

# **FINAL EXAM ANSWER**

Dibuat untuk memenuhi Ujian Akhir Semester

Mata Kuliah IS429 – Big Data Analytics



**Disusun oleh:**

Bintang Muhammad Ramdhan (00000082200)

**Dosen Pengampu:**

Iwan Prasetiawan, S. Kom., M.M.

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA  
TANGERANG  
2023/2024**

# Jurnal Perusahaan Tani: Optimasi Produktivitas dan Efisiensi melalui Big Data Analytics

## BAB 1. Latar Belakang



Tani adalah perusahaan pertanian yang fokus pada penelitian luas budidaya, volume keluaran, dan hasil berbagai tanaman, termasuk beras, gandum, dan millet. Untuk menciptakan metode yang mengoptimalkan produktivitas dan efisiensi, Tani menggunakan Big Data Analytics untuk menganalisis hasil panen mereka. Tani beroperasi di tingkat negara bagian dan distrik, dengan fokus pada berbagai cara untuk menangani informasi dan data regional. Perusahaan ini berupaya untuk mempromosikan keragaman pertanian melalui tanaman pokok dan komoditas seperti sorgum, kacang-kacangan, tebu, dan kapas, untuk mendorong pembangunan berkelanjutan dan memberdayakan petani dengan solusi inovatif.

Tani bertujuan untuk mendapatkan wawasan komprehensif tentang kumpulan data yang disediakan dari Kaggle.com yang mencakup informasi tentang produksi tanaman, luas budidaya, hasil panen, dan variabel relevan lainnya di berbagai negara bagian dan distrik.

### 1.1. Perumusan Masalah

Dikarenakan perusahaan ini menjual banyak komoditas, Tani sedang mencari peluang dalam supply beras dan gandum.

### 1.2. Tujuan Penelitian

Tujuan dari jurnal ini adalah mencari negara bagian terbaik yang memiliki produksi beras dan gandum terbaik.

### 1.3. Manfaat Penelitian

1. Menentukan negara bagian dengan potensi produksi tertinggi untuk beras dan gandum.
2. Memberikan wawasan bagi petani tentang praktik terbaik dan daerah yang paling produktif.

## BAB 2. Telaah Literatur

### 2.1 Forest

Algoritma Forest digunakan untuk membuat keputusan klasifikasi atau regresi dengan membangun banyak pohon keputusan dan menggabungkan hasilnya. Algoritma ini efektif dalam menangani data yang kompleks dan bervariasi.

### 2.2 Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi yang bekerja dengan mencari hyperplane yang memisahkan kelas data dengan margin maksimal. SVM sering digunakan untuk masalah klasifikasi dengan dimensi tinggi.

### 2.3 Linear Regression

Linear Regression digunakan untuk memprediksi nilai kontinu berdasarkan hubungan linear antara variabel independen dan dependen. Algoritma ini cocok untuk analisis prediktif dan inferensial.

### 2.4 Logistic Regression

Logistic Regression adalah model statistik untuk klasifikasi yang digunakan untuk memprediksi probabilitas kejadian dari suatu

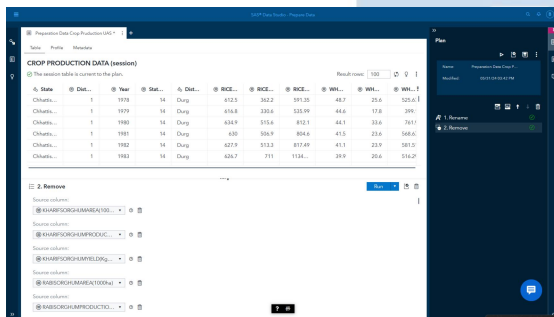
variabel biner. Model ini cocok untuk analisis data yang bersifat kategoris.

## BAB 3. Eksplorasi Dataset

### 3.1 Eksplorasi Dataset

Dataset ini diperoleh dari Kaggle.com dengan link [Crop Production Data \(kaggle.com\)](https://www.kaggle.com/datasets/rajatdas10/crop-production). Dataset mencakup informasi tentang produksi tanaman, luas budidaya, hasil panen, dan variabel relevan lainnya di berbagai negara bagian dan distrik di India.

### 3.2 Data Preparation

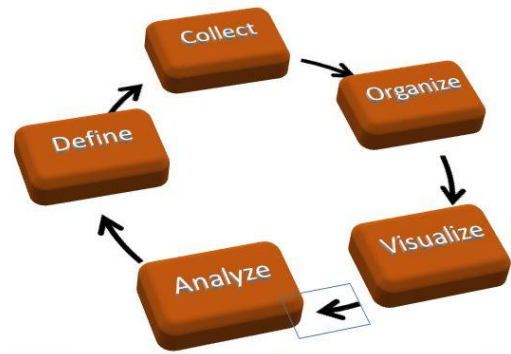


State	Year	Crop	Area	Yield	Production
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35
Chhattisgarh	1979	Dry	412.5	362.2	151.35

Sebelum data divisualisasikan, ada beberapa langkah persiapan data yang dilakukan untuk memastikan analisis yang akurat dan efisien:

1. **Penggantian Nama Kategori:** Kolom StateName diganti menjadi State untuk memudahkan pembacaan dan analisis data.
2. **Pemeriksaan Null Values:** Dataset ini tidak memiliki nilai kosong (null values), sehingga tidak perlu dilakukan penghapusan atau standardisasi data.
3. **Penghapusan Kolom yang Tidak Penting:** Kolom yang tidak relevan dengan tujuan penelitian dihapus, sehingga hanya kolom yang berhubungan dengan produksi beras dan gandum yang dipertahankan. Hal ini dilakukan untuk fokus pada analisis supply terbaik dalam rice dan wheat.

### 3.2 DCOVA & I



Metode DCOVA & I adalah pendekatan yang digunakan dalam analisis data untuk memastikan bahwa data yang dikumpulkan dan dianalisis dapat menghasilkan wawasan yang berguna dan dapat diimplementasikan. DCOVA & I terdiri dari beberapa langkah kunci: Define, Collect, Organize, Visualize, Analyze, dan Interpret. Berikut adalah penjelasan masing-masing langkah dalam konteks perusahaan Tani:

#### 1. Define (Menentukan)

Langkah pertama adalah menentukan tujuan dan masalah penelitian. Dalam konteks perusahaan Tani, tujuannya adalah untuk mengidentifikasi negara bagian terbaik yang memiliki produksi beras dan gandum terbaik. Tani ingin memahami faktor-faktor yang mempengaruhi produksi tanaman untuk mengoptimalkan produktivitas dan efisiensi.

- **Tujuan:** Mengidentifikasi negara bagian dengan produksi beras dan gandum terbaik.
- **Masalah:** Mencari peluang dalam supply beras dan gandum di berbagai negara bagian dan distrik.

#### 2. Collect (Mengumpulkan)

Langkah berikutnya adalah mengumpulkan data yang relevan. Perusahaan Tani menggunakan dataset yang tersedia di Kaggle.com yang mencakup informasi tentang produksi tanaman, luas budidaya, hasil panen, dan variabel relevan lainnya di berbagai negara bagian dan distrik.

- **Sumber Data:** Kaggle.com, dataset "[Crop Production Data \(kaggle.com\)](https://www.kaggle.com/datasets/rajatdasgupta/crop-production-data)".
- **Variabel yang Dikumpulkan:** Produksi beras, produksi gandum, luas area budidaya, hasil per hektar, tahun, negara bagian, dan distrik.

### 3. Organize (Mengorganisir)

Setelah data dikumpulkan, langkah selanjutnya adalah mengorganisir data tersebut agar mudah dianalisis. Dalam perusahaan Tani, data diatur berdasarkan negara bagian, distrik, dan tahun untuk memungkinkan analisis yang lebih terstruktur.

- **Pengorganisasian Data:** Data disusun dalam tabel dengan kolom untuk setiap variabel yang relevan.
- **Preprocessing:** Menghilangkan kolom yang tidak relevan, mengisi nilai yang hilang, dan memastikan konsistensi data.

### 4. Visualize (Memvisualisasikan)

Visualisasi data membantu dalam memahami pola dan tren dalam data. Perusahaan Tani menggunakan berbagai jenis grafik, seperti barchart dan pie chart, untuk memvisualisasikan produksi beras dan gandum di berbagai negara bagian.

- **Barchart Rice:** Menampilkan luas area penanaman dan produksi beras per negara bagian.
- **Barchart Wheat:** Menampilkan luas area penanaman dan produksi gandum per negara bagian.
- **Pie Chart:** Menampilkan perbandingan hasil produksi beras dan gandum per 1000-ton berdasarkan negara bagian.

### 5. Analyze (Menganalisis)

Langkah analisis melibatkan penerapan berbagai metode statistik dan algoritma untuk mengekstraksi wawasan dari data. Perusahaan Tani menggunakan beberapa

algoritma seperti Random Forest, Support Vector Machine (SVM), Linear Regression, dan Logistic Regression untuk menganalisis data produksi tanaman.

- **Forest Analysis:** Menganalisis produksi beras dan gandum menggunakan Random Forest untuk menentukan akurasi dan faktor penting.
- **SVM Analysis:** Menganalisis data produksi menggunakan SVM untuk klasifikasi dan prediksi.
- **Regression Analysis:** Menggunakan Linear dan Logistic Regression untuk memprediksi hasil produksi berdasarkan variabel yang relevan.

### 6. Interpret (Menginterpretasi)

Langkah terakhir adalah menginterpretasikan hasil analisis untuk mengambil keputusan yang berbasis data. Perusahaan Tani menginterpretasikan hasil dari berbagai model untuk menentukan negara bagian terbaik untuk produksi beras dan gandum, serta untuk mengembangkan strategi pertanian yang lebih efektif.

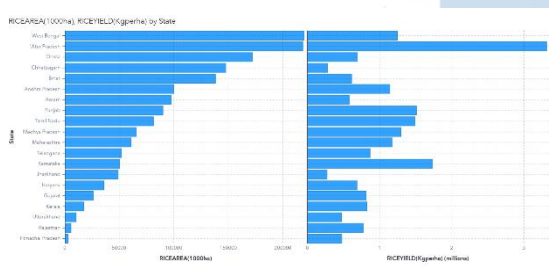
- **Interpretasi Hasil Forest:** Uttar Pradesh diidentifikasi sebagai negara bagian terbaik berdasarkan analisis Random Forest dengan nilai KS (Youden) yang tinggi.
- **Interpretasi Hasil SVM:** Menunjukkan bahwa Uttar Pradesh memiliki efisiensi produksi yang lebih baik dibandingkan dengan negara bagian lain.
- **Interpretasi Hasil Regression:** Menunjukkan bahwa Hissar memiliki nilai R-Square tertinggi, tetapi tidak memenuhi kriteria lain untuk analisis lebih lanjut.

### Integrasi dengan Perusahaan Tani

Dengan menerapkan metode DCOVA & I, perusahaan Tani dapat mengelola dan menganalisis data produksi tanaman secara

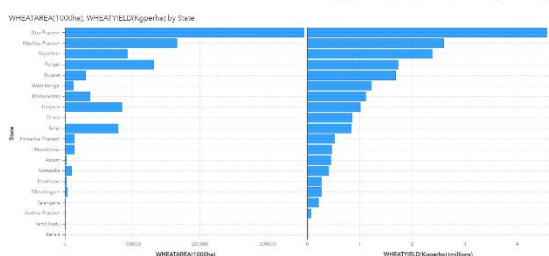
efektif. Metode ini membantu Tani dalam mengidentifikasi negara bagian yang paling produktif, memahami faktor-faktor yang mempengaruhi produksi, dan mengembangkan strategi untuk meningkatkan efisiensi dan produktivitas pertanian. Hasil analisis ini juga dapat digunakan untuk memberikan rekomendasi kepada petani mengenai praktik terbaik dan keputusan agronomi yang tepat.

### 3.3 Barchart Rice



Berdasarkan visualisasi ini, West Bengal memiliki area penanaman padi terbesar dengan 219.286 hektar, tetapi hanya memproduksi 1.247.390 kg per hektar. Sementara itu, Uttar Pradesh memiliki luas area penanaman padi kedua terbesar dengan 218.328 hektar, tetapi memproduksi 3.312.240 kg per hektar. Hal ini menunjukkan bahwa Uttar Pradesh memiliki efisiensi produksi yang lebih tinggi dibandingkan dengan West Bengal.

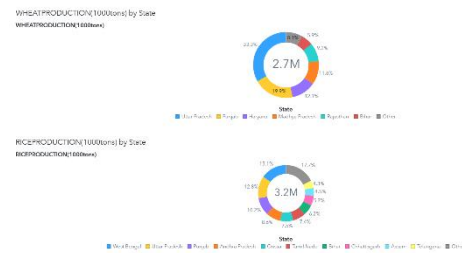
### 3.4 Barchart Wheat



Berdasarkan visualisasi ini, Uttar Pradesh menempati tempat tertinggi dalam area dan produksi gandum. Uttar Pradesh memiliki area penanaman gandum sebesar 353.548 hektar dengan produksi tertinggi mencapai 4.545.901 kg per hektar, menempati

peringkat tertinggi dibandingkan negara bagian lainnya.

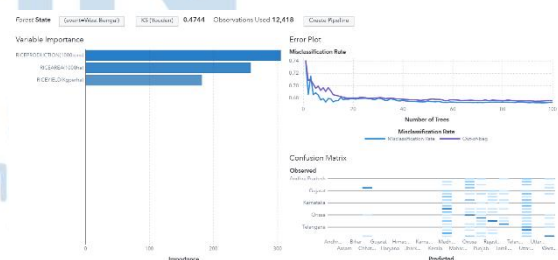
### 3.5 Perbandingan Rice dan Wheat



Berdasarkan visualisasi Pie chart ini, kita dapat melihat perbandingan antara hasil produksi beras per 1000-ton dan hasil produksi gandum per 1000-ton berdasarkan negara bagian. Visualisasi menunjukkan bahwa West Bengal memiliki hasil produksi beras terbesar dengan 15.1% dari total hasil produksi beras. Sedangkan Uttar Pradesh memiliki hasil produksi gandum terbesar dengan 33.3% dari total hasil produksi gandum. Hal ini menunjukkan bahwa Uttar Pradesh memiliki pertanian yang besar dan bervariasi dibandingkan dengan negara bagian lainnya.

## BAB 4. Implementasi dan Hasil Algoritma

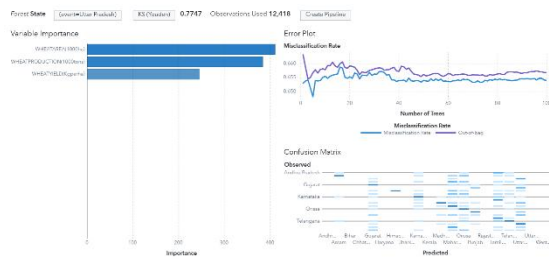
### 4.1 Hasil Forest Rice



Pemodelan Forest menggunakan data hasil produksi beras, luas area beras, dan produksi beras kg per hektar di negara bagian West Bengal memiliki nilai KS (Youden) akurasi 0.4744, yang menunjukkan tingkat akurasi yang rendah.

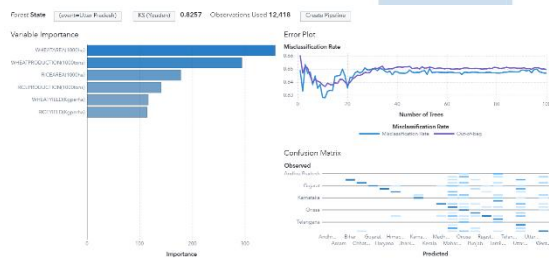
### 4.2 Hasil Forest Wheat





Pemodelan Forest dilakukan di negara bagian Uttar Pradesh, karena memiliki data tertinggi dalam area gandum, hasil produksi gandum, dan produksi gandum kg per hektar. Pemodelan ini menunjukkan hasil yang cukup tinggi dengan nilai KS (Youden) 0.7747.

#### 4.3 Hasil Forest Rice dan Wheat



Pemodelan menggunakan Forest dengan menggabungkan data gandum dan beras di negara bagian Uttar Pradesh menunjukkan bahwa negara bagian ini memiliki nilai KS (Youden) 0.8257, menunjukkan ketersediaan yang baik untuk kedua jenis tanaman tersebut.

#### 4.4 Hasil SVM Rice



Pemodelan SVM dengan data hasil produksi beras, luas area beras, dan produksi beras kg per hektar di negara bagian West Bengal memiliki nilai KS (Youden) akurasi 0.5549, yang artinya kurang akurat.

#### 4.5 Hasil SVM Wheat



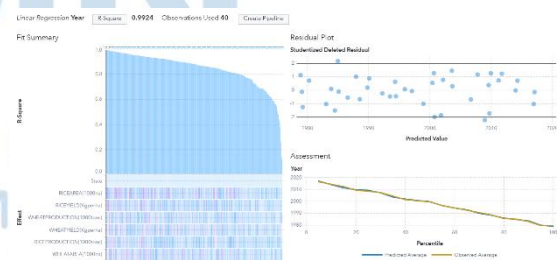
Pemodelan SVM dengan data hasil produksi gandum, luas area gandum, dan produksi gandum kg per hektar di negara bagian Uttar Pradesh memiliki nilai KS (Youden) akurasi 0.5891, yang juga menunjukkan akurasi yang kurang memuaskan.

#### 4.6 Hasil SVM Rice dan Wheat



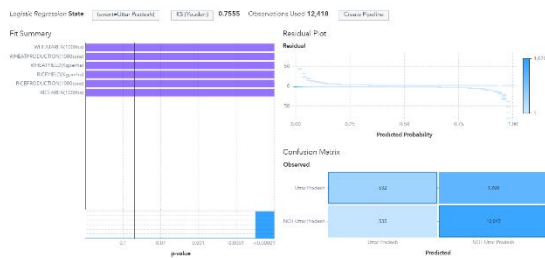
Pemodelan SVM dengan data beras dan gandum di negara bagian Uttar Pradesh memiliki akurasi yang lumayan dengan nilai KS (Youden) 0.7694.

#### 4.7 Hasil Linear Regression Rice dan Wheat



Menggunakan model Linear Regression, negara bagian Hissar memiliki nilai R-Square tertinggi dengan 0.9924 dari 40 observasi. Namun, negara bagian Hissar tidak memenuhi kriteria hasil produksi per ton untuk dijadikan patokan dalam analisis ini.

#### 4.8 Hasil Logistic Regression Rice dan Wheat



Menggunakan model Logistic Regression dengan data gandum dan beras yang difokuskan pada negara bagian Uttar Pradesh menunjukkan nilai KS (Youden) 0.7555, yang menunjukkan akurasi yang cukup tinggi.

#### 4.9 Hasil yang Dicapai

Berdasarkan analisis yang dilakukan, berikut adalah hasil yang dicapai dari penelitian ini:

##### 1. Visualisasi Data:

**Barchart Rice:** Visualisasi menunjukkan bahwa West Bengal memiliki luas area penanaman padi terbesar, tetapi Uttar Pradesh memiliki produksi per hektar yang lebih tinggi. Hal ini menunjukkan efisiensi produksi padi di Uttar Pradesh yang lebih baik dibandingkan West Bengal.

**Barchart Wheat:** Visualisasi menunjukkan bahwa Uttar Pradesh menempati tempat tertinggi dalam area dan produksi gandum, dengan hasil produksi tertinggi per hektar dibandingkan negara bagian lainnya.

**Perbandingan Rice dan Wheat:** Visualisasi Pie chart menunjukkan bahwa West Bengal memiliki hasil produksi beras terbesar sementara Uttar Pradesh memiliki hasil produksi gandum terbesar, menegaskan variasi dan keunggulan pertanian di Uttar Pradesh.

##### 2. Analisis Algoritma:

###### Forest Analysis:

- Hasil model Random Forest menunjukkan bahwa Uttar Pradesh memiliki nilai KS (Youden) tertinggi untuk produksi gandum, menunjukkan akurasi dan efisiensi yang tinggi dalam produksi gandum.
- Kombinasi data produksi beras dan gandum dengan algoritma Random Forest menunjukkan bahwa Uttar Pradesh memiliki nilai KS (Youden) tertinggi, mengindikasikan bahwa negara bagian ini memiliki ketersediaan dan efisiensi yang baik dalam produksi kedua tanaman tersebut.

###### SVM Analysis:

- Model SVM untuk produksi beras dan gandum menunjukkan bahwa Uttar Pradesh memiliki nilai akurasi yang lebih tinggi dibandingkan negara bagian lainnya, meskipun tidak setinggi model Random Forest.

###### Regression Analysis:

- Model Linear Regression menunjukkan bahwa Hissar memiliki nilai R-Square tertinggi, namun tidak memenuhi kriteria lain yang penting dalam analisis ini.
- Model Logistic Regression menunjukkan bahwa Uttar Pradesh memiliki nilai KS (Youden) yang lumayan tinggi, mendukung hasil dari model Random Forest.

## BAB 5. Kesimpulan

Dari penelitian ini, dapat disimpulkan bahwa negara bagian terbaik dengan produksi beras dan gandum terbaik adalah Uttar Pradesh. Negara bagian ini tidak hanya memiliki luas area penanaman yang signifikan, tetapi juga menunjukkan efisiensi produksi yang tinggi, baik untuk beras maupun gandum. Hasil analisis menggunakan berbagai algoritma (Random Forest, SVM, dan Regression) konsisten menunjukkan bahwa Uttar Pradesh

memiliki keunggulan dalam hal ketersediaan dan produktivitas tanaman.

Kesimpulan ini memberikan dasar bagi perusahaan Tani untuk fokus pada pengembangan strategi pertanian di Uttar Pradesh, mengoptimalkan produksi dan efisiensi, serta mengimplementasikan praktik agronomi yang dapat meningkatkan hasil panen secara keseluruhan. Dengan demikian, Tani dapat memberdayakan petani di negara bagian ini dengan solusi inovatif yang berdampak pada pembangunan berkelanjutan dan kesejahteraan komunitas agraris.

## Reference

- [1] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access*, 5, 16568-16575. <https://doi.org/10.1109/CSE-EUC.2017.99>.
- [2] Genuer, R., Poggi, J., Tuleau-Malot, C., & Villa-Vialaneix, N. (2015). Random Forests for Big Data. *Big Data Res.*, 9, 28-46. <https://doi.org/10.1016/j.bdr.2017.07.003>.
- [3] Muthusi, J., Mwalili, S., & Young, P. (2019). %svy\_logistic\_regression: A generic SAS macro for simple and multiple logistic regression and creating quality publication-ready tables using survey or non-survey data. *PLoS ONE*, 14. <https://doi.org/10.1371/journal.pone.0214262>.
- [4] Hossen, J., H, M., & Sayeed, S. (2018). Modifying Cleaning Method in Big Data Analytics Process using Random Forest Classifier. *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, 208-213. <https://doi.org/10.1109/ICCCE.2018.8539254>.
- [5] Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *J. Mach. Learn. Res.*, 15, 3133-3181. <https://doi.org/10.5555/2627435.2697065>.
- [6] Deshmukh, M., Ghate, A., Mathe, P., Dhote, A., Patte, P., & Mange, V. (2023). Diabetes Prediction using Machine Learning. *International Journal of Advanced Research in Science, Communication and Technology*. <https://doi.org/10.48175/ijarsct-9556>.
- [7] Dhanda, S., Yadav, A., Yadav, D., & Chauhan, B. (2022). Emerging Issues and Potential Opportunities in the Rice-Wheat Cropping System of North-Western India. *Frontiers in Plant Science*, 13. <https://doi.org/10.3389/fpls.2022.832683>.
- [8] Abousalh-Neto, N., & Kazgan, S. (2012). Big data exploration through visual analytics. , 285-286. <https://doi.org/10.1109/VAST.2012.6400514>.
- [9] Konasani, V., & Kadre, S. (2015). Practical Business Analytics Using SAS. . <https://doi.org/10.1007/978-1-4842-0043-8>.
- [10] Lebanon, G. (2010). Linear Regression. . <https://doi.org/10.12746/swrccc2014.0206.077>.

### Link Visualisasi SAS:

<https://v4e086.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2F432a41c5-47a1-4d6f-89ca-d8fbde7d0b45&page=vi2059>