MODEL PREDIKSI DAN DESKRIPSI BERDASARKAN DATA PENJUALAN TIKET PESAWAT UNTUK BISNIS TOURISMWAVE

PREDICTION AND DESCRIPTION MODELS BASED ON AIRPLANE TICKET SALES DATA FOR TOURISMWAVE BUSINESS

Handito Muhammad Septiadi[1], Andry Alamsyah[2]

[1,2]Prodi S1 Manajemen Bisnis Telekomunikasi dan Informatika, Fakultas Ekonomi dan Bisnis, Universitas Telkom
handito.mseptiadi@hotmail.com[1], andry.alamsyah@gmail.com[2]

Abstrak

Jumlah keberangkatan baik penumpang dan kargo di bandara Indonesia pada tahun 1999-2013 mengalami peningkatan yang signifikan. Tidak hanya keberangkatan domestik atau dalam negeri, tetapi juga keberangkatan luar negeri. Hal ini bermakna bahwa minat masyarakat Indonesia akan melancong atau bepergian meningkat sejak tahun 1999.

Pemain dalam industri *intermediary* atau perantara sebagai penyedia jasa pemesanan tiket pesawat secara online di Indonesia meningkat seiring dengan meningkatnya jumlah keberangkatan penumpang di bandara Indonesia. Beberapa penyedia jasa yang ternama antara lain adalah Traveloka, Tiket.com dan Skyscanner. Diantara ketiga penyedia jasa tersebut memiliki kemiripan fitur, yakni mencari tiket pesawat secara manual dengan preferensi kota asal, kota tujuan, tanggal penerbangan dan jumlah penumpang. Proyek TourismWave ini akan menawarkan perbedaan yakni dengan fitur rekomendasi destinasi pariwisata berdasarkan *budget*. Dengan adanya fitur ini, hasil dari penelitian ini akan memiliki suatu nilai tambah bagi TourismWave yaitu fitur unik yang belum dimiliki oleh kompetitor.

Dari data yang didapatkan dari TamaTour, setelah dilakukan *preprocessing* atribut data yang terkandung adalah Bulan Transaksi, Metode Pembayaran, Maskapai, Rute Penerbangan, Harga dan Gender. Setelah itu dilakukan pengolahan dengan metode *data mining* klasifikasi dan deskripsi. *Pengolahan* model deskripsi menggunakan bantuan perangkat lunak Tableau, sedangkan model prediksi menggunakan SPSS dengan algoritma CHAID. Dari penelitian dihasilkan 6 model deskripsi, dan model prediksi yang berbentuk *decision tree* atau pohon keputusan.

Kata kunci: *Data Mining*, tiket pesawat, *decision tree*

Abstract

The amount of passenger and cargo departures at Indonesian airport in 1999-2013 has experienced a significant increase. Not only domestic departures, also international departures. This means that the people of Indonesia's interest of travelling has increased since 1999.

Players in intermediary industry or the ones that provides online ticketing service is growing because the increased amount of passenger departures in Indonesian airports. A few of the most known service providers are Traveloka, Tiket.com, and Skyscanner. There is a dominant feature among those three, which is searching airplane tickets manually using city of departure, destination, date of flight and amount of passengers. This TourismWave project offers difference which is destination recommendation based on budget. As is this feature, the result of this research will give TourismWave a competitive advantage.

From the data that obtained from Tama Tour, after preprocessing phase had been done, the attributes are month of transaction, payment method, airline, flight route, price and gender. After that, the process of the data which used data mining classification and description. The process is by using Tableau software for description models, and SPSS software with CHAID algorithm for prediction model. The results of this research are 6 description models and a decision tree-based prediction model.

Key word: Data mining, airplane ticketing, decision tree

1. Pendahuluan

1.1 Latar Belakang

Berdasarkan tabel yang peneliti dapatkan dari Badan Pusat Statistik tahun 2015 mengenai jumlah keberangkatan penumpang dan barang di bandara Indonesia pada tahun 1999-2013 [1], jumlah keberangkatan penumpang dan kargo di bandara Indonesia mengalami peningkatan yang signifikan. Peningkatan jumlah keberangkatan penumpang tidak hanya untuk keberangkatan dalam negeri atau domestik, namun keberangkatan luar negeri juga mengalami peningkatan. Hal ini bermakna bahwa minat masyarakat Indonesia akan *travelling* atau melancong meningkat.

Jumlah dari pemain industri *intermediary* atau perantara sebagai penyedia jasa pemesanan tiket pesawat secara online di Indonesia meningkat seiring dengan meningkatnya jumlah keberangkatan penumpang di bandara Indonesia. Beberapa penyedia jasa yang ternama antara lain ialah Traveloka, Tiket.com, dan Skyscanner. Diantara ketiga penyedia jasa tersebut memiliki kemiripan fitur, yakni mencari tiket pesawat secara manual dengan preferensi kota asal, kota tujuan, tanggal penerbangan, dan jumlah penumpang. Proyek TourismWave ini akan menawarkan perbedaan yakni dengan fitur rekomendasi destinasi pariwisata berdasarkan *budget*. Dengan adanya fitur ini, hasil dari proyek TourismWave akan memiliki suatu nilai tambah yaitu fitur unik yang belum dimiliki oleh kompetitor.

Dalam proyek TourismWave dibutuhkan model data prediktif untuk dijadikan dasar algoritma untuk fitur rekomendasi destinasi pariwisata. Seperti yang sudah dikatakan sebelumnya tentang fitur unik yakni mengalokasikan *budget* tiket pesawat sesuai dengan preferensi pribadi. Model data dapat diperoleh dengan cara mengolah data sebelumnya untuk dijadikan prediksi dan dijadikan *overview* dari suatu periode. Dari data yang didapatkan, kita dapat mencari model prediksi yang berbentuk *decision tree* atau pohon keputusan dan model deskripsi untuk dijadikan dasar algoritma rekomendasi dari TourismWave.

1.2 Perumusan Masalah

Berdasarkan latar belakang dan judul penelitian dapat diketahui perumusan masalah pada penelitian ini adalah sebagai berikut:

1. Kombinasi skenario model prediksi dan deskripsi apa saja yang bisa didapat dari data penjualan tiket pesawat pada tahun 2014?
2. Adakah model yang dapat diimplementasikan untuk kebutuhan TourismWave?

2. Dasar Teori dan Metodologi

2.1 Big Data

Big data adalah data yang melebihi kapasitas pengolahan dari database konvensional. Entah data tersebut terlalu besar, bergerak terlalu cepat, atau data tersebut tidak cocok dengan struktur *database*. Untuk memperoleh nilai dari data ini, diperlukan cara alternatif untuk memproses data tersebut [2]. Soares dalam Sathi [3] telah mengidentifikasikan 5 tipe dari *big data*: web dan media sosial, *machine-to-machine* (M2M), data transaksi besar, *biometrics*, dan data yang dihasilkan manusia.

Menurut Sathi [3] alasan *big data* berbeda dengan jenis data lain yang dipakai sebelumnya karena terdapat empat buah "V" yang mengkarakterkan data ini, yakni: *volume, velocity, variety,* dan *veracity*. Jika diartikan ke dalam bahasa Indonesia menjadi volume, kecepatan, variasi dan akurasi.

a. *Volume*

Sathi [3] menjelaskan bahwa menurut majalah *Fortune*, mereka menciptakan 5 *exabytes* digital data terhitung sampai Juni 2003, *Fortune* sendiri telah berdiri sejak tahun 1929. Sedangkan pada tahun 2011, jumlah yang sama diciptakan dalam waktu 2 hari. Dekade lalu, kebanyakan organisasi mengitung penyimpanan untuk data dalam *terabyte*, namun sekarang kebanyakan aplikasi membutuhkan penyimpanan dalam *petabyte* yang mana 1 *petabyte = 1000 terabytes*. Menurut Sathi [3] dari diskusi beliau dengan salah satu perusahaan *Communications Service Provider* (CSP) dengan 100 juta pelanggan yang menciptakan data 50 *petabytes* per hari, perusahaan membuang sebagian besar data tersebut karena kekurangan tempat untuk penyimpanan. Jadi lebih kurang 25 *petabytes* data terbuang sia-sia tanpa ada usaha untuk membudidayakan nya.

b. *Velocity*

Sathi [3] menjelaskan bahwa ada dua aspek dalam kecepatan, yang pertama adalah tingkat produksi data, dan yang kedua adalah *latency* atau interval waktu antara stimulasi dan respon. Jumlah *mobile data* global diperkirakan akan mencapai 10.8 *exabytes* perbulan pada 2016 seiring konsumen berbagi gambar dan video, hal ini adalah tingkat produksi data. Sedangkan untuk *latency*, jika awalnya analisis data berbentuk laporan yang berasal dari data hari sebelumnya. Sekarang, sebagai contoh turn.com melakukan analisis data dalam waktu 10 milidetik atau *miliseconds* untuk melakukan penempatan iklan online.

c. *Variety*

Menurut Sathi [3] *big data* telah meluaskan wawasan kita dengan cara mengintegrasi data dan teknologi analisis. Yang dimaksudkan dengan integrasi data adalah sumber data yang berasal dari teks, suara dan video. Contoh adalah *InfoSphere Streams* yang dimiliki oleh IBM, yang mengolah bermacam jenis sumber data baik untuk analisis *real-time* maupun pengambilan keputusan. Termasuk instrumen medis untuk analisis *neonatal*, data seismik, tag RFID, pola lalu lintas, data cuaca, dll.

d. *Veracity*

Sathi [3] menjelaskan tentang kebenaran atau akurasi dari *big data* karena tidak seperti data internal pemerintah, kebanyakan *big data* berasal dari sumber diluar kehendak yang mengakibatkan masalah akurasi yang signifikan. Oleh karena itu, menemukan korelasi yang tepat untuk suatu model data mempunyai peran penting untuk mendapatkan data akurat. Karena *veracity* mewakilkan baik kredibilitas dan kesesuaian data untuk *target audience*.

Menurut Dumbill [2] terdapat 2 kategori dari manfaat *big data* yang didapatkan oleh sebuah perusahaan:

1. Penggunaan analisis, analisis *big data* dapat mengungkap pengetahuan yang tersembunyi sebelumnya karena biaya untuk mengolah data terlalu mahal. Contohnya, *peer influence* antar konsumen yang diungkapkan dengan menganalisa transaksi pembeli, data sosial dan data geografis.
2. Penciptaan produk baru. Contohnya, mengamati apa yang konsumen lakukan dengan konsumen lain pada media sosial serta mengambil sinyal untuk kemudian diolah menjadi model prediksi yang berguna untuk menciptakan produk baru.

2.2 Data Mining

Ahlemeyer-Stubbe dan Coleman [4] menjelaskan *data mining* berarti mengambil informasi dari data-data yang diciptakan setiap saat dalam hidup kita. Ternyata dalam data yang kita pakai setiap hari mempunyai makna lebih dalam dan mempunyai pola tertentu. Ahlemeyer-Stubbe dan Coleman [4] menjelaskan bahwa *data mining* mencakup berbagai kegiatan dan berusaha untuk menjawab pertanyaan seperti:

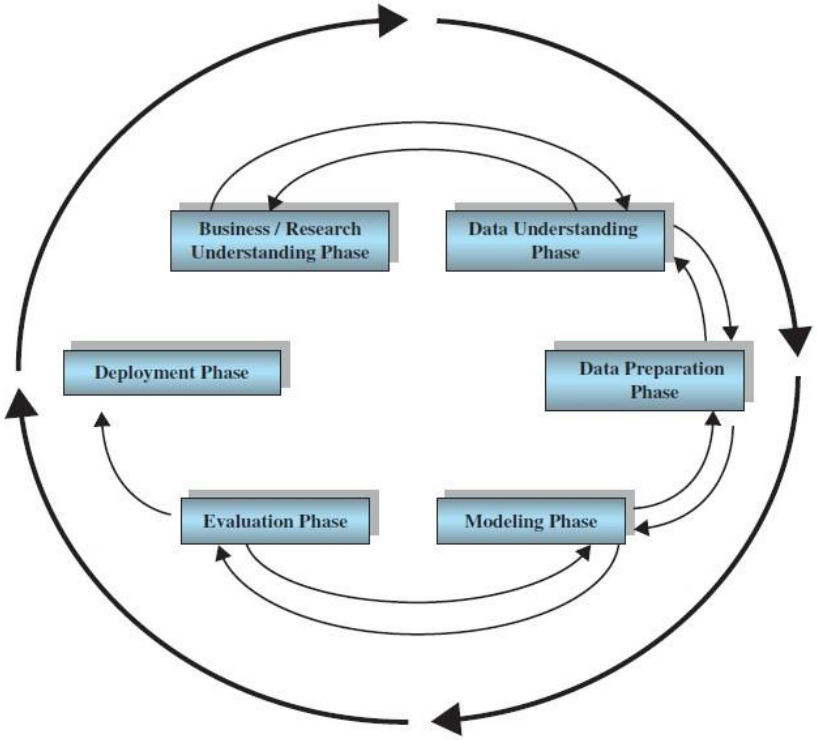
1. Apa yang terkandung dalam data?
2. Jenis pola apa saja yang bisa didapatkan dari data tersebut?
3. Bagaimana bisa semua data ini digunakan untuk masa depan?

Zhao [5] juga menyebutkan *data mining* banyak digunakan dalam berbagai bidang seperti, *retail*, keuangan, telekomunikasi dan media sosial. Menurut Larose [6] *data mining* dapat melakukan kegiatan sebagai berikut:

1. *Description*, atau menggambarkan pola dan kecenderungan yang terdapat dalam data.
2. *Estimation*, atau mengestimasikan nilai dari target variabel yang berbentuk numerik dengan menggunakan variabel prediksi.
3. *Classification*, mirip dengan *estimation*, yang membedakannya adalah target variabelnya berbentuk kategorikal.
4. *Prediction*, hampir sama dengan *classification* dan *estimation*, yang membedakan *prediction* adalah hasilnya terdapat pada masa depan. Contoh: memprediksi harga dari saham 3 bulan kedepan, atau memprediksi persentase kematian pada kecelakaan lalu lintas tahun berikutnya jika batas kecepatan ditingkatkan.
5. *Clustering*, yang dimaksud dengan *clustering* adalah pengelompokkan dari catatan, observasi atau kasus kedalam kelas atau objek yang sama.
6. *Association*, adalah mencari atribut mana yang berhubungan. Model ini adalah model paling lazim ditemui pada dunia bisnis, dimana tugas *association* atau asosiasi berusaha untuk menemukan dan mengukur hubungan antara 2 atribut atau lebih.

2.3 CRISP-DM

CRISP-DM merupakan metode yang bebas untuk dipakai bagi siapa saja untuk meningkatkan keberhasilan *data mining*. Terdapat enam tahap dalam metode ini, dan setiap tahap berjenis adaptif yang artinya tahap selanjutnya bergantung pada hasil tahap sebelumnya. Misal, proyek sedang berada dalam tahap *modeling*, bergantung dari karakteristik dari modelnya sendiri, proyek dapat kembali ke tahap preparasi untuk penyempurnaan sebelum melanjutkan ke tahap evaluasi. Untuk selengkapnya dapat dilihat pada gambar 1.[6]



Gambar 1. 6 Tahap CRISP-DM

2.4 Pohon Keputusan (*Decision Tree*)

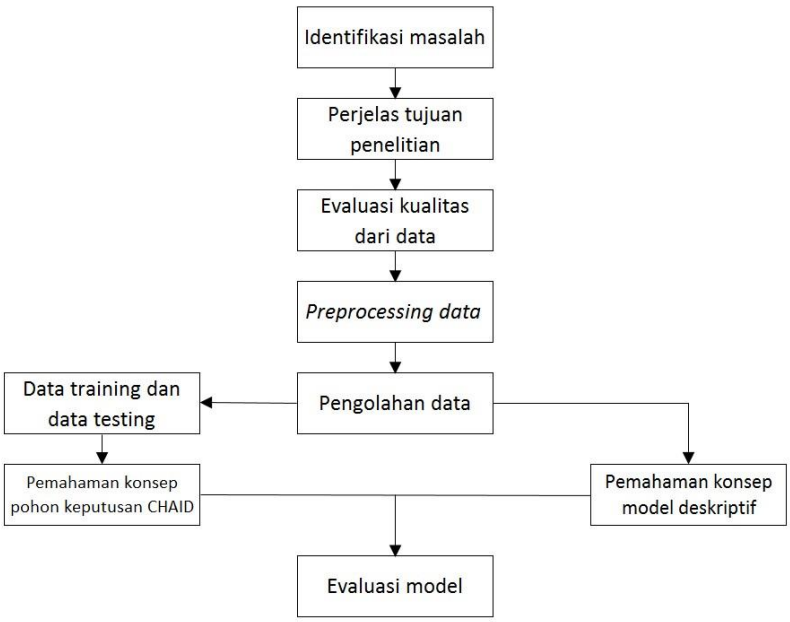
Menurut Rokach dan Maimon [7] pohon keputusan adalah sebuah model prediktif yang dapat digunakan sebagai model klasifikasi dan model regresi. Di dalam pemakaian penelitian pohon keputusan mengacu kepada model hierarki dari sebuah keputusan dan dari hasil masing-masing keputusan. Sang pengambil keputusan memakai pohon keputusan untuk mengidentifikasikan strategi mana yang akan mencapai target. Saat dimana pohon keputusan dipakai sebagai fungsi klasifikasi, akan disebut sebagai *classification tree*, sedangkan jika digunakan sebagai fungsi regresi, maka akan disebut *regression tree*. Menurut Rokach dan Maimon [7] beberapa keunggulan dari pohon keputusan adalah:

1. Pohon keputusan bersifat *self-explanatory*, dan mudah diikuti.
2. Pohon keputusan dapat mengolah baik atribut numerik maupun nominal.
3. Representasi pohon keputusan sudah cukup mewakili nilai diskrit.
4. Pohon keputusan dapat mengolah dataset yang mengandung error.
5. Pohon keputusan dapat mengolah dataset yang mengandung *missing values*.
6. Pohon keputusan tidak memasukkan asumsi tentang distribusi dan struktur klasifikasi.

2.5 Metode Penelitian

Peneliti menggunakan metode *data mining* untuk melakukan *Knowledge Discovery in Database* (KDD), menurut Fayyad et al dalam Rokach dan Maimon [7] sebagai proses *nontrivial* dari mengidentifikasi valid, baru, berpotensi berguna dan memiliki pola yang dapat dimengerti dalam data. Friedman dalam Rokach dan Maimon [7] menganggap

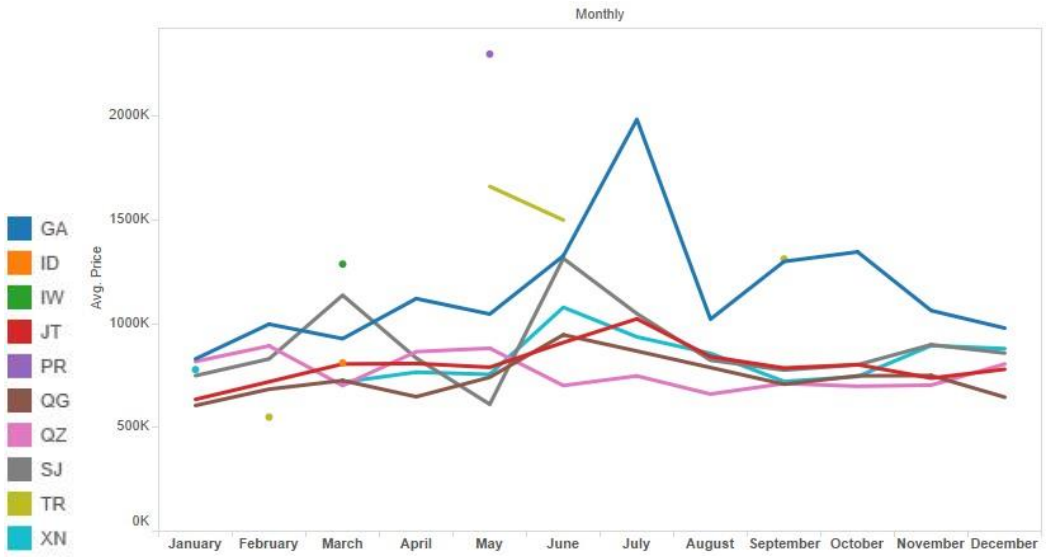
proses KDD sebagai data eksploratori otomatis dari *database* besar. Peneliti menggunakan teori CRISP-DM dari Larose [6] dan KDD dari Han et al [8] sebagai dasar dari tahapan penelitian, dan juga menggunakan teori lain yang sudah ada. Tahapan penelitian dapat dilihat pada gambar 3.



Gambar 3. Tahapan Penelitian

3. Pembahasan

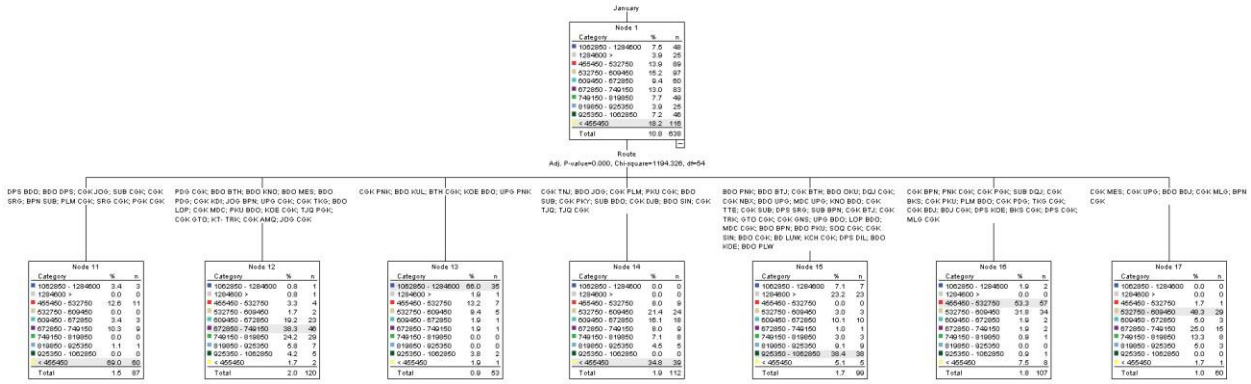
Hasil dari penelitian ini adalah model deskripsi dan prediksi. Beberapa model deskripsi yang didapatkan antara lain model frekuensi destinasi, model frekuensi keberangkatan, model transaksi perbulan, model frekuensi rute, model rata-rata harga maskapai, model rata-rata harga dengan jenis pembayaran dan model rata-rata harga dengan gender. Model yang terdapat pada gambar 4 adalah model rata-rata harga maskapai.



Gambar 4. Rata-rata Harga Maskapai

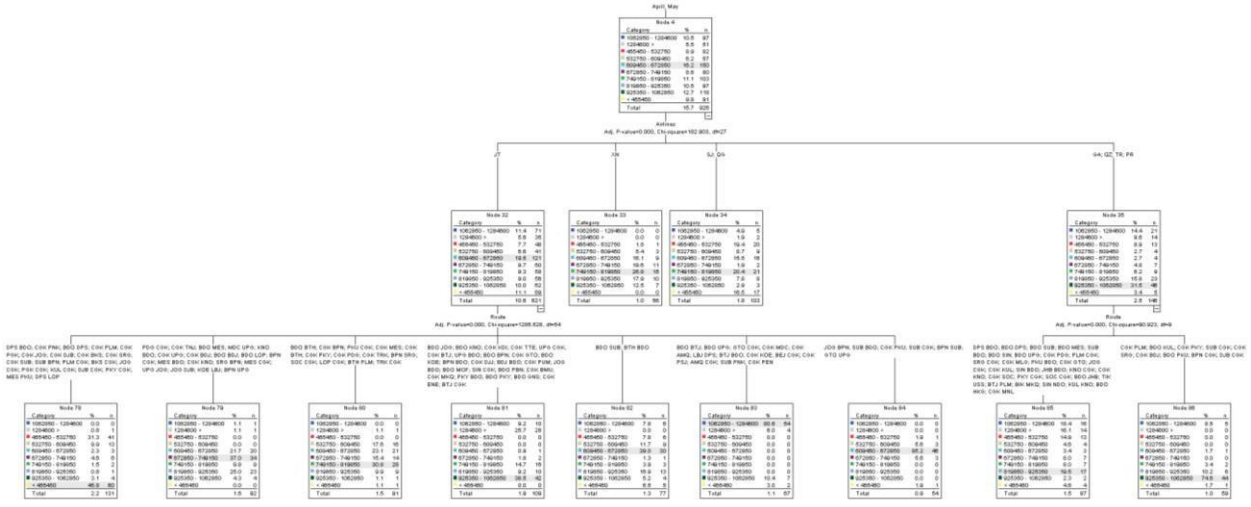
Dari gambar 4 digambarkan rata-rata harga dari tiap maskapai perbulannya di tahun 2014. Terlihat bahwa rata-rata harga tertinggi diduduki oleh Garuda Indonesia dengan kode GA. Sedangkan Lion Air dengan kode JT termasuk dalam rata-rata menengah, ini adalah salah satu alasan mengapa Lion Air menjadi maskapai yang paling sering dipakai untuk bepergian menurut dari data penjualan Tama Tour pada tahun 2014.

Selain model deskripsi, penelitian ini juga menghasilkan model prediksi berbentuk pohon keputusan yang terdiri dari 87 *node*. Model didapatkan menggunakan bantuan perangkat lunak SPSS dan algoritma CHAID. Berikut adalah bagian dari model prediksi yang telah didapatkan:



Gambar 5. Bagian Januari dari Model Prediksi

Gambar 5 adalah *Parent Node* dari bulan Januari, terdapat *Child Node* Rute. Dari model diketahui bahwa mayoritas kemungkinan harga yang terdapat pada bulan Januari adalah < 455450 sebesar 18.2%. Januari mempunyai *child node* sebanyak 7 yakni node 11, 12, 13, 14, 15, 16 dan 17. Untuk interpretasinya diatas tabel node terdapat sejumlah rute, dan didalam tabel terdapat kategori harga beserta persentase. Jadi jika calon pembeli mempunyai *budget* sebesar 500.000 Rupiah dan ingin bepergian pada bulan Januari, jika dilihat dari *Child Node* yang dipunyai kemungkinan terbesar rute yang tersedia terdapat pada node 16. Karena hanya node 16 yang mempunyai persentase kategori harga 455450 – 532750 terbesar di bulan Januari yakni 53.3%. Untuk interpretasi model prediksi bagian yang lain sejenis dengan interpretasi pada gambar 5. Berikut adalah contoh lain dari sebagian model prediksi:



Gambar 6. Bagian April dan Mei dari Model Prediksi

Untuk model dari tiap *Parent Node* interpretasinya kurang lebih sama, tetapi pada gambar 6 *Parent Node* ke 11 yakni bulan April dan Mei selain *Child Node* Rute, juga terdapat maskapai. Ini disebabkan karena pada bulan April dan Mei, selain Rute yang mempengaruhi harga penjualan juga terdapat faktor Maskapai. Untuk interpretasi khusus bulan April dan Mei jika *budget* yang dimiliki oleh calon pembeli pertama akan disesuaikan dengan maskapai terlebih dahulu kemudian rute. Misal calon pembeli memiliki *budget* 1.000.000 Rupiah maka calon pembeli dapat memilih baik maskapai yang terdapat dalam node 32, 33, 34 dan 35. Namun mayoritas kategori harga untuk 1.000.000 Rupiah berada pada node 35 yakni maskapai Garuda Indonesia, Citilink, Tiger Airways atau Philippine Airlines. Dan rute yang tersedia dengan kemungkinan terbesar yakni 74.6% terdapat didalam node 86.

4. Kesimpulan

Berdasarkan hasil yang telah didapatkan dalam bab yang sebelumnya, dan untuk menjawab pertanyaan penelitian, skenario model prediksi dan deskripsi yang didapatkan dari data penjualan tiket pesawat oleh Tama Tour pada tahun 2014 adalah sebagai berikut:

1. Model Frekuensi Destinasi
2. Model Frekuensi Keberangkatan
3. Model Transaksi Perbulan
4. Model Frekuensi Rute
5. Model Rata-rata Harga Maskapai
6. Model Rata-rata Harga Dengan Jenis Pembayaran
7. Model Rata-rata Harga Dengan Gender
8. Model Prediksi Berbentuk Pohon Keputusan

Seluruh model deskripsi dapat digunakan sebagai wawasan tambahan untuk kebutuhan strategis perusahaan. Dan model prediksi dapat diimplementasikan untuk menjadi algoritma dasar fitur rekomendasi destinasi pariwisata berdasarkan *budget*. Model prediksi yang dihasilkan berasal dari pengolahan data penjualan tiket oleh Tama Tour pada tahun 2014, menggunakan bantuan perangkat lunak SPSS dan algoritma CHAID maka dihasilkan model prediksi berbentuk pohon keputusan yang memiliki 87 *node*. Model ini diharapkan dapat memberikan *competitive advantage* bagi TourismWave dalam fitur unik untuk bersaing dalam industri *intermediary*.

Daftar Pustaka

[1] Badan Pusat Statistik. (2013). *Jumlah Keberangkatan Penumpang dan Barang di Bandara Indonesia Tahun 1999 - 2013* [*Online*]. Tersedia: http://www.bps.go.id/linkTabelStatis/view/id/1404 [26 Juli 2015]

[2] Dumbill, Edd. (2012). *Planning for Big Data: A CIO's Handbook to the Changing Data Landscape*. Sebastopol: O'Reilly Media Inc.

[3] Sathi, Arvind. (2012). *Big Data Analytics: Disruptive Technologies for Changing the Game*. Boise: MC Press.

[4] Ahlemeyer-Stubbe, Andrea., dan Coleman, Shirley. (2014). *A Practical Guide to Data Mining for Business and Industry*. West Sussex: John Wiley & Sons Ltd.

[5] Zhao, Yangchang. (2013). *R and Data Mining, Examples and Case Studies.* Elsevier Inc.

[6] Larose, Daniel T., dan Larose, Chantal D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining Second Edition*. New Jersey: John Wiley & Sons Inc.

[7] Rokach, Lior., dan Maimon, Oded. (2015). *Data Mining with Decision Trees: Theory and Applications 2nd Edition.* Singapore: World Scientific Publishing.

[8] Han, Jiawei., Kamber, Micheline., dan Pei, Jian. (2012). *Data Mining Concepts and Techniques Third Edition*. Waltham: Elsevier.