**FINAL EXAM ANSWER**

Dibuat untuk memenuhi Ujian Akhir Semester

Mata Kuliah IS429 – Big Data Analytics

**Logo, company name

Description automatically generated**

**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**

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

## Managing Data

A screenshot of a computer

Description automatically generated

Pertama-tama dataset yang dipakai adalah Crop Production Data yang berasal dari Kaggle ([Crop Production Data (kaggle.com)](https://www.kaggle.com/datasets/zsinghrahulk/crop-production-data/data))

**A screenshot of a computer

Description automatically generated**

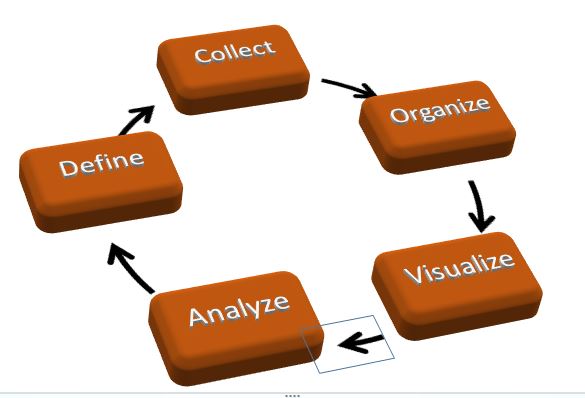
Diawal kita import data dari Kaggle yang sudah didownload untuk diupload kedalam SAS.

A screenshot of a computer

Description automatically generated

Setelah itu kita ke manage data untuk kita load ke memory, buka Crop Production Data yang sudah diload memorynya dan lakukan action prepare data biar langsung ke seksi prepare data didalam SAS.

## Exploratory Data Analysis (EDA)



Untuk metode DCOVA & I adalah pendekatan yang digunakan dalam analisis data dengan Langkah-langkah Define, Collect, Organize, Visualize, Analyze, dan Interpret. Pengimplementasian DCOVA & I dalam jurnal ini bisa dilihat dari aspek-aspek:

1. **Define (Menentukan)**

Langkah pertama adalah menentukan tujuan dan masalah penelitian.

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

1. **Collect (Mengumpulkan)**

Langkah berikutnya adalah mengumpulkan data yang relevan sesuai tujuan.

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

1. **Organize (Mengorganisir)**

Setelah data dikumpulkan, langkah selanjutnya adalah mengorganisir data tersebut agar mudah dianalisis

* **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.

1. **Visualize (Memvisualisasikan)**

Visualisasi data membantu dalam memahami pola dan tren dalam data.

* **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.

1. **Analyze (Menganalisis)**

Langkah analisis melibatkan penerapan berbagai metode statistik dan algoritma untuk mengekstraksi wawasan dari data.

* **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.

1. **Interpret (Menginterpretasi)**

Langkah terakhir adalah menginterpretasikan hasil analisis untuk mengambil keputusan yang berbasis data.

* **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.

**Barchart Rice**

A graph of a bar chart

Description automatically generated with medium confidence

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.

**Barchart Wheat**

A graph with blue and white bars

Description automatically generated

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.

**Perbandingan Rice dan Wheat**

A pie chart with numbers and a number

Description automatically generatedA circular chart with numbers and a number of text

Description automatically generated with medium confidence

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.

A screenshot of a computer

Description automatically generated

Didalam prepare data tidak banyak hal yang dilakukan dikarenakan dataset ini tidak memiliki null jadi datanya tidak perlu dibersihkan dan tidak ada yang perlu distandardize juga.

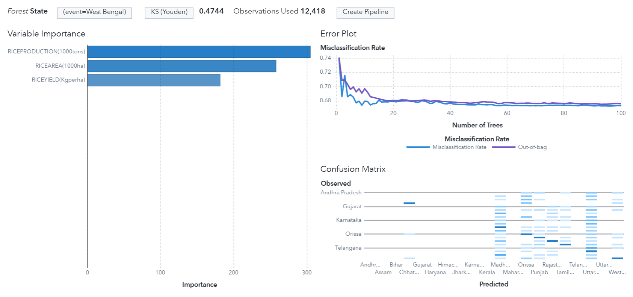
A screenshot of a computer

Description automatically generated

Prepare data yang dilakukan hanya penggantian nama “StateName” Menjadi “State” Agar lebih mudah dibaca dan penghapusan kolom yang tidak digunakan dikarenakan tujuan dari jurnal. Jadi hanya dipakai kolom *rice* dan *wheat* beserta kategori-kategorinya seperti luas area hektar, Kg per hektar dan per 1000-ton.

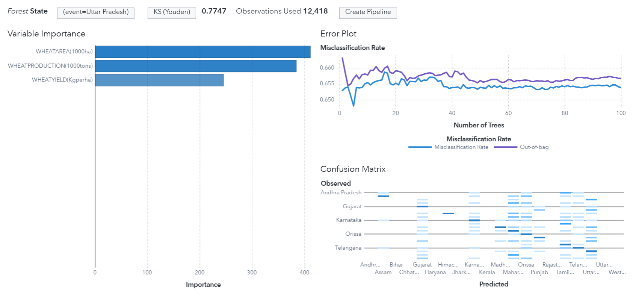
## Analytical Model

**Hasil Forest Rice**



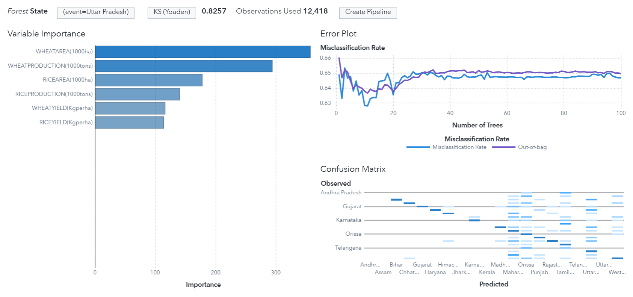
Pemodelan Forest menggunakan data hasil produksi beras, luas area beras, dan produksi beras kg per hektar di negara bagian West Bengal memiliki nilai KS (Youden) akurasi 0.4744, yang menunjukkan tingkat akurasi yang rendah.

**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.

**Hasil Forest Rice dan Wheat**



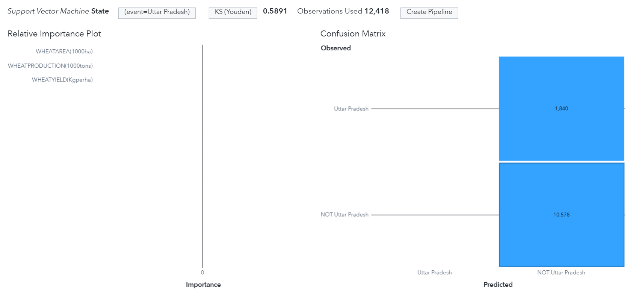
Pemodelan menggunakan Forest dengan menggabungkan data gandum dan beras di negara bagian Uttar Pradesh menunjukkan bahwa negara bagian ini memiliki nilai KS (Youden) 0.8257, menunjukkan ketersediaan yang baik untuk kedua jenis tanaman tersebut.

**Hasil SVM Rice**



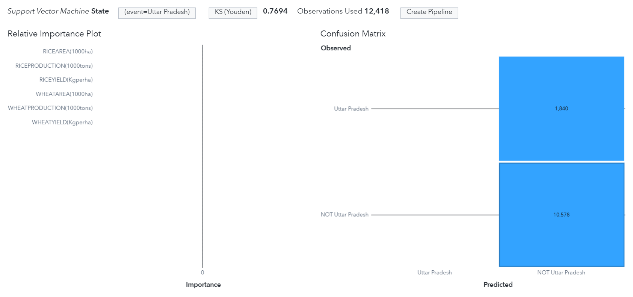
Pemodelan SVM dengan data hasil produksi beras, luas area beras, dan produksi beras kg per hektar di negara bagian West Bengal memiliki nilai KS (Youden) akurasi 0.5549, yang artinya kurang akurat.

**Hasil SVM Wheat**



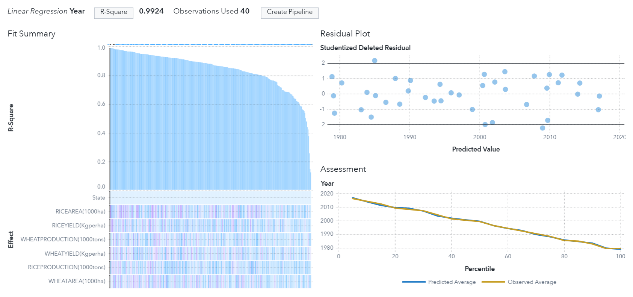
Pemodelan SVM dengan data hasil produksi gandum, luas area gandum, dan produksi gandum kg per hektar di negara bagian Uttar Pradesh memiliki nilai KS (Youden) akurasi 0.5891, yang juga menunjukkan akurasi yang kurang memuaskan.

**Hasil SVM Rice dan Wheat**



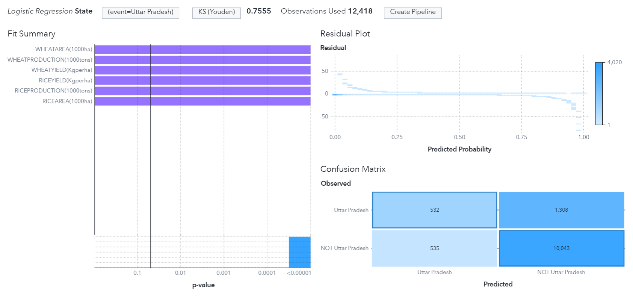
Pemodelan SVM dengan data beras dan gandum di negara bagian Uttar Pradesh memiliki akurasi yang lumayan dengan nilai KS (Youden) 0.7694.

**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.

**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.

## Perbandingan Seluruh Model

Dari hasil algoritma yang ada bisa disimpulkan lewat tabel bahwa:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritma | Produksi Beras – West Bengal | Produksi Gandung – Uttar Pradesh | Produksi Gabungan Beras dan Gandum - Uttar Pradesh | Mencari kriteria negara bagian terbaik |
| Forest | KS (Youden): 0.4744 | KS (Youden): 0.7747 | KS (Youden): 0.8257 | - |
| SVM | KS (Youden): 0.5549 | KS (Youden): 0.5891 | KS (Youden): 0.7694 | - |
| Linear Regression | - | - | - | R-Square: 0.9924 (Hissar) |
| Logistic Regression | - | - | KS (Youden): 0.7555 | - |

Berdasarkan perbandingan hasil di atas, dapat disimpulkan bahwa:

1. **Random Forest** adalah algoritma terbaik untuk analisis ini, khususnya dalam memprediksi efisiensi dan produksi gabungan beras dan gandum di Uttar Pradesh, dengan nilai KS (Youden) tertinggi (0.8257).
2. **SVM** memiliki akurasi yang baik tetapi lebih rendah dibandingkan dengan Random Forest, menunjukkan bahwa algoritma ini kurang tepat untuk data gabungan beras dan gandum.
3. **Linear Regression** menunjukkan hasil yang sangat baik untuk negara bagian Hissar dengan nilai R-Square tinggi, tetapi tidak sesuai dengan kriteria analisis produksi per ton.
4. **Logistic Regression** menunjukkan akurasi yang lumayan tinggi untuk data gabungan di Uttar Pradesh tetapi masih lebih rendah dibandingkan dengan Random Forest.