Data Mining

GEMASTIK 8

Analisis Profit Divergence pada Data Pemesanan Taksi Menggunakan Metode Kulback Leibler Divergence dengan Isotonic Regression dan K-Means Clustering

Disusun oleh:
Joshua Casey Darian Gunawan
Muhammad Zaky Khairuddin
Zamil Majdy

ZJZ
FAKULTAS ILMU KOMPUTER
UNIVERSITAS INDONESIA
2015

Daftar Isi

1. Latar Belakang

Informasi adalah hal yang sangat penting. Informasi memungkinkan seseorang untuk bertindak dengan lebih relevan dengan keadaan saat itu. Dengan adanya informasi, keputusan dapat diambil dengan lebih tepat sehingga diperoleh hasil yang optimal. Salah satu contoh pentingnya informasi adalah dalam bidang bisnis. Agar dapat menguasai dunia bisnis, kita harus mengetahui dan menguasai fenomena-fenomena yang terjadi di pasar. Dengan memiliki data tentang hal-hal yang terjadi di pasar, mengolahnya, dan mengambil informasi yang terdapat dalam data itu, kita dapat mengetahui kejadian di pasar, memprediksi apa yang mungkin terjadi, mengetahui keberadaan anomali dalam bisnis, sehingga dapat mengambil langkah yang tepat.

Dewasa ini, bentuk konkrit dari informasi yang beredar adalah data. Data beredar dalam jumlah yang sangat banyak sehingga tidak mungkin dapat dilakukan pemrosesan data tersebut satu persatu. Padahal, informasi yang terkandung dalam data tersebut bisa jadi sangatlah bermanfaat. Oleh karena itu, diperlukan metode khusus untuk mengolah data sehingga informasi-informasi yang bermanfaat tersebut dapat diperoleh.

Dalam studi ilmu komputer, terdapat bidang yang khusus mempelajari pengolahan data dalam jumlah besar, yakni *data mining*. Teknik-teknik dalam *data mining* memungkinkan pengolahan data besar secara optimal sehingga informasi-informasi yang bermanfaat dapat diperoleh dari data tersebut. Salah satu contoh aplikasi *data mining* yang akan kami bahas dalam makalah ini adalah pengolahan terhadap dataset pemesanan taksi di daerah Portugal pada tanggal 1 Juli 2013 hingga 30 Juni 2014.

2. Tujuan dan Manfaat

Analisis yang kami lakukan bertujuan mencari dan menganalisis fenomena serta anomali keuntungan pasar pada penggunaan layanan taksi dengan pembagian berdasarkan daerah. Anomali yang dicari adalah anomali divergensi keuntungan, yakni perbedaan drastis keuntungan pada suatu daerah dengan skala waktu per bulan. Perbedaan drastis ini dapat berupa penurunan maupun kenaikan yang drastis. Keuntungan per daerah ini dihitung dengan memperhitungkan perjalanan-perjalanan dengan daerah tersebut sebagai titik awal maupun titik tujuan perjalanan.

Manfaat yang dapat diperoleh dari analisis ini adalah identifikasi daerah yang memiliki tingkat permintaan yang berubah drastis. Dengan mengetahui daerah-daerah yang mengalami perubahan tingkat permintaan yang tinggi, dapat dilakukan analisis bisnis yang memadai untuk memaksimalkan keuntungan, misalnya menambah armada di sekitar daerah yang permintaannya menaik drastis. Analisis ini akan sangat berguna apabila dilakukan *stream* data baru secara terus menerus perbulannya, agar hasil analisis ini dapat digunakan untuk mengambil kebijakan bisnis secara lebih *real time*. Analisis dari

segi bisnis secara khusus tidak dibahas dalam makalah ini.

3. Batasan

Pada penelitian ini penulis melakukan pengamatan terhadap data yang memiliki *daytype* A, yakni hari kerja. Hal ini dilakukan karena *daytype* A paling umum ditemui di antara ketiga *daytype* yang ada, sementara setiap *daytype* tentu memiliki distribusi dan karakteristik yang sangat berbeda, sehingga tidak mungkin melakukan evaluasi yang sama terhadap ketiga *daytype* tersebut. Karena data yang harus diproses relatif banyak, algoritma yang digunakan adalah algoritma-algoritma yang memiliki kompleksitas rendah, misalnya *k-means clustering algorithm* dan *isotonic regression* yang memiliki kompleksitas linear terhadap banyaknya data.

4. Metode

4.1 Perangkat Lunak

4.1.1 WEKA Visualization and Learning Library

WEKA adalah *learning* dan *visualization library* yang dikembangkan menggunakan bahasa Java. *Library* ini memiliki banyak implementasi pemrosesan data serta algoritma learning yang cukup beragam seperti *Support Vector Machine*, *Artificial Neural Network*, dan tentu algoritma utama yang dipakai pada proyek ini yaitu *K-Means* serta *Isotonic Regression* (algoritma ini ada pada Weka Additional Package). Kelebihan *library* Java dibandingkan library lain adalah *nature* bahasa *Java* yang bisa dioptimasi dan cukup cepat dibandingkan *library* lain (dalam kasus ini adalah *library python* [1]). Metode analisis pada kasus ini membutuhkan data secara utuh (tidak melalui metode *sampling*) sehingga pemrosesan data dengan cepat sangat dibutuhkan pada kasus ini. Oleh karena itu, *library* ini merupakan pilihan yang cukup tepat.

4.1.2 ChartJS

ChartJS merupakan *library* dengan bahasa *Javascript* yang digunakan untuk melakukan visualisasi data dalam bentuk *chart*. *ChartJS* ini dipilih karena *library* ini merupakan *library* yang cukup ekstensif dan dapat digunakan untuk berbagai pemodelan data.

[1] http://benchmarksgame.alioth.debian.org/u64q/python.html

4.1.3 Google Fusion Tables

Google Fusion Tables adalah sebuah aplikasi web untuk visualisasi dan sharing data. *Google Fusion Tables* dapat membuat visualisasi berbentuk peta dengan bantuan *Google Maps*. Aplikasi ini digunakan dalam pembuatan *heatmap*. Keunggulan dari aplikasi ini adalah visualisasi yang dihasilkan cukup jelas dan sesuai dengan kebutuhan pengolahan data dalam analisis ini.

4.1.4 Python (Scikit-Learn & GGplot)

Python pada kasus ini digunakan untuk melakukan *prototyping*, untuk mengevaluasi data *sample* yang berjumlah kecil untuk menguji model algoritma yang diberikan pada data. *Python* dipilih karena kemudahan implementasinya dan dapat dikembangkan dengan cepat. Semua algoritma yang dikembangkan di sini diimplementasikan ulang menggunakan *Java* untuk digunakan pada data yang utuh.

4.2 Dataset

Dataset yang digunakan adalah data *taxi service trip* yang merupakan data penggunaan jasa taksi yang tercatat pada (hari Senin, 01 Juli 2013 00:00:58 GMT sampai dengan Senin, 30 Juni 2014 19:39:07 GMT). Terdapat 1.710.670 penggunaan jasa taksi pada interval waktu tersebut, hanya 10 penggunaan yang tidak memiliki keterangan data yang lengkap.

Dataset memiliki 9 atribut, yaitu *trip_id*, *call_type*, *origin_call*, *origin_stand*, *taxi_id*, *timestamp*, *day_type*, *missing*, dan *polyline*. Penjelasan untuk masing-masing atribut adalah sebagai berikut:

- *trip_id* : ID untuk setiap trip penggunaan taksi
- *call_type* : Tipe penggunaan jasa taksi (A: permintaan langsung ke pusat, B: permintaan langsung ke supir taksi, C: penggunaan jasa taksi di tengah jalan)
- *origin_call* : ID nomor telepon pemesan (hanya untuk call_type A)
- *origin_stand* : ID stand taksi (hanya untuk call_type B)
- *taxi_id* : ID supir taksi
- *timestamp* : Unix Timestamp (dalam detik) yang menandakan waktu mulai perjalanan.
- *day_type* : Tipe hari pemesanan (A: hari kerja, B: hari libur, C: hari sebelum hari libur)
- *missing* : Boolean (False jika data GPS utuh, True jika data GPS ada yang hilang)
- *polyline* : Data koordinat GPS (WGS84 format) perjalanan taksi setiap 15 detik

4.3 Algoritma dan Teknik

4.3.1 K-Means

K-means adalah salah satu metode *clustering* yang populer digunakan dalam penerapan *data mining*. *K-means* dibangun dari ide bahwa 2 buah data yang memiliki kemiripan akan dikelompokkan ke dalam sebuah *cluster*. Kemiripan diukur dari jarak Euclidean (*Euclidean distance*) dari 2 buah data. Dalam kasus ini, jarak Euclidean dihitung dengan cara jarak Euclidean antara *centroid* data perjalanan yang tercatat oleh *GPS* pada data.

Euclidean distance antara titik a dan b dihitung sebagai berikut:

$$dist(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

Sebanyak k buah *centroid* (titik tengah) diambil secara acak, kemudian dilakukan iterasi hingga konvergen.

4.3.2 Davis Bouldin Index

Permasalahan terbesar dari *k-means* adalah menentukan nilai k terbaik. Sulit untuk mengetahui berapa *cluster* yang dapat terbentuk dari sebuah dataset. Untuk itu, penulis akan melakukan percobaan dengan beberapa buah nilai k dan menentukan mana yang terbaik dengan menghitung *Davies Bouldin Index* dari masing-masing *cluster*.

Davies Bouldin Index (DBI) adalah indeks pengukuran validitas model *clustering* berdasarkan kemiripan dan perbedaan dari *cluster-cluster*. Model *clustering* dinilai semakin baik apabila anggota-anggota *cluster*-nya saling berdekatan dan jarak antar *cluster* semakin jauh. Nilai DBI yang lebih rendah menandakan model *clustering* yang lebih baik.

DBI dihitung dengan cara:

$$DB_k = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1,\ldots,k, i \neq j} \frac{diam(c_i+) + diam(c_j)}{dist(z_i, z_j)}$$

$$diam(c_i) = \sqrt{\frac{1}{n_i} \sum_{x \in c_i} dist(x, z_i)^2}$$

- c_i adalah *cluster*
- diam(c_i) adalah diameter *cluster* c_i
- n adalah banyaknya titik

- z_i adalah *centroid cluster* c_i
- dist(a,b) adalah *euclidean distance* antara titik a dan b

4.3.3 Haversine Formula

Haversine Formula adalah salah satu formula trigonometri yang digunakan untuk menghitung *great circle distance*, yakni jarak terdekat antara 2 titik pada permukaan bola dengan melalui permukaan bola tersebut. *Haversine formula* menghitung *great circle distance* dari 2 titik pada permukaan bola dengan parameter *latitude* dan *longitude* dari kedua titik tersebut.

Haversine formula didefinisikan sebagai berikut:

$$haversin(\frac{d}{r}) = haversin(\phi_2 - \phi_1) + cos(\phi_1)cos(\phi_2)haversin(\lambda_2 - \lambda_1)$$

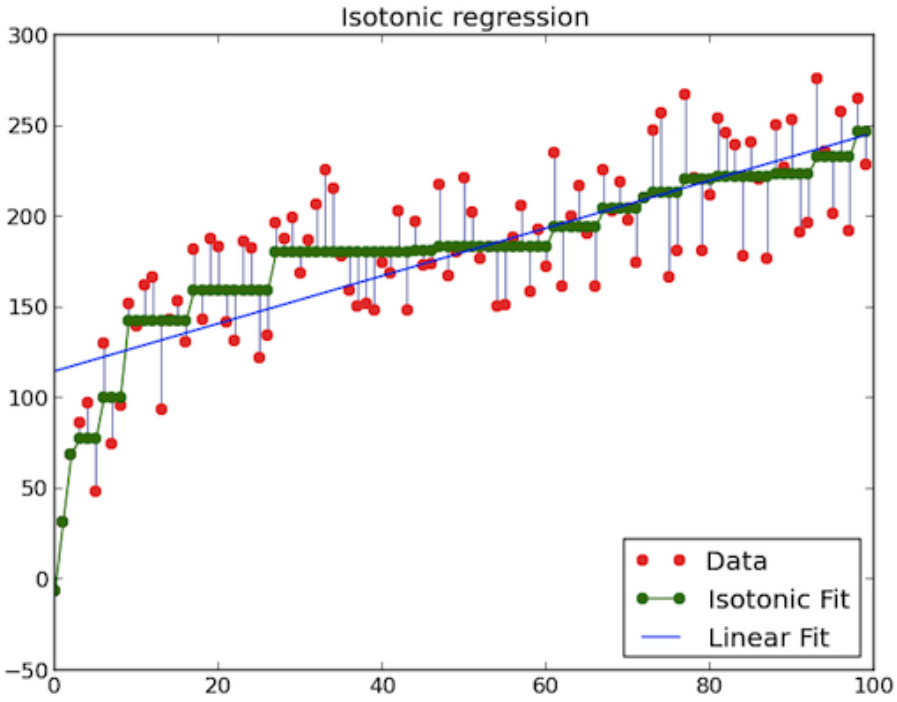
- d adalah *great circle distance* antara tiitk 1 dan 2
- r adalah jari-jari bola
- ϕ adalah *latitude*
- λ adalah *longitude*
- *haversin* adalah fungsi trigonometri $haversin(\theta) = sin^2(\frac{\theta}{2}) = \frac{1-cos(\theta)}{2}$

4.3.4 Isotonic Regression

Isotonic regression merupakan varian dari algoritma regresi yang mempertahankan arah trend dari *perdictor* regresi (fungsi regresi dipastikan monoton menaik atau menurun). Secara matematis, definisi *isotonic regression* adalah sebagai berikut:

Diberikan data dengan nilai a1,a2,...,an, regresi F(x) merupakan fungsi monoton ($F(i) <= F(j) | i <= j$) dengan nilai $\sum_{i=1}^{n}(f(i) - a_i)^2$ seminimal mungkin.

Sifat monotonik ini dimanfaatkan untuk melakukan pendekatan nonlinear yang lebih akurat akan tetapi tidak membuat aproksimasinya menjadi *overfit*. Pada data yang bersangkutan, regresi ini digunakan untuk mengaproksimasi perkembangan divergensi *profit* pemesanan taksi.



Gambar 1: Contoh *isotonic regression*

4.3.5 Kullback Leibler Divergence

Kullback Leibler Divergence merupakan salah satu teknik penghitungan divergensi dari distribusi dua nilai (pada umumnya distribusi probabilitas). *Kullback Leibler Divergence* pada distribusi data P dan Q didefinisikan sebagai $D_{Kl}(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)}$ dengan P(i) merupakan nilai tendensi dari data ke-i pada dataset-P dan Q(i) merupakan nilai tendensi dari data ke-i pada dataset-Q.

Pada kasus ini, *KL Divergence* digunakan untuk menghitung divergensi dari dua nilai *estimated score* yaitu estimasi probabilitas (persentase) *profit* dari suatu daerah pada *cluster* tertentu. *Estimated score* ini yang akan menjadi atribut utama untuk membandingkan divergensi *profit* yang akan dianalisis.

4.4 Teknik

Dalam penelitian ini, teknik-teknik yang digunakan antara lain:

4.4.1 Preprocessing

Sebelum dataset diproses, perlu dilakukan langkah-langkah tertentu terlebih dahulu agar dataset tersebut lebih mudah untuk diproses dan sesuai dengan kriteria yang diinginkan. Langkah-langkah yang diambil dalam *preprocessing* antara lain:

- Data Filtering

  Data filtering adalah proses pembuangan data-data yang tidak memenuhi batasan yang ditentukan. Pembuangan data yang tidak memenuhi batasan dilakukan agar hasil penelitian relevan dengan batasan tersebut dan proses evaluasi dan analisis data menjadi lebih mudah dan cepat karena tidak ada data sampah yang terlibat dalam komputasi-komputasi yang dilakukan.

- Data Standardization dan Meta Data Customization

  Standardisasi data termasuk mengatur kembali banyaknya baris atau kolom pada dataset dan mengubah nilai yang ada menjadi kisaran tertentu, misalnya data nominal dijadikan numerik, boolean, atau lainnya. Dalam penelitian ini, penulis melakukan standardisasi data, yaitu mengubah nilai-nilai numerik menjadi nilai probabilitasnya (persentase).

- Feature Extraction

  Feature extraction adalah pengurangan atribut pada dataset apabila ukuran dataset terlalu besar atau ada atribut yang berulang (*redundant*). Dalam penelitian ini, kami mengabaikan atribut *call type*, *origin call*, *origin stand*, dan *taxi id*, karena tidak relevan dengan tujuan analisis yang dilakukan.

4.4.2 Modeling

Setelah melalui tahap *preprocessing*, data akan dimodelkan untuk menggambarkan distribusi data tersebut dan hubungan antara data yang satu dengan yang lain. Pada penelitian ini, teknik *modeling* yang digunakan antara lain:

- Clustering

  Clustering adalah teknik mengelompokkan data yang memiliki kemiripan karakteristik dan menjadikan data yang memiliki kesamaan karakteristik tersebut menjadi 1 kelompok.

- Regression Line Fitting

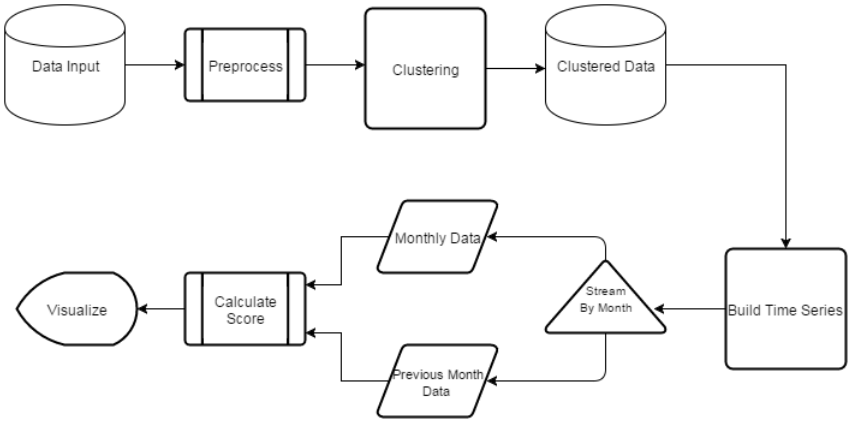
  Regression line fitting adalah teknik mengaproksimasi trend data dengan membentuk fungsi regresi yang mendekati distribusi data asli dan melakukan prediksi terhadap fluktuasi data dengan menggunakan fungsi regresi tersebut.

4.4.3 Inference

Inferensi (penarikan kesimpulan) dilakukan terhadap hasil *modeling* pada langkah sebelumnya dengan menerapkan perhitungan dan konsep matematis dengan pendekatan statistik sehingga diperoleh informasi yang berkaitan dengan tujuan pengolahan data yang telah dipaparkan pada bagian sebelumnya.

5. Desain dan Implementasi

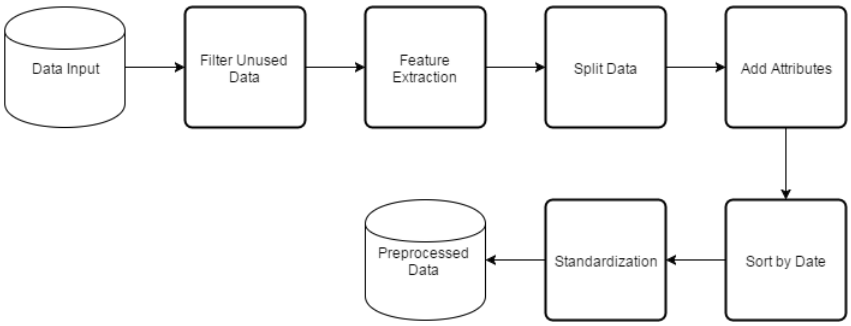
5.1 Desain



Gambar 2: Rancangan pengolahan dataset

Langkah pertama yang dilakukan pada data adalah *preprocessing* sehingga data tersebut siap untuk melalui langkah-langkah berikutnya. Langkah kedua adalah melakukan *clustering* terhadap data tersebut. Langkah berikutnya adalah pembuatan *time series*, yaitu data berisi pendapatan masing-masing *cluster* yang diurutkan berdasarkan waktu. Setelah itu, data tersebut akan dibaca dengan terbagi setiap bulannya. Data per bulan tersebut akan melalui sistem *scoring* kemudian divisualisasi.

5.2 Preprocessing



Gambar 3: Teknik *preprocessing* dataset

Sebelum diolah lebih lanjut, dataset yang diterima akan diproses terlebih dahulu. Langkah awal pemrosesan awal ini adalah membuang data yang tidak termasuk dalam batasan analisis ini. Setelah itu, kami melakukan pemisahan data. Tiap data kami pecah menjadi dua data, dengan data yang pertama memuat titik awal perjalanan dan data yang kedua memuat titik akhir perjalanan. Terakhir, atribut-atribut baru ditambahkan pada data.

5.2.1 Filter Unused Data

Sebelum memproses data lebih lanjut, langkah pertama yang dilakukan adalah membuang data yang tidak sesuai. Data yang kami buang adalah data yang memiliki *stream GPS* yang tidak lengkap, karena membuat data secara keseluruhan tidak akurat, sementara jumlahnya yang sangat sedikit dibandingkan dengan keseluruhan data hampir tidak menimbulkan pengaruh ketika data dianulasi. Selain itu, data yang kami buang adalah data yang tipe harinya bukan hari kerja, karena tidak sesuai untuk analisis dalam batasan yang telah ditentukan.

5.2.2 Feature Extraction

Agar data lebih mudah diproses, atribut-atribut yang tidak sesuai dengan tujuan dari analisis ini dihilangkan dari dataset. Atribut-atribut tersebut antara lain *daytype* (karena hanya digunakan saat filtering), *call type*, *origin stand*, *origin call*, *taxi id*, dan *trip id*.

5.2.3 Split Data

Masing-masing data yang telah difilter dipecah menjadi 2 data yang berbeda. Data x akan dipecah menjadi data x_1 dan x_2, dimana x_1 memuat koordinat awal perjalanan dari x dan x_2 memuat koordinat akhir perjalanan dari x. Hal ini dilakukan karena dalam batasan yang ditentukan, data yang diperlukan adalah *traffic* dari suatu daerah secara keseluruhan, baik dari maupun ke daerah tersebut.

5.2.4 Add Attributes

Setelah dilakukan pemecahan data, atribut baru yang diperlukan untuk komputasi ditambahkan ke dalam data tersebut. Atribut-atribut tersebut antara lain:

- Total Distance

  Total distance adalah jarak yang dilewati oleh taksi tersebut selama perjalanan. Total Distance dihitung dengan total *great circle distance* antar titik pada *GPS stream*. Masing-masing *great circle distance* tersebut dihitung dengan menggunakan formula Haversine.

- Gross Profit Estimation

  Gross Profit Estimation adalah estimasi total keuntungan yang diperoleh sebuah layanan taksi pada perjalanan tersebut. *Gross profit* ini digunakan untuk mempermudah representasi data. Estimasi ini dihitung menggunakan formula yang tidak mengubah proporsi rasio data pemesanan terhadap data lain yang diperoleh berdasarkan *time-rate* dan *distance-rate*. Formula yang digunakan adalah

$$f(x) = \begin{cases} baseRate, & \text{jika } dist \leq minDist \\ baseRate + mileRate(dist - minDist), & \text{jika } dist \leq minDist.duration \\ baseRate + duration.minutesRate, & \text{kondisi lain} \end{cases}$$

  dengan baseRate adalah ongkos minimum taksi, minDist jarak minimum perjalanan, mileRate biaya per mil, dist jarak perjalanan, duration lama perjalanan, dan minutesRate biaya perjalanan per menit.

5.2.5 Sort by Date

Untuk keperluan pengolahan data selanjutnya, data akan diurutkan berdasarkan *timestamp*.

5.2.6 Standardisasi

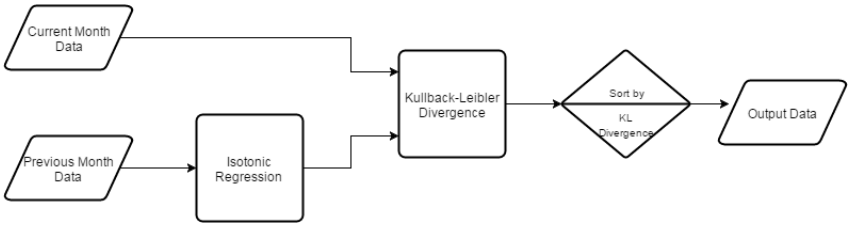
Untuk mempermudah membandingkan dua nilai pada distribusi yang berbeda, penulis melakukan standardisasi nilai dengan menghitung *profit probability* (persentasi keuntungan pada suatu bulan) pada dataset. Standarisasi dengan pendekatan probabilitas ini menjamin total dari semua nilai pada suatu distribusi menjadi tepat = 1. Penulis juga menggunakan teknik *laplace-smoothing* untuk menjamin data terhindar dari pembagian dengan 0 dan memperhalus trend distribusi profit. Berikut adalah cara menghitung *profit probability* untuk setiap data per bulan:

$$P(i) = \frac{Profit(i)}{\sum_{j \in M_i} Profit(j)}$$

$$P(i) = \frac{Profit(i) + \epsilon}{\sum_{j \in M_i} Profit(j) + n(M)\epsilon} \text{(Laplace Smoothing)}$$

- P(i) adalah estimasi keuntungan pada hari ke i
- M adalah himpunan hari dalam satu bulan

5.3 Scoring



Gambar 4: Teknik perhitungan *score* data

Sistem *scoring* adalah sistem yang bertujuan memberikan skor terhadap data dalam suatu bulan. Sistem *scoring* ini menggunakan nilai *Kullback-Leibler Divergence* sebagai skor terhadap data bulan tersebut. Nilai *Kullback-Leibler Divergence* ini diperoleh dari perbandingan antara regresi dari nilai bulan sebelumnya dengan nilai-nilai data pada bulan yang dievaluasi.

5.4 Input

5.4.1 Input Preprocess

Input terhadap *preprocess* yang kami lakukan adalah dataset asli berupa *comma separated value* dengan atribut antara lain ID taksi, tipe panggilan, asal panggilan, stand asal, waktu mulai perjalanan, tipe hari perjalanan, ada tidaknya data yang hilang, dan koordinat GPS setiap 15 detik perjalanan.

5.4.2 Input Clustering

Input yang digunakan pada proses *clustering* adalah dataset hasil *preprocess* yang telah dilakukan sebelumnya, yakni *comma separated value* dengan atribut-atribut antara lain waktu mulai perjalanan, posisi *latitude*, posisi *longitude*, biaya perjalanan, dan jarak yang ditempuh.

5.4.3 Input Scoring

Input yang digunakan pada proses *scoring* adalah dataset 2 bulan dari dataset hasil *clustering*, yakni dataset yang digunakan untuk analisis "bulan ini" dan dataset "bulan lalu".

5.5 Eksperimen

Dalam proses *clustering*, untuk menentukan banyak *cluster* yang paling sesuai, dilakukan eksperimen dengan melakukan proses *clustering* tersebut dengan banyak *cluster* k = 10 hingga k = 19. Batasan k tersebut diambil berdasarkan pengamatan manual pada peta. Terdapat cukup banyak daerah-daerah terseparasi pada data, tetapi tidak terlalu banyak, sehingga k hanya dibatasi sampai k = 19. Untuk setiap nilai k, *cluster* akan di-*generate* dan untuk memilih nilai k terbaik dilakukan perhitungan *Davies Bouldin Index* untuk setiap nilai k tersebut.

Hasil dari perhitungan *Davies Bouldin Index* adalah sebagai berikut:

- 10 Cluster : 25.39045107804417
- 11 Cluster : 25.422136153311417
- 12 Cluster : 24.83355270788093
- 13 Cluster : 24.894371593983927
- 14 Cluster : 24.930897636400257
- 15 Cluster : 20.077528269926834
- 16 Cluster : 20.338767774901303
- 17 Cluster : 24.997422365531364

- 18 Cluster : 25.019546784641772
- 19 Cluster : 25.048188086470475

Dari eksperimen ini, diperoleh banyak cluster yang paling representatif, yakni k = 15 karena memiliki nilai DBI yang paling rendah.

5.6 Output

5.6.1 Output Preprocess

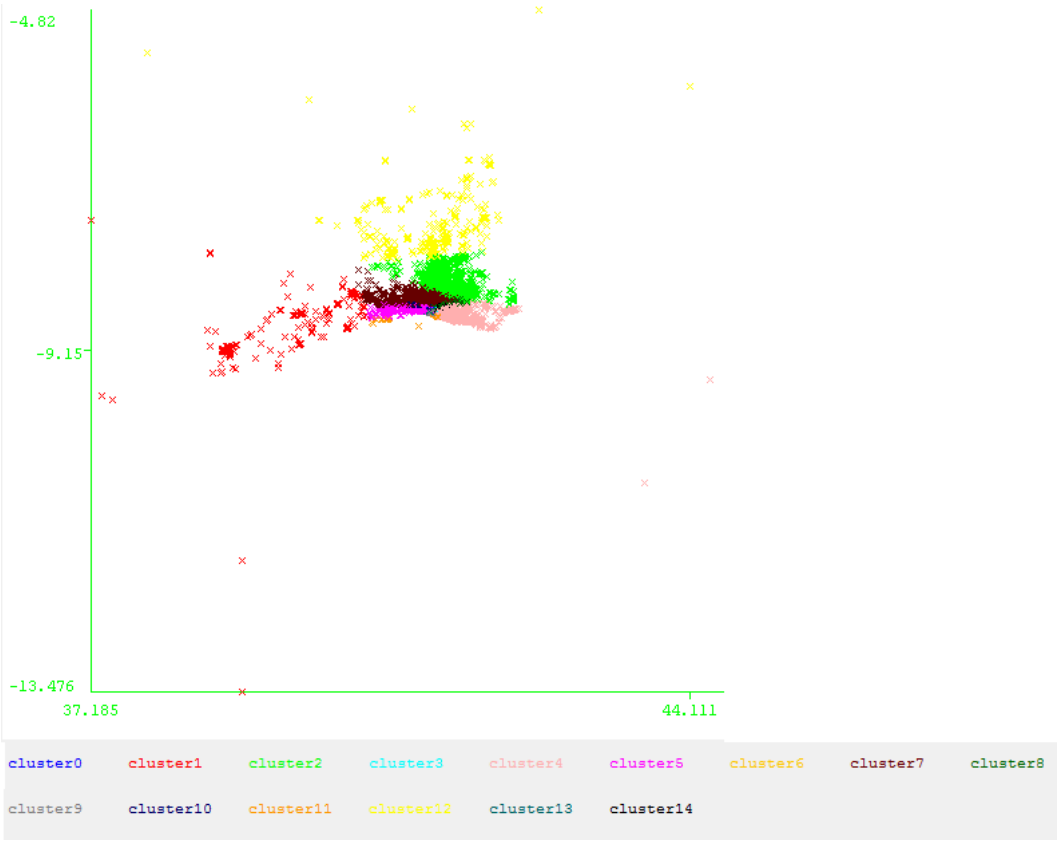
Output dari *preprocess* terhadap dataset asli adalah dataset yang sudah sesuai dengan batasan-batasan yang telah ditentukan dalam analisis ini. Dataset yang sesuai tersebut memiliki atribut antara lain waktu awal perjalanan, posisi *latitude*, posisi *longitude*, biaya perjalanan, dan jarak yang ditempuh, yang merupakan hasil pembagian dua tiap data yang sesuai pada dataset asli.

5.6.2 Output Clustering

Output dari proses *clustering* adalah dataset yang setiap datanya telah diberi label *cluster index*, yaitu *cluster* letak data tersebut. Karena DBI terkecil diperoleh untuk k=15, maka banyak *cluster* yang digunakan untuk analisis adalah 15. Berikut adalah detail *cluster* yang digunakan:

Tabel 1: Detail cluster

Cluster ID	Number of Data	Centroid Latitude	Centroid Longitude	Region Name
0	52570	41.1767	-8.5408	Rua Doutor Ral Chagas 77, 4435-124 Rio Tinto
1	3265	40.909	-8.5939	Travessa da Estrada Nova 101, 3885-062 Arada
2	476	41.129	-7.592	EM512 5, 5120
3	171187	41.1747	-8.653	Rua Conde Covilh 1460, 4100 Porto
4	229332	41.1826	-8.6051	Alameda Professor Hernni Monteiro 813, 4200 Porto
5	705847	41.1444	-8.6145	Travessa do Ferraz 2, 4050-141 Porto
6	2402	41.341	-8.3149	CM1128 404, 4620
7	453884	41.1611	-8.6275	Avenida da Frana 352, 4050-278 Porto
8	287113	41.155	-8.6448	IC23, 4150-172 Porto
9	476618	41.1601	-8.5832	Rua Jos Monteiro da Costa, 4350-307 Porto
10	76135	41.2426	-8.6709	Rua da Caralinda 259, 4470-558 Vila Nova da Telha
11	189885	41.1576	-8.6697	Rua Afonso Baldaia 368, 4150-002 Porto
12	595941	41.1528	-8.6052	Rua de Santa Catarina 753, 4000-425 Porto
13	80643	41.1782	-8.6877	Rua Dom Joo i 394, 4450-163 Matosinhos
14	182	39.2218	-8.9714	IC2, 2065 Alcoentre

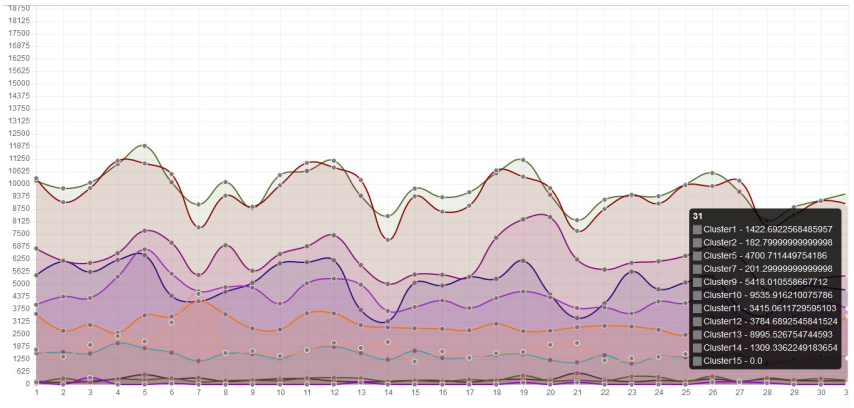


Gambar 5: Distribusi cluster daerah

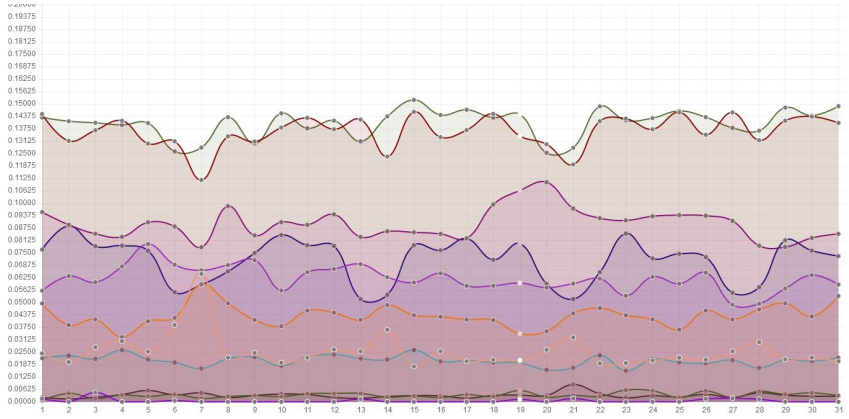
6. Analisis

Pada eksperimen yang dipaparkan pada bagian sebelumnya, banyak cluster yang memiliki nilai *Davies Bouldin Index* terkecil adalah 15 cluster. Hasil detail dari *clustering* dengan banyak k = 15 telah disajikan pada bagian output *clustering* yang telah dibahas sebelumnya.

Seluruh hasil komputasi statistik pada data yang diklasifikasikan dari *cluster-cluster* tersebut dapat divisualisasikan menjadi sebagai berikut:



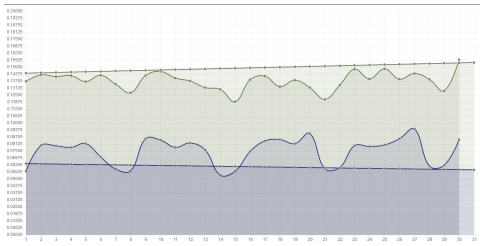
Gambar 6: Data *total cost* setiap cluster



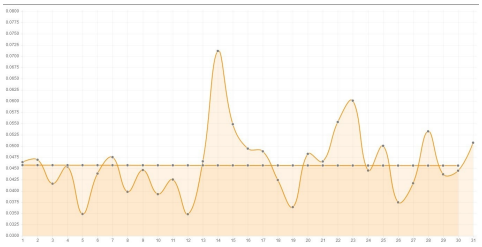
Gambar 7: Data *total cost* setelah dikonversi menjadi probabilitas

Dari hasil tersebut dilakukan *plotting* data dengan perkembangan *profit* untuk setiap bulannya, hasil visualisasi tersebut adalah sebagai berikut:

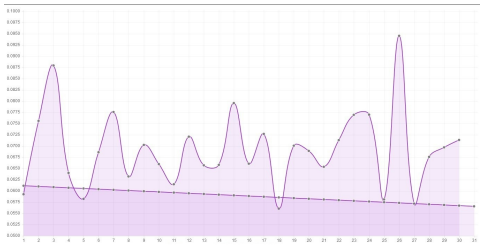
Gambar 8: Visualisasi Bulan Kedua
$KLDivergence = 0.015632698168624958$



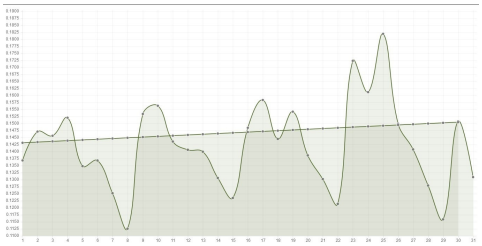
Gambar 9: Visualisasi Bulan Ketiga
$KLDivergenceTop = 0.044998303321150374$
$KLDivergenceBot = 0.026714006810494257$



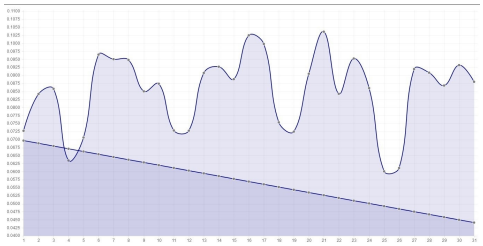
Gambar 10: Visualisasi Bulan Keempat
$KLDivergence = 0.03152148657242697$



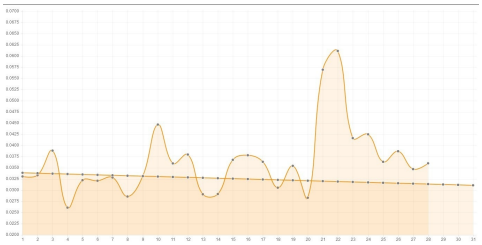
Gambar 11: Visualisasi Bulan Kelima
$KLDivergence = 0.0472788377243259$



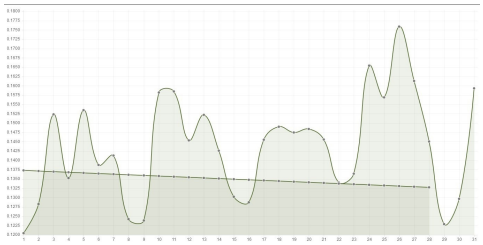
Gambar 12: Visualisasi Bulan Keenam
$KLDivergence = 0.036093264517592136$



Gambar 13: Visualisasi Bulan Ketujuh
$KLDivergence = 0.07018660930043484$

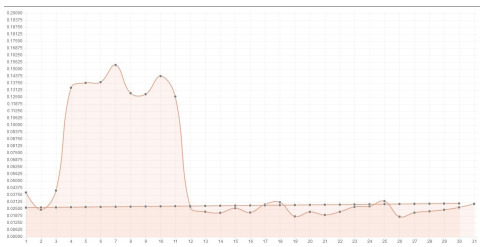


Gambar 14: Visualisasi Bulan Kedelapan
$KLDivergence = 0.039705311189136054$

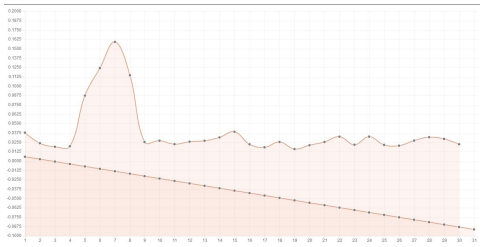


Gambar 15: Visualisasi Bulan Ketiga
$KLDivergence = 0.04904454504562369$

Gambar 16: Visualisasi Bulan Kesepuluh
$KLDivergence = 0.026295761592674213$



Gambar 17: Visualisasi Bulan Kesebelas
$KLDivergence = 0.26752695868379334$



Gambar 18: Visualisasi Bulan Keduabelas
$KLDivergence = 0.07129923008134906$

Dari hasil keseluruhan tersebut fenomena anomali dengan perbedaan keuntungan paling tinggi (nilai divergen paling tinggi) terdapat pada bulan 10, yaitu pada cluster yang terpusat pada daerah Matosinhos. Pada tanggal 4 sampai 11 terdapat kenaikan *profit* yang sangat signifikan dan kembali pada distribusi sebelumnya setelah interval waktu tersebut. Gambar 19 hingga 22 di bawah ini merupakan *plotting* frekuensi pemesanan taksi pada peta berdasarkan lokasi *GPS* pada data. Daerah Matosinhos mengalami kenaikan frekuensi pemesanan taksi dalam jumlah besar.

Gambar 19: Plot Frekuensi Penggunaan Taksi Bulan Kesembilan

Gambar 20: Plot Frekuensi Penggunaan Taksi Bulan Kesepuluh

Gambar 21: Plot Frekuensi Penggunaan Taksi Bulan Kesebelas

Gambar 22: Plot Frekuensi Penggunaan Taksi Bulan Keduabelas

7. Kesimpulan

Dengan menggunakan *k-means clustering* dengan pemilihan jumlah *cluster* yang tepat, daerah-daerah yang berdekatan dapat dikelompokkan menjadi kelompok-kelompok yang terpisah dan modular. *Centroid* setiap *cluster* juga memudahkan penentuan representasi *cluster-cluster* tersebut.

Implementasi model regresi yang tepat pada suatu data dapat memengaruhi hasil estimasi data dan seberapa cocok regresi tersebut terhadap data asli. Seperti pada kasus ini, prediksi menggunakan regresi linier pada data monoton (*isotonic regression*) memberikan hasil yang lebih representatif dibandingkan data yang tidak monoton.

Penghitungan divergensi profit menggunakan *Kullback Leibler Divergence* mampu mendeteksi perkembangan keuntungan penggunaan jasa taksi secara cepat dan akurat.