktor 4 (berdasarkan pendidikan SMA/SMK/Aliyah/sejenisnya)

No	Nama	Nilai C1	Nilai C2	Hasil
2	RESP00001	201.56	111.80	*Offline*
3	RESP00002	235.85	285.04	*Online*
4	RESP00003	353.55	416.08	*Online*
5	RESP00004	419.08	375.83	*Offline*
6	RESP00005	382.43	350.89	*Offline*
8	RESP00006	125.00	141.42	*Online*
10	RESP00007	152.07	176.78	*Online*
11	RESP00008	382.43	350.89	*Offline*
12	RESP00009	55.90	150.00	*Online*
13	RESP000010	477.62	535.02	*Online*

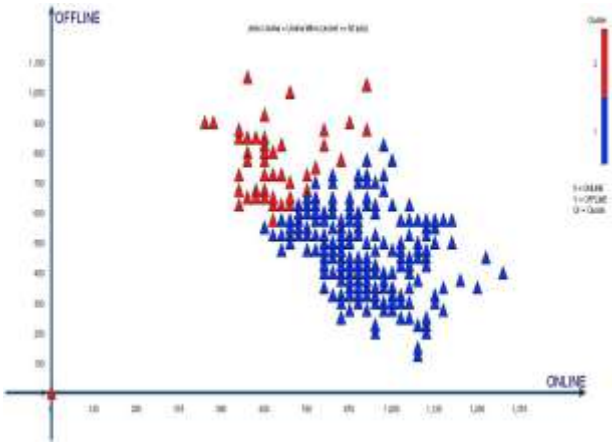


Gambar 13. Grafik hasil output simulasi faktor 4 (berdasarkan pendidikan SMA/SMK/Aliyah/Pesantren sederajat)

Gambar 13 menampilkan grafik hasil output dari simulasi faktor keempat yang difilter berdasarkan Pendidikan terakhir yaitu Pendidikan SMA/SMK/ Aliyah/Pesantren Sederajat. Pada iterasi pertama sebagian besar masuk kedalam *cluster online*, yakni terlihat pada banyaknya segitiga berwarna biru yang mendekati pusat *cluster online*. Begitu juga pada gambar di atas, segitiga berwarna biru yang mendekati pusat *cluster online* terlihat lebih dominan. Iterasi ini dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai. Sehingga dapat disimpulkan bahwa responden dengan pendidikan SMA/SMK/Aliyah/Pesantren cenderung masuk ke *cluster online*. Dari 238 koresponden, sebanyak 169 koresponden masuk ke *cluster online* dan 69 koresponden masuk ke *cluster offline*.

Tabel 11. Hasil output simulasi faktor 5 (berdasarkan unit kerja sains dan teknologi)

No	Nama	Nilai C1	Nilai C2	Hasil
1	RESP0001	395.28	442.30	*Online*
2	RESP0002	55.90	50.00	*Offline*
3	RESP0003	512.96	570.09	*Online*
4	RESP0004	246.22	282.84	*Online*
5	RESP0005	502.49	498.12	*Offline*
6	RESP0006	313.25	254.95	*Offline*
7	RESP0007	419.08	375.83	*Offline*
8	RESP0008	282.84	347.31	*Online*
9	RESP0009	25.00	127.48	*Online*
10	RESP00010	195.26	269.26	*Online*



Gambar 14. Grafik hasil output simulasi faktor 5 (berdasarkan jenis usaha mikro)

Gambar 14 menampilkan grafik hasil output dari simulasi faktor kelima yang difilter berdasarkan Jenis Usaha yaitu Usaha Mikro. Pada iterasi pertama sebagian besar masuk ke dalam *cluster online*, yakni terlihat pada banyaknya segitiga berwarna biru yang mendekati pusat *cluster online*. Namun pada gambar di atas yang menunjukkan iterasi kedua terlihat adanya penambahan pada *cluster offline* yang terlihat dari banyaknya belah ketupat berwarna merah yang mendekati pusat *cluster offline*. Akan tetapi pada iterasi kedua *cluster online* masih terlihat lebih dominan. Iterasi ini dilakukan sebanyak dua kali dikarenakan pada iterasi kedua, nilai sudah stabil dan konvergen atau tidak ada perubahan nilai. Sehingga dapat disimpulkan bahwa responden jenis usaha mikro cenderung masuk ke *cluster online*. Dari 154

koresponden, sebanyak 124 koresponden masuk ke *cluster online* dan 30 koresponden masuk ke *cluster offline*.

Sehingga dapat disimpulkan bahwa dari kelima simulasi yang dilakukan, keseluruhan hasilnya didapatkan bahwa sebagian besar responden yakni cenderung masuk ke dalam *cluster online.*

V. KESIMPULAN

Big data menjadi istilah populer untuk menggambarkan pertumbuhan eksponensial serta ketersediaan data, baik struktural maupun unstruktur. Maka dari itu, dibutuhkan analisis *big data* secara akurat dan juga real time untuk menghasilkan sebuah keputusan yang lebih tepat. Salah satu cara analisa *big data* yang sangat mudah adalah menggunakan HDFS (*Hadoop File Distributed File System*). Selain itu, ada juga pengolahan *big data* menggunakan metode *machine learning.* Metode *Machine Learning* (ML) merupakan cara pengolahan data yang bertumpu pada dasar sains serta pertanyaan *engineering*. Dalam penelitian ini menerapkan metode *unsupervised learning* terhadap data yang sudah dikumpulkan melalui kuesioner yang didapatkan secara manual dan kuesioner *online*. Proses analisa yang dilakukan pada penelitian ini menggunakan salah satu algoritma *unsupervised learning* yakni *K-Mean algorithm* dengan melakukan *cluster*ing terhadap data tentang cara pemasaran produk oleh pengusaha mikro.

Output pertama akan memberikan informasi tentang jumlah data pada tiap *cluster* dan nilai titik pusat pada iterasi pertama dan kedua dalam bentuk teks. Output kedua menjabarkan data koresponden, jarak masing-masing data terhadap pusat *cluster* dan *cluster* masing-masing data dalam bentuk tabel. Output ketiga menampilkan pola populasi data dan pusat *cluster* pada iterasi pertama dan kedua dalam bentuk grafik.

Dari total 650 koresponden, sebanyak 474 koresponden termasuk dalam *cluster* sistem penjualan secara *online* dan 176 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 85%. Pengujian pertama dengan melakukan data *filtering* pada jenis kelamin perempuan, dimana sebanyak 240 koresponden termasuk dalam *cluster online* dan 82 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 85%.

Pengujian kedua dengan melakukan data *filtering* pada usia 15 sampai dengan 25 tahun, dimana sebanyak 113 koresponden termasuk dalam *cluster online* dan 35 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 87%. Pengujian ketiga dengan melakukan data filtering pada jumlah *asset* kurang dari 300 juta, dimana sebanyak 273 koresponden termasuk dalam *cluster online* dan 49 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 96%.

Pengujian keempat dengan melakukan data filtering pada pendidikan SMA/SMK dan sederajat, dimana sebanyak 169 koresponden termasuk dalam *cluster online* dan 69 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 82%.

Pengujian kelima dengan melakukan data *filtering* pada jenis usaha mikro, dimana sebanyak 124 koresponden termasuk dalam *cluster online* dan 30 koresponden termasuk dalam *cluster offline* dan memiliki tingkat akurasi sebesar 88%. Rata-rata akurasi pada kelima pengujian di atas adalah 85%. Akurasi tersebut tergolong cukup besar, sehingga hasil penelitian ini dapat digunakan sebagai *Decision Support System* (DSS), di antaranya dapat digunakan dalam mengembangkan suatu website atau *marketplace* untuk menjadi media promosi bagi para pengusaha mikro.

DAFTAR PUSTAKA

[1] Brownlee, Jason. 2016. *Supervised and Unsupervised Machine Learning Algorithm*. *https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/* (diakses 1 November 2017)

[2] Hurwitz, Judith., Et. al. 2013. *Big Data for Dummies*. New Jersey: John Wiley & Sons, Inc.

[3] Metisen, Melpa. Latipa, Herlina. 2015. *Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan produk pada Swalayan Fadhila*. Jurnal media Infotama, Vol 11. No. 2.

[4] Wijaya, Arim. 2010. *Analisis Algoritma K-Means untuk Sistem Pendukung Keputusan Penjurusan Siswa di MAN Binong Subang*. Skripsi. Bandung: Universitas Komputer Indonesia.

[5] Apriyanti, Nur Ridha., Nugroho, Radityo Adi., Soesanto, Oni. 2015. *Algoritma K-Means Clustering dalam Pengolahan Citra Digital Landsat*. Kalimantan: Universitas Lambung Mangkurat.

[6] Simon Hudson, Li Huang, Martin S. Roth, Thomas J. Madden. 2015. *The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors.* ScienceDirect Intern. J. of Research

[7] Hartati, Sri dan Adi Nugroho. 2012. *MongoDB: Implementasi VLDB (Very Large Database) Untuk Sistem Basis Data Tersebar (Distributed Database)*. Jurnal Teknik Informatika.

[8] Bacquet, Carlos; Gumus, Kubra; Tizer, Dogukan; Zincir-Heywood, A. Nur, and Heywood, Malcolm I. 2009. *A Comparison of Unsupervised Learning Techniques for Encrypted Traffic Identification*. Dalhousie University.

[9] Anonim. 2016. Netbeans IDE 8.2 Information. https://netbeans.org/community/releases/82/, (diakses 3 Juli 2017)

[10] Aryan, P. 2010. *Algoritma K-Means Clustering.* *http://pebbie.wordpress.com/2008/11/13/algoritma-kmeansclustering.Html*. Diakses 1 November 2017.

[11] MacQueen, J. B. 1967. *Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

[12] C. Bacquet, A.N. Zincir-Heywood, and M.I. Heywood. Sep 2009. *An Investigation of Multiobjective Genetic algorithms for Encrypted Traffic Identification. In Computational Intelligence in Security for Information Systems: Cisis' 09*, 2nd International Workshop Burgos, Spain, pages 93–100. Springer.

[13] Jordan, M. I. and Mitchel, T. M. 2015. *Machine Learning: Trends, Perspectives, and Prospects*. American Association for the Advancement of Science.