BAB II

TINJAUAN PUSTAKA

2.1 Usaha

Usaha adalah setiap jenis usaha baik perseorangan maupun persekutuan yang berdasarkan atas hukum denda ataupun persetujuan memakai atau menguasai suatu benda tak bergerak untuk keperluan menjalankan kerja nafkahnya atau perusahaannya, yang untuk mendirikannya atau memperluasnya menurut peraturan perundang-undangan (Perda Kota Pekanbaru, 2012).

2.2 Izin Gangguan (HO)

Gangguan adalah perbuatan dan/atau kondisi yang tidak menyenangkan atau mengganggu kesehatan, keselamatan, ketentraman dan/atau kesejahteraan terhadap kepentingan umum.

Izin gangguan adalah pemberian izin tempat usaha/kegiatan kepada orang pribadi atau badan di lokasi tertentu yang dapat menimbulkan bahaya, kerugian dan gangguan.

Surat izin gangguan atau biasa disebut dengan HO (*Hinder Ordonnanntie*) adalah surat yang menyatakan tidak adanya keberatan dan gangguan atas lokasi usaha yang kita jalankan. Salah satu syarat umum untuk mendapatkan surat ini adalah tidak adanya pencemaran lingkungan atau tidak ada dampak negative terhadap lingkungan dari usaha yang kita lakukan.

Retribusi izin gangguan adalah suatu pungutan yang harus dibayar/dilunasi oleh pemegang izin yang mendirikan dan atau keperluan tempat usaha.

2.3 Tata Cara dan Persyaratan Perizinan

Untuk memperoleh izin gangguan dilakukan dengan cara mengajukan permohonan tertulis menurut formulir dan daftar isian. Adapun persyaratan yang harus dipenuhi oleh pemohon adalah :

1. Pas photo 3 x 4 berwarna 2 lembar.
2. Menunjukkan KTP dan melampirkan fotocopy KTP.
3. Skema lokasi tempat usaha.
4. Fotocopy Akta Perusahaan (apabila berbadan hukum).
5. Surat Keterangan pemeriksaan alat pemadam kebakaran.
6. Surat Bukti Pemilikan Tanah/ Bangunan.
7. Fotocopy surat perjanjian sewa menyewa (bila menyewa atau kontrak).
8. Fotocopy surat Izin Mendirikan Bangunan (Hotel, Bangunan Skala Besar).
9. Surat keterangan Fiskal Daerah (lunas pajak reklame dan PBB tahun terakhir).
10. Surat rekomendasi camat setempat (kecuali untuk perkantoran dan pertokoan).
11. Khusus untuk HO Hiburan Umum harus ada Rekomendasi RT dan RW.
12. Foto tempat usaha.
13. Pertimbangan teknis/ rekomendasi dari instansi teknis sesuai dengan jenis usaha (jika dianggap perlu).

2.4 Struktur dan Besarnya Tarif Retribusi

Retribusi dihitung didasarkan atas perkalian Luas Tempat Usaha, Indeks Gangguan, Indeks Lokasi Jalan Satuan Retribusi Gangguan. Luas Tempat Usaha adalah luas ruang kantor, ruang penjualan, ruang toko, ruang gudang, ruang penimbunan, pabrik, ruang terbuka dan ruang lainnya yang digunakan untuk penyelenggaraan usaha.

Indeks gangguan sebagaimana yang telah ditetapkan adalah sebagai berikut :

Tabel 2.1 Indeks Gangguan

No.	Kategori Intensitas Gangguan	Indeks Gangguan
1.	Jenis Usaha dengan Intensitas Gangguan Kecil	1
2.	Jenis Usaha dengan Intensitas Gangguan Sedang	1.5
3.	Jenis Usaha dengan Intensitas Gangguan Besar	2

Indeks lokasi sebagaimana yang telah ditetapkan adalah sebagai berikut :

Tabel 2.2 Indeks Lokasi

No.	Kategori Intensitas Gangguan	Indeks Gangguan
1.	Indeks lokasi jalan Lingkungan	1
2.	Indeks lokasi jalan Kolektor	1.5
3.	Indeks lokasi jalan Arteri	2

Rumus untuk mendapatkan jumlah retribusi adalah :

Luas Tempat Usaha x Indeks Gangguan x Indeks Lokasi x Tarif Retribusi

Tarif retribusi nya adalah sebagai berikut :

- Luas tempat usaha 1 s/d 100 m² = Rp.8000/m²
- Luas tempat usaha 100 s/d 200 m² = Rp.7000/m²
- 201 m² keatas = Rp.1000/m² *(Setelah di dapat perkalian 200 m²)*

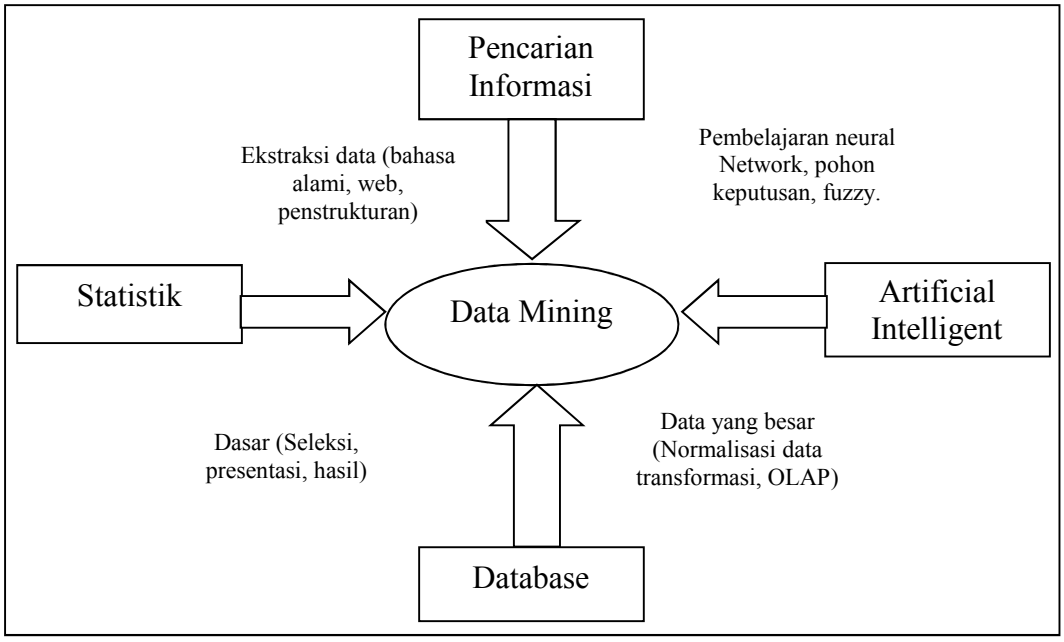
Dalam pengurusan izin gangguan ini, BPTPM akan mengeluarkan surat izin paling lama 5 hari kerja.

2.5 Pengertian Data Mining

Data mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode *data mining* ini dapat digunakan untuk mengambil keputusan di masa depan. *Data mining* ini juga dikenal dengan istilah *pattern recognition (Santosa, 2007).*

Ada banyak definisi dari *data mining* yang dapat diperoleh dari buku maupun jurnal yang ada. Diantaranya adalah sebagai berikut:

1. Menurut Turban, dkk (2005) *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengektraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar.
2. Menurut Larose (2005) *Data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecendrungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti statistic dan matematika.
3. Menurut Pramudiono (2006) *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.
4. Menurut Gorunescu (2011) *Data mining* adalah perpaduan dari statistic, *Artificial Intelligent* dan *Database*.



Gambar 2.1 Bidang Ilmu Data Mining

2.6 Tahapan Data Mining

Berikut tahapan proses KDD secara garis besar menurut Kusrini dan Luthfi (2009) :

1. *Data Selection*

   Seleksi dari sekumpulan data sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining* disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing / Cleaning*

   Proses *cleaning* yang menjadi fokus KDD mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses"memperkaya"data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD,seperti data atau informasi eksternal

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih,sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

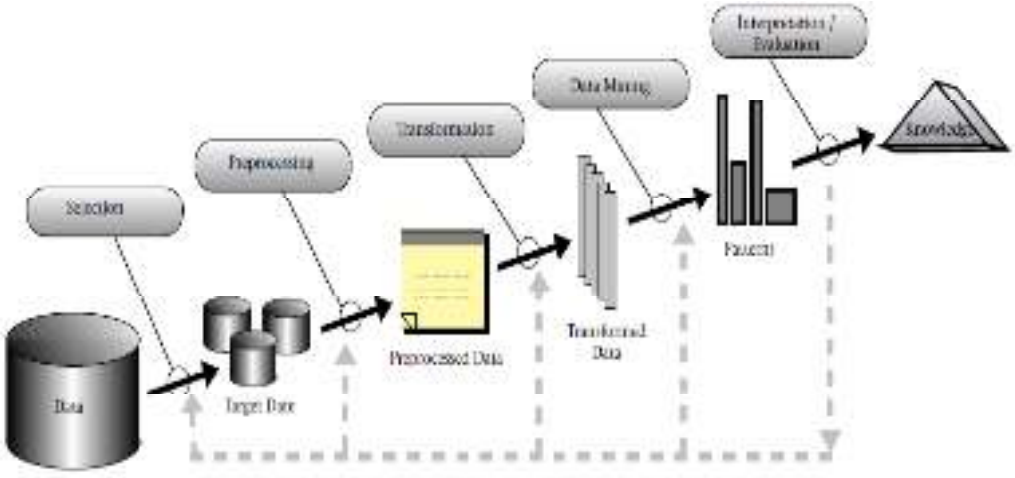
4. *Data Mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation / Evaluation*

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan

interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.



Sumber : Fayyad, dkk (1996)

Gambar 2.2 Tahapan dalam Proses *Data Mining*

2.7 Metode dalam Data Mining

Secara umum dalam *data mining* ada beberapa metode yang dapat digunakan untuk mendapatkan hasil yang diinginkan, diantaranya (Larose, 2005):

1. Deskripsi

   Deskripsi merupakan metode yang digunakan untuk menyajikan pola-pola dan *trend* yang ada dalam data.

2. Estimasi

   Metode estimasi didasarkan pada kelengkapan data yang terdiri dari *target variable* dan *predictor variable*.*Target variable* adalah variabel yang diestimasi. Sedangkan *predictor variable* adalah variabel yang dijadikan pertimbangan untuk mengestimasi *target variable*. Pada metode estimasi *target variable* haruslah berupa data numerik.

3. Klasifikasi

   Klasifikasi bertujuan untuk membagi data ke dalam kelas-kelas tertentu berdasarkan data di masa lalu. Dalam klasifikasi terdapat 2 tipe variabel yaitu *target variable* dan *predictor variable*.*Target variable* berbentuk kategorikal yang nantinya akan menjadi kelas-kelas dalam klasifikasi.*Predictor variable* adalah variabel yang menjadi dasar atau acuan untuk mengklasifikasikan data ke dalam kelas-kelas yang ada.

4. Kluster

   Metode kluster (*clustering*) adalah metode yang mengelompokkan objek atau atribut yang sama ke dalam suatu grup atau kelas. Meskipun kelihatannya kluster mirip dengan klasifikasi, tetapi pada proses nya kluster berbeda dengan klasifikasi. Perbedaan yang paling mendasar adalah kluster tidak memiliki *target variable* seperti klasifikasi. Metode kluster bertujuan untuk menghimpun data yang memiliki kesamaan dalam suatu kelas dan data yang berbeda dihimpun dalam kelas lain. Umumnya hasil dari kluster dapat digunakan sebagai *input* untuk metode lain.

5. Asosiasi

   Metode asosiasi memiliki tujuan untuk menemukan aturan-aturan yang berkaitan antar dua variabel atau lebih. Salah satu contoh penggunaan asosiasi dalam hal mengetahui barang apa yang dibeli secara bersamaan oleh pelanggan supermarket.

6. Prediksi

   Metode prediksi memiliki jangka waktu. Prediksi bertujuan untuk memperkirakan kejadian di masa akan datang. Contoh kasus prediksi adalah memprediksi harga saham 5 bulan mendatang, prediksi persentase tingkat pertumbuhan penduduk 10 tahun kemudian.

2.8 Proses Data Mining

Secara sistematis, ada tiga langkah utama dalam data mining (Gonunesce,2011) :

1. Eksplorasi / pemrosesan awal data

   Eksplorasi / pemrosesan awal data terdiri dari 'pembersihan' data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subsest fitur, dan sebagainya.

2. Membangun model dan melakukan validasi terhadapnya

   Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis cluster, deteksi anomaly, analisis asosiasi, analisis pola sekuensial, dan sebagainya. Dalam beberapa referensi, deteksi anomaly juga masuk dalam langkah eksplorasi. Akan tetapi, deteksi anomaly juga dapat digunakan sebagai algoritma utama, terutama untuk mencari data-data yang special.

3. Penerapan

   Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

2.9 Analisis Clustering

Analisis kluster (*clustering*) adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa group berdasarkan kesamaan-kesamaan yang telah ditentukan sebelumnya (Gorunescu, 2011).

Clustering merupakan pekerjaan yang memisahkan data/vector kedalam sejumlah kelompok (*cluster*) menurut karakteristiknya masing-masing. Data-data yang mempunyai kemiripan karakteristik akan berkumpul dalam *claster* yang sama, dan data-data dengan karakteristik berbeda akan terpisah dalam *claster* yang berbeda. Tidak diperlukan label kelas untuk setiap data yang diproses dalam *clustering* karena

nantinya label baru bisa diberikan ketika *claster* sudah terbentuk. Karena tidak adanya target label untuk setiap data, maka *Clustering* sering disebut juga pembelajaran tidak terbimbing (*unsupervised learning*).

Karena tidak ada label kelas yang digunakan dalam prosesnya, *clustering* sangat cocok untuk melakukan *clustering* data yang label kelasnya memang sulit didapatkan pada saat pembagian fitur. Pada *clustering,* segera setelah *cluster* terbentuk, maka label kelas untuk setiap data dapat diberikan dengan mengamati hasil *cluster*. Pekerjaan tidak terbimbing (*unsupervised*) seperti *clustering* juga sering digunakan untuk mengeksplorasi dan mengkarakteristikkan set data sebelum menjalankan pekerjaan yang terbimbing (*supervised*). Karena *Clustering* tidak membutuhkan label kelas, yang perlu dicatat adalah kemiripan (*similarity*) harus didefinisikan berdasarkan pada atribut objek. Definisi kemiripan dan metode dalam data yang dikelompokkan berbeda tergantung pada algoritma *clustering* yang diterapkan. Algoritma *clustering* yang 'bagus' digunakan tergantung pada penerapan set data yang diproses.

Adapun tujuan dari data *clustering* ini adalah untuk meminimalisasikan *objective function* yang diset dalam proses *clustering*, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalisasikan variasi antar *cluster*.

Secara garis besar, terdapat beberapa metode *clustering* data. Pemilihan metode *clusterisasi* bergantung pada tipe data dan tujuan *clustering* itu sendiri. Metode-metode beserta algoritma yang termasuk didalamnya meliputi (Baskoro, 2010):

1. *Partitioning Method*

   Membangun berbagai partisi dan kemudian mengevaluasi partisi tersebut dengan beberapa kriteria, yang termasuk metode ini meliputi algoritma K-Means, K- Medoid, PROCLUS, CLARA, CLARANS, dan PAM.

2. *Hierarchical Methods*

   Membuat suatu penguraian secara hierarkikal dari himpunan data dengan menggunakan beberapa kriteria. Metode ini terdiri atas dua macam, yaitu *Agglomerative* yang menggunakan strategi *bottom-up* dan *Disisive* yang menggunakan strategi *top-down*. Metode ini meliputi algoritma BIRCH, AGNES, DIANA, CURE, dan CHAMELEON.

3. *Density-based Methods*

   Metode ini berdasarkan konektivitas dan fungsi densitas. Metode ini meliputi algoritma DBSCAN, OPTICS, dan DENCLU.

4. *Grid-based Methods*

   Metode ini berdasarkan suatu struktur granularitas multi-level. Metode clusterisasi ini meliputi algoritma STING, WaveCluster, dan CLIQUE.

5. *Model-based Methods*

   Suatu model dihipotesakan untuk masing-masing cluster dan ide untuk mencari best fit dari model tersebut untuk masing-masing yang lain. Metode clusterisasi ini meliputi pendekatan statitik, yaitu algoritma COBWEB dan jaringan syaraf tiruan, yaitu SOM.

2.10 Algoritma K-Means

K-means termasuk dalam *partitioning clustering* yaitu setiap data harus masuk dalam *cluster* tertentu dan memungkinkan bagi setiap data yang termasuk dalam *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke *cluster* yang lain. K-means memisahkan data ke *k* daerah bagian yang terpisah, dimana *k* adalah bilangan integer positif.

Algoritma K-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi set data ke dalam sejumlah K *cluster* yang sudah diterapkan di awal. Algoritma K-Means sederhana untuk diimplementasikan dan dijalankan, relative cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining (Wu dan Kumar, 2009).

K-Means adalah suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity)* antara satu data dengan data yang lain. Sehingga setiap klaster akan berisi data yang saling mirip (Han, 2006). *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan *unsupervised.* Maksudnya, metode ini tanpa adanya latihan *training* serta tidak memerlukan target output. Didalam *data mining* terdapat dua jenis metode *clustering* yang digunakan untuk pengelompokan data, yaitu *Hierarchical clustering* dan *Non-hierarchical clustering* (Han, 2006).

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan megelompokkan dua atau lebih objek yang memiliki kesamaan yang lebih dekat. Kemudian diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya hingga membentuk pohon antar objek. Sedangkan *Non-hierarchical clustering* dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan. Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini disebut dengan K-*means clustering.*

Algoritma K-*means clustering* merupakan salah satu metode *Non-hierarchical clustering* yang mengelompokan data dalam bentuk satu atau lebih *cluster/*kelompok. Data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster*/kelompok. Dan data yang memiliki karakteristik yang berbeda dikelompokkan dengan *cluster/*kelompok yang lain. Sehingga *data* yang berada dalam satu *cluster* /kelompok memiliki tingkat variasi yang kecil (Han, 2006).

Data *cluster*ing menggunakan metode *K-means* ini secara umum dilakukan dengan algoritma dasar sebagai berikut (Agusta, 2007):

1. Tentukan jumlah *cluster* (k).
2. Alokasikan data ke dalam *cluster* secara random.
3. Hitung *centroid*/ rata-rata dari data yang ada di masing-masing *cluster.*
4. Alokasikan masing-masing data ke *centroid*/ rata-rata terdekat.
5. Kembali ke Step 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang

ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan.

Dalam praktiknya, peneliti sering dihadapkan pada fitur dengan nilai yang terletak dalam jangkauan nilai berbeda. Akibatnya, fitur dengan nilai atau jangkauan yang besar mempunyai pengaruh yang lebih besar dalam fungsi biaya daripada fitur dengan nilai kecil atau jangkauan kecil. Untuk menanggapi masalah ini, bisa digunakan teknik normalisasi fitur sehingga semua fitur akan berada dalam jangkauan yang sama. Untuk menskalakan dalam jangkauan [0,1] dapat digunakan persamaan berikut (Prasetyo, 2014).

$$X_{ik} = \frac{X_{ik} - MIN(X_k)}{MAX(X_k) - MIN(X_k)} \quad (2.1)$$

Ada beberapa cara yang digunakan untuk mengukur jarak data kepusat kelompok, diantaranya adalah *L1* (*Manhattan/City Block*) *distance space*, *L2* (*Euclidean*) *distance space*, dan *Lp (Minkowski) distance space* (Agusta, 2007). Pada penelitian ini akan menggunakan euclidean untuk menghitung jarak *centroid*.
Persamaan euclidean adalah sebagai berikut :

$$D(X_1, X_2) = \| X_2 - X_1 \| = \sqrt{\sum_{j-1}^{p} |X_{2j} - X_{1j}|^2} \quad (2.2)$$

p : Dimensi data

|.| : Nilai absolut

Untuk menentukan nilai *k* terbaik digunakan *sum of squered eror* (SSE).
SSE didefinisikan sebagai berikut (Baehaki, 2014).

$$SSE = \sum_{i=1}^{k} \sum_{x \in Ci} dist(C_i X)^2 \quad (2.3)$$

Dengan mengetahui nilai SSE dari tiap *k* maka dapat diketahui *clustering* yang menghasilkan nilai kesamaan atau kemiripan terbaik. *Clustering* yang memiliki nilai SSE terkecil adalah *clustering* dengan hasil terbaik.

2.11 Principal Component Analysis

Algoritma *Principal component analysis* (PCA) merupakan salah satu teknik analisa dalam ilmu statistik yang digunakan untuk menyederhanakan suatu data, sehingga terbentuk sistem koordinat baru dengan varians maksimum dan digunakan untuk pengelompokan data berdasarkan kemiripan data (Saleh, 2013). Perkembangan PCA diperkenalkan pertama kali oleh Karl Pearson pada tahun 1901 seiring perkembangan teknologi komputer dan kemajuan bidang matematika.

Langkah langkah dalam algoritma *principal component analysis* adalah sebagai berikut (Soleh, 2013) :

a. Identifikasi data

   Menurut Masnurulyani (2008) dalam Soleh (2013), Dalam identifikasi data ini ada beberapa proses untuk mendapatkan variabel asli yang akan ditransformasi ke dalam variabel baku, yaitu :

   1. Menentukan jumlah variabel yang akan digunakan sebagai pembanding (kriteria penilaian).
   2. Buat tabel variabel asli dengan jumlah variabel tersebut sebagai jumlah kolomnya dan jumlah sample sebagai jumlah barisnya.
   3. Setelah semua sample dimasukkan ke dalam tabel, dicari rata-rata dari setiap variabel ($\bar{X}$) dengan rumus :

$$\bar{X} = \frac{\sum_{i=n}^{n} x_i}{(n)} \quad \text{.....} \quad (2.4)$$

   4. Setelah mendapatkan rata-rata ($\bar{X}$) dari setiap variabel, dicari Simpangan Baku dari setiap variabel dengan rumus :

$$S = \sqrt{S^2} \text{ atau } \sqrt{\frac{\sum_{i=n}^{n}(X_i - \bar{X})^2}{(n-1)}} \quad \text{.....} \quad (2.5)$$

b. Transformasi variabel asli kedalam bentuk variabel standar (standarisasi data). Transformasi ini bertujuan untuk membuat variabel baku yang lebih sederhana dengan rumus :

$$\overline{Z_i} = \frac{X_{ik} - \overline{X_k}}{SX_i} \quad \text{.....} \quad (2.6)$$

c. Menghitung nilai *covariance*

Mencari nilai *covariance* dari masing-masing variabel dengan menggunakan rumus berikut :

$$\text{Cov (X,X)} = \frac{\sum_{i=n}^{n}(X_i - \bar{X})^2}{(n-1)} \quad \text{.....} \quad (2.7)$$

dan

$$\text{Cov (X,Y)} = \frac{\sum_{i=n}^{n}(X_i - \bar{X})\ (X_j - \bar{X})}{(n-1)} \quad \text{.....} \quad (2.8)$$

Atau

$$\text{A} = \frac{1}{(n-1)} Y^t Y \quad \text{.....} \quad (2.9)$$

Setelah dapat keseluruhan covariance dari seluruh variabel, tampilkan data tersebut dalam bentuk matriks R. setelah itu cari R2 dengan mengkalikan matriks tersebut dengan dirinya sendiri.

d. Menghitung vektor ciri (*eigen vektor*)

Untuk mencari vektor ciri, perlu dilakukan langkah-langkah berikut :

1. Cari Vektor Awal (a'0)

   Vektor awal didapat dengan memperhatikan jenis bilangan pada baris pertama dari matriks R2. Jika bilangannya bernilai positif maka nilainya adalah 1, jika bilangannya bernilai negatif maka nilainya adalah -1.

2. Cari Vektor Matriks (*a'0 Rn)*

   Vektor matriks didapat dari perkalian matriks dan vektor awal.

3. Cari Iterasi

   Iterasi didapat dari pembagian Elemen terbesar dari vektor matriks dengan seluruh anggota dari vektor matriks tersebut.

4. Lakukan langkah b dan c sampai hasil iterasi terakhir sama dengan hasil iterasi sebelumnya.
5. Cari Vektor Ciri (Eigen Vector)

Setelah didapat hasil akhir iterasinya, normalkan dengan rumus berikut:

$$A_{ij} = \frac{a}{a_i^2 + \cdots + a_j^2} \quad \text{..... (2.10)}$$

e. Penentuan Komponen Utama (*principal Component*)

yi = yhi = ai zh, ….. yhk = ak zn.....

dimana *zn* merupakan vektor skor baku dari variabel yang diamati pada obyek pengamatan ke-h, *yhi* adalah skor komponen ke-i dari obyek pengamatan ke-h, *yhk* adalah skor komponen ke-k dari obyek pengamatan ke-h dan n adalah ukuran contoh. Setelah dapat hasil dari yhi, maka data bisa dikelompokkan menjadi 3 bagian dengan aturan sebagai berikut :

Tinggi: jika $yhi > \bar{yi} + Sy1$ ..... (2.11)

Tinggi: jika $\bar{yi} - Sy1 < yhi < \bar{yi} + Sy1$ ..... (2.12)

Tinggi: jika $yhi < \bar{yi} - Sy1$ ..... (2.13)

Mengenai layak atau tidaknya analisis faktor, maka perlu dilakukan uji *Kaiser mayer Olkin* (KMO) dan *Barlett Test.* jika nilai KMO kurang dari 0,5 maka analisis faktor tidak layak dilakukan. Sedangkan *barlett test* digunakan untuk menguji benar variabel-variable yang dilibatkan berkorelasi.

Hipotesis :

H0: tidak ada korelasi antar variabel bebas

H1: ada korelasi antar variabel bebas

Kriteria uji dengan melihat *p*-value (signifikan): terima h0 jika sig. >0,05 atau tolak H0 jika sig.< 0,05.

Keuntungan penggunaan *Principal Component Analysis* (PCA) dibandingkan metode lain :

1. Dapat menghilangkan korelasi secara bersih (korelasi = 0) sehingga masalah multikolinearitas dapat benar-benar teratasi secara bersih.
2. Dapat digunakan untuk segala kondisi data / penelitian
3. Dapat dipergunakan tanpa mengurangi jumlah variabel asal
4. Walaupun metode Regresi dengan PCA ini memiliki tingkat kesulitan yang tinggi akan tetapi kesimpulan yang diberikan lebih akurat dibandingkan dengan pengunaan metode lain.

2.12 Validitas Cluster

Dalam klasifikasi, evaluasi system sudah menjadi bagian penting dalam proses pembangunan model klasifikasi, ukuran dan metode untuk mengevaluasi, seperti akurasi, *precision* dan *recall*, validasi silang, dan sebagainya. Model dibangun menggunakan set data latih dengan sejumlah parameter yang diminta oleh metodenya. Selanjutnya set data latih diberikan untuk menguji kinerja system, ada label kelas asli yang dibandingkan dengan label kelas yang didapatkan dari proses prediksi. Terakhir, dilakukan evaluasi terhadap sistem klasifikator yang dibuat (Prasetyo, 2014).

Pengujian ini dilakukan untuk melihat apakah kombinasi algoritma PCA dengan *K-Means* menghasilkan pengelompokan data yang lebih baik jika dibandingkan dengan metode *k-means* dengan *centroid random.* Adapun pengujian yang dilakukan adalah sebagai berikut (Prasetyo, 2014):

1. *Davies-Bouldin Index*

   Matriks Davies-Bouldin Index (DBI) diperkenalkan oleh David L. Davis dan Donald W. Bouldin yang digunakan untuk mengevaluasi *cluster*. Validitas internal yang dilakukan adalah seberapa baik *cluster* sudah dilakukan dengan menghitung kuantitas dan fitur turunan dari set data. Nilai DBI didapatkan dengan persamaan sebagai berikut:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j}(R_{i,j}) \quad \text{.....(2.14)}$$

K adalah jumlah *cluster* yang digunakan. Berdasarkan perhitungan diatas, dapat diamati bahwa semakin kecil nilai SSW maka hasil *Clustering* yang didapat juga lebih baik. Secara esensial, DBI menginginkan nilai sekecil (non-negatif ≥ 0) mungkin untuk menilai baiknya *cluster* yang dididapat. Indeks tersebut didapat dari rata-rata semua indeks *cluster,* dan nilai yang didapat bisa digunakan sebagai pendukung keputusan untuk menilai jumlah *cluster* yang paling cocok digunakan. DBI juga banyak digunakan untuk membantu *k-means* dalam menentukan jumlah *cluster* yang tepat untuk dugunakan karena biasanya *k-means* belum bisa mengetahui beberapa *cluster* yang digunakan untuk *clustering* data.

2. *Silhouette Index*

   Jika DBI digunakan untuk mengukur validasi seluruh *cluster* dalam set data, maka *Silhouette Index* (SI) dapat digunakan untuk memvalidasi baik sebuah data, *cluster* tunggal (satu *cluster* dari sejumlah *cluster*) atau bahkan keseluruhan *cluster*. Untuk menghitung nilai SI dari sebuah data ke-*i*, ada 2 komponen yaitu a_i dan b_i. a_i adalah rata-rata jarak data ke-*i* terhadap semua data lainnya dalam satu *cluster*, sedangkan b_i didapat dengan menghitung jarak rata- rata data ke-*i* terhadap semua data dari *cluster* yang tidak lain dalam satu *cluster* dengan data ke-*i,* kemudian diambil yang terkecil.

Berikut untuk menghitung a_i^j :

$$a_i = \frac{1}{M_j - 1}\sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_i^j, x_r^j) \; i = 1,2,\ldots . m_j \quad \text{.....} \quad (2.15)$$

$d(x_i^j, x_r^j)$ adalah jarak data ke-*i* dengan data ke-*r* dalam satu cluster *j*, sedangkan m_j adalah jumlah data cluster ke-*j*.

Berikut untuk menghitung b_i^j :

$$b_i = \text{Min}\left\{\frac{1}{m}\sum_{\substack{r=1 \\ r\neq i}}^{m_i} d(x_i^j x_r^n\right\}\ i = 1,2,\ldots m_n \quad \text{.....} \quad (2.16)$$

Untuk mendapatkan nilai *Silhouette Index* (SI) data ke-*i* menggunakan persamaan berikut :

$$SI_j = \frac{1}{m_j}\sum_{i=1}^{m_j} SI_i^j \text{.....} \quad (2.17)$$

nilai a_i mengukur seberapa tidak mirip sebuah data dengan cluster yang diikutinya, nilai yang semakin kecil menandakan semakin tepatnya data tersebut berada dalam cluster. Nilai b_i yang besar menandakan seberapa jelek data terhadap cluster yang lain. Nilai SI yang didapat dalam rentang [-1, +1]. Nilai SI yang mendekati 1 menandakan bahwa data tersebut semakin tepat berada dalam cluster. Nilai SI negative ($a_i > b_i$) menandakan bahwa data tersebut tidak tepat berada dalam cluster. SI bernilai 0 berarti data tersebut posisinya berada di perbatasan di antara dua cluster.

Untuk nilai SI dari sebuah cluster didapatkan dengan menghitung rata-rata nilai SI semua data yang bergabung dalam cluster tersebut, seperti pada persamaan berikut:

$$SI_j = \frac{1}{m_j}\sum_{i=1}^{m_j} SI_i^j \text{.....} \quad (2.18)$$

Untuk nilai SI secara global didapat dengan menghitung rata-rata nilai SI dari semua *cluster* seperti persamaan berikut:

$$SI = \frac{1}{K}\sum_{j=1}^{k} SI_j \text{.....} \quad (2.19)$$

3. *Dunn Index*

   Dunn Index (DI) diperkenalkan oleh J. C. Dunn (1973) sebagai metrik untuk validitas cluster. DI menghitung validitas cluster menggunakan diameter cluster (kohesi) dan jarak antara dua cluster (separasi). Untuk mendapatkan diameter sebuah cluster ke-*i* dilakukan dengan menghitung

jarak pasangan dua data dalam sebuah cluster, kemudian diambil yang terbesar, seperti dinyatakan oleh persamaan berikut :

$$\Delta_i = \max_{x,y \in Ci}\{d(x,y)\} \quad \text{(2.20)}$$

x dan y adalah data yang berada dalam cluster C_i.

Untuk menghitung jarak dua *cluster* digunakan persamaan berikut:

$$\partial_{i,j} = \text{Min}_{x,y \in C}\{d\,(x,y)\} = \partial_{sl(i,j)} \quad \text{(2.21)}$$

Dunn Index (DI) didapatkan dari persamaan berikut:

$$\text{DI} = min_{1 \le i \le k} \left\{ Min_{\substack{1 \le j \le k \\ j \ne 1}} \left\{ \frac{\partial_{i,j}}{\max\{\Delta\}} \right\} \right\} \quad \text{(2.22)}$$

k adalah jumlah *cluster*. Nilai DI yang semakin besar menandakan hasil *cluster*ing yang semakin baik.

4. *Validitas Fuzzy Clustering*

   Ketiga metode validitas sebelumnya hanya bisa diterapkan pada metode pengelompokan berbasis partisi seperti K-Means, K-Harmonic Means, dan sejenisnya. Sementara untuk metode pengelompokan yang menggunakan konsep *fuzzy*, sebuah data bisa menjadi anggota di semua cluster dengan nilai derajat keanggotaan yang dimilikinya. Semakin tinggi nilai derajat keanggotaan pada sebuah cluster maka semakin besar kecendrungannya menjadi anggota cluster tersebut.

   Bezdek (1981) mengusulkan validitas dengan menghitung koefisien partisi atau *partition coefficient* (PC) sebagai evaluasi nilai keanggotaan data setiap cluster. Nilai PC Index (PCI) hanya mengevaluasi nilai derajat keanggotaan, tanpa memandang nilai vector (data) yang biasanya mengandung informasi geometric (sebaran data). Nilainya dalam rentang [0,1], nilai yang semakin besar (mendekati 1) mempunyai arti bahwa kualitas cluster yang didapat semakin baik. PCI didapatkan dari persamaan berikut:

$$PCI = \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^{2}\right)$$.....(2.23)

Bezdek (1974a,b) juga mengusulkan validitas dengan menghitung entropi partisi atau *partition entropy* (PE). Nilai PE Index (PEI) mengevaluasi keteracakan data dalam cluster. Nilainya dalam rentang [0,1], nilai yang semakin kecil (mendekati 0) mempunyai arti bahwa kualitas cluster yang didapat semakin baik. PEI didapatkan dari persamaan berikut:

$$PEI = -\frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}\ x\ log_{2}u_{ij}\right)$$.....(2.23)

Kedua metric PCI dan PEI memiliki kecendrungan monotonic terhadap K. modifikasi nilai PCI (MPCI) dilakukan oleh Dave (1996) terhadap kecendrungan monotonic tersebut. MPCI didapatkan dari persamaan berikut:

$$MPCI = 1 - \frac{K}{K-1}(1 - PCI)$$.....(2.24)

2.13 Penelitian Terkait

Peneliti memerlukan berbagai referensi dalam melalukan penelitian, untuk dapat membantu peneliti dalam membangun penelitian kearah yang lebih baik, peneliti menggunakan berbagai referensi, salah satunya adalah penelitian terdahulu yang mempunyai kesamaan metode dengan peneliti.

Penggunaan metode K-Means terbukti dapat mempermudah Dinas Kesehatan Provinsi Riau dalam mengambil keputusan kecamatan-kecamatan yang sering dan jarang terkena penyakit (Ahmad Muhajir Siregar, 2013).

Penelitian menggunakan metode PCA K-Means dapat digunakan untuk mereduksi dimensi *dataset* tanpa harus kehilangan banyak informasi dan menggunakan metode PCA terbukti dapat meningkatkan kualitas *cluster* yang dihasilkan oleh algoritma *K-Means* (Ahmad Izzudin, 2015).

Tabel 2.3 Penelitian Terdahulu

Peneliti	Topik	Hasil
Ahmad Muhajir Siregar, 2013.	Analisa Data Mining Menggunakan Metode *Clustering K-Means* Dalam Pengelompokan Daerah Berdasarkan Kesehatan.	Aplikasi *clustering k-means* ini dapat mempermudah Dinas Kesehatan Provinsi Riau dalam mengambil keputusan kecamatan-kecamatan yang sering dan jarang terkena penyakit.
Fionda, 2013	Analisa Penjualan Barang Menggunakan Metode Clustering K-Means Untuk Perencanaan Penjualan Pada Swalayan Hawaii.	Penerapan aplikasi *clustering k-means* ini membantu dalam melakukan analisa yang dapat memberikan informasi berupa barang yang laris dan tidak laris pada *event* tertentu.
Ahmad Izzuddin, 2015	Optimasi *Cluster* pada Algoritma *K-Means* dengan Reduksi Dimensi *Dataset* Menggunakan *Principal Component Analysis* untuk Pemetaan Kinerja Dosen.	metode PCA dapat digunakan untuk mereduksi dimensi *dataset* tanpa harus kehilangan banyak informasi. Reduksi dimensi *dataset* menggunakan metode PCA terbukti dapat meningkatkan kualitas *cluster* yang dihasilkan oleh algoritma *K-Means*.