# DAFTAR TABEL

# BAB I PENDAHULUAN

## 1.1. Latar Belakang

Rumah merupakan salah satu bentuk investasi yang menarik. Permintaan akan tempat tinggal di Kabupaten Bogor masih terus meningkat akibat pertumbuhan penduduk akibat urbanisasi dan migrasi dari luar kota. Pertumbuhan penduduk dan kepadatan penduduk adalah dua faktor yang mempengaruhi pembangunan perumahan di Kabupaten Bogor, dan sekali lagi semua harga menjadi masalah. Perumahan dan tempat tinggal juga merupakan kebutuhan dasar manusia. Selain itu, rumah juga merupakan kebutuhan dasar manusia untuk perbaikan martabat, kualitas hidup, penghidupan. Dan sebagai refleksi dalam upaya meningkatkan taraf hidup dan pembentukan kepribadian kebangsaan dan karakter.

Menurut data di atas, dengan melihat jumlah penduduk di Kabupaten Bogor yang semakin bertambah seiring waktu, masyarakat tentu membutuhkan tempat tinggal yang layak dan nyaman untuk ditinggali. Dengan laju pertumbuhan penduduk yang cukup cepat maka dibutuhkan banyak tempat tinggal yang disiapkan untuk beberapa tahun ke depan. Bersamaan dengan data kurs fluktuatif yang pada tahun 2018, 1 USD (United States Dollar) sama dengan 15.233 Rupiah, pada 2019, 1 USD (United States Dollar) sama dengan 14.212 Rupiah, dan pada 2020, 1 USD (United States Dollar) sama dengan 14.034 Rupiah.

Penelitian ini bertujuan untuk menguji performa dua model, yaitu *clustering* dengan menggunakan *K-Means* dan *Gaussian Mixture Model* dan akan dibandingkan hasil proses dari kedua metode tersebut. *Clustering* merupakan *proses machine learning* yang berfungsi untuk mengelompokkan sekumpulan data yang memiliki kesamaan karakteristik menjadi satu cluster. Sebagai contoh Pada penelitian sebelumnya, ada penelitian yang meggunakan *Gaussian Mixture Model* untuk mengidentifikasi kepadatan kendaraan di jalan tol, dan menghasilkan tingkat akurasi sebesar 91%[15]. Lalu ada identifikasi area kanker ovarium pada citra CT Scan,yang memperoleh hasil tidak cukup bagus dengan persentasi *true positive* sebesar 45% lebih kecil dari *false positive* sebesar 55%[16]. Hasil dari penelitian

tersebut memberi informasi tentang prediksi jumlah total kasus COVID-19 di seluruh dunia dan prediksi tanggal berakhirnya pandemic COVID-19. Lalu terdapat penelitian yang melakukan perbandingan *clustering cloud workloads* dengan menggunakan dua metode yaitu *K-Means* dan *Gaussian Mixture Model*[11]. Dari penelitian tersebut menghasilkan *Gaussian Mixture Model* memberikan *cluster* yang lebih baik, dan lebih rinci daripada *K-Means,* namun dibutuhkan waktu yang lebih lama dibanding proses *clustering* pada K-Means.

## 1.2. Perumusan Masalah

Rumusan masalah yang dihadapi dalam proposal tugas akhir ini secara umum :

1. Bagaimana performa dan kualitas clustering harga rumah menggunakan K-Means dan GMM?

## 1.3. Tujuan

Berdasarkan rumusan masalah, tujuan penelitian Tugas Akhir ini adalah sebagai berikut :

1. Menguji performa dan kualitas hasil clustering harga rumah dengan menggunakan *Gaussian Mixture Model* dan *K-Means.*

## 1.4. Batasan Masalah

Adapun batasan masalah yang menjadi lingkup pada tugas akhir ini adalah sebagai berikut:

1. Data rumah yang dipakai pada penelitian kali ini adalah data rumah di Kabupaten Bogor yang dihimpun dari website olx.co.id
2. Atribut yang digunakan untuk penelitian adalah harga rumah, karena tujuan penelitian ini hanya menguji performansi dari dua model, bukan untuk membuat informasi data rumah secara detail.
3. Pembanding yang digunakan adalah nilai *Silhouette Score*.

## 1.5. Metode Penelitian

Penelitian ini akan dilakukan dengan kegiatan sebagai berikut:

1. Bimbingan dengan dosen pembimbing

   Kegiatan bimbingan dengan dosen pembimbing dilakukan secara daring kepada dosen-dosen pembimbing untuk mendiskusikan hal yang berkaitan dengan penelitian ini.

2. Studi Literatur

   Kegiatan ini meliputi pengumpulan literatur terkait yang berhubungan dengan tugas akhir ini. Beberapa diantaranya mengenai ***Big data, machine learning, K-Means, dan Gaussian Mixture Model*** dari jurnal, buku, dan penelitian-penelitian sebelumnya.

3. Pengumpulan dan Pra-Processing Data

   Melakukan pengumpulan data spefisikasi dan harga rumah melalui proses *scraping* pada *website* olx.co.id. Lalu data yang telah dikumpulkan melalui proses scraping akan dilakukan proses penghapusan variabel-variabel yang tidak digunakan.

4. Proses Clustering

   Pada tahap ini dilakukan proses clustering pada dataset yang telah dibuat, dengan menggunakan model K-Means dan dibandingkan dengan Gaussian Mixture Model.

5. Analisis Perbandingan Hasil Clustering

   Setelah dilakukan proses clustering menggunakan model K-Means, selanjutkan akan dilakukan perbandingan hasil clustering yang menggunakan model K-means dengan hasil clustering menggunakan Gaussian Mixture Model.

6. Penulisan Laporan

   Setelah melakukan tahap-tahap pengujian, lalu dilakukan penyusunan laporan tugas akhir sesuai sistematika penulisan yang telah ditetapkan.

# BAB II KAJIAN PUSTAKA

## 2.1. Big Data

Menurut Dumbill (2012), *big data* merupakan kumpulan data yang ukurannya sangat besar dan alurnya sangat cepat yang melebihi kapasitas atau tidak sesuai dengan struktur arsitektur dari sistem *database* yang ada [1]. Sedangkan Eaton, Dirk, Tom, George, dan Paul (2015), menuturkan *big data* adalah istilah yang disematkan untuk informasi yang tidak dapat dianalisa atau diproses dengan alat tradisional [2]. Dari pendapat para ahli yang telah disebutkan, dapat diambil kesimpulan bahwa *big data* merupakan data yang volumenya sangat besar dan atau alurnya sangat cepat yang melebihi dari kapasitas dan tidak sesuai dengan struktur arsitektur dari sistem *database* yang ada, yang tidak dapat diproses menggunakan alat yang tradisional.

Mengumpulkan data dengan jumlah besar bukan merupakan upaya yang baru-baru ini terjadi. Itu telah terjadi sudah sejak lama, sebagai contoh dahulu para peneliti sejarah mengumpulkan data-data tentang peninggalan sejarah di berbagai negara yang menjadi situs keajaiban dunia. namun data-data tersebut masih banyak yang terbuka untuk didebatkan, karena tidak jelas apakah sebagian besar kumpulan *big data* tersebut memenuhi kriteria agar bisa disebut *big data*[3]. Mengacu pada 5V dari *big data*, yaitu *Volume, Variety, Velocity, Variability, Veracity,* berikut penjelasan dari 5V yang membuat *big data* memiliki keunikan[4]:

1. Volume: *Big data* sering melebihi kapasitas penyimpanan yang tersedia. selama beberapa periode terakhir, kapasitas penyimpanan meningkat secara pesat menyebabkan *big data* yang sebelumnya dianggap ancaman kini tidak lagi dianggap seperti itu; namun tetap saat ini sangat banyak informasi yang tersedia dan tersebar dimanapun.
2. Variety: *Big data* saat ini juga dihasilkan dari berbagai macam aspek, tipe, dan format; terstruktur maupun yang tidak terstruktur. beberapa

aspek itu seperti budaya, ekonomi, politik, sosial, psikologi, biologi, fisika, geospasial, dan masih banyak lagi.

3. Velocity: Bukan hanya kecepatan pada jumlah data besar yang dihasilkan, tapi juga kecepatan data-data yang harus diperoleh dan diproses. Juga ada *big data* yang penting untuk waktu yang sangat singkat, seperti jadwal pesawat yang tertunda, fluktuasi harga saham; itu membutuhkan cara kompleks untuk menambah laju data agar data dapat terproses dengan baik.
4. Variability: Ketika dilihat, *velocity* dan *volume big data* tampak stabil; namun sebenarnya mereka agak bervariasi dikarenakan terdapat inkonsistensi pada laju aliran *big data*, seperti contoh kasus tren pada google atau twitter, dengan cepat bisa muncul dengan adanya interaksi pada masyarakat pengguna internet.
5. Veracity: *Big data* seringkali tidak berbentuk format yang mudah dieksplorasi atau terhubung satu sama lain.

## 2.2. Clustering

*Data Mining* adalah metode yang digunakan untuk mengekstraksi informasi prediktif tersembunyi pada database, ini adalah teknologi yang sangat potensial bagi perusahaan dalam memberdayakan data wirehouse[5]. *Data Mining* dibagi menjadi 2 kategori, *Descriptive Mining* dan *Predictive Mining. Deskriptive mining* adalah proses menemukan karakteristik penting dari data dalam satu basis data. Contoh teknik data mining yang termasuk bagian dari *descriptive mining* adalah *clustering*, *asosiation*, dan *sequential mining*. *Predictive Mining* adalah proses untuk menemukan pola dari data dengan menggunakan beberapa variable lain di masa depan. Contoh teknik data mining yang termasuk bagian dari *predictive mining* adalah klasifikasi.

*Clustering* adalah metode pengelompokkan data yang tidak memiliki pengawasan, dalam artian metode ini menggunakan *dataset* yang belum diklasifikasikan ke kelompok apapun dan tidak mempunyai kelas maupun atribut. Metode *Clustering* bertujuan untuk mendeskripsikan data yang belum memiliki kelas dan atribut [6]. *Clustering* akan membagi data yang belum memiliki atribut

ke dalam kelompok-kelompok berdasarkan kemiripan kriteria *cluster* tersebut. Namun objek data mungkin saja berada di lebih dari *cluster.* Pada tugas akhir ini, *clustering* akan mengelompokkan rumah berdasarkan harga dengan melihat dari beberapa aspek seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan tahun berapa bangunan tersebut didirikan.

Tujuan utama metode analisis *clustering* adalah mengelompokkan beberapa objek berdasarkan kemiripan karakteristik antar objek-objek tersebut. Objek bisa berupa produk (barang dan jasa), benda, serta orang.

Clustering merupakan Teknik machine learning yang mengelompokkan data yang belum memiliki kelas ke dalam kelompok-kelompok berdasarkan kemiripan karakteristik antar data satu dengan yang lain.

## 2.3. Gaussian Mixture Model

*Gaussian Mixture Model* merupakan teknik clustering yang tidak memiliki pengawasan yang akan membentuk cluster berdasarkan probabilitas *density* dengan menggunakan algoritma *Expectation-Maximization*. *Mean* dan Covariance yang lebih baik dibanding model *K*-Means memberikan GMM kemampuan untuk menghasilkan perhitungan kuantitatif pada *fitness* per jumlah cluster yang lebih baik[11]. *Gaussian Mixture* dengan *Expectation-Maximization* ini memiliki kemampuan untuk mencari dari kumpulan data yang tidak memiliki kelas dengan melakukan iterasi pada setiap *instance* untuk mencari nilai likelihood tertinggi. Pada pemodelan data menggunakan GMM, diperlukan 3 parameter yaitu mean, kovarians, dan koefisien. Dan nilai parameter data diambil dari proses Algoritma E-M yang akan menghasilkan nilai likelihood yang optimal pada parameter.

Algoritma E-M memiliki dua tahap, yaitu:

1. Tahap pertama yaitu Expectation, di tahap ini fungsi akan menghitung salah satu data yang menjadi acuan untuk ekspektasi.
2. Tahap kedua yaitu Maximization, pada tahap ini setelah didapatkan nilai ekspektasinya lalu parameter akan dihitung nilai optimalnya yang akan dijadikan acuan untuk mendapatkan nilai *likelihood* tertinggi terhadap nilai ekspektasi.

Persamaan yang digunakan pada Gaussian Mixture Model ini, yaitu:

$$p(X) = \sum_{k=1}^{K} \pi_k \, N(X|\mu_k, \sum_k)$$

Keterangan:

K = Jumlah komponen

$\mu$ = Mean

$\pi$ = Koefisien / Bobot

$\sum$ = Varians

## 2.4. K-Means

K-Means adalah metode clustering yang digunakan untuk mengelompokkan data dengan cara menemukan pusat cluster atau centroid. Kemiripan data dengan data lain didasarkan pada seberapa dekat data tersebut dengan pusat atau centroid. Proses *K-Means* akan berhenti ketika jumlah iterasi yang ditentukan telah tercapai dan data telah berhasil dikelompokkan. Secara umum pengelompokkan data pada *K-Means* menggunakan jarak Euclidian kuadrat sebagai ukuran kesamaan untuk keanggotaan cluster[13]:

$$d_{sq} = \sum_{i=1}^{D} (x_i - y_i)^2$$

Keterangan:

x,y = titik data

D = ruang dimensi

Untuk memproses data algoritma K-means Clustering , data dimulai dengan kelompok pertama centroid yang dipilih secara acak, yang digunakan sebagai titik awal untuk setiap cluster, dan kemudian melakukan perhitungan berulang untuk mengoptimalkan posisi centroid. Proses ini berhenti atau telah selesai dalam mengoptimalkan cluster ketika centroid telah stabil tidak ada perubahan dalam nilai-nilai mereka karena pengelompokan telah berhasil, dan ketika jumlah iterasi yang ditentukan telah tercapai.