**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS KRISTEN PETRA**

**USULAN SKRIPSI**

|  |  |  |
| --- | --- | --- |
| Nama | : | Andre Gunawan |
| Nrp | : | 26414004 |
| Program | : | Sistem Informasi Bisnis |
| Judul Skripsi | : | Aplikasi Rekomendasi Metode Analisis Sesuai dengan Karakter Data |
| Pembimbing 1 | : | Henry Novianus Palit, Ph.D. |
| Pembimbing 2 | : | Andreas Handojo, M.MT. |
| Dilaksanakan | : | Semester Gasal 2017 / 2018 |

|  |  |  |
| --- | --- | --- |
|  |  | Surabaya, 21 April 2017 |
|  |  | Yang mengusulkan |
|  |  |  |
|  |  |  |
|  |  | Andre Gunawan |
|  |  |  |
|  | Mengetahui : |  |
| Pembimbing I |  | Pembimbing II |
|  |  |  |
|  |  |  |
| Henry Novianus Palit, Ph.D |  | Andreas Handojo, M.MT. |
|  |  |  |
|  | Koordinator Skripsi |  |
|  |  |  |
|  |  |  |
|  | Silvia Rostianingsih, M.MT |  |

1. **Judul Skripsi**

Aplikasi Rekomendasi Metode Analisis Sesuai dengan Karakter Data

1. **Latar Belakang Masalah**

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. Contoh pemanfaatan data menjadi informasi:

Dalam website *Forbes* dikatakan bahwa salah satu *retail mart*  ingin melakukan strategi penjualan di mana perusahaan tersebut men-target ibu hamil sebagai target pembeli. Ketika ibu hamil maka para ibu membutuhkan keperluan untuk ibu hamil tetapi beberapa waktu kemudian para ibu membutuhkan keperluan untuk bayi yang dilahirkan. *Retail mart* ini melakukan analisa untuk memprediksi ibu hamil dari data transaksi yang dilakukan pembeli. (Kashmir, 2012)

Walmart, salah satu perusahaan retail terbesar di Amerika telah memiliki banyak cabangan yang tersebar di seluruh Amerika. Perusahaan retail ini harus memenuhi *demand* atau kebutuhan pelanggan dengan cabang yang buka. Salah satu hal yang mempengaruhi pembelian produk pada Walmart adalah cuaca. Walmart melakukan analisa data penjualan yang digabungkan dengan data cuaca di setiap cabang perusahaan dan menemukan pola penjualan yang dapat membantu mengambil keputusan barang apa yang sebaiknya dijual dengan jumlah banyak sebelum memasuki musin angin topan. (Provost & Fawcett, 2013)

Berdasarkan contoh yang diberikan, data menjadi salah satu asset yang berharga karena jika diolah dengan baik, data dapat menjadi informasi yang sangat baik dalam pengambilan keputusan. Dikutip dari website *Wired* “*When your business is growing, more and more people have opinions about which steps need to be taken. It helps to work with a ‘good data beats opinion’ philosophy. Almost everything can be tested, measured and improved.*” (Joris Toonders, 2014), data mengalahkan opini. Data dapat memberikan kenyataan apa yang sedang terjadi, hampir semua pengambilan keputusan yang dilakukan berdasarkan data yang ada, oleh karena itu dapat mengolah data dengan baik dapat memberikan dampak yang baik dalam mengembangkan bisnis terutama dalam bidang *Business Intelligence*.

Skripsi ini dibuat untuk memberikan sebuah rekomendasi kepada user. Rekomendasi yang diberikan adalah metode analisa yang sesuai dengan karakteristik data yang ingin dianalisa user. Aplikasi akan menerima input berupa *file* dengan format CSV (*Comma Seperated Value*) dan akan memberikan rekomendasi metode analisa yang baik digunakan sesuai dengan input *file*. Metode analisa yang digunakan adalah *Classification* dan *Regression*, kemudian hasil dari proses ini adalah *predictive model.* *Model* dapat disimpan kedalam sebuah *file*.

Pada skripsi ini aplikasi menggunakan beberapa data yaitu *West Nile Virus*, merupakan data yang berisi mengenai data pengambilan sample nyamuk pada kota Chicago dan indikasi adanya penyakit *West Nile Fever* pada kota tersebut. House Price, merupakan data yang berisi harga rumah dengan fitur – fitur yang ada pada rumah tersebut. Human Resource, merupakan data yang berisi hasil survey kepuasan pegawai, shift kerja pegawai tersebut, dan cacatan apakah pegawai itu keluar dari perusahaan atau tidak. Insurance, merupakan data yang berisi data pelanggan asuransi mobil beserta harga *purchase* masing – masing pelanggan. Bakery, merupakan data yang berisi transaksi sebuah toko roti beserta produk yang dijual.

Semua data digunakan untuk meanalisa metode yang akan direkomendasikan. Lima data ini akan dijadikan input, kemudian user dapat melakukan *pre-processing* pada aplikasi dan meminta rekomendasi metode pada aplikasi ini. Aplikasi akan memberikan rekomendasi metode dan user dapat memilih untuk mempercayai rekomendasi tersebut atau tidak. Hasil akhir dari aplikasi ini adalah *file* yang berisi model hasil proses analisa.

1. **Perumusan Masalah**

Masalah yang akan timbul:

1. Bagaimana mengetahui metode yang baik menurut user, contoh: kecepatan atau akurasi.
2. Bagaimana melakukan pemilihan metode – metode *Machine Learning* yang sesuai dengan karakteristik data.
3. Bagaimana melakukan *sampling* untuk melakukan analisa awal untuk menentukan model yang baik
4. Bagaimana menentukan cara validasi model pada analisa awal yang dilakukan
5. **Tujuan Skripsi**

Tujuan dari penelitian ini adalah membuat sebuah aplikasi analisis data dengan fitur sugesti/rekomendasi metode yang sebaiknya digunakan.

1. **Ruang Lingkup**

Skripsi ini dibuat dengan batasan:

1. Data yang akan dijadikan input adalah data tabular dengan format CSV (*Comma Seperated Value*) yang memiliki kolom lebih dari 2.
2. Banyak  *record* atau jumlah baris pada data tabular lebih dari 1000 baris / *record.*
3. Data CSV sebagian besar diambil dari website http://www.kaggle.com
4. Terdapat 5 Kasus yang akan digunakan untuk analisa, yaitu:
   1. West Nile Virus (https://www.kaggle.com/c/predict-west-nile-virus)
   2. House Price (https://www.kaggle.com/c/house-prices-advanced-regression-techniques)
   3. Human Resource (https://www.kaggle.com/ludobenistant/hr-analytics)
   4. Insurance (https://www.kaggle.com/c/allstate-purchase-prediction-challenge/data)
   5. Bakery (https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery)
5. Sebelum masuk kedalam proses, user dapat memilih menggunakan beberapa kolom saja untuk melakukan analisa
6. Aplikasi dapat menambakan baris pada data
7. Aplikasi dapat menambahkan kolom pada data dengan nama kolom diinputkan
8. Aplikasi dapat *update* atau merubah data sesuai dengan data yang dipilih oleh user.
9. Aplikasi dapat *delete* atau menghapus data sesuai dengan data yang dipilih oleh user.
10. Aplikasi dapat menggambarkan plot atau grafik
11. Proses pada data menggunakan metode:
    1. *Classification*

Proses meklasifikasikan data pada suatu kelas seperti (Beli, Tidak Beli)

* 1. *Regression*

Proses menghitung berapa besar nilai yang akan keluar berikutnya seperti berapa besar harga rumah.

1. Proses Rekomendasi, merupakan proses dimana aplikasi akan memberikan rekomendasi metode analisa yang sebaiknya digunakan. Proses ini akan mengambil data *sample* kemudian data tersebut akan dijalankan kepada beberapa metode yang berbeda dan setelah itu melakukan evaluasi untuk menentukan metode yang akan digunakan.
2. User dapat memilih untuk mempercayai hasil rekomendasi atau menolak hasil rekomendasi.
3. Proses *Model Creation* dimana aplikasi akan melakukan analisa dengan motode yang diinginkan user (dari rekomendasi, atau user memilih sendiri)
4. Hasil dari analisa atau output adalah sebuah keterangan atau *result* dari model yang dibuat. Hasil merupakan validasi *accuracy* model.
5. User dapat menyimpan model kedalam bentuk file *pickle (object serialization)*
6. Aplikasi adalah sebuah website
7. **Tinjauan Pustaka**
   1. Data-Driven Decision Making

*Data-Driven Decision Making* atau disingkat DDD, merupakan sebuah praktik pengambilan kesimpulang berdasarkan analisis data yang ada. Sebagai contoh seorang *marketer* dapat mengambil keputusan berdasarkan pengalaman ia bekerja atau dengan analisa data dari perilaku konsumen. (Provost & Fawcett, 2013)

* 1. Data yang digunakan

Data yang digunakan dalam penelitian ini sebagian besar berasal dari website http://www.kaggle.com.

1. West Nile Virus (https://www.kaggle.com/c/predict-west-nile-virus)

* Id: the id of the record
* Date: date that the WNV test is performed
* Address: approximate address of the location of trap. This is used to send to the GeoCoder.
* Species: the species of mosquitos
* Block: block number of address
* Street: street name
* Trap: Id of the trap
* AddressNumberAndStreet: approximate address returned from GeoCoder
* Latitude, Longitude: Latitude and Longitude returned from GeoCoder
* AddressAccuracy: accuracy returned from GeoCoder
* NumMosquitos: number of mosquitoes caught in this trap
* WnvPresent: whether West Nile Virus was present in these mosquitos. 1 means WNV is present, and 0 means not present.

1. House Price (https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* Alley: Type of alley access
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to main road or railroad
* Condition2: Proximity to main road or railroad (if a second is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Overall material and finish quality
* OverallCond: Overall condition rating
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 1stFlrSF: First Floor square feet
* 2ndFlrSF: Second floor square feet
* LowQualFinSF: Low quality finished square feet (all floors)
* GrLivArea: Above grade (ground) living area square feet
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* WoodDeckSF: Wood deck area in square feet
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* PoolQC: Pool quality
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

1. Human Resource (https://www.kaggle.com/ludobenistant/hr-analytics)

* Satisfaction Level
* Last evaluation
* Number of projects
* Average monthly hours
* Time spent at the company
* Whether they have had a work accident
* Whether they have had a promotion in the last 5 years
* Departments
* Salary
* Whether the employee has left

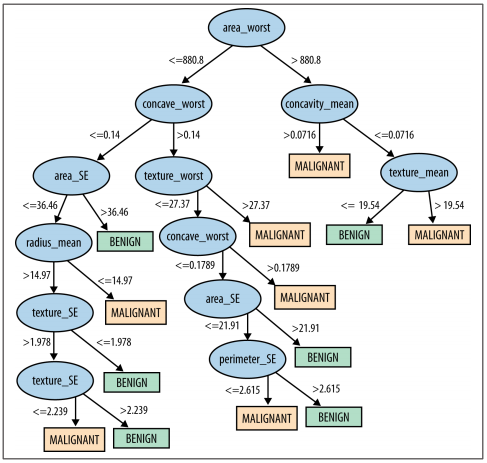
1. Insurance (https://www.kaggle.com/c/allstate-purchase-prediction-challenge/data)

* customer\_ID - A unique identifier for the customer
* shopping\_pt - Unique identifier for the shopping point of a given customer
* record\_type - 0=shopping point, 1=purchase point
* day - Day of the week (0-6, 0=Monday)
* time - Time of day (HH:MM)
* state - State where shopping point occurred
* location - Location ID where shopping point occurred
* group\_size - How many people will be covered under the policy (1, 2, 3 or 4)
* homeowner - Whether the customer owns a home or not (0=no, 1=yes)
* car\_age - Age of the customer’s car
* car\_value - How valuable was the customer’s car when new
* risk\_factor - An ordinal assessment of how risky the customer is (1, 2, 3, 4)
* age\_oldest - Age of the oldest person in customer's group
* age\_youngest - Age of the youngest person in customer’s group
* married\_couple - Does the customer group contain a married couple (0=no, 1=yes)
* C\_previous - What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4)
* duration\_previous - how long (in years) the customer was covered by their previous issuer
* A,B,C,D,E,F,G - the coverage options
* cost - cost of the quoted coverage options

1. Bakery (https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery)

* Goods: list of pastries and drinks on the menu
  + Item (Id of Item)
  + Flavor
  + Category (Type of Cake, e.g. Éclair, Cake)
  + Price
  + Type (Food or Drink)
* Location: list of bakery locations
  + City
  + State
  + Zip
  + Street
  + StoreNum (ID of Store)
* Receipts: list of sales (receipts)
  + ReceiptNumber
  + Date
  + Weekend (1 = yes, 0 = no)
  + Cash
  + Employee
  + StoreNum
* Items: list of purchased items (each receipt can have more than one)
  + ReceiptNumber
  + Qty
  + Item (ID of Item)
  1. Classification  
     Merupakan metode yang berusaha memprediksi, untuk setiap individual dalam populasi berada atau termasuk pada kelas yang telah ditentukan. Contoh dari permasalahan yang dapat dipecahkan dengan *classification* adalah menentukan apakah seorang akan menerima *offer* atau tawaran yang akan diberikan? (Provost & Fawcett, 2013)

Metode yang digunakan untuk *classification* adalah *decision tree, decision tree* merupakan sebuah metode yang melakukan segmentasi dan mencari indivudal termasuk dalam segmen yang mana. (Provost & Fawcett, 2013)

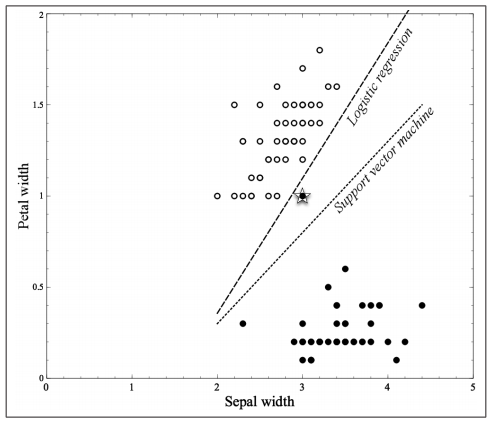


Gambar 6.1 Contoh Decision Tree (Provost & Fawcett, 2013)

Metode lainya yang digunakan untuk *classification* adalah Naïve Bayes. Naïve Bayes merupakan metode *classification* sederhana yang menghemat *memory* dan waktu komputasi. (Provost & Fawcett, 2013) Rumus Naïve bayes adalah:

P(c | **E**) =

Metode *classification* lainya adalah *Support Vector Machine* metode ini adalah metode yang digunakan untuk menambah atau memaksimalkan margin. Margin adalah jarak antara garis dengan kelompok data.



Gambar 6.2 Contoh Support Vector Machine (Provost & Fawcett, 2013)

* 1. Regression  
     *Regression* merupakan sebuah method yang mencoba untuk melakukan estimasi atau prediksi pada nilai *numerical* yang dimiliki individu tersebut. Contoh dari permasalahan yang dijawab oleh *Regression* adalah seberapa besar atau *how much* pelanggan yang akan menggunakan layanan yang disediakan? (Provost & Fawcett, 2013)

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

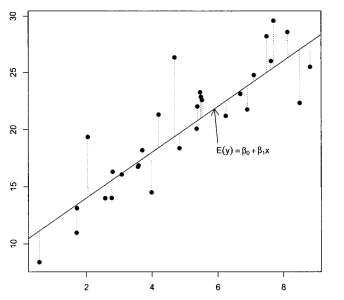
**Y2 = X21 β1 + X22 β2 + … + X2k βk + ε2**

**…**

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

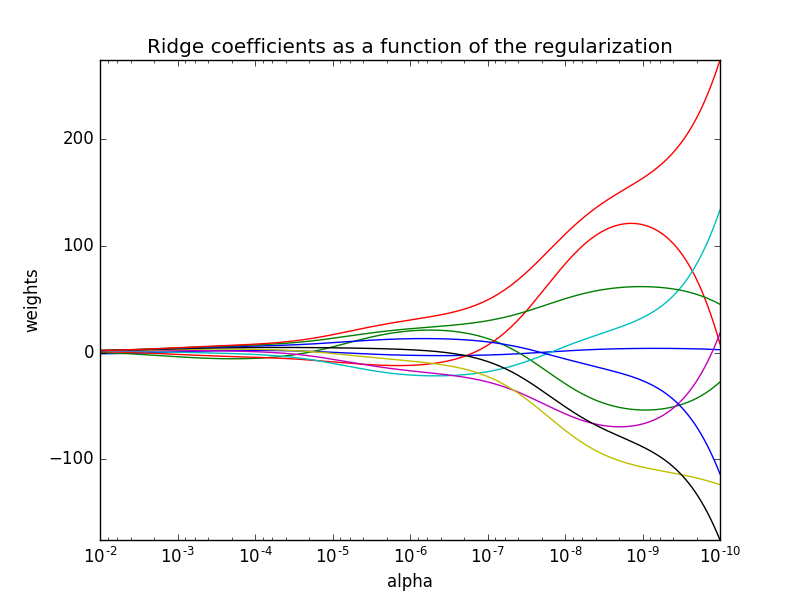
Rumus dasar *Regression Analysis*.

*Regression* merupakan metode yang melihat korelasi antara 1 variable yang disebut sebagai *dependent variable* (y-axis) dengan 1 atau lebih *independent variable* (x-axis)



Gambar 6.3 Contoh Linear Regression (Chatterjee & Simonoff, 2013)

Selain dengan metode *linear regression* diatas, metode regresi lainya yang digunakan adalah *Ridge Regression*. *Ridge Regression* merupakan metode yang melakukan regulasi pada *coefficient* rumus dasar regresi (β).



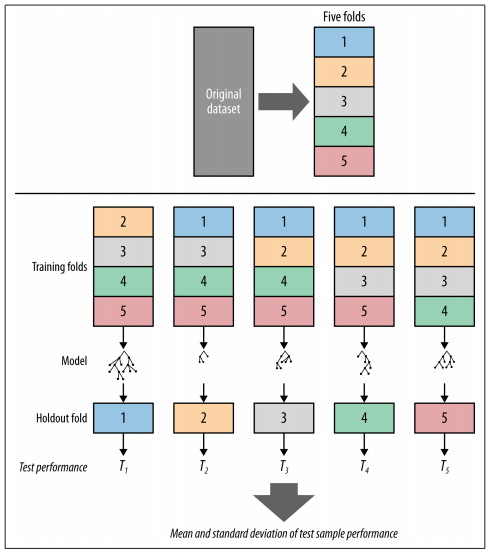
Gambar 6.1 Contoh Coefficient Ridge Regression (http://scikit-learn.org/stable/modules/linear\_model.html)

Metode regresi yang terakhir adalah *Lasso Regression*. Metode ini memperkirakan koefisien yang jarang. Hal ini berguna dalam beberapa konteks karena kecenderungannya untuk memilih solusi dengan nilai parameter yang lebih sedikit. *Lasso Regression*  memiliki keunggulan dimana metode ini dapat mengurangi dimensi yang digunakan regresi dan mengurangi *multicolinearity*.

* 1. Cross Validation

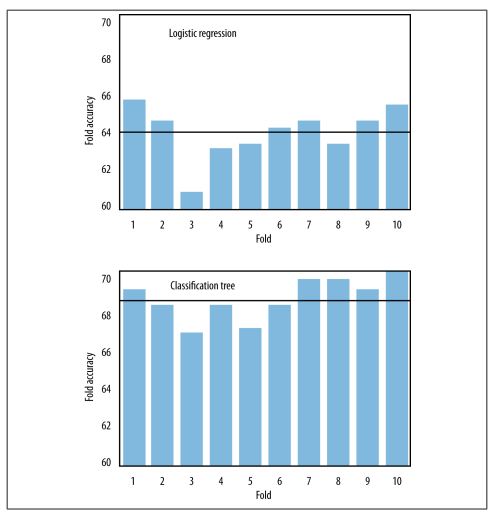
*Cross Validation* merupakan sebuah prosedur dalam melakukan *testing* untuk melihat apakah sebuah *model* sudah memiliki performa yang baik. *Cross Validation* menggunakan metode statistik dalam melakukan testing sehingga kita dapat mengetahui informasi lebih dari hasil testing terhadap model. (Provost & Fawcett, 2013)

Ilustrasi *Cross Validation*:



Gambar 6.3 Ilustrasi *Cross Validation* (Provost & Fawcett, 2013)

*Cross Validation* dilakukan dengan membagi *dataset* yang digunakan untuk membuat model menjadi n potongan / bagian. Pembagian berapa banyak potongan / bagian biasanya dilakukan diantara 5 – 10. Terdapat 1 potongan yang di sebut *holdout fold* dan n – 1 potongan disebut sebagai *training fold*. *Training fold* akan dijadikan model sedangkan *holdout fold*  akan dijadikan data *testing*.



Gambar 6.4 Hasil Plot *Cross Validatio*n menggunakan *Bar Chart* (Provost & Fawcett, 2013)

Setelah melakukan pembagian / pemotongan, setiap model dihitung seberapa besar *Standard Error* yang dihasilkan setiap *testing*. Setiap *Standart Error* akan divisualisasikan dengan *bar chart* seperti pada gambar di atas. Perhitungan statistik yang dapat dilakukan:

* Mean / Rata – rata, menenrukan tingkat tingginya akurasi yang dimiliki oleh model.
* Standart Deviasi, menentukan seberapa konsisten model dalam melakukan prediksi.

1. **Metodologi Penelitian**

Langkah – langkah dalam pembuatan skripsi:

1. Studi literatur mengenai:
   1. Python – Flask
   2. Cross Validation
2. Desain sistem untuk melakukan proses data beserta rekomendasi metode analisis*.*
3. Analisa Metode untuk menentukan metode yang digunakan dalam sistem rekomendasi
4. Membuat desain sistem untuk menyimpan konfigurasi aplikasi dan konfigurasi user.
5. Membuat tampilan / desain antarmuka.
6. Implementasi Sistem.
7. Evaluasi dan validasi model yang memiliki potensi dengan input data yang ada
8. Pengambilan Kesimpulan
   1. Rekomendasi yang diberikan memiliki *support* yang baik dan membantu pengambilan keputusan*.*
   2. Metode yang digunakan tidak menimbukan *overfitting*.
9. **Relevansi**

Manfaat dari kegiatan analisa ini adalah menghasilkan sebuah sistem rekomendasi untuk memberikan sugesti metode analisis apa yang sebaiknya digunakan untuk membantu kegiatan analisa data. Proses rekomendasi dapat berkembang dimana aplikasi dapat memikirkan sendiri metode apa yang paling baik dalam menyelesaikan masalah. Dengan adanya sistem rekomendasi ini kegiatan analisa menjadi lebih mudah dan tidak perlu melakukan banyak *testing* terhadap berbagai macam metode untuk mencari yang terbaik.

1. **Jadwal Kegiatan**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Kegiatan | Bulan 1 | | | | Bulan 2 | | | | Bulan 3 | | | | Bulan 4 | | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Desain Sistem |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing Method |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Analisa Sistem penyimpanan konfigurasi user |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementasi Sistem |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Pengujian & Analisa |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Pengambilan Kesimpulan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Pembuatan Laporan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. **Daftar Pustaka**

Chatterjee, S., & Simonoff, S. J. (2013). *Handbook of Regression Analysis.* Hoboken: John Wiley & Sons Inc.

Joris Toonders, Y. (2014, 07 01). *Data Is the New Oil of the Digital Economy*. Retrieved from Wired: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/

Kashmir, H. (2012, 2 16). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. Retrieved from Forbes: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#39701a176668

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big Data into Big Money.* New Jersey: John Wiley & Sons, Inc.

Provost, F., & Fawcett, T. (2013). *Data Science for Business.* Sebastopol: O'Reilly Media Inc.