**APLIKASI REKOMENDASI METODE ANALISIS SESUAI DENGAN KARAKTER DATA**

Andre Gunawan, Henry Novianus Palit, Ph.D, Andreas Handojo, M.MT

Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Kristen Petra

Jl. Siwalankerto 121 – 131 Surabaya 60236

Telp. (031) – 2983455, Fax. (031) – 8417658

E-Mail: agun111295@gmail.com1, hnpalit@petra.ac.id2, andreas@petra.ac.id3

**ABSTRAK**

Data telah menjadi asset yang sangat berharga dalam perusahaan, namun yang lebih berhaga dari data adalah informasi yang dapat kita olah dari data tersebut. Informasi yang kita dapatkan dari data memiliki banyak manfaat. Dalam pengambilan keputusan bisnis data memiliki peran yang besar untuk menentukan apakah tindakan bisnis yang diambil dapat menimbulkan dapak baik pada perusahaan.Perusahaan – perusahaan yang mendasari keputusan yang dibuat berdasarkan data yang ada (*Data-Driven Decision Making*) disebut sebagai *Data-Driven Company*, perusahaan ini pada uumumnya memiliki sebuah divisi yang bergerak dalam bidang *Business Intelligence*.

*Business Intelligence* adalah sebuah sistem terintegrasi yang memberikan fakta / informasi dari data yang telah diolah untuk kepentingan pengambilan keputusan. Sistem ini bertugas untuk memberikan dukungan keputusan (Decision Suport) untuk tujuan yang spesifik pada sebuah proses bisnis. Fondasi dari Business Intelligence adalah data yang telah diolah menjadi informasi yang berguna untuk proses pendukungan keputusan. *Business Intelligence* menggunakan bergabai metode untuk melakukan ekstraksi informasi dari data yang ada. *Business Intelligenc*e memberikan informasi pada waktu yang tepat, kepada orang yang tepat dalam bentuk yang mudah dipahami pula.

**Kata Kunci**

*Business Intelligece*, *Classification*, *Regression*

***ABSTRACT***

*Data has become a very valuable asset in the company, but the more powerful than the data is information that we can though from the data. The information we get from the data has many benefits. In business decision making the data has a big role to determine whether the business action taken can lead to a good foot in the company. Companies that underlie data-driven decision making (Data-Driven Decision Making) are referred to as Data-Driven Company this in general has a division that is engaged in Business Intelligence.*

*Business Intelligence is an integrated system that provides facts / information from data that has been processed for the benefit of decision making. This system is in charge of providing decision support for the specific purpose of a business process. The foundation of Business Intelligence is data that has been processed into useful information for the decision support process. Business Intelligence uses a multitude of methods to extract information from existing data. Business Intelligence provides information at the right time, to the right person in easy-to-understand form*

***Keywords***

*Business Intelligece*, *Classification*, *Regression*

# PENDAHULUAN

Data telah menjadi asset yang sangat berharga dalam perusahaan, namun yang lebih berhaga dari data adalah informasi yang dapat kita olah dari data tersebut. Informasi yang kita dapatkan dari data memiliki banyak manfaat. Dalam pengambilan keputusan bisnis data memiliki peran yang besar untuk menentukan apakah tindakan bisnis yang diambil dapat menimbulkan dapak baik pada perusahaan.Perusahaan – perusahaan yang mendasari keputusan yang dibuat berdasarkan data yang ada (*Data-Driven Decision Making*) disebut sebagai *Data-Driven Company*, perusahaan ini pada uumumnya memiliki sebuah divisi yang bergerak dalam bidang *Business Intelligence*.

*Business Intelligence* adalah sebuah sistem terintegrasi yang memberikan fakta / informasi dari data yang telah diolah untuk kepentingan pengambilan keputusan. Sistem ini bertugas untuk memberikan dukungan keputusan (Decision Suport) untuk tujuan yang spesifik pada sebuah proses bisnis. Fondasi dari Business Intelligence adalah data yang telah diolah menjadi informasi yang berguna untuk proses pendukungan keputusan. *Business Intelligence* menggunakan bergabai metode untuk melakukan ekstraksi informasi dari data yang ada. *Business Intelligenc*e memberikan informasi pada waktu yang tepat, kepada orang yang tepat dalam bentuk yang mudah dipahami pula.

Semakin berkembangnya Business Intelligence, kemampuan kebutuhan untuk pengolohan data juga akan meningkat, tetapi tenaga kerja yang dapat melaksanakan pekerjaan tersebut tidak dapat memenuhi kebutuhan tersebut. Skripsi ini dibuat untuk menyediakan sebuah alat untuk melakukan analisa secara interaktif, dan membantu pengguna dalam melakukan analisa. Salah satu proses yang membutuhkan bantuan dalam bentuk rekomendasi adalah ketika pengguna ingin membuat model prediksi. Banyak metode yang disediakan untuk melakukan prediksi tetapi tidak semua sesuai dengan data yang sekarang dimiliki pengguna.

# DASAR TEORI

## Business Intelligence Task

Untuk mencapai tujuan dalam melakukan analisa ada beberapa tugas / *task* yang harus dilalui. Beberapa tugas / *task* itu adalah:

1. Data *Task*

Tugas ini adalah bagaimana mendapatkan data dan apa pengertian / informasi mengenai data tersebut. Informasi yang didapatkan saat melakukan *Business Intelligence* semua berasal dari data, maka tugas ini perlu dilakukan untuk mengetahui sumber data.

1. *Business and Data Understanding*

Tugas ini dilakukan untuk memahami dan mengerti apa yang dimaksut dengan data yang dimiliki. Selain itu pengertian mengenai proses bisnis juga diperlukan untuk memahami bagaimana data tergenerasi.

1. *Modeling Task*

Tugas ini dilakukan untuk mengetahui model analisa yang ingin dilakukan. Model analisa ini digunakan untuk menghasilkan informasi yang dibutuhkan.

1. *Analysis Task*

Tugas ini adalah implementasi dari *Modeling Task* dimana algoritma digunakan untuk menghasilkan informasi. Tugas ini adalah tugas utama dari *Business Intelligence* untuk menghasilkan fakta / pengetahuan mengenai bisnis.

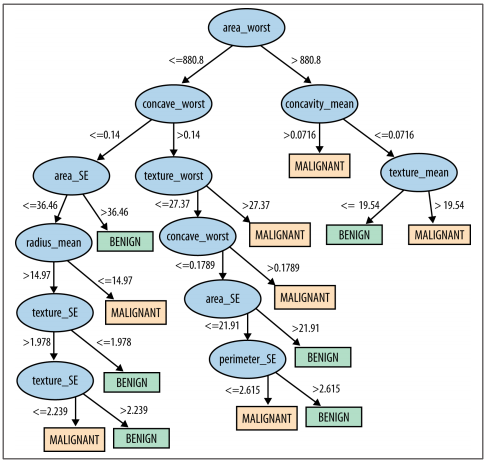
1. Evaluation & Reporting Task

Tugas ini adalah tugas terakhir dalam *Business Intelligence* tugas ini dilakukan untuk melakukan evaluasi mengenai performa yang dihasilkan model yang dibuat. Pada tahap ini visualisasi dari data juga diperlukan untuk proses  *reporting.* Visualisasiyang baik dapat menghasilkan informasi yang mudah untuk dipahami

## Metode Prediksi

### Classification

*Classification* merupakan metode yang mencoba untuk melakukan prediksi dengan melakukan klasifikasi data. Klasifikasi mencari data tersebut milik *class* yang mana. Salah satu contoh pertanyaan *classification* adalah “Dari semua *customer* siapa yang akan merespon penawaran yang akan diberikan?”. Dalam kasus ini terdapat 2 *class* yaitu respon dan tidak respon. Salah satu contoh metode yang dapat melakukan *classification* adalah *decision tree. Decision tree* merupakan metode yang memodelkan data menjadi bentuk sebuah *tree* dimana data baru yang yang masuk akan berjalan dan menghasilkan sebuah alur yang menyatakan pada *class* mana data tersebut terklasifikasi.[1]



**Gambar 1. Contoh Metode Classification [1]**

### Metode Regression

Regression atau value estimation mencoba untuk melakukan estimasi atau meprediksi sebuah nilai numeric variabel pada sekumpulan data. Salah satu contoh pertanyaan *regression* adalah “Berapa banyak customer yang akan menggunakan layanan kita”. Properti variabel yang ingin diprediksi adalah jumlah penggunaan layanan dan model analisa dapat terbuat dengan melihat variable lain yang sejenis.

Regression berhubungan dengan Classification tapi bukan merupakan sesuatu hal yang sama. Classification mencoba untuk memprediksi apakah sesuatu akan terjadi atau tidak, sedangkan Regression mencoba untuk memprediksi berapa banyak yang akan dihasilkan.[1]

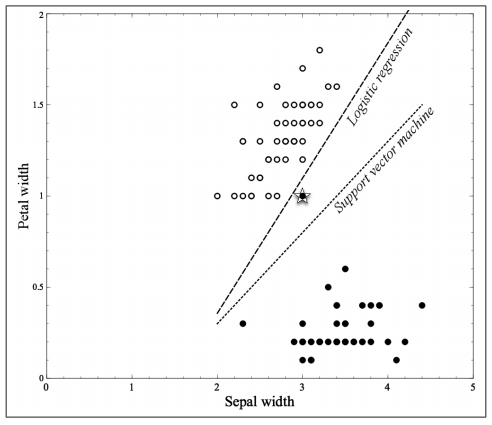
Rumus *Regression* adalah:

**Y1 = X11 β1 + X12 β2 + … + X1k βk + ε1**

**Y2 = X21 β1 + X22 β2 + … + X2k βk + ε2**

**…**

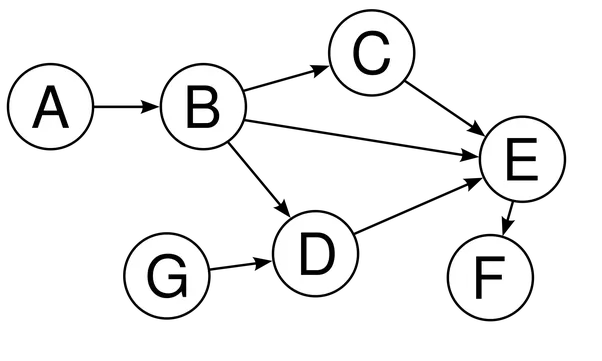
**Yn = Xn1 β1 + Xn2 β2 + … + Xnk βk + εn**



Gambar 2. Contoh Metode Regresi [1]

## Directed Acyclic Graph (DAG)

*Directed Acyclic Graph* merupakan *graph* yang tidak memiliki *cycle* atau kondisi dimana sebuah *node* kembali ke *node* sebelumnya.[6]



**Gambar 3. Aristektur Mapreduce pada Hadoop**

DAG digunakan untuk membuat tampilan *interface data flow* yang interaktif. Tampilan ini merupakan bentuk antar muka yang antara *user* dengan aplikasi, dimana *user* melakukan *pre-processing* data.

## Flask

*Flask* merupakan *microframework* yang dibangun dengan menggunakan bahasa pemrograman *Python*. *Flask* digunakan untuk me-*develop* sebuah aplikasi web. *Flask* merupakan *microframework* yang artinya *flask* membuat sebuah pengerjaan aplikasi web menjadi mudah dan *simple* karena dapat menjalankan sebuah web hanya dengan menggunakan 1 file *Python*. *Flask* membuat susunan kerja yang ringan, dan mudah tetapi juga dapat dikembangkan dengan mudah.[7]

## Data Kaggle

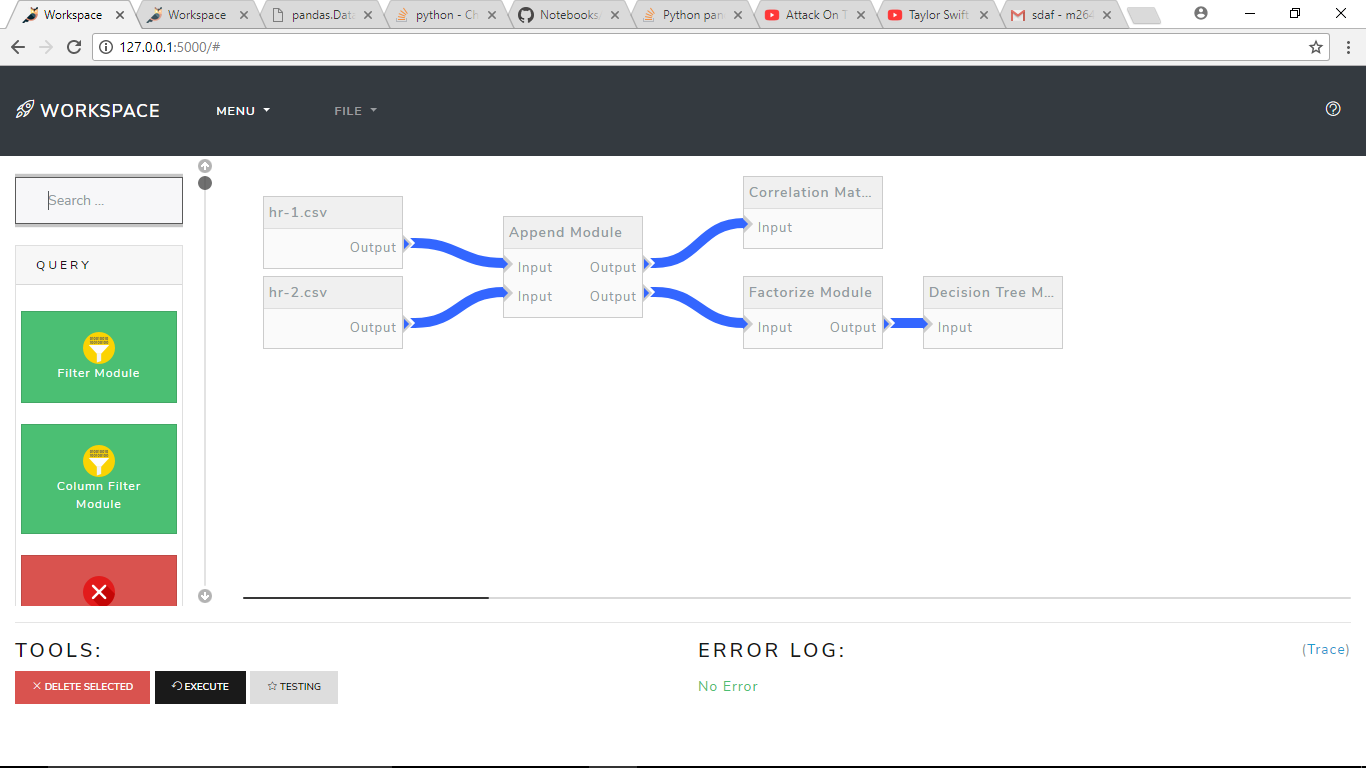
Dataset yang digunakan dalam penelitian ini adalah data – data yang berasal dari website kaggle. Data – data tersebut antara lain:

1. *Human Resource* [2]
2. *House Price [3]*
3. *West Nile Virus [4]*
4. *Bakery [5]*

# DESAIN SISTEM

## *Data Flow*

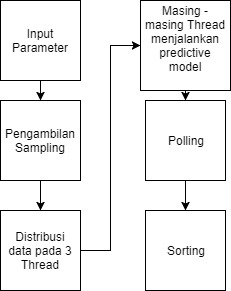
Data Flow merupakan tampilan antar muka yang memudahkan *user* untuk melakukan proses *pre-processing*. Tampilan ini dibuat berdasarkan DAG, yang mana masing – masing *node* memiliki parameter / data yang diinputkan *user.*



Gambar 4. Tampilan Data Flow

## Proses Rekomendasi

Proses ini dilakukan secara multithreading dengan membagi thread sebanyak metode prediksi yang ingin dijalankan.



Gambar . Proses Rekomendasi Secara Umum

Proses dimulai dengan pemisahan data menjadi n bagian sesuai dengan banyak metode prediksi yang ingin dijalankan bersama. Setelah itu data pada masing – masing thread akan di olah sesuai dengan metode prediksi tiap thread. Setelah itu terjadi proses polling dimana aplikasi akan menunggu hasil dari semua thread, jika semua hasil tiap thread sudah didapatkan, maka selanjutnya adalah proses sorting. Proses sorting memiliki 2 mode, yaitu *accuracy* atau *time.* Proses sorting akan melakukan sort berdasarkan mode yang dipilih pengguna.

# IMPLEMENTASI SISTEM

Implementasi sistem dilakukan pada computer dengan spesifikasi:

* RAM: 8GB, DDR3
* HDD: 500 GB
* CPU: Core i5
* OS Ubuntu Server 16.06 LTS

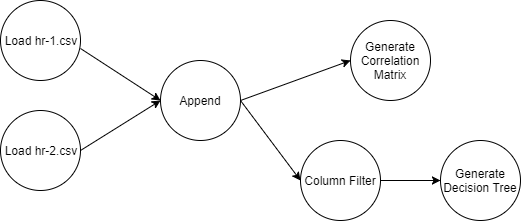
# ANALISA DAN PENGUJIAN

## *Data Flow*

*Data flow* diuji dengan hasil berjalanya proses, apakah hasilnya sesuai dengan *module - module* yang di inginkan.

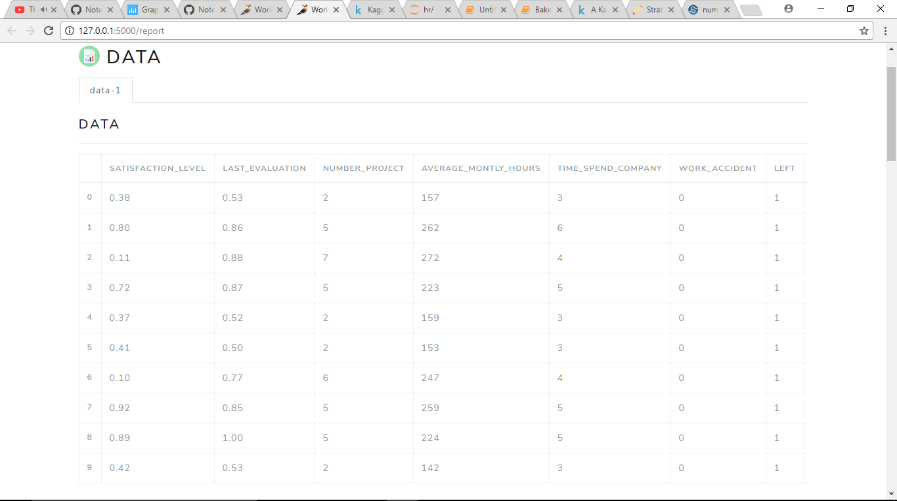
### Bentuk DAG

Bentuk DAG yang diuji adalah sebagai berikut



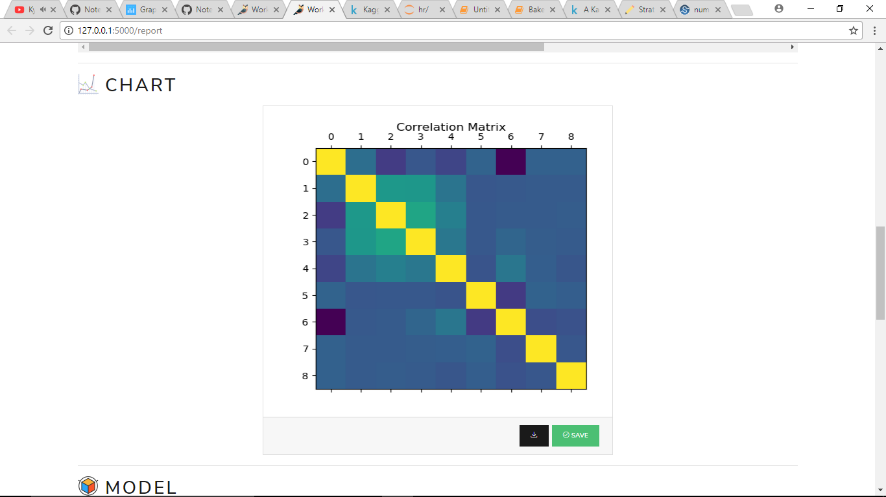
**Gambar 7. DAG yang diuji**

Data Flow diatas akan menghasilkan sebuah model dan sebuah *correlation matrix* dalam bentuk *chart*. Setelah di jalankan aplikasi menghasilkan:



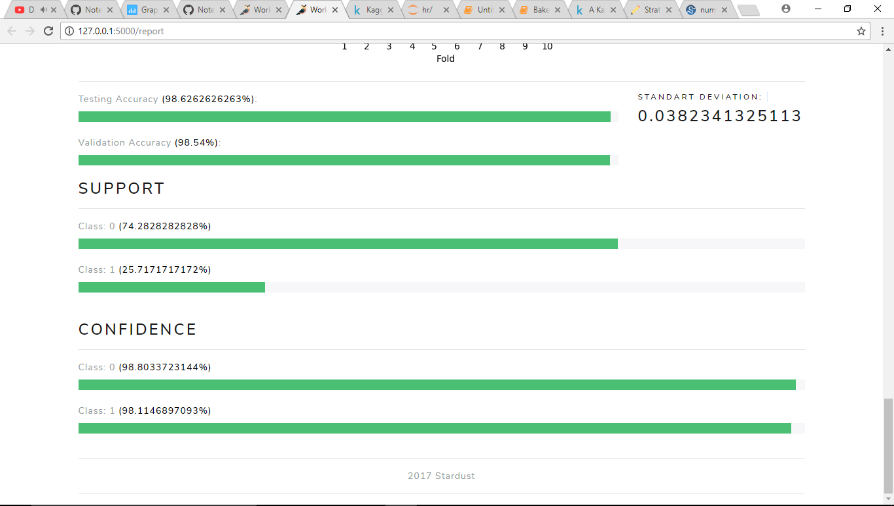
**Gambar 8. Data Hasil DAG**

Data dengan bentuk yang sama sesuai dengan module yang digunakan.



**Gambar 9. Chart Hasil DAG**

Chart yang diinginkan sesuai dengan *module* yang digunakan.



**Gambar 10. Model Hasil DAG**

Statistik sederhana dari model yang dipilih pada DAG.

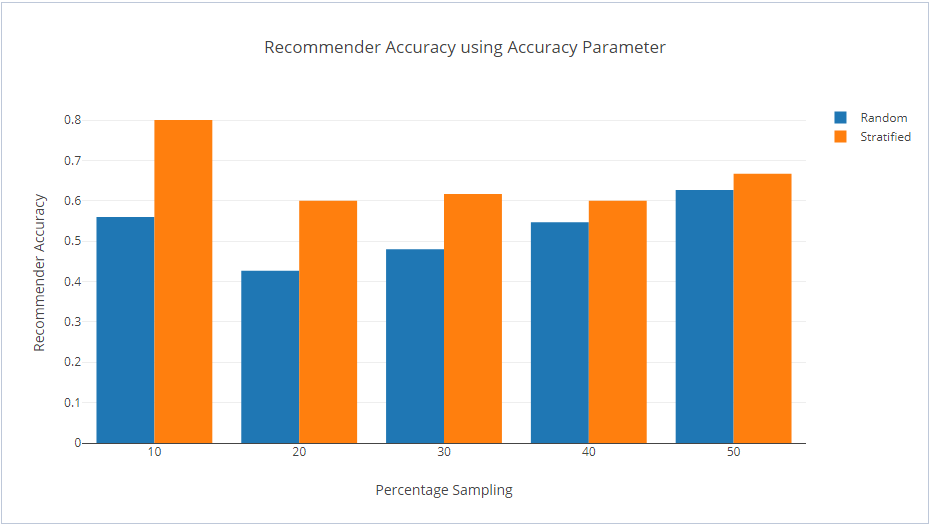
## Pengujian Sistem Rekomendasi

Pengujian sistem rekomendasi dilakukan dengan 4 data, dari setiap data proses akan berjalan 5 kali dimana setiap iterasi aplikasi akan melakukan sampling dari 10% sampai 50% dengan rentan 10%. Sampling yang dilakukan ada 2 yaitu random sampling dan stratified sampling. Perhitungan akurasi rekomendasi dilakukan dengan membandingkan 2 urutan kemudian melakukan preseentasi berapa banyak yang memiliki urutan yang sama Hasil rata – rata dari pengujian ini adalah:

|  |  |  |
| --- | --- | --- |
|  | Random Sampling | Stratified Sampling |
| 10% | 0.56 | 0.8 |
| 20% | 0.426666667 | 0.6 |
| 30% | 0.48 | 0.616666667 |
| 40% | 0.546666667 | 0.6 |
| 50% | 0.626666667 | 0.666666667 |
| **AVG** | 0.528 | 0.656666667 |

**Tabel 1. Hasil Sistem Rekomendasi, Parameter Accuracy**

Terlihat bahwa rata – rata *accuracy recommender* lebih baik ketika *sampling* yang digunakan adalah *Stratified Sampling.*

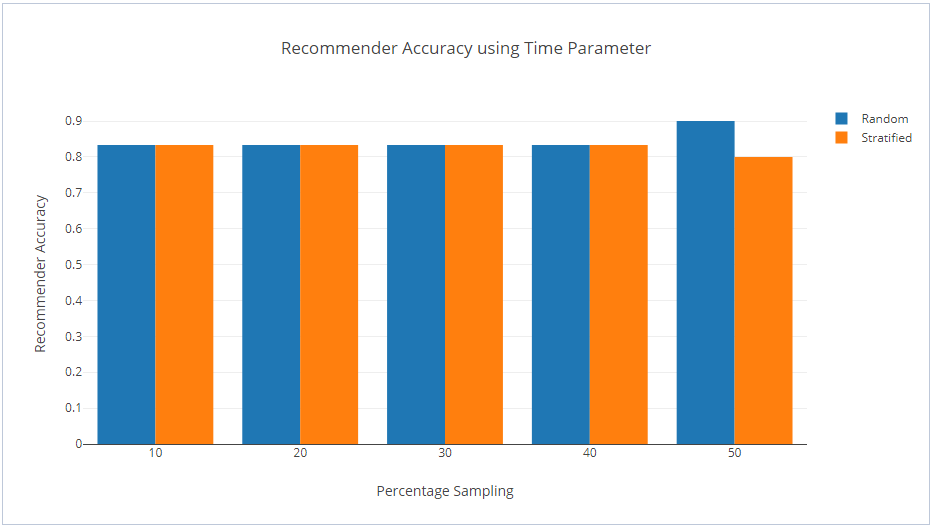


**Gambar 11. Chart Hasil Sistem Rekomendasi Parameter Accuracy**

Selanjutnya pengujian akan dilakukan terhadap parameter *time.* Pengujian ini dilakukan dengan cara yang sama dengan sebelumnya namun ketika proses sorting dilakukan, aplikasi akan melakukan cort berdasarkan waktu tercepat dalam eksekusi.

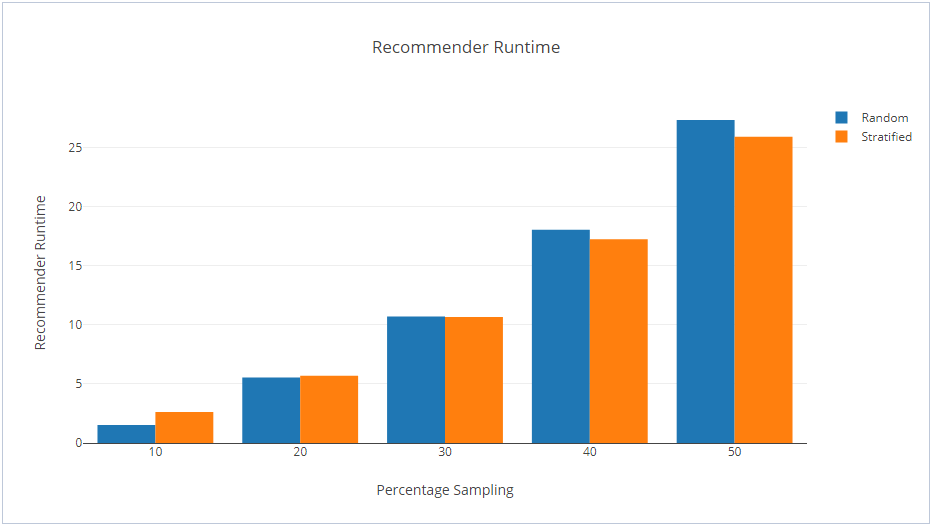
|  |  |  |
| --- | --- | --- |
|  | Random Sampling | Stratified Sampling |
| 10 | 0.833333333 | 0.833333333 |
| 20 | 0.833333333 | 0.833333333 |
| 30 | 0.833333333 | 0.833333333 |
| 40 | 0.833333333 | 0.833333333 |
| 50 | 0.9 | 0.8 |
| **AVG** | 0.846666667 | 0.826666667 |

**Tabel 2. Hasil Sistem Rekomendasi, Parameter Time**



**Gambar 12. Chart Hasil Sistem Rekomendasi Parameter Time**

Pengujian yang terakhir adalah dengan melihat waktu berjalanya sistem rekomendasi



**Gambar 13. Chart Lama Runtime Sistem Rekomendasi**

# KESIMPULAN

Setelah dilakukan perancangan sistem, pengimplementasian, dan pengujian terhadap aplikasi yang telah dibuat, dapat ditarik kesimpulan sebagai berikut:

* Aplikasi memiliki 2 fitur utama, yaitu *data flow* *interface* dan sisten rekomendasi metode analisis.
* *Interface Data Flow* dibuat berdasarkan *Directed Acyclic Graph* yang setiap *node* memiliki parameter yang berbeda – beda dan tersimpan dalam 2 *metadata* yaitu *parameters* dan *schema* dengan format JSON*.*
* Aplikasi meartikan *metadata* *paramters* yang berbentuk JSON dengan menggunakan metode BFS (Breath First Search) yang dikombinasikan dengan *shared resource*.
* *Interface Data Flow* dapat diartikan menjadi proses yang dimengerti komputer sehingga metode yang diusulkan dapat dikembangkan
* Sistem rekomendasi metode analisa dapat membantu dengan memberikan sugesti metode yang sebaiknya digunakan dengan kondisi data yang ada.
* Sistem rekomendasi akan memberikan rekomendasi yang berbeda untuk data yang berbeda, dan data yang sama namun telah melalui proses *pre-processing* yang berbeda
* Sistem rekomendasi memiliki 2 mode, yaitu *Random Sampling* dan *Stratified Sampling* dimana *Random Sampling* memberikan rekomendasi secara cepat, sedangkan *Stratified Sampling* memberikan rekomendasi yang lebih konsisten.
* Sistem rekomendssi sangat dipengaruhi kondisi data yang sedang dianalisa, seberapa bersih, normal dan padat data yang di analisa membuat sistem rekomendasi memiliki akurasi yang beragam (deviasi tinggi)
* Sistem keseluruhan dibuat dalam linkungan server Ubuntu 16.06 LTS, dimana aplikasi merupakan aplikasi *web-based* dengan teknologi *cloud.* Aplikasi web yang berbasis cloud memudahkan penggunaan karena dapat diakses dimanapun dan kapanpun.

Saran untuk pengembangan kedepannnya adalah:

* Penambahan module pada halaman *workspace*.
* Pengembangan *error log* yang lebih interaktif untuk pengguna terutama ketika *error* yang terjadi merupakan *runtime error*.
* Pengembangan fitur untuk memberikan *hyper parameter tuning* seperti *GridSearch* atau *GridSearchCV* untuk meningkatkan sistem rekomendasi metode analisa beserta parameternya.
* Integrasi dengan infrastruktur *Big Data* seperti Hadoop, Spark, Hive, Kafka.
* Peningkatan infrastruktur *cloud* yang terdistribusi sesusai dengan point sebelumnya, dengan tujuan untuk meningkatkan kekuatan *computing* dan infrastruktur memiliki sifat skalabilitas

# DAFTAR PUSTAKA

1. Provost, F & Fawcett, T, 2013. Data Science for Business. O’Reilly.
2. Kaggle. 2015.; Human Resources Analytics. URI = https://www.kaggle.com/ludobenistant/hr-analytics/data
3. Kaggle. 2015.; House Prices: Advanced Regression Techniques. URI = https://www.kaggle.com/c/house-prices-advanced-regression-techniques
4. Kaggle. 2015.; West Nile Virus Prediction. URI = https://www.kaggle.com/c/predict-west-nile-virus
5. Trac. 2015.; Extended BAKERY dataset. URI = https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery
6. Monash. 1999 Directed Acyclic Graphs. URI = http://users.monash.edu/~lloyd/tildeAlgDS/Graph/DAG/
7. Flask Overview | Flask. URI = http://flask.pocoo.org/