



***CLUSTERING* TENAGA KERJA SEKTOR PERTANIAN TERHADAP  
SOSIO-EKONOMI DAN EFISIENSI ENERGI PADA NEGARA ANGGOTA  
G20**

**LAPORAN KERJA PRAKTEK**

Diajukan sebagai Laporan Pelaksanaan Kerja Praktek Program  
Studi Informatika

Oleh :

Fransiscus Kristian Susanto

40621100012



**SK BADAN AKREDITASI NASIONAL PERGURUAN TINGGI (BAN-PT)**

**Nomor :2143/SK/BAN-PT /Ak-PPJ/S/III/2022**

**PROGRAM STUDI INFORMATIKA  
FAKULTAS TEKNIK UNIVERSITAS WIDYATAMA  
BANDUNG  
2025**

**LEMBAR PENGESAHAN**  
***CLUSTERING* TENAGA KERJA SEKTOR PERTANIAN TERHADAP**  
**SOSIO-EKONOMI DAN EFISIENSI ENERGI PADA NEGARA ANGGOTA**  
**G20**

Kerja Praktek

Program Studi Informatika

Fakultas Teknik

Universitas Widyatama

Oleh :

Fransiscus Kristian Susanto

40621100012

Telah disetujui dan disahkan di Bandung, Tanggal \_\_\_\_\_ 2025

Pembimbing Kampus,

Pembimbing Lapangan,

Dr. Feri Sulianta, S.T., M.T.

NIP. 1130814321

Fitri Hestikarani, S.Si.

NIP. -

## SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama : Fransiscus Kristian Susanto  
NPM : 40621100012  
Tempat dan Tanggal Lahir : Bandung, 21 Juni 2003

Menyatakan bahwa :

Judul : *CLUSTERING* TENAGA KERJA  
SEKTOR PERTANIAN TERHADAP  
SOSIO-EKONOMI DAN EFISIENSI  
ENERGI PADA NEGARA ANGGOTA  
G20  
Tempat Kerja Praktek : Startup Campus (Yayasan Bakti  
Achmad Zaky)

Merupakan hasil pekerjaan sendiri. Apabila terbukti Laporan Kerja Praktek ini bukan hasil saya sendiri, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian surat pernyataan ini saya buat dengan sebagaimana mestinya dan benar apa adanya.

Bandung, \_\_\_\_\_ 2025

Penulis,

Fransiscus Kristian Susanto

## ABSTRAK

Perkembangan data yang pesat di berbagai sektor kehidupan membuat penerapan *data science* menjadi sangat penting untuk menghasilkan wawasan dan solusi bagi berbagai permasalahan global. Penelitian ini bertujuan untuk menganalisis tenaga kerja sektor pertanian di negara-negara G20 dalam kaitannya dengan dampaknya terhadap sosio-ekonomi dan efisiensi energi untuk mendukung transisi menuju ekonomi hijau. Fokus utama dari penelitian ini adalah mengeksplorasi bagaimana sektor pertanian dapat beradaptasi dengan prinsip ekonomi hijau untuk memperkuat ketahanan pangan global, serta menarik minat generasi muda untuk terlibat dalam sektor ini. Penelitian ini menggunakan dataset dari OECD dan Bank Dunia dengan rentang waktu 2001 hingga 2023. Metode yang diterapkan dalam penelitian ini adalah teknik *clustering* menggunakan t-SNE dan K-Means untuk mengelompokkan negara-negara G20 berdasarkan ketergantungan mereka pada sektor pertanian dan efisiensi energi. Hasil analisis mengidentifikasi empat *cluster* utama, yang masing-masing mencerminkan negara dengan karakteristik tenaga kerja sektor pertanian dan efisiensi energi yang berbeda. Negara-negara dalam *Cluster 0* (Argentina, Brasil, Tiongkok, India, Indonesia, Meksiko, Rusia, Arab Saudi, Afrika Selatan, dan Turki), yang sangat bergantung pada sektor pertanian namun memiliki efisiensi energi rendah cenderung menghasilkan emisi karbon yang tinggi. Di sisi lain, negara-negara dalam *Cluster 1* (Australia, Kanada, dan Amerika Serikat), yang lebih maju dengan sektor pertanian kecil menunjukkan efisiensi energi yang lebih tinggi dan tingkat kemakmuran yang lebih baik. Beberapa saran dari analisis ini termasuk pengembangan teknologi ramah lingkungan di negara-negara yang sangat bergantung pada sektor pertanian, penerapan kebijakan keberlanjutan sektor pertanian di negara maju, serta membangun kerjasama internasional untuk berbagi teknologi dan pengetahuan dalam menciptakan sistem pertanian yang lebih efisien dan ramah lingkungan. Penelitian ini diharapkan dapat memberikan kontribusi signifikan bagi negara-negara G20 dalam merumuskan kebijakan yang lebih tepat dalam sektor pertanian dan energi untuk mendukung keberlanjutan global.

**Kata kunci:** *Data Science*, *Clustering*, Sektor Pertanian, Sosio-Ekonomi, Efisiensi Energi.

## ABSTRACT

*The rapid development of data in various sectors of life makes the application of data science very important to generate insights and solutions to various global problems. This study aims to analyze the agricultural sector workforce in G20 countries in relation to its socio-economic impact and energy efficiency to support the transition to a green economy. The main focus of this study is to explore how the agricultural sector can adapt to the principles of a green economy to strengthen global food security, as well as attract the interest of the younger generation to be involved in this sector. This study uses datasets from the OECD and the World Bank with a time span of 2001 to 2023. The method applied in this study is the clustering technique using t-SNE and K-Means to group G20 countries based on their dependence on the agricultural sector and energy efficiency. The results of the analysis identified four main clusters, each of which reflects countries with different characteristics of the agricultural sector workforce and energy efficiency. Countries in Cluster 0 (Argentina, Brazil, China, India, Indonesia, Mexico, Russia, Saudi Arabia, South Africa, and Türkiye), which are highly dependent on the agricultural sector but have low energy efficiency, tend to produce high carbon emissions. On the other hand, countries in Cluster 1 (Australia, Canada, and the United States), which are more developed with small agricultural sectors, show higher energy efficiency and better levels of prosperity. Some suggestions from this analysis include the development of environmentally friendly technologies in countries that are highly dependent on the agricultural sector, the implementation of agricultural sector sustainability policies in developed countries, and building international cooperation to share technology and knowledge in creating more efficient and environmentally friendly agricultural systems. This research is expected to provide significant contributions to G20 countries in formulating more appropriate policies in the agricultural and energy sectors to support global sustainability.*

**Keywords:** Data Science, Clustering, Agricultural Sector, Socio-Economic, Energy Efficiency.

## KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena dengan rahmat dan karuniaNya, penulis dapat menyelesaikan laporan kerja praktek yang berjudul “**Clustering Tenaga Kerja Sektor Pertanian Terhadap Sosio-Ekonomi Dan Efisiensi Energi Pada Negara Anggota G20**”. Laporan Kerja Praktek ini disusun dalam rangka memenuhi salah satu syarat untuk menyelesaikan mata kuliah Kerja Praktek pada Program Studi Informatika Fakultas Teknik Universitas Widyatama.

Dalam menyelesaikan laporan kerja praktek ini penulis mendapatkan banyak bantuan, masukan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Orang tua dan kakak tercinta yang memberikan perhatian dan kasih sayang kepada penulis.
2. Dr. Feri Sulianta, S.T., M.T. selaku dosen pembimbing yang telah memberi masukan dan waktunya dalam membimbing penulis.
3. Fitri Hestikarani, S.Si. selaku pembimbing lapangan yang telah membantu penulis dalam pengerjaan kerja praktek.
4. Semua pihak yang telah memberikan bantuan dan motivasi kepada penulis baik secara langsung maupun tidak langsung.

Penulis menyadari bahwa penyusunan laporan kerja praktek ini masih banyak kekurangan sehingga kritik dan saran yang membangun sangat membantu penulis guna perbaikan laporan kerja praktek ini. Penulis berharap laporan kerja praktek ini dapat bermanfaat bagi semua pihak.

Bandung, \_\_\_\_\_ 2025

Fransiscus Kristian Susanto  
40621100012

## DAFTAR ISI

<b>LEMBAR PENGESAHAN .....</b>	<b>1</b>
<b>SURAT PERNYATAAN .....</b>	<b>2</b>
<b>ABSTRAK .....</b>	<b>3</b>
<b>ABSTRACT .....</b>	<b>4</b>
<b>KATA PENGANTAR.....</b>	<b>5</b>
<b>DAFTAR ISI.....</b>	<b>6</b>
<b>DAFTAR GAMBAR.....</b>	<b>9</b>
<b>DAFTAR TABEL .....</b>	<b>10</b>
<b>DAFTAR LAMPIRAN .....</b>	<b>11</b>
<b>BAB I PENDAHULUAN.....</b>	<b>12</b>
1.1 Latar Belakang .....	12
1.2 Rumusan Masalah .....	12
1.3 Batasan Masalah.....	13
1.4 Tujuan Penelitian.....	13
1.5 Manfaat Penelitian.....	14
1.6 Metode Penelitian dan Metode Pengembangan Platform .....	14
1.6.1 Metode Penelitian Kuantitatif dan Pengumpulan Data .....	14
1.6.2 <i>Data Preparation</i> .....	15
1.6.3 <i>Modeling</i> .....	15
1.6.4 <i>Evaluation</i> .....	15
1.7 Sistematika Penulisan.....	16
<b>BAB II KEADAAN UMUM PERUSAHAAN.....</b>	<b>18</b>
2.1 Deskripsi Program Studi Independen Bersertifikat.....	18
2.2 Profil Instansi .....	18
2.3 Struktur Organisasi.....	19
2.4 Deskripsi Kerja dan Uraian Penugasan .....	19
2.4.1 <i>Foundation for Data Science</i> .....	19
2.4.2 <i>Data Preprocessing</i> .....	20
2.4.3 <i>Machine Learning</i> .....	21
2.4.4 <i>Data Visualization</i> .....	22

2.4.5 <i>Final Project</i> .....	22
<b>BAB III LANDASAN TEORI</b> .....	<b>24</b>
3.1 <i>Tenaga Kerja Pertanian</i> .....	24
3.2 <i>Group of 20 (G20)</i> .....	25
3.3 <i>Data Science</i> .....	26
3.4 <i>Data Mining</i> .....	29
3.5 <i>Cross-Industry Standard Process for Data Mining (CRISP-DM)</i> .....	30
3.6 <i>Feature Scaling</i> .....	32
3.7 <i>Machine Learning</i> .....	33
3.8 <i>Clustering</i> .....	35
3.9 <i>K-Means Clustering</i> .....	36
3.10 <i>Agglomerative Clustering</i> .....	37
3.11 <i>Reduksi Dimensionalitas</i> .....	39
3.12 <i>Python</i> .....	43
3.13 <i>Google Colaboratory</i> .....	44
<b>BAB IV ANALISIS DAN PERANCANGAN</b> .....	<b>45</b>
4.1 <i>Alur Kerja</i> .....	45
4.1.1 <i>Business Understanding</i> .....	45
4.1.2 <i>Data Understanding</i> .....	46
4.1.3 <i>Data Preparation</i> .....	46
4.1.4 <i>Modeling dan Evalution</i> .....	47
4.2 <i>Analisis Dataset</i> .....	48
4.3 <i>Platform Google Colaboratory</i> .....	51
<b>BAB V IMPLEMENTASI DAN PENGUJIAN</b> .....	<b>52</b>
5.1 <i>Kebutuhan Platform</i> .....	52
5.1.1 <i>Kebutuhan Perangkat Keras</i> .....	52
5.1.2 <i>Kebutuhan Perangkat Lunak</i> .....	52
5.2 <i>Implementasi Platform</i> .....	53
5.2.1 <i>Business Understanding</i> .....	53
5.2.2 <i>Data Understanding</i> .....	53
5.2.3 <i>Data Preparation</i> .....	54
5.2.4 <i>Modeling dan Evaluation</i> .....	60



5.3	Pengujian Platform.....	66
5.3.1	Perbandingan Algoritma <i>Clustering</i> .....	66
5.3.2	Perbandingan Teknik Reduksi Dimensi .....	68
5.3.3	Visualisasi Hasil <i>Clustering</i> .....	69
<b>BAB VI PENUTUP .....</b>		<b>71</b>
6.1	Kesimpulan.....	71
6.2	Saran.....	72
<b>DAFTAR PUSTAKA .....</b>		<b>74</b>
<b>LAMPIRAN .....</b>		<b>76</b>

## DAFTAR GAMBAR

<b>Gambar 2.1</b> Struktur Organisasi Startup Campus.....	19
<b>Gambar 3.1</b> Ilmu Data dalam Konteks Proses.....	28
<b>Gambar 4.1</b> Diagram Kotak <i>Data Preparation</i> .....	46
<b>Gambar 4.2</b> Diagram Kotak <i>Modeling</i> dan <i>Evaluation</i> .....	47
<b>Gambar 5.1</b> Impor <i>Dataset</i> OECD .....	53
<b>Gambar 5.2</b> Impor <i>Dataset</i> Bank Dunia .....	54
<b>Gambar 5.3</b> <i>Feature Selection</i> Data OECD .....	54
<b>Gambar 5.4</b> <i>Feature Selection</i> Data Bank Dunia.....	55
<b>Gambar 5.5</b> Data Jadi pada <i>Dataset</i> OECD.....	55
<b>Gambar 5.6</b> Data Jadi pada <i>Dataset</i> Bank Dunia .....	56
<b>Gambar 5.7</b> Memastikan Nama Negara pada Dua <i>Dataset</i> .....	56
<b>Gambar 5.8</b> Menggabungkan Dua <i>Dataset</i> .....	57
<b>Gambar 5.9</b> Memperbaiki <i>Missing Values</i> .....	57
<b>Gambar 5.10</b> <i>Data Filtering</i> .....	58
<b>Gambar 5.11</b> <i>Statistical Summaries</i> pada EDA .....	58
<b>Gambar 5.12</b> <i>Bivariate Analysis</i> .....	59
<b>Gambar 5.13</b> <i>Multivariate Analysis</i> .....	60
<b>Gambar 5.14</b> <i>Feature Selection</i> .....	60
<b>Gambar 5.15</b> <i>Feature Scaling</i> .....	61
<b>Gambar 5.16</b> <i>Dimensionality Reduction</i> dan <i>Clustering Model Selection</i> .....	62
<b>Gambar 5.17</b> Iterasi Model <i>Clustering</i> .....	62
<b>Gambar 5.18</b> Pemilihan Teknik Reduksi Dimensi dan Model Terbaik .....	63
<b>Gambar 5.19</b> Hasil Iterasi Model <i>Clustering</i> .....	64
<b>Gambar 5.20</b> Pemilihan Hasil Evaluasi Model Terbaik.....	65
<b>Gambar 5.21</b> Statistik Deskriptif per <i>Cluster</i> .....	65
<b>Gambar 5.22</b> Pengelompokan Negara berdasarkan <i>Cluster</i> .....	66
<b>Gambar 5.23</b> Teknik Reduksi Dimensi PCA .....	68
<b>Gambar 5.24</b> Teknik Reduksi Dimensi t-SNE.....	69
<b>Gambar 5.25</b> Visualisasi Hasil <i>Cluster</i> .....	69

## DAFTAR TABEL

<b>Tabel 3.1</b> Metode <i>Agglomerative Hierarchical Clustering</i> Standar .....	38
<b>Tabel 4.1</b> Fitur Dataset OECD.....	49
<b>Tabel 4.2</b> Fitur Dataset Bank Dunia .....	50
<b>Tabel 5.1</b> Kebutuhan Perangkat Keras.....	52
<b>Tabel 5.2</b> Kebutuhan Perangkat Lunak.....	52
<b>Tabel 5.3</b> Algoritma <i>Clustering</i> dengan Teknik Reduksi PCA .....	66
<b>Tabel 5.4</b> Algoritma <i>Clustering</i> dengan Teknik Reduksi t-SNE.....	67

## **DAFTAR LAMPIRAN**

<b>Lampiran 1 Kartu Bimbingan Kerja Praktek .....</b>	<b>76</b>
<b>Lampiran 2 Kuesioner Kerja Praktek .....</b>	<b>77</b>

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Perkembangan data saat ini mengalami perkembangan yang pesat. Miliaran data dihasilkan di berbagai sektor kehidupan dalam setiap hari. Untuk mengolah data tersebut, dibutuhkan penerapan *data science* agar dapat memperoleh wawasan dan solusi untuk mengatasi berbagai masalah. Sebagai peserta program yang berfokus pada pengembangan keterampilan digital, diharapkan peserta mampu menyelesaikan masalah dalam proyek akhir seperti seorang *data scientist*.

Peserta diminta untuk membuat proyek akhir dengan nilai bisnis terkait *green economy* dan *sustainability*, mulai dari pemilihan *study case* hingga narasi dengan *dataset* yang diberikan. Dalam *final project* ini dipilih topik *clustering* terkait tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara anggota G20.

Tujuan utama dari penelitian ini adalah untuk memahami kontribusi sektor pertanian dalam mendukung transisi menuju ekonomi hijau di negara anggota G20, serta menyusun rekomendasi kebijakan strategis guna memastikan sektor ini berkelanjutan dan berkontribusi pada ketahanan pangan global. Berdasarkan tantangan regenerasi petani, terutama dalam menghadapi urbanisasi dan transformasi ekonomi, analisis ini mencoba menggali bagaimana sektor pertanian dapat lebih menarik minat generasi muda, yang sejalan dengan prinsip-prinsip ekonomi hijau.

### 1.2 Rumusan Masalah

Adapun rumusan masalah dalam kerja praktek ini adalah sebagai berikut.

1. Bagaimana cara melakukan *data preparation* pada *dataset* tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara anggota G20?
2. Bagaimana cara merancang *modeling* pada *dataset* tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara anggota G20?

3. Bagaimana wawasan yang dapat diambil pada *dataset* tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara anggota G20?

### 1.3 Batasan Masalah

Adapun batasan masalah dalam kerja praktek ini adalah sebagai berikut.

1. *Dataset* utama yang digunakan bersumber pada *dataset* OECD dan *dataset* tambahan yang bersumber pada *dataset* Bank Dunia.
2. Data diambil dengan memilih fitur-fitur *dataset* yang relevan dengan sektor pertanian, serta mencakup rentang waktu dari 2001 hingga 2022.
3. Teknik *modeling* yang dipilih ialah *clustering* di mana teknik ini digunakan untuk mengelompokkan data ke dalam kelompok atau *cluster* berdasarkan kesamaan karakteristik tertentu.
4. Bahasa pemrograman yang digunakan adalah Python dengan menggunakan sebuah platform *notebook* berbasis *cloud* yang bernama *Google Colaboratory*.
5. Proses standar untuk *data mining* akan menggunakan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sampai tahap *Evaluation*.

### 1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijabarkan, adapun tujuan penelitian adalah sebagai berikut.

1. Meningkatkan kontribusi sektor pertanian dalam mendukung transisi menuju ekonomi hijau di negara anggota G20.
2. Mengembangkan rekomendasi kebijakan strategis yang memastikan sektor pertanian berkontribusi pada ketahanan pangan global.
3. Meningkatkan peluang dalam tantangan sektor pertanian melalui pendekatan sosial, ekonomi, dan efisiensi energi sejalan dengan prinsip ekonomi hijau.

## 1.5 Manfaat Penelitian

Adapun manfaat dalam penelitian ini adalah sebagai berikut.

1. Memberikan rekomendasi konkret mengenai cara mengatasi masalah regenerasi petani serta menarik minat generasi muda untuk terlibat dalam sektor pertanian.
2. Meningkatkan pemahaman masyarakat mengenai pentingnya peran sektor pertanian dalam mencapai ekonomi hijau dan ketahanan pangan global.
3. Meningkatkan pemahaman peserta dalam memanfaatkan data besar untuk pengambilan keputusan yang lebih baik dalam konteks ekonomi dan keberlanjutan sektor pertanian.

## 1.6 Metode Penelitian dan Metode Pengembangan Platform

Dalam melakukan penelitian ini, adapun beberapa metode penelitian, pengumpulan data, dan metode pengembangan *notebook* yang digunakan penulis dalam mengembangkan *clustering* tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara anggota G20 ini sebagai berikut.

### 1.6.1 Metode Penelitian Kuantitatif dan Pengumpulan Data

Penelitian ini menggunakan metode kuantitatif, di mana pendekatan ini memungkinkan peneliti untuk mengubah kompleksitas dunia nyata menjadi angka-angka yang dapat dianalisis, memberikan peluang untuk pengembangan pengetahuan dan pemecahan masalah. Dengan metode ini, hubungan antarvariabel dapat dieksplorasi, pola-pola dapat diidentifikasi, dan generalisasi yang kuat dapat dibuat untuk mendukung temuan penelitian [1].

Tahap awal adalah mengumpulkan data pada *dataset* berikut.

1. *Dataset* utama yang bersumber pada data OECD pada tautan <https://data-explorer.oecd.org>.
2. *Dataset* tambahan yang bersumber pada data Bank Dunia pada tautan <https://databank.worldbank.org>.

Dalam *dataset* utama dibebaskan untuk menentukan “*Reference Area*” setidaknya satu negara untuk dianalisis dan diperbolehkan untuk mencari sumber pendukung, seperti dari BPS dan lembaga/badan negara dengan catatan wajib memberikan *credit/reference*.

### **1.6.2 Data Preparation**

Data yang telah dikumpulkan akan melalui tahapan pembersihan data dan eksplorasi data (EDA). Pada tahap ini dilakukan penanganan terhadap *missing values*, duplikat, dan *outliers* untuk memastikan kualitas data. Selain itu, dilakukan pemilihan fitur yang relevan (*feature selection*) agar analisis lebih fokus pada atribut yang signifikan sesuai dengan kondisi *dataset*.

### **1.6.3 Modeling**

Dalam tahap ini, *modeling* menggunakan teknik *clustering* yang ditujukan untuk mengelompokkan data ke dalam kelompok atau *cluster* berdasarkan kesamaan mereka. *Clustering* dapat membantu dalam mengidentifikasi pola yang mungkin tidak terlihat secara langsung dan dapat digunakan untuk membuat segmen data berdasarkan perilaku atau karakteristik tertentu.

### **1.6.4 Evaluation**

Langkah selanjutnya adalah menyediakan *evaluation metric* dan nilai performa yang sesuai untuk setiap model yang bertujuan untuk memastikan bahwa kinerja model sesuai dengan harapan dan menghasilkan hasil yang akurat.



## **1.7 Sistematika Penulisan**

Sistematika penulisan yang akan digunakan dalam penulisan laporan kerja praktek ini adalah sebagai berikut.

### **Bab I Pendahuluan**

Dalam bab ini dijelaskan mengenai latar belakang, rumusan masalah, basatasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian, pengumpulan data, metode pengembangan aplikasi dan sistematika penulisan dalam laporan kerja praktek pembuatan *clustering* tenaga kerja sektor pertanian terhadap sosio-ekonomi dan efisiensi energi pada negara G20.

### **Bab II Keadaan Umum Perusahaan**

Bab ini memaparkan profil umum dari Startup Campus khususnya pada program *Data Science: Greener Future with Data Driven Solution* di mana penulis melaksanakan kerja praktek.

### **Bab III Landasan Teori**

Berisi penjelasan mengenai teori-teori yang digunakan pada penelitian dan hal-hal yang berkaitan dengan perancangan *Clustering* Tenaga Kerja Sektor Pertanian Terhadap Sosio-Ekonomi Dan Efisiensi Energi Pada Negara Anggota G20.

### **Bab IV Analisis dan Perancangan**

Analisis dan Perancangan pada bab ini membahas mengenai kebutuhan sistem dan pengembangan kode dalam *Clustering* Tenaga Kerja Sektor Pertanian Terhadap Sosio-Ekonomi Dan Efisiensi Energi Pada Negara Anggota G20.

### **Bab V Pengujian dan Implementasi**

Pengujian dan Implementasi pada bab ini menjelaskan skenario dalam mengeksekusi *notebook* dalam *Clustering* Tenaga Kerja Sektor Pertanian Terhadap Sosio-Ekonomi Dan Efisiensi Energi Pada Negara Anggota G20.

## **Bab VI Penutup**

Pada bab ini menjelaskan terkait kesimpulan dan saran dari keseluruhan perancangan dan pengembangan *Clustering* Tenaga Kerja Sektor Pertanian Terhadap Sosio-Ekonomi Dan Efisiensi Energi Pada Negara Anggota G20 yang dibuat.

## BAB II

### KEADAAN UMUM PERUSAHAAN

#### 2.1 Deskripsi Program Studi Independen Bersertifikat

Startup Campus *Batch 7 - Data Science: Greener Future with Data Driven Solution* merupakan Studi Independen Bersertifikat yang diselenggarakan oleh Yayasan Bakti Achmad Zaky dan didukung penuh oleh Kemdikbud Ristekdikti di bawah naungan inisiatif KAMPUS MERDEKA. Batch 7 diselenggarakan dari September hingga Desember 2024. Program ini dirancang sepenuhnya secara daring untuk mempersiapkan talenta digital - *Data Science: Greener Future with Data Driven Solution* yang mampu bersaing di era digital.

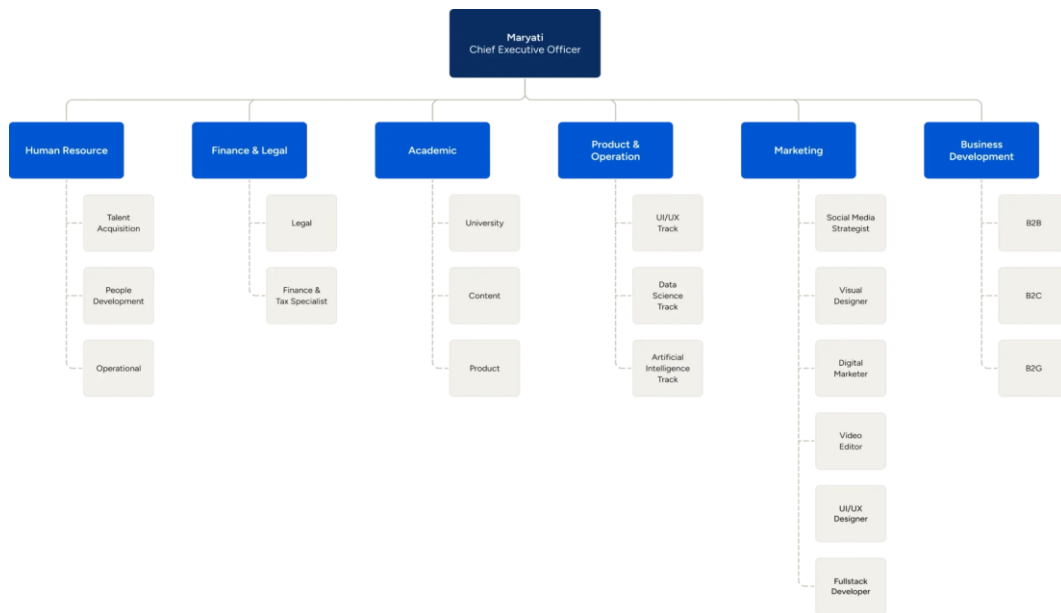
*Data Science: Greener Future with Data-Driven Solutions* juga berfokus pada keberlanjutan (*Solutions For A Sustainable Future*). *Sustainability* (keberlanjutan) adalah konsep yang menekankan pentingnya menjaga dan melestarikan keseimbangan antara kebutuhan manusia, lingkungan, dan ekonomi untuk generasi sekarang dan mendatang. Sedangkan *Green Economy* (Ekonomi Hijau) adalah konsep pembangunan ekonomi yang berkelanjutan dan ramah lingkungan. Yang termasuk ekonomi hijau antara lain *Agriculture, Renewable Energy, Forestry, Tourism, Transportation, Waste, Water Scarcity, Climate Change*, dan sebagainya.

#### 2.2 Profil Instansi

Nama Mitra	: Startup Campus (Yayasan Bakti Achmad Zaky)
Nama Program	: Data Science Greener Future with Data Driven Solution
Penanggung Jawab	: Fitri Hestikarani, S.Si.
Alamat Korespondensi	: Jl. Sunan Giri No.1, Rawamangun, Kec. Pulo Gadung, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta - 13220
Durasi Program	: 6 September - 31 Desember 2024

## 2.3 Struktur Organisasi

Struktur organisasi adalah sebuah sistem penugasan yang bersifat formal, yang menggambarkan pembagian tugas dan tanggung jawab setiap anggota perusahaan, serta hubungan antara pihak-pihak dalam organisasi yang bekerja sama untuk mencapai tujuan bersama. Struktur organisasi dalam pelaksanaan program Startup Campus adalah sebagai berikut.



**Gambar 2.1** Struktur Organisasi Startup Campus

## 2.4 Deskripsi Kerja dan Uraian Penugasan

Selama menjalani program ini, peserta menyelesaikan beberapa tugas berdasarkan materi-materi yang diberikan melalui Learning Management System (LMS). Tugas ini terdiri dari tugas individu dan tugas tim. Lingkup pembelajaran dan proyek meliputi: *Foundation for Data Science*, *Data Preprocessing*, *Machine Learning*, *Data Visualization*, dan *Final Project*.

### 2.4.1 *Foundation for Data Science*

Pada *chapter Foundation for Data Science*, peserta akan dibekali dengan keterampilan dasar yang sangat penting bagi seorang *data scientist*. Peserta akan mengenal konsep dasar dalam *data science* serta mengaplikasikan bahasa pemrograman SQL untuk mengolah data dan Python sebagai dasar dalam *data science*. Materi yang akan dipelajari mencakup

*Introduction to Data Science, SQL Foundation, Basic Python Foundation, dan Advanced Python Foundation.* Dalam sesi ini, peserta akan melakukan banyak kegiatan praktis menggunakan aplikasi seperti *Google Colab* dan *Google BigQuery*. Pembelajaran ini dilengkapi dengan materi asinkronus berupa modul, video pembelajaran, latihan, kuis, dan referensi lainnya yang dapat diakses di *Learning Management System*. Selain itu, sesi sinkronus yang meliputi pemaparan materi, diskusi, dan tanya jawab di kelas (*live session*) serta *mentoring* juga turut memperkaya pengalaman pembelajaran. Selama proses pembelajaran, peserta akan mengerjakan berbagai tugas dan *assignment* untuk mengasah kemampuan analisis dan evaluasi mengenai pondasi utama dalam *data science*. Dengan total durasi 180 jam, peserta akan memperoleh akses ke berbagai sumber daya, termasuk modul dari Startup Campus, rekaman video pembelajaran dari para ahli, *mentoring*, serta video dan referensi lainnya. Metode penilaian akan dilakukan melalui tugas individu dan/atau kelompok, keaktifan dalam kelas, serta kuis.

#### **2.4.2 Data Preprocessing**

Pada *chapter Data Preprocessing*, peserta akan mempelajari tahapan-tahapan penting dalam mengolah data mentah menjadi data yang siap digunakan untuk analisis lebih lanjut. Materi yang akan dipelajari mencakup persiapan data, mulai dari transformasi, pembersihan, penggabungan data, hingga statistik untuk *data science*, *Exploratory Data Analysis* (EDA), dan *Feature Engineering* menggunakan *Google Colab*. Peserta akan diajarkan cara mengolah data agar menjadi kaya akan fitur-fitur yang berguna bagi seorang *data scientist*. Pembelajaran ini dilengkapi dengan materi asinkronus berupa modul, video pembelajaran, latihan, kuis, dan referensi lainnya yang dapat diakses melalui *Learning Management System*. Selain itu, sesi sinkronus yang mencakup pemaparan materi, diskusi, dan tanya jawab di kelas (*live session*) serta *mentoring* akan memperkaya pengalaman peserta. Proses pembelajaran akan diperkaya dengan berbagai tugas dan *assignment* untuk melatih kemampuan analisis dan evaluasi dalam mempersiapkan data mentah. Dengan durasi pembelajaran 180 jam, peserta juga akan

mendapatkan akses ke berbagai sumber daya, termasuk modul dari Startup Campus, rekaman video pembelajaran dari para ahli, *mentoring*, serta video dan referensi lainnya. Metode penilaian akan dilakukan melalui tugas individu dan/atau kelompok, keaktifan dalam kelas, serta kuis.

### **2.4.3 Machine Learning**

Pada *chapter Machine Learning*, peserta akan mempelajari berbagai konsep *machine learning*, baik yang bersifat *supervised* maupun *unsupervised*, serta bagaimana mengimplementasikannya untuk menyelesaikan berbagai masalah dalam *data science*. Materi yang akan dipelajari meliputi *Introduction to Machine Learning*, *Unsupervised Learning (Clustering)*, *Supervised Learning (Regression)*, serta teknik *Supervised Learning* lainnya seperti *Decision Trees*, *Random Forest*, *Support Vector Machine*, dan *Forecasting*. Selain itu, peserta juga akan diajarkan mengenai evaluasi model dengan menggunakan teknik seperti *Cross Validation*, *Bootstrapping*, dan *Learning Curves*, serta cara mengoptimalkan model melalui *Hyperparameter Tuning*. Dalam *chapter* ini, peserta akan banyak melakukan sesi praktikal untuk memperdalam pemahaman mereka. Pembelajaran asinkronus dilengkapi dengan modul, video pembelajaran, latihan, kuis, dan referensi lain yang tersedia di *Learning Management System*. Pembelajaran sinkronus dilakukan melalui pemaparan materi, diskusi, dan sesi tanya jawab dalam kelas (*live session*) serta *mentoring*. Proses ini juga akan diperkaya dengan berbagai tugas dan *assignment* untuk mengasah kemampuan analisis dan evaluasi terkait penerapan *machine learning* dalam pemodelan data. Dengan total durasi 180 jam, peserta mendapatkan akses ke sumber daya seperti modul dari Startup Campus, rekaman video dari para ahli, *mentoring*, serta referensi lainnya. Metode penilaian dalam *chapter* ini meliputi tugas individu dan/atau kelompok, keaktifan di kelas, serta kuis.

#### **2.4.4 Data Visualization**

Pada *chapter Data Visualization*, peserta akan mempelajari cara memvisualisasikan dan menyampaikan hasil olah data dengan efektif menggunakan berbagai *Business Intelligence Tools*, seperti *Looker Studio* dan *Tableau*. Materi yang dipelajari mencakup *Data Visualization: Working with BI Tools* serta *Data Storytelling and Communication*, yang akan membekali peserta dengan kemampuan untuk membuat visualisasi yang jelas dan menarik, serta menyampaikan pesan dari data dengan cara yang mudah dipahami. Pembelajaran asinkronus akan mencakup modul, video pembelajaran, latihan, kuis, dan referensi lainnya yang dapat diakses melalui *Learning Management System*. Sementara itu, pembelajaran sinkronus dilakukan melalui pemaparan materi, diskusi, dan sesi tanya jawab dalam kelas (*live session*) serta *mentoring*. Proses pembelajaran ini juga dilengkapi dengan berbagai tugas dan *assignment* untuk mengasah keterampilan analisis dan evaluasi dalam memvisualisasikan data secara efektif. Dengan durasi pembelajaran 135 jam, peserta akan mendapatkan berbagai sumber daya, termasuk modul dari Startup Campus, rekaman video pembelajaran dari para ahli, *mentoring*, serta video dan referensi lainnya. Metode penilaian akan dilakukan melalui tugas individu dan/atau kelompok, keaktifan dalam kelas, serta kuis.

#### **2.4.5 Final Project**

Pada tahap *Final Project*, peserta akan berkolaborasi dalam tim untuk menyelesaikan masalah bisnis menggunakan pemodelan kuantitatif dan teknik analisis data yang telah dipelajari sebelumnya. Peserta diberi kebebasan untuk memilih topik dan berkreasi dalam mengembangkan produk *data science* yang relevan dengan minat mereka, menggunakan pendekatan yang telah diajarkan di kelas. Proyek ini akan berbasis pada studi kasus yang terkait dengan *real dataset* di bidang *green economy* sesuai dengan ketentuan yang diberikan. Peserta akan melakukan analisis data, merancang eksperimen berdasarkan hipotesis untuk fitur baru, dan menyusun temuan hingga menghasilkan kesimpulan akhir. Selama pengerjaan proyek, peserta akan

didampingi oleh mentor dan ahli di bidangnya. Proyek ini juga akan dilombakan, dan di akhir program, setiap tim akan mempresentasikan hasil kerja mereka di depan para ahli dan peserta lainnya. Dengan durasi pembelajaran 225 jam, peserta akan mendapatkan berbagai sumber daya, termasuk modul dari Startup Campus, rekaman video dari para ahli, *mentoring*, serta referensi lainnya. Penilaian akhir akan dilakukan melalui tugas individu dan/atau kelompok, keaktifan dalam kelas, serta kuis.



## **BAB III**

### **LANDASAN TEORI**

#### **3.1 Tenaga Kerja Pertanian**

Tenaga kerja adalah faktor penting dalam proses produksi sebagian besar tanaman pangan dan seluruh sistem pertanian-pangan. Dengan demikian, pemahaman mengenai ketersediaan tenaga kerja di sektor pertanian sangatlah penting dalam penelitian ekonomi pertanian serta analisis kebijakan [2]. Dalam perkembangan ekonomi suatu negara, kontribusi tenaga kerja di sektor pertanian sangat signifikan terhadap PDB sektor tersebut. Kenaikan jumlah penduduk dan kebutuhan tenaga kerja, khususnya yang bekerja di sektor pertanian, dianggap sebagai faktor yang mendukung peningkatan Produk Domestik Bruto (PDB) di sektor pertanian [3].

Sektor pertanian di Indonesia hingga saat ini tetap memainkan peran penting, sejajar dengan sektor lainnya, terutama industri. Meskipun kontribusinya terhadap pendapatan negara semakin menurun, mayoritas penduduk Indonesia masih bergantung pada sektor ini untuk mata pencahariannya. Tingginya penyerapan tenaga kerja di sektor pertanian belum diimbangi dengan kebijakan yang memadai dari pemerintah untuk mendukung perkembangannya. Selain itu, sektor pertanian semakin terpinggirkan oleh sektor lain akibat alih fungsi lahan pertanian yang semakin meluas dan bertambahnya lahan kritis [4].

Sudah menjadi pengetahuan umum bahwa di kawasan perdesaan, petani umumnya merupakan individu berusia di atas 50 tahun, yang kini merasa kebingungan memikirkan keberlanjutan usaha tani mereka, karena hampir tidak ada generasi muda yang tertarik untuk melanjutkan pekerjaan yang telah mereka tekuni dan warisi turun-temurun. Beberapa alasan utama yang menyebabkan menurunnya minat tenaga kerja muda di sektor pertanian adalah citra sektor ini yang kurang prestisius dan tidak mampu memberikan imbalan yang memadai. Hal ini disebabkan oleh terbatasnya luas lahan pertanian yang dimiliki.

Selain itu, cara pandang dan gaya hidup generasi muda telah berubah seiring perkembangan masyarakat pascamodern saat ini. Bagi anak muda di pedesaan, sektor pertanian semakin kehilangan daya tarik, tidak hanya karena secara ekonomi sektor ini semakin tidak menjanjikan, tetapi juga karena keengganan mereka untuk bertani dipengaruhi oleh subkultur baru yang berkembang di era digital sekarang ini [5].

Sektor pertanian merupakan pilar utama perekonomian suatu negara. Namun, dalam beberapa tahun terakhir, terjadi pergeseran signifikan dalam penyerapan tenaga kerja dari sektor pertanian ke sektor-sektor lain, terutama di kalangan Generasi Z. Hal ini menunjukkan bahwa dalam sepuluh tahun terakhir, jumlah tenaga kerja yang terserap di sektor pertanian menurun lebih dari setengahnya. Penurunan ini mencerminkan pergeseran aliran tenaga kerja dari sektor pertanian ke sektor lainnya. Oleh karena itu, perlu ada pemahaman yang lebih mendalam tentang penyerapan tenaga kerja di sektor pertanian agar dapat mempertahankan dan bahkan meningkatkan jumlah tenaga kerja di sektor tersebut [6].

### **3.2 *Group of 20 (G20)***

Pembentukan G20 dimulai pada tahun 1999 setelah krisis keuangan Asia 1997-1998, dengan tujuan menciptakan forum informal bagi Menteri Keuangan dan Gubernur Bank Sentral dari negara-negara industri dan berkembang yang paling berpengaruh untuk membahas stabilitas ekonomi dan keuangan global. Awalnya, forum ini lebih fokus pada isu ekonomi makro yang besar, namun seiring waktu, agendanya berkembang mencakup berbagai topik seperti perdagangan, perubahan iklim, pembangunan berkelanjutan, kesehatan, pertanian, energi, lingkungan, perubahan iklim, dan pemberantasan korupsi.

G20 terdiri dari 19 negara (Argentina, Australia, Brasil, Kanada, Tiongkok, Prancis, Jerman, India, Italia, Jepang, Republik Korea, Meksiko, Rusia, Arab Saudi, Afrika Selatan, Turki, Inggris, dan Amerika Serikat), serta dua badan regional, yaitu Uni Eropa (UE) dan Uni Afrika (AU). Negara anggota G20 mencakup sekitar 85% dari Produk Domestik Bruto global, lebih dari 75% perdagangan internasional, dan sekitar dua pertiga populasi dunia.

Peningkatan G20 ke tingkat Kepala Negara/Pemerintahan terjadi setelah krisis ekonomi dan keuangan global tahun 2007. Pada 2009, disadari bahwa koordinasi krisis yang efektif hanya bisa dilakukan di tingkat politik tertinggi. Sejak saat itu, para Pemimpin G20 rutin mengadakan pertemuan, menjadikan G20 forum utama untuk kerja sama ekonomi internasional [7].

Sebagai anggota G20, Indonesia memiliki kesempatan untuk mendapatkan informasi dan wawasan lebih awal tentang perkembangan ekonomi global, potensi risiko yang mungkin muncul, serta kebijakan ekonomi yang diterapkan oleh negara lain khususnya negara maju. Hal ini memungkinkan Indonesia untuk merumuskan kebijakan ekonomi yang lebih tepat dan efektif. Selain itu, Indonesia dapat memperjuangkan kepentingan nasionalnya dengan dukungan internasional yang diperoleh melalui forum ini. Partisipasi dalam G20 juga meningkatkan pengakuan terhadap nama dan prestasi Indonesia di dunia internasional menjadikannya lebih dikenal oleh berbagai organisasi dan forum global [8].

### **3.3 Data Science**

Istilah “*science*” menunjukkan pada pengetahuan yang didapat melalui studi sistematis. Dalam satu definisi, sains adalah usaha sistematis yang membangun dan mengatur pengetahuan dalam bentuk penjelasan dan prediksi yang dapat diuji. Oleh karena itu, *data science* mungkin dimaksudkan kepada fokus yang melibatkan data dan sebagai perluasan, statistik, atau studi sistematis tentang organisasi, properti, dan analisis data serta perannya dalam kesimpulan, termasuk keyakinan pengguna terhadap kesimpulan tersebut.

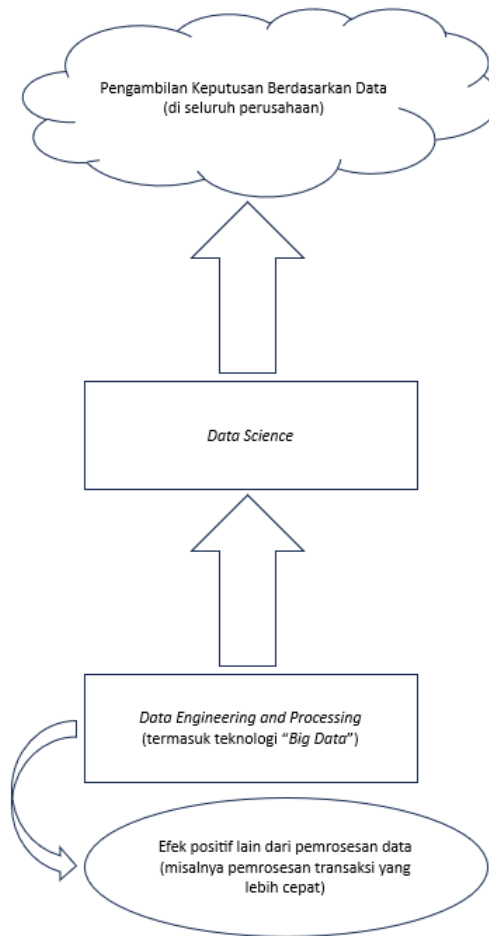
*Data science* berbeda dari statistik dan disiplin ilmu lain yang sudah ada dalam beberapa hal penting. Bagian “data” dari ilmu data, semakin bervariasi dan tidak terstruktur, seperti teks, gambar, dan video yang sering kali berasal dari jaringan dengan hubungan yang kompleks antara kehadirannya. Perubahan cepat pada ilmu data juga menimbulkan tantangan serius terkait cara organisasi mengelola ilmuwan data mereka. Selain mengenali dan mengembangkan keterampilan yang tepat, hal ini membutuhkan perubahan pola pikir menuju pengambilan keputusan berbasis data untuk menggantikan atau menambah naluri dan praktik masa lalu.

Dari sudut pandang pengambilan keputusan, pengguna tengah bergerak ke era *big data* di mana untuk banyak jenis masalah komputer secara bawaan merupakan pengambil keputusan yang lebih baik daripada manusia, di mana banyak hal yang lebih baik dapat ditentukan dalam hal biaya, akurasi, dan skalabilitas. Pergeseran ini telah terjadi di dunia keuangan yang sangat bergantung pada data, di mana komputer membuat sebagian besar keputusan investasi saat informasi baru tersedia. Hal yang sama berlaku di bidang kontrol lalu lintas udara, perutean pengiriman paket, dan banyak jenis tugas perencanaan yang memerlukan skala, kecepatan, dan keakuratan secara bersamaan. Ini akan menjadi sebuah tren yang kemungkinan akan meningkat dalam beberapa tahun ke depan [9].

*Data science* saling terkait erat dengan konsep penting lainnya yang juga semakin penting dan mendapat perhatian. Program akademik ilmu data yang sedang dikembangkan dalam lingkungan akademis mempunyai batasannya. Namun, agar ilmu data dapat melayani bisnis secara efektif, penting untuk mengetahui hubungannya dengan konsep-konsep penting dan terkait erat lainnya, dan mulai memahami apa saja prinsip dasar yang mendasari ilmu data.

Ilmu data adalah prinsip dasar yang membantu ekstraksi informasi dan pengetahuan dari data. Ilmu data juga diterapkan untuk manajemen hubungan pelanggan umum untuk menganalisis perilaku pelanggan guna mengelola pengurangan dan memaksimalkan nilai pelanggan yang diharapkan. Banyak perusahaan telah membedakan diri secara strategis dengan ilmu data, terkadang sampai pada titik berevolusi menjadi perusahaan penambangan data. Namun, ilmu data melibatkan lebih dari sekadar algoritma penambangan data.

Ilmuwan data harus dapat melihat masalah bisnis dari perspektif data. Terdapat struktur fundamental untuk pemikiran analitik data, dan prinsip dasar yang harus dipahami. Ada juga bidang-bidang tertentu di mana naluri, kreativitas, akal sehat, dan pengetahuan tentang aplikasi tertentu harus digunakan. Perspektif ilmu data menyediakan struktur dan prinsip bagi para praktisi, yang memberikan kerangka kerja bagi ilmuwan data untuk secara sistematis menangani masalah dalam mengekstraksi pengetahuan yang berguna dari data. Ilmu data dalam konteks proses lain yang terkait dengan data di organisasi terdapat pada gambar berikut.



**Gambar 3.1** Ilmu Data dalam Konteks Proses

Ilmu data melibatkan prinsip, proses, dan teknik untuk memahami fenomena melalui analisis data secara otomatis. Adapun tujuan akhir ilmu data adalah meningkatkan pengambilan keputusan karena hal ini umumnya menjadi perhatian utama bagi bisnis. Pengguna harus mencermati industri ini dan industri sejenisnya untuk mencari tanda-tanda kemajuan dalam *big data* dan ilmu data yang selanjutnya akan diadopsi oleh industri lain. Salah satu aspek terpenting dari ilmu data adalah dukungan pemikiran analitis data. Keterampilan dalam berpikir analitis data penting tidak hanya bagi ilmuwan data tetapi juga bagi seluruh organisasi. Investor dalam usaha ilmu data perlu memahami prinsip-prinsip fundamental agar dapat menilai peluang investasi secara akurat.

Secara lebih umum, bisnis semakin didorong oleh analisis data dan mempunyai keuntungan profesional yang besar dalam hal kemampuan untuk berinteraksi secara kompeten dengan dan di dalam bisnis tersebut. Memahami konsep fundamental, dan memiliki kerangka kerja untuk mengatur pemikiran analisis data, tidak hanya akan memungkinkan seseorang untuk berinteraksi secara kompeten, tetapi juga akan membantu untuk membayangkan peluang untuk meningkatkan pengambilan keputusan berdasarkan data atau untuk menghadapi ancaman kompetitif yang berorientasi pada data [10].

### 3.4 *Data Mining*

*Data mining* mengarah pada proses analisis sejumlah besar data untuk menemukan pola dan mengungkapkan pengetahuan. Banyak teknik dalam *data mining* mempunyai kesamaan dengan *machine learning* atau metode statistik. *Data mining* ini merupakan cabang dalam ilmu komputer yang bertujuan untuk mengekstraksi informasi dari data besar dengan cara yang cerdas. Proses *data mining* ini melibatkan beragam teknik, mulai dari pengumpulan data mentah hingga analisis data.

Aplikasi *data mining* sangat beragam, termasuk dalam pengumpulan, ekstraksi, penyimpanan, dan analisis data, serta penerapannya pada bidang kecerdasan buatan. *Data mining* mempunyai sejumlah teknik dapat dibagi ke dalam enam kategori utama sebagai berikut.

1. Deteksi anomali berfokus pada identifikasi *outlier* atau deviasi perubahan yang tidak biasa.
2. Regresi bertujuan untuk menemukan fungsi yang menggambarkan hubungan antara variabel respon dan variabel prediktor.
3. *Clustering* digunakan untuk menemukan kelompok atau struktur dalam data yang memiliki kesamaan tertentu.
4. Klasifikasi melibatkan penerapan pola yang telah diketahui untuk mengklasifikasikan data baru.
5. *Association rule learning* mencari pola hubungan antara variabel.
6. *Summarization* menyajikan data dalam bentuk yang lebih ringkas dan mudah dipahami [11].

Ada berbagai istilah lain yang sering muncul dalam artikel dan dokumen yang memiliki makna serupa atau sedikit berbeda, seperti *knowledge mining from databases*, *knowledge extraction*, *data archaeology*, *data dredging*, *data analysis*, dan sebagainya. Dengan penemuan pengetahuan dalam basis data, informasi yang menarik, pola, atau informasi tingkat tinggi dapat diekstraksi dari kumpulan data relevan dalam basis data dan dianalisis dari berbagai perspektif. Dengan cara ini, basis data yang besar berfungsi sebagai sumber yang kaya dan andal untuk pembuatan dan verifikasi pengetahuan.

Pengetahuan yang ditemukan ini dapat diterapkan dalam manajemen informasi, pemrosesan kueri, pengambilan keputusan, kontrol proses, dan berbagai aplikasi lainnya. Peneliti dari berbagai disiplin ilmu, seperti sistem basis data, sistem basis pengetahuan, kecerdasan buatan, pembelajaran mesin, akuisisi pengetahuan, statistik, basis data spasial, dan visualisasi data, menunjukkan ketertarikan yang besar terhadap *data mining*. Selain itu, beberapa aplikasi yang muncul dalam layanan penyediaan informasi, seperti layanan daring dan *World Wide Web*, juga memerlukan berbagai teknik *data mining* untuk lebih memahami perilaku pengguna, meningkatkan layanan, dan memperbesar peluang bisnis [12].

### **3.5 Cross-Industry Standard Process for Data Mining (CRISP-DM)**

CRISP-DM adalah sebuah model proses yang tidak bergantung pada industri tertentu untuk penambangan data. Model ini terdiri dari enam tahap yang bersifat iteratif, mulai dari *business understanding* hingga *deployment*. Adapun fase dan deskripsi model proses CRISP-DM adalah sebagai berikut.

#### **1. Business understanding**

Untuk mendapatkan gambaran yang jelas tentang sumber daya yang tersedia dan yang dibutuhkan, evaluasi situasi bisnis sangat penting. Menentukan tujuan *data mining* merupakan langkah kunci dalam fase ini. Hal pertama yang perlu dilakukan adalah menjelaskan jenis *data mining* yang akan digunakan (misalnya, klasifikasi) serta kriteria yang akan mengukur keberhasilan *data mining* (seperti tingkat presisi). Rencana proyek yang jelas juga harus disusun.

## 2. *Data understanding*

Mengumpulkan data dari berbagai sumber, menganalisis, dan memeriksa kualitas data adalah tugas utama dalam fase ini. Agar lebih konkret, CRISP-DM disarankan menggunakan analisis statistik untuk mendeskripsikan data dan menentukan atribut serta hubungan antar data.

## 3. *Data preparation*

Pemilihan data dilakukan dengan cara menetapkan kriteria inklusi dan eksklusi. Data yang memiliki kualitas buruk perlu dibersihkan. Bergantung pada model yang dipilih (yang didefinisikan dalam fase pertama), atribut turunan juga harus dibuat. Berbagai metode dapat digunakan dalam langkah ini, tergantung pada model yang diterapkan.

## 4. *Modeling*

Fase pemodelan mencakup pemilihan teknik pemodelan yang tepat, pembuatan kasus uji, dan pengembangan model. Semua teknik *data mining* dapat digunakan, dengan pilihan yang bergantung pada masalah bisnis dan data yang ada. Yang lebih penting adalah menjelaskan alasan di balik pemilihan teknik tersebut. Untuk membangun model, parameter tertentu harus diatur, dan untuk mengevaluasi model, sebaiknya menggunakan kriteria yang ditentukan untuk memilih model terbaik.

## 5. *Evaluation*

Pada fase ini, hasil evaluasi dibandingkan dengan tujuan bisnis yang sudah ditetapkan. Oleh karena itu, hasil yang diperoleh harus diinterpretasikan dengan baik, dan langkah-langkah selanjutnya harus ditentukan. Selain itu, proses secara keseluruhan perlu ditinjau.

## 6. *Deployment*

Fase penerapan mencakup perencanaan penerapan, pemantauan, dan pemeliharaan, yang dijelaskan lebih lanjut dalam panduan pengguna. Hasil dari fase ini bisa berupa laporan akhir atau komponen perangkat lunak.

Tahap *business understanding* dijelaskan dengan cara yang sangat beragam, karena CRISP-DM diterapkan di berbagai bidang. Pada umumnya, deskripsi tujuan bisnis dan manfaat *data mining* untuk kasus tertentu ditemukan dalam bentuk teks. *Data understanding* pada CRISP-DM dapat dilakukan dengan menyebutkan



sumber data yang spesifik dan menjelaskan dari mana data dikumpulkan. Statistik deskriptif adalah metode utama untuk memahami data, sementara metode lainnya meliputi visualisasi data atau mewawancarai para ahli untuk mendapatkan pemahaman yang lebih baik.

Tugas umum dalam persiapan data, seperti pemilihan data, transformasi, dan pembersihan data. Penggunaan beberapa model dan algoritma memungkinkan perbandingan hasil yang berbeda selama evaluasi dengan menggunakan metrik untuk mengevaluasi kualitas model yang dilatih. Metrik ini juga divisualisasikan untuk menggambarkan hasil, misalnya menggunakan matriks kebingungan. Visualisasi model yang dilatih, seperti pohon keputusan, digunakan untuk menjelaskan dan mengevaluasi model [13].

### 3.6 *Feature Scaling*

*Feature scaling* yang juga dikenal sebagai standardisasi adalah langkah dalam proses *data pre-processing*. Tujuan dari penskalaan ini adalah untuk menormalkan data ke dalam rentang tertentu. *Feature scaling* mempermudah perhitungan dalam algoritma. Kumpulan data yang digunakan dalam *feature scaling* ini terdiri dari variabel dengan skala yang berbeda sehingga *feature scaling* dilakukan untuk menyesuaikan vektor fitur agar lebih sesuai dengan teknik *deep learning*. Ada beberapa metode penskalaan yang dapat digunakan dan yang paling umum adalah `StandardScaler()`, `MinMaxScaler()`, dan `RobustScaler()`.

`StandardScaler()` mengubah data sehingga distribusi hasilnya memiliki rata-rata nol dan deviasi standar satu. Proses ini dilakukan dengan cara mengurangi nilai rata-rata dari nilai asli dan membaginya dengan deviasi standar. Rumus untuk transformasi ini adalah sebagai berikut. Adapun  $z$  adalah nilai fitur yang telah ditransformasikan,  $x$  adalah nilai asli,  $\mu$  adalah rata-rata, dan  $\sigma$  adalah deviasi standar.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

`MinMaxScaler()` melakukan penskalaan data sehingga semua nilai dalam kumpulan data berada dalam rentang antara 0 dan 1. Adapun  $t$  adalah nilai fitur yang telah ditransformasikan,  $X$  adalah nilai asli, dan  $X_{min}$  serta  $X_{max}$  adalah nilai minimum dan maksimum dari fitur  $X$ .

$$t = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

RobustScaler() menghapus median dan melakukan penskalaan berdasarkan rentang kuartil, yaitu antara kuartil ke-25 dan kuartil ke-75. Data yang telah ditransformasikan memiliki rentang nilai yang lebih besar dibandingkan dengan metode penskalaan sebelumnya. Nilai dalam kumpulan data diubah agar berada dalam rentang  $[-2, 3]$ . Metode ini mirip dengan MinMaxScaler, namun menggunakan rentang antar kuartil daripada nilai minimum dan maksimum. Adapun  $t$  adalah nilai fitur yang telah diubah,  $x$  adalah nilai asli, dan  $Q_1(x)$  dan  $Q_3(x)$  adalah rentang antara kuartil pertama dan ketiga [14].

$$t = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \quad (3)$$

### 3.7 Machine Learning

Algoritma *machine learning* adalah sebuah proses komputasi yang menggunakan data input untuk menyelesaikan tugas yang diinginkan tanpa diprogram secara eksplisit untuk menghasilkan hasil tertentu. Algoritma dianggap “*soft coded*” karena secara otomatis menyesuaikan atau mengadaptasi strukturnya melalui pengalaman sehingga semakin efektif dalam menyelesaikan tugas tersebut. Proses adaptasi ini dikenal sebagai pelatihan di mana data masukan beserta hasil yang diinginkan disediakan. Algoritma kemudian mengoptimalkan diri sehingga tidak hanya dapat menghasilkan hasil yang diinginkan saat menghadapi data pelatihan, tetapi juga dapat digambarkan untuk menghasilkan hasil yang serupa pada data baru yang belum pernah dilihat sebelumnya.

Pelatihan ini adalah bagian dari “*learning*” dalam *machine learning* yang tidak terbatas pada adaptasi awal dalam jangka waktu tertentu. Ada berbagai cara bagi algoritma untuk menyesuaikan diri dalam merespons pelatihan. Data masukan dapat dipilih dan diberi bobot untuk menghasilkan hasil yang lebih menentukan. Algoritma dapat memiliki parameter numerik yang dapat diubah melalui proses *iterative optimization*. Algoritma juga bisa memiliki jaringan jalur komputasi yang dapat diatur untuk menghasilkan hasil optimal. Selain itu, algoritma dapat menentukan distribusi probabilitas dari data input dan menggunakannya untuk meramalkan hasil.

Tujuan utama *machine learning* adalah meniru cara manusia (dan makhluk berakal lainnya) memproses sinyal sensorik (*input*) untuk mencapai tujuan tertentu. Tujuan ini bisa berupa tugas pengenalan pola di mana pelajar berusaha membedakan apel dari jeruk. Meskipun setiap apel dan jeruk itu unik, manusia masih dapat membedakan keduanya. Ketimbang memprogram mesin dengan berbagai representasi apel dan jeruk yang tepat, mesin dapat diprogram untuk belajar membedakan keduanya melalui pengalaman berulang dengan apel dan jeruk yang sesungguhnya. Ini merupakan contoh *supervised learning* di mana setiap contoh data input (seperti warna, bentuk, bau, dll.) dipasangkan dengan label klasifikasi yang sudah diketahui (misalnya, apel atau jeruk). Dengan demikian, pengguna dapat memahami persamaan dan perbedaan saat objek yang diklasifikasikan memiliki banyak sifat yang bervariasi dalam kelasnya namun tetap memiliki ciri khas yang mendasarinya. Penting bagi pengguna yang berhasil harus dapat mengenali apel atau jeruk yang belum pernah dilihat sebelumnya.

Jenis *machine learning* lainnya adalah *unsupervised learning* yang tujuannya bisa jadi untuk melempar anak panah ke sasaran. Perangkat atau manusia memiliki berbagai tingkat kebebasan dalam mekanisme yang mengendalikan jalur anak panah. Ketimbang memprogram kinematika secara tepat terlebih dahulu, pengguna berlatih melempar anak panah, dan untuk setiap percobaan kebebasan kinematik disesuaikan agar anak panah semakin mendekati sasaran. Ini disebut *unsupervised learning* karena pelatihan tidak menghubungkan konfigurasi kinematik tertentu dengan hasil tertentu. Algoritma menemukan jalannya sendiri berdasarkan data masukan pelatihan. Pelempar anak panah yang terlatih dapat menyesuaikan kinematika yang dipelajari untuk menampung perubahan posisi target.

Adapun juga *semi-supervised learning* di mana sebagian data diberi label dan sebagian lainnya tidak. Dalam skenario ini, bagian yang diberi label membantu dalam mempelajari bagian yang tidak diberi label. Pendekatan ini mirip dengan cara manusia mengembangkan keterampilan mereka dan lebih sesuai dengan banyak proses alami. Algoritma *machine learning* lainnya yang menarik adalah *reinforcement learning* di mana orang mencoba mengambil serangkaian tindakan untuk memaksimalkan *cumulative reward* contohnya memenangkan permainan catur. Pendekatan ini sangat berguna dalam aplikasi pembelajaran daring [15].

### 3.8 Clustering

*Clustering* adalah pengorganisasian sekumpulan pola (yang biasanya direpresentasikan sebagai vektor pengukuran, atau titik dalam ruang multidimensi) ke dalam *cluster* berdasarkan kesamaan. Secara intuitif, pola-pola dalam *cluster* yang sah lebih mirip satu sama lain daripada dengan pola yang termasuk dalam *cluster* yang berbeda. Berbagai teknik untuk merepresentasikan data, mengukur kedekatan (kesamaan) antar elemen data, dan mengelompokkan elemen data menghasilkan beragam metode *clustering* yang beragam.

Penting untuk memahami perbedaan antara *clustering* (*unsupervised classification*) dan *discriminant analysis* (*supervised classification*). Dalam *supervised classification*, pengguna diberikan sekumpulan pola berlabel (*preclassified*) yang mempunyai masalah dalam memberi label pada pola baru yang belum diberi label. Biasanya, pola berlabel (*training*) yang diberikan digunakan untuk mempelajari deskripsi kelas yang kemudian digunakan untuk memberi label pola baru. Sementara itu, masalah dalam *clustering* adalah mengelompokkan sekumpulan pola yang belum diberi label ke dalam *cluster* yang bermakna. Dalam hal ini, label memang diasosiasikan dengan *cluster*, namun label kategori ini didorong oleh data di mana label tersebut diperoleh hanya dari data itu sendiri.

*Clustering* berguna dalam beberapa situasi *exploratory pattern-analysis*, pengelompokan, pengambilan keputusan, dan *machine-learning situations*, termasuk *machine learning*, pengambilan dokumen, segmentasi gambar, dan klasifikasi pola. Namun, dalam banyak masalah semacam ini, informasi sebelumnya tentang data hampir tidak tersedia, dan pengambil keputusan harus membuat asumsi sesedikit mungkin tentang data. Dalam batasan-batasan ini, metodologi *clustering* sangat sesuai untuk eksplorasi hubungan antar titik data untuk membuat penilaian sementara tentang struktur mereka [16].

### 3.9 K-Means Clustering

*K-Means Clustering* merupakan salah satu algoritma *unsupervised learning* yang paling sederhana untuk menyelesaikan masalah pengelompokan yang umum. Prosedur ini mengikuti pendekatan yang sederhana dan mudah untuk mengklasifikasikan sekumpulan data tertentu ke dalam beberapa *cluster* (dengan jumlah *cluster* yang telah ditentukan sebelumnya). Konsep utamanya adalah menentukan  $k$  *centroid*, satu untuk setiap *cluster*. Peletakkan *centroid* ini harus dilakukan dengan hati-hati karena lokasi yang berbeda dapat menghasilkan hasil yang berbeda. Oleh karena itu, pilihan terbaik adalah menempatkan *centroid* sejauh mungkin satu sama lain.

Langkah selanjutnya adalah mengaitkan setiap titik dalam kumpulan data ke *centroid* terdekat. Ketika tidak ada titik yang tersisa untuk diproses, langkah pertama selesai, dan pengelompokan awal tercapai. Pada tahap ini, *centroid* baru perlu dihitung ulang sebagai pusat *cluster* berdasarkan hasil langkah sebelumnya. Setelah *centroid* baru dihitung, pengelompokan ulang dilakukan antara titik data yang sama dan *centroid* terdekat yang baru. Proses ini akan berulang dalam sebuah iterasi. Selama iterasi ini,  $k$  *centroid* mungkin berpindah tempat langkah demi langkah hingga tidak ada lagi perubahan posisi yang terjadi. Dengan kata lain, *centroid* berhenti bergerak. Tujuan akhir dari algoritma ini adalah meminimalkan fungsi objektif, yaitu fungsi kesalahan kuadrat. Adapun fungsi objektif adalah sebagai berikut.

$$W(S, C) = \sum_{k=1}^K \sum_{i \in S_k} \|y_i - c_k\|^2 \quad (4)$$

Fungsi  $S$  ini menggambarkan partisi  $k$ -cluster dari himpunan entitas yang direpresentasikan oleh vektor  $y_i$  ( $i \in I$ ) ( $i \in I$ ) dalam ruang fitur  $M$ -dimensional, yang terdiri dari *cluster*  $S_k$  yang tidak kosong dan tidak tumpang tindih, masing-masing dengan *centroid*  $c_k$  ( $k = 1, 2, \dots, K$ ). Algoritma ini terdiri dari langkah-langkah berikut:

1. Tentukan  $k$  titik dalam ruang yang diwakili oleh objek yang akan dikelompokkan. Titik-titik ini mewakili *initial group centroids*.
2. Tetapkan setiap objek ke grup yang memiliki *centroid* terdekat.
3. Setelah semua objek ditetapkan, hitung ulang posisi  $k$  *centroid*.

4. Ulangi langkah 2 dan 3 hingga posisi *centroid* tidak lagi berubah.

Ada berbagai pendekatan yang diusulkan dalam literatur untuk memilih nilai  $K$  yang tepat setelah menjalankan K-Means beberapa kali. Salah satunya akan fokus pada pendekatan dengan memilih  $k$  menggunakan *silhouette*. Sejumlah pendekatan menggunakan indeks yang membandingkan jarak antar objek dalam *cluster* dengan jarak antar *cluster* di mana semakin besar perbedaannya, maka semakin baik kecocokannya.

Dua indeks yang digunakan adalah: (a) korelasi titik-biserial, yaitu koefisien korelasi antara matriks jarak entitas-ke-entitas dan matriks partisi biner yang menetapkan pasangan entitas 1 jika mereka berada dalam *cluster* yang sama, dan 0 jika tidak, serta (b) versi ordinal. Koefisien yang seimbang (lebar siluet) telah terbukti efektif dalam eksperimen. Konsep lebar siluet melibatkan perbedaan antara keketatan dalam *cluster* dan pemisahan antar *cluster*. Secara khusus, lebar siluet  $s(i)$  untuk entitas  $i \in I$  didefinisikan sebagai berikut.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

Adapun  $a(i)$  adalah jarak rata-rata antara  $i$  dan semua entitas lain dalam *cluster* yang sama, dan  $b(i)$  adalah jarak rata-rata minimum antara  $i$  dan entitas-entitas di *cluster* lainnya. Nilai lebar siluet berada dalam rentang -1 hingga 1. Jika nilai lebar siluet suatu entitas mendekati nol, ini menunjukkan bahwa entitas tersebut bisa saja dimasukkan ke dalam *cluster* lain. Jika nilai lebar siluet mendekati -1, artinya entitas tersebut salah diklasifikasikan. Jika semua nilai lebar siluet mendekati 1, itu berarti data telah ter-*cluster* dengan baik. Pengelompokan dapat dijelaskan dengan lebar siluet rata-rata entitas individu. Lebar siluet rata-rata terbesar di antara berbagai nilai  $K$  menunjukkan jumlah *cluster* terbaik [17].

### 3.10 Agglomerative Clustering

*Agglomerative Clustering* adalah metode yang penting dan telah terbukti efektif dalam *unsupervised machine learning*. Skema *agglomerative clustering* dimulai dengan membagi set data menjadi simpul tunggal dan kemudian secara bertahap menggabungkan pasangan simpul yang paling dekat menjadi simpul baru, sampai hanya tersisa satu simpul yang mencakup seluruh data. Berbagai skema

pengelompokan menggunakan prosedur ini sebagai dasar umum, namun memiliki perbedaan dalam cara memperbarui pengukuran perbedaan antar-*cluster* setelah setiap langkah. Tujuh metode yang paling banyak digunakan di antaranya adalah hubungan tunggal, lengkap, rata-rata (UPGMA), tertimbang (WPGMA, McQuitty), Ward, centroid (UPGMC), dan median (WPGMC) [18].

**Tabel 3.1** Metode *Agglomerative Hierarchical Clustering* Standar

Metode	Nama Alternatif	Penggunaan Umum	Definisi Jarak antar <i>Cluster</i>
<i>Single linkage Sneath</i>	<i>Nearest neighbour</i>	Kesamaan atau jarak	Jarak terkecil antara dua objek, satu berada di satu <i>cluster</i> dan yang lainnya di <i>cluster</i> berbeda.
<i>Complete linkage Sorensen</i>	<i>Furthest neighbour</i>	Kesamaan atau jarak	Jarak terbesar antara dua objek, satu berada di satu <i>cluster</i> dan yang lainnya di <i>cluster</i> yang berbeda.
<i>(Group) Average linkage Sokal and Michener</i>	UPGMA	Kesamaan atau jarak	Jarak rata-rata antara dua objek, di mana satu berada di satu <i>cluster</i> dan yang lainnya di <i>cluster</i> yang berbeda.
<i>Centroid linkage Sokal and Michener</i>	UPGMC	Jarak (membutuhkan data mentah)	Kuadrat jarak <i>Euclidean</i> antara vektor rata-rata (titik pusat).
<i>Weighted average linkage McQuitty</i>	WPGMA	Kesamaan atau jarak	Jarak rata-rata antara dua objek, di mana satu berada di satu <i>cluster</i> dan yang lainnya di <i>cluster</i> yang berbeda.

<i>Median linkage Gower</i>	WPGMC	Jarak (membutuhkan data mentah)	Kuadrat jarak <i>Euclidean</i> antara titik pusat yang berbobot.
<i>Ward's method Ward</i>	Jumlah kuadrat minimum	Jarak (membutuhkan data mentah)	Penambahan jumlah kuadrat dalam <i>cluster</i> setelah penggabungan dihitung untuk semua variabel.

Prosedur aglomeratif berkemungkinan merupakan metode hierarkis yang paling sering digunakan. Metode ini menghasilkan serangkaian pembagian data di mana yang pertama terdiri dari  $n$  'cluster' dengan satu anggota masing-masing dan yang terakhir berupa satu *cluster* yang mencakup semua  $n$  individu. Operasi dasar dari semua metode ini serupa dan akan dijelaskan melalui dua contoh spesifik, yaitu *single linkage* dan *centroid linkage*. Pada setiap langkah, metode ini menggabungkan individu atau kelompok individu yang paling dekat atau mirip. Perbedaan antara metode terletak pada cara mendefinisikan jarak atau kesamaan antara individu dan kelompok yang terdiri dari beberapa individu, atau antara dua kelompok individu [19].

### 3.11 Reduksi Dimensionalitas

Reduksi dimensionalitas adalah teknik dalam *machine learning* yang mengurangi jumlah dimensi data dengan meminimalkan kehilangan informasi sebelum analisis lebih lanjut. Teknik ini mengubah data berdimensi tinggi menjadi ruang berdimensi lebih rendah yang tetap bermakna. Representasi yang tereduksi harus mencerminkan dimensi intrinsik data, yaitu jumlah parameter minimum yang diperlukan untuk menggambarkan sifat data yang dapat diamati. Tujuan utama dari reduksi dimensionalitas adalah untuk mengurangi redundansi dan mengidentifikasi struktur intrinsik data.

Teknik ini juga digunakan dalam ekstraksi fitur, visualisasi data, komputasi, dan aplikasi *machine learning* lainnya. Reduksi dimensionalitas sering dianggap sebagai langkah *data pre-processing*. Tantangan utamanya adalah melakukan transformasi dengan kehilangan informasi minimal sembari mempertahankan



struktur data. Berbagai teknik telah dikembangkan untuk mengubah fitur yang ada menjadi serangkaian fitur baru yang tereduksi atau memilih subset dari fitur yang ada.

Teknik reduksi dimensionalitas umumnya dibagi menjadi teknik linier dan nonlinier. Reduksi dimensionalitas linier mengubah data ke ruang berdimensi rendah melalui kombinasi linier dari variabel asli, yang berlaku ketika data terletak dalam subruang linier, menggantikan variabel asli dengan kumpulan variabel dasar yang lebih kecil. Sementara itu, reduksi dimensionalitas nonlinier diterapkan ketika data berdimensi tinggi memiliki hubungan nonlinier, dengan tujuan menjaga jarak asli antar titik data meskipun dalam dimensi yang lebih rendah. Beberapa teknik reduksi dimensionalitas linier dan nonlinier yang umum digunakan ialah *Principal Component Analysis* (PCA) dan *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [20].

PCA adalah algoritma transformasi linier *unsupervised* yang menghasilkan fitur baru, yang disebut *Principal Components* (PCs) dengan cara memaksimalkan varians data. PCA memproyeksikan data berdimensi tinggi ke dalam subruang baru di mana sumbu ortogonal atau PCs dianggap sebagai arah dengan varians data maksimum. Selama proses transformasi, PC pertama memiliki varians terbesar dan PCs berikutnya memiliki varians yang semakin kecil.

PCA membangun *d* × *k*-dimensional transformation matrix *W* yang memetakan *d*-dimensional space *X* asli ke *k*-dimensional space *Y* (*k* ≤ *d*). Metode dekomposisi eigen linier diterapkan pada matriks kovariansi (*X.X<sup>T</sup>*) untuk menghasilkan vektor eigen (PCs) dan nilai eigen. Vektor eigen menunjukkan arah data sementara nilai eigen menunjukkan besaran data.

Nilai eigen digunakan untuk mengurutkan kolom dalam matriks *W* di mana setiap kolom merupakan vektor eigen. Proses dekomposisi eigen dapat dijelaskan sebagai berikut di mana matriks kovariansi pertama kali didekomposisi menjadi tiga matriks lainnya.

$$X.X^T \rightarrow B.D.B^T \quad (6)$$

$B$  adalah matriks persegi ( $d \times d$ ) yang berisi vektor eigen dan  $D$  adalah matriks diagonal ( $d \times d$ ) di mana elemen-elemen selain elemen diagonal utama bernilai nol, serta elemen diagonalnya adalah nilai eigen masing-masing  $B^T$  adalah matriks  $B$  yang ditransposisikan.

PCA memiliki beberapa kelebihan seperti tidak memerlukan pengulangan sehingga lebih efisien waktu, mengurangi risiko *overfitting*, serta dapat digunakan sebagai teknik *denoising* dan kompresi data. Namun, PCA juga memiliki beberapa kekurangan seperti terbatas pada proyeksi linier sehingga tidak cocok untuk menangani data nonlinier dengan baik. Standar data harus dilakukan sebelum menerapkan PCA karena tanpa hal tersebut pencarian *PCs* optimal akan terhambat akibat ketergantungan terhadap skala fitur dan jika pemilihan *PCs* tidak hati-hati, dapat terjadi kehilangan informasi. Adapun skema algoritma PCA adalah sebagai berikut.

1. Mempunyai masukan  $X \in R^{n \times d}$  dan keluaran  $Y \in R^{n \times k}$ .
2. Merancang matriks kovariansi ( $X.X^T$ ).
3. Menerapkan dekomposisi eigen linier pada  $X.X^T$  untuk mendapatkan nilai dan vektor eigen.
4. Mengurutkan nilai eigen dalam urutan menurun untuk mengurutkan vektor eigen.
5. Membangun matriks  $W(d \times k)$  dengan  $k$  vektor eigen teratas.
6. Mentransformasikan  $X$  menggunakan  $W$  untuk memperoleh subruang baru  $Y = X.W$ .

t-SNE adalah metode ekstraksi fitur nonlinier *unsupervised* berbasis *manifold* yang memetakan data berdimensi tinggi ke dalam ruang berdimensi rendah (biasanya 2 atau 3 dimensi) dengan tetap mempertahankan struktur penting dari data asli. t-SNE umumnya digunakan untuk eksplorasi dan visualisasi data. Dengan kata lain, t-SNE membantu memberikan gambaran tentang bagaimana data terstruktur dalam ruang berdimensi tinggi. Meskipun banyak metode ekstraksi fitur memiliki kinerja yang kuat, sebagian besar tidak berhasil dalam memvisualisasikan data berdimensi tinggi dan tidak mampu mempertahankan struktur lokal dan global data. t-SNE sangat berguna untuk memvisualisasikan data berdimensi tinggi sambil tetap mempertahankan struktur penting dari data.

Proses t-SNE dimulai dengan menerapkan *Stochastic Neighbor Embedding* (SNE) pada data, yang mengubah jarak *Euclidean* dalam dimensi tinggi menjadi probabilitas bersyarat yang menggambarkan kesamaan antar pasangan data. Kesamaan antara data  $x_a$  dan  $x_b$  direpresentasikan oleh probabilitas bersyarat  $p_{a|b}$ , yang dihitung menggunakan persamaan berikut.

$$p_{a|b} = \frac{\exp \frac{-\|x_b - x_a\|^2}{2\sigma^2}}{\sum_{a \neq k} \frac{-\|x_k - x_a\|^2}{2\sigma^2}} \quad (7)$$

Persamaan ini mengukur seberapa dekat titik data  $x_a$  dengan titik data lainnya  $x_b$ , dengan mempertimbangkan distribusi Gaussian di sekitar  $x_b$  dengan varians tertentu  $\sigma^2$ . Varians ini bervariasi untuk setiap titik data dan dipilih sedemikian rupa sehingga data yang terletak di area padat memiliki varians lebih kecil daripada data yang berada di area jarang.

Selanjutnya, ketimbang menggunakan distribusi Gaussian, “*Student t-distribution*” dengan satu derajat kebebasan (mirip dengan distribusi Cauchy) digunakan untuk menghitung himpunan probabilitas kedua ( $Q_{a|b}$ ) dalam ruang berdimensi rendah. Jika data berdimensi tinggi  $x_a$  dan  $x_b$  dipetakan dengan benar ke dalam data berdimensi rendah  $y_a$  dan  $y_b$ , maka kesamaan antara  $P_{a|b}$  dan  $Q_{a|b}$  akan menjadi sama. Oleh karena itu, t-SNE meminimalkan perbedaan antara kedua probabilitas ini antara ruang dimensi rendah dan tinggi. Perbedaan tersebut diukur dengan mengoptimalkan fungsi biaya ( $\emptyset$ ) melalui penjumlahan divergensi Kullback–Leibler, seperti yang dijelaskan dalam persamaan berikut.

$$\emptyset = \sum_a \sum_b P_{a|b} \log \frac{P_{a|b}}{Q_{a|b}} \quad (8)$$

Adapun skema algoritma t-SNE adalah sebagai berikut.

1. Mempunyai masukan  $X \in R^{n \times d}$  dan keluaran  $Y \in R^{n \times k}$ .
2. Menerapkan SNE pada  $X$  untuk menghitung probabilitas kondisional  $P_{a|b}$  dan  $Q_{a|b}$ .
3. Petakan  $X$  ke  $Y$  dengan meminimalkan perbedaan antara  $P_{a|b}$  dan  $Q_{a|b}$  berdasarkan fungsi biaya  $\emptyset$  [21].

### 3.12 Python

Python adalah bahasa pemrograman *interpreted* yang memungkinkan pengguna dengan mudah menguji cara kerja operator atau fungsi menggunakan interpreter baris perintah. Interpreter Python dilengkapi dengan modul bantuan bawaan yang sangat berguna untuk mempercepat pemahaman tentang berbagai aspek bahasa ini. Python dengan pustakanya yang luas sangat cocok untuk membuat prototipe perangkat lunak yang kemudian bisa diterjemahkan ke dalam bahasa yang lebih rendah jika diperlukan.

Keunggulan Python sangat jelas sehingga penggunaan bahasa ini sebagai bahasa utama untuk mempelajari pemrograman dapat mempercepat pemahaman ilmu komputer secara keseluruhan. Python tidak memiliki atribut keamanan (seperti *public/private/protected*) yang membuat program lebih sederhana, pendek, dan mudah dipahami. Selain itu, Python sangat dinamis yang memungkinkan pembuatan kolom atau atribut dengan cepat. Polimorfisme dalam Python dapat diterapkan pada fungsi dan metode kelas berbeda dengan C++ yang menggunakan fungsi virtual atau non-virtual. *Overloading* operator di Python memberikan fleksibilitas lebih pada objek yang memungkinkan ekspresi alami yang tidak mungkin dilakukan di JAVA karena sintaksnya yang terbatas.

Indentasi dalam Python memiliki peran penting dalam struktur program yang membuat program yang ditulis dengan Python lebih mudah dibaca dan dipahami. Sebaliknya, dalam JAVA atau C++, banyak pengguna mencoba menyederhanakan program dengan menggabungkan semuanya dalam satu baris atau menggunakan indentasi yang tidak konsisten yang justru membuat kode sulit dibaca dan dipahami. Python juga dilengkapi dengan algoritma-algoritma hebat dalam pustakanya. Hal ini berarti siswa tidak perlu menguasai aritmatika rumit untuk melakukan perhitungan besar karena tipe data panjang sudah tersedia di Python. Ada pula berbagai alat untuk *sort*, *find*, *slice*, dan *join* urutan data apapun.

Python memiliki cara unik dalam menangani variabel sehingga programmer tidak perlu merumuskan tipe variabel secara eksplisit, yaitu tipe variabel akan ditentukan berdasarkan nilai yang disimpannya. Walaupun penting untuk memahami bagaimana tipe data disimpan dalam memori, untuk praktik pemrograman, akan lebih mudah jika informasi tersebut dikesampingkan pada

tahap awal. Python juga memiliki berbagai kata kunci dan perintah yang intuitif dan adaptif yang sangat membantu siswa dalam mempelajari pemrograman [22].

### 3.13 Google Colaboratory

Sebagai layanan gratis dari Google, Colaboratory (atau Colab) menawarkan antarmuka Jupyter Notebook yang terhubung dengan perangkat keras Google. Notebook ini dijalankan pada *virtual machines* (VMs) berbasis Linux yang disediakan dan dikelola oleh Google di mana memungkinkan komputasi dilakukan menggunakan *central processing units* (CPUs) atau dapat dipercepat dengan *graphical processing units* (GPUs) dan *tensor processing units* (TPUs).

Selain menyediakan sumber daya komputasi, VMs berbasis *cloud* yang mendukung *notebook* Colab sudah dilengkapi dengan paket AI umum seperti *numpy*, *torch*, dan *tensorflow*. Meskipun ada alternatif seperti Anaconda untuk mengelola paket, hal tersebut sering kali menjadi beban tambahan bagi pengguna yang sudah menghadapi tantangan kognitif.

Karena *notebook* Colab dihosting sebagai Jupyter Notebooks, ia mengikuti alur kerja umum yang memadukan penjelasan naratif dengan demonstrasi berbasis kode yang interaktif. Selain digunakan dalam pendidikan AI, alur kerja berbasis buku catatan ini juga menjadi media populer di kalangan peneliti AI dan ilmu data untuk berbagi pendekatan serta hasil temuan mereka [23].

## BAB IV

### ANALISIS DAN PERANCANGAN

#### 4.1 Alur Kerja

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah metodologi yang paling banyak digunakan dalam proyek *data mining*. Proses ini dirancang untuk memastikan bahwa proyek analisis data dilakukan secara terstruktur dan terorganisir. Alur kerja CRISP-DM terdiri dari enam tahap yang saling berkesinambungan, dimulai dari *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, hingga *deployment*. Setiap tahap memiliki peran yang krusial dalam mengonversi data mentah menjadi *wisdom* yang dapat diterapkan. Dikarenakan proses standar untuk data mining hanya sampai tahap *evaluation*, adapun penjelasan alur kerja sebagai berikut.

##### 4.1.1 *Business Understanding*

Proyek ini bertujuan untuk menganalisis tenaga kerja sektor pertanian di negara anggota G20 dengan fokus pada pengidentifikasian kelompok-kelompok (*cluster*) yang memiliki karakteristik serupa berdasarkan faktor-faktor sosio-ekonomi dan efisiensi energi. Negara anggota G20 memiliki perbedaan signifikan dalam hal tenaga kerja di sektor pertanian, kondisi sosial ekonomi yang terlibat, serta produktivitas energi dan material non-energi.

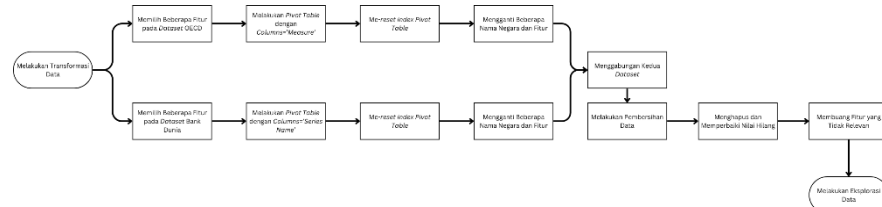
Oleh karena itu, memahami pola-pola yang ada akan memungkinkan pengambilan kebijakan yang lebih tepat, baik dalam hal pengembangan kapasitas tenaga kerja di sektor pertanian maupun dalam peningkatan efisiensi energi yang digunakan dalam kegiatan pertanian. Dengan menggunakan teknik *clustering*, kita dapat mengelompokkan negara-negara tersebut berdasarkan kemiripan-kemiripan tertentu, seperti tenaga kerja sektor pertanian, kondisi sosial ekonomi, dan produktivitas energi dan material non-energi, yang pada akhirnya bisa membantu merancang kebijakan yang lebih sesuai dengan kebutuhan masing-masing negara.

#### 4.1.2 Data Understanding

Untuk mencapai tujuan tersebut, data yang diperlukan meliputi berbagai informasi mengenai kondisi tenaga kerja sektor pertanian, sosio-ekonomi, serta data terkait efisiensi energi di negara anggota G20. Data dapat diperoleh dari sumber internasional yang terpercaya, seperti OECD dan Bank Dunia, yang menyediakan data komprehensif mengenai tenaga kerja sektor pertanian, kondisi sosio-ekonomi, serta produktivitas energi dan material non-energi.

Data yang diambil dari *dataset* OECD dilakukan dengan mengambil data dari URL API menggunakan *request* dan membaca CSV menjadi *pandas DataFrame*. Data juga yang diambil dari *dataset* Bank Dunia dilakukan dengan membaca data dari *file* Excel, kemudian mengonversi data ke dalam *DataFrame*. Lalu, memberikan kode fungsi *head()* dan *info()* dilakukan untuk memberikan gambaran ringkas tentang data, memeriksa data telah dimuat dengan benar dan untuk mempersiapkan analisis lebih lanjut.

#### 4.1.3 Data Preparation

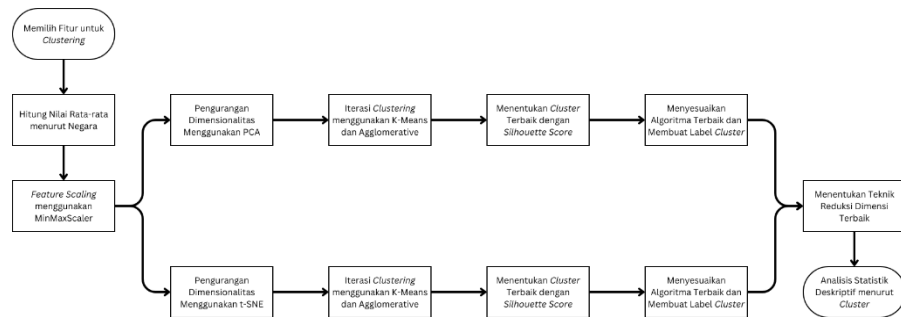


**Gambar 4.1** Diagram Kotak *Data Preparation*

Tahap persiapan data merupakan langkah yang sangat penting dalam memastikan bahwa analisis *clustering* dapat dilakukan dengan hasil yang optimal. Hal yang dilakukan pertama kali ialah *data transformation* yang dilakukan untuk perubahan nama fitur, *pivoting data*, dan beberapa nama negara. Kemudian, data OECD dan Bank Dunia akan diatur berdasarkan negara dan digabung menjadi satu *DataFrame*. Selanjutnya, ada proses *data cleaning* akan dilakukan untuk mengatasi masalah data yang hilang (*missing values*), data yang duplikat, *outlier* yang bisa mengganggu akurasi model, dan memeriksa *imbalance data*.

Setelah data dibersihkan dan dipersiapkan, kita juga akan melakukan seleksi fitur (*data filtering*) untuk memilih variabel yang paling relevan, seperti pekerjaan di bidang pertanian, kepadatan penduduk, harapan hidup saat lahir, produk domestik bruto, pendapatan pajak tenaga kerja, paritas daya beli, intensitas energi per kapita, total pasokan energi, pasokan energi terbarukan, pembangkitan listrik terbarukan, dan konsumsi energi. Dan terakhir dilakukan *Exploratory Data Analysis* (EDA) dengan mengeksplorasi beberapa hal, seperti *statistical summaries*, distribusi data, dan *bivariate & multivariate analysis*.

#### 4.1.4 Modeling dan Evaluation



**Gambar 4.2** Diagram Kotak *Modeling* dan *Evaluation*

Pada tahap pemodelan akan memilih algoritma *clustering* yang sesuai untuk mengelompokkan data. Hal yang dilakukan pertama kali adalah *feature selection* dengan memilih fitur-fitur yang akan dikelompokkan. Lalu, *feature scaling* perlu dilakukan dengan menggunakan `MinMaxScaler()`. Adapun metode yang digunakan adalah *K-Means* dan *Agglomerative Clustering*. Analisis *clustering* dilakukan dengan tujuan untuk menemukan kombinasi terbaik antara metode reduksi dimensi dan algoritma *clustering* berdasarkan *Silhouette Score*.

Tahap pertama yang dilakukan adalah memilih fitur yang akan digunakan untuk *clustering* dengan mengabaikan kolom negara dan mengambil fitur yang ada di dalam *feature selection* sebelumnya. Kemudian, dua metode reduksi dimensi (PCA dan t-SNE) dan dua metode *clustering* (*K-Means* dan *Agglomerative Clustering*) disiapkan dalam bentuk *dictionary*.



Selanjutnya, untuk setiap kombinasi metode reduksi dimensi dan *clustering* akan dilakukan evaluasi untuk berbagai jumlah *cluster* (dari 2 hingga 10) dengan menghitung *silhouette score*. Setiap iterasi mencatat hasil terbaik dengan *silhouette score* tertinggi dan mengupdate model terbaik yang ditemukan.

*Evaluation* dilakukan untuk memastikan kualitas dan efektivitas *cluster* yang dihasilkan. Dalam evaluasi ini pengguna dapat menggunakan *Silhouette Score* untuk mengukur seberapa baik tiap *data point* dikelompokkan dengan skor tinggi menunjukkan bahwa data tersebut cocok berada dalam model terbaik dan *cluster* yang ditentukan. Berdasarkan hasil evaluasi ini dapat mengartikan masing-masing *cluster* beserta analisis statistik per *cluster*, seperti kelompok negara dengan sosio-ekonomi yang baik atau negara dengan produktivitas energi bersih yang baik, serta memberikan rekomendasi kebijakan yang lebih spesifik sesuai dengan profil tiap *cluster*.

Evaluasi ini diharapkan dapat memberikan gambaran untuk anggota negara G20 berdasarkan fitur yang dipilih. Hasil evaluasi ini divisualisasikan dalam grafik yang menunjukkan perubahan *Silhouette Score* terhadap jumlah *cluster*, dengan garis vertikal untuk menunjukkan jumlah cluster terbaik. Setelah menemukan kombinasi terbaik, model tersebut diterapkan pada *dataset* penuh dan hasil *clustering* terbaik disimpan dalam kolom baru '*Cluster*'. Terakhir, kode menampilkan hasil terbaik beserta visualisasi untuk memudahkan pemahaman tentang *clustering* yang diperoleh.

## 4.2 Analisis Dataset

Penelitian ini menggunakan beberapa *dataset* yang didapatkan dari OECD dan Bank Dunia adalah sebagai berikut.

1. Data Sosio-Ekonomi dan Efisiensi Energi

*Dataset* OECD merupakan kumpulan data yang diperoleh dari *open dataset* pada [data-explorer.oecd.org](https://data-explorer.oecd.org). Adapun fitur-fitur yang terdapat dalam *dataset* ini adalah sebagai berikut.

**Tabel 4.1** Fitur *Dataset* OECD

No	Fitur	Jenis
1	STRUCTURE	Kategorik
2	STRUCTURE_ID	Kategorik
3	STRUCTURE_NAME	Kategorik
4	ACTION	Kategorik
5	REF_AREA	Kategorik
6	Reference area	Kategorik
7	FREQ	Kategorik
8	Frequency of observation	Kategorik
9	MEASURE	Kategorik
10	Measure	Kategorik
11	UNIT_MEASURE	Kategorik
12	Unit of measure	Kategorik
13	ACTIVITY	Kategorik
14	Economic activity	Kategorik
15	TIME_PERIOD	Numerik
16	Time period	Numerik
17	OBS_VALUE	Numerik
18	Observation value	Numerik
19	OBS_STATUS	Kategorik
20	Observation status	Kategorik
21	OBS_STATUS_2	Numerik
22	Observation status 2	Numerik
23	UNIT_MULT	Numerik
24	Unit multiplier	Kategorik
25	PRICE_BASE	Kategorik
26	Price base	Kategorik
27	BASE_PER	Numerik
28	Base period	Numerik
29	TIMELINESS	Numerik

30	Timeliness	Kategorik
31	DECIMALS	Numerik
32	Decimals	Kategorik

## 2. Data Tenaga Kerja Sektor Pertanian

*Dataset* Bank Dunia merupakan kumpulan data yang diperoleh dari *open dataset* pada [databank.worldbank.org](http://databank.worldbank.org). Adapun fitur-fitur yang terdapat dalam *dataset* ini adalah sebagai berikut.

**Tabel 4.2** Fitur *Dataset* Bank Dunia

No	Fitur	Jenis
1	Series Name	Kategorik
2	Series Code	Kategorik
3	Country Name	Kategorik
4	Country Code	Kategorik
5	2001 [YR2001]	Numerik
6	2002 [YR2002]	Numerik
7	2003 [YR2003]	Numerik
8	2004 [YR2004]	Numerik
9	2005 [YR2005]	Numerik
10	2006 [YR2006]	Numerik
11	2007 [YR2007]	Numerik
12	2008 [YR2008]	Numerik
13	2009 [YR2009]	Numerik
14	2010 [YR2010]	Numerik
15	2011 [YR2011]	Numerik
16	2012 [YR2012]	Numerik
17	2013 [YR2013]	Numerik
18	2014 [YR2014]	Numerik
19	2015 [YR2015]	Numerik
20	2016 [YR2016]	Numerik
21	2017 [YR2017]	Numerik

22	2018 [YR2018]	Numerik
23	2019 [YR2019]	Numerik
24	2020 [YR2020]	Numerik
25	2021 [YR2021]	Numerik
26	2022 [YR2022]	Kategorik
27	2023 [YR2023]	Kategorik

### 4.3 Platform Google Colaboratory

Pengembangan platform Google Colaboratory terdapat pada tautan <https://drive.google.com/view?usp=sharing>.

## **BAB V**

### **IMPLEMENTASI DAN PENGUJIAN**

#### **5.1 Kebutuhan Platform**

Kebutuhan platform menjelaskan berbagai macam kebutuhan yang diperlukan dalam platform yang akan dikembangkan dan diuji. Tujuannya adalah untuk memahami kebutuhan yang cukup dalam pengembangan dan pengujian platform, serta memberikan informasi mengenai spesifikasi perangkat yang harus dimiliki agar platform dapat berjalan dengan lancar. Dalam penerapan ini juga diperlukan perangkat keras dan perangkat lunak pendukung seperti berikut.

##### **5.1.1 Kebutuhan Perangkat Keras**

Perangkat keras yang digunakan dalam pengembangan tenaga kerja sektor pertanian terkait dengan sosio-ekonomi dan efisiensi energi di negara anggota G20 adalah sebagai berikut.

**Tabel 5.1** Kebutuhan Perangkat Keras

<i>Processor</i>	Intel Celeron N4020 1.1 GHz
<i>Memory</i>	4 GB
<i>Storage</i>	256 GB

##### **5.1.2 Kebutuhan Perangkat Lunak**

Perangkat lunak yang digunakan dalam pengembangan tenaga kerja sektor pertanian terkait dengan sosio-ekonomi dan efisiensi energi di negara anggota G20 adalah sebagai berikut.

**Tabel 5.2** Kebutuhan Perangkat Lunak

Sistem Operasi	Windows 11 64-bit
Bahasa Pemrograman	Python
Platform Notebook	Google Colaboratory
<i>Browser</i>	Google Chrome

## 5.2 Implementasi Platform

Implementasi program menjelaskan tentang alur proyek sains data dari platform sesuai dengan perancangan yang telah dibuat. Hasil dari implementasi ini adalah sebuah platform *notebook* yang siap untuk diuji coba dan digunakan. Berikut adalah gambaran mengenai alur implementasi platform pada proyek sains data yang telah siap digunakan.

### 5.2.1 Business Understanding

Sektor pertanian merupakan bagian penting dalam perekonomian banyak negara, khususnya dalam konteks negara-negara berkembang dan negara dengan sektor agraris yang besar. Selain itu, sektor ini juga berkontribusi pada dampak lingkungan, seperti emisi karbon dan konsumsi energi. Banyak negara G20 memiliki kebijakan yang mengarah pada peningkatan efisiensi energi dan keberlanjutan dalam sektor pertanian, namun hasilnya beragam.

Untuk memahami lebih lanjut bagaimana sektor pertanian mempengaruhi perekonomian dan efisiensi energi, penting untuk mengelompokkan negara-negara berdasarkan karakteristik yang berkaitan dengan sektor pertanian, ekonomi, dan energi. Pengelompokan ini bisa memberikan wawasan yang berharga mengenai bagaimana negara-negara berbeda dalam menerapkan kebijakan terkait pertanian dan energi, serta efektivitasnya dalam mendukung pertumbuhan ekonomi yang berkelanjutan.

### 5.2.2 Data Understanding

	STRUCTURE	STRUCTURE_ID	STRUCTURE_NAME	ACTION	REF_AREA	Reference area	FREQ	Frequency of observation	MEASURE	Measure
0	DATAFLOW	OECD.ENV.EPI.DSD_GG@DF_GREEN_GROWTH(1,1)	Green Growth	I	BRA	Brazil	A	Annual	NRGC	Energy consumption
1	DATAFLOW	OECD.ENV.EPI.DSD_GG@DF_GREEN_GROWTH(1,1)	Green Growth	I	BRA	Brazil	A	Annual	NRGC	Energy consumption
2	DATAFLOW	OECD.ENV.EPI.DSD_GG@DF_GREEN_GROWTH(1,1)	Green Growth	I	BRA	Brazil	A	Annual	NRGC	Energy consumption
3	DATAFLOW	OECD.ENV.EPI.DSD_GG@DF_GREEN_GROWTH(1,1)	Green Growth	I	BRA	Brazil	A	Annual	NRGC	Energy consumption
4	DATAFLOW	OECD.ENV.EPI.DSD_GG@DF_GREEN_GROWTH(1,1)	Green Growth	I	BRA	Brazil	A	Annual	NRGC	Energy consumption

5 rows × 32 columns

**Gambar 5.1** Impor *Dataset* OECD

*Dataset* utama yang bersumber pada data OECD pada tautan <https://data-explorer.oecd.org>. *Dataset* ini diakses data dari API SDMX milik OECD dan data diunduh dalam format CSV berdasarkan parameter yang ditentukan dalam URL. Hal selanjutnya ialah memuat data CSV tersebut ke dalam *pandas DataFrame* yang memudahkan analisis lebih lanjut dan mendeskripsikan data/fitur dalam *dataset* tersebut.

Series Name	Series Code	Country Name	Country Code	2001 [YR2001]	2002 [YR2002]	2003 [YR2003]	2004 [YR2004]	2005 [YR2005]	2006 [YR2006]	...	2014 [YR2014]	2015 [YR2015]	2016 [YR2016]
Employment in agriculture (% of total employe...	SL.AGR.EMPL.ZS	Argentina	ARG	1.065900	1.061939	1.021523	0.969933	0.919946	0.873533	...	0.677366	0.625514	0.608956
Employment in agriculture (% of total employe...	SL.AGR.EMPL.ZS	Australia	AUS	4.754306	4.382115	3.913121	3.737893	3.579841	3.411950	...	2.798319	2.636508	2.621225
Employment in agriculture (% of total employe...	SL.AGR.EMPL.ZS	Brazil	BRA	15.324849	15.407717	15.566475	16.043453	15.516146	14.495949	...	9.308502	9.332391	9.797930

**Gambar 5.2** Impor *Dataset* Bank Dunia

*Dataset* tambahan yang bersumber pada data Bank Dunia pada tautan <https://databank.worldbank.org>. *Dataset* ini bermula dari unduhan *file* Excel yang diunggah ke repositori GitHub penulis. Hal selanjutnya ialah memuat data XLSX tersebut ke dalam *pandas DataFrame* yang memudahkan analisis lebih lanjut dan mendeskripsikan data/fitur dalam *dataset* tersebut.

### 5.2.3 Data Preparation

```
# memilih beberapa feature dan disimpan dalam variabel baru yang bernama data_temp_oecd
data_temp_oecd = data_oecd[['Reference area', 'Measure', 'Unit of measure', 'TIME_PERIOD', 'OBS_VALUE']]

# menampilkan beberapa feature pertama dari DataFrame
data_temp_oecd.head()
```

	Reference area	Measure	Unit of measure	TIME_PERIOD	OBS_VALUE
0	Brazil	Energy consumption	Percentage of energy consumption	2001	37.05
1	Brazil	Energy consumption	Percentage of energy consumption	2002	38.13
2	Brazil	Energy consumption	Percentage of energy consumption	2003	39.27
3	Brazil	Energy consumption	Percentage of energy consumption	2004	39.54
4	Brazil	Energy consumption	Percentage of energy consumption	2005	39.34

**Gambar 5.3** *Feature Selection* Data OECD

Pada data OECD dilakukan pemilihan kolom untuk membantu menyediakan data yang diperlukan, menyederhanakan dataset, dan menghindari penggunaan kolom yang tidak relevan.

```
# membuat daftar kolom berdasarkan pola nama kolom
columns_to_select = ['Series Name', 'Country Name'] + [f'YR{year}' for year in range(2001, 2024)]

# memilih kolom yang diperlukan dan menyimpannya dalam data_temp_world_bank
data_temp_world_bank = data_world_bank[columns_to_select]

# menampilkan beberapa feature pertama dari DataFrame
data_temp_world_bank.head()
```

	Series Name	Country Name	2001 [YR2001]	2002 [YR2002]	2003 [YR2003]	2004 [YR2004]	2005 [YR2005]	2006 [YR2006]	2007 [YR2007]	2008 [YR2008]	...	2014 [YR2014]
0	Employment in agriculture (% of total employe...	Argentina	1.065900	1.061939	1.021523	0.969933	0.919946	0.873533	0.832650	0.800602	...	0.677366
1	Employment in agriculture (% of total employe...	Australia	4.754306	4.382115	3.913121	3.737893	3.579841	3.411950	3.320811	3.223327	...	2.798319

**Gambar 5.4** Feature Selection Data Bank Dunia

Pada data Bank Dunia dilakukan pemilihan kolom untuk membantu menyediakan data yang diperlukan, menyederhanakan dataset, dan menghindari penggunaan kolom yang tidak relevan.

```
# menampilkan beberapa instance pertama dari DataFrame
df_oecd.head()
```

	Measure	Year	Country	Consumption of biomass (Percentage of domestic material consumption)	Consumption of metals (Percentage of domestic material consumption)	Consumption of non-metallic minerals (Percentage of domestic material consumption)	Energy consumption (Percentage of energy consumption)	Energy intensity per capita (Tonnes of oil equivalent per person)	Energy productivity, GDP per unit of TES (US dollars per tonne of oil equivalent)	Female population (Percentage of population)	GDP deflator (Index)
0		2001	Argentina	76.94	9.34	13.72	19.998	1.54	16132.17	50.70	7.69
1		2001	Australia	43.30	36.84	19.86	20.000	5.49	7216.75	50.37	67.26
2		2001	Brazil	71.57	4.89	23.54	19.998	1.09	10409.30	50.34	32.88
3		2001	Canada	22.89	23.26	53.86	20.000	8.01	4884.65	50.48	75.53
4		2001	China	26.18	6.22	67.59	20.000	0.92	4221.28	48.91	59.32

**Gambar 5.5** Data Jadi pada *Dataset* OECD

Transformasi data pada *dataset* OECD dilakukan dengan melakukan *pivot table* untuk membantu dalam pengonversian data mentah menjadi format yang lebih mudah dibaca dan memungkinkan analisis yang lebih mendalam. Me-reset *index* adalah langkah penting dalam menyederhanakan struktur *DataFrame* terutama setelah melakukan operasi *pivot* yang mengubah beberapa kolom menjadi *index*. Kemudian, beberapa kolom



diganti namanya, seperti tulisan ‘TIME\_PERIOD’ menjadi ‘Year’, tulisan ‘Reference area’ menjadi ‘Country’, dan nama-nama negara menggunakan nama resmi yang sederhana.

```
# menampilkan beberapa instance pertama dari DataFrame
df_world_bank.head()
```

Series Name	Country Name	Agricultural land (% of land area)	Agricultural land (sq. km)	Agricultural raw materials exports (% of merchandise exports)	Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	Employment in agriculture (% of total employment) (modeled ILO estimate)	Year
0	Argentina	46.993266	1286060.0	1.61216	0.869	1.0659	2001
1	Australia	59.318173	4557000.0	6.235404	1.3771	4.754306	2001
2	Brazil	27.329908	2284272.0	4.163496	4.6632	15.324849	2001
3	Canada	6.838162	613081.608	5.970256	2.1065	2.160145	2001
4	China	55.757872	5234668.444	0.903333	24.6966	50.0	2001

**Gambar 5.6** Data Jadi pada *Dataset* Bank Dunia

Transformasi data pada *dataset* Bank Dunia dilakukan dengan melakukan *pivot table* untuk membantu dalam pengonversian data mentah menjadi format yang lebih mudah dibaca dan memungkinkan analisis yang lebih mendalam. Me-reset *index* adalah langkah penting dalam menyederhanakan struktur *DataFrame* terutama setelah melakukan operasi *pivot* yang mengubah beberapa kolom menjadi *index*. Kemudian, beberapa kolom diganti namanya, seperti tulisan ‘Reference area’ menjadi ‘Country’, lambang ‘%’ menjadi tulisan ‘Percentage’, nama-nama negara menggunakan nama resmi yang lebih sederhana, dan mengganti tipe data selain fitur ‘Country’ dan ‘Year’ menjadi *float*.

```
# men-setting negara-negara df_oecd dan df_world_bank agar sama dan disimpan di variabel common_countries
common_countries = set(oecd_countries) & set(world_bank_countries)

# mencetak jumlah negara
print(f"Number of common countries: {len(common_countries)}")

# mengurutkan negara-negara berdasarkan abjad
print("Common countries:")
sorted(list(common_countries))
```

```
Number of common countries: 20
Common countries:
['Argentina',
 'Australia',
 'Brazil',
 'Canada',
 'China',
 'European Union',
 'France',
 'Germany',
 'India',
 'Indonesia',
```

**Gambar 5.7** Memastikan Nama Negara pada Dua *Dataset*

Sebelum menggabungkan *dataset*, penulis akan mencari negara yang ada di kedua dataset, menghitung jumlahnya, dan mengurutkan daftar negara tersebut. Hal ini sangat berguna untuk membandingkan dua sumber data dan mengonfirmasi kesesuaian negara yang tercatat dalam masing-masing *dataset*.

```
# mengubah tipe data feature 'Year' menjadi integer pada kedua DataFrame
df_oecd['Year'] = df_oecd['Year'].astype(int)
df_world_bank['Year'] = df_world_bank['Year'].astype(int)

# melakukan inner join kedua DataFrame berdasarkan feature 'Country' dan 'Year'
df = pd.merge(df_world_bank, df_oecd, on=['Country', 'Year'], how='inner')

# menampilkan beberapa instance pertama dari DataFrame
df.head()
```

	Country	Agricultural land (Percentage of land area)	Agricultural land (sq. km)	Agricultural raw materials exports (Percentage of merchandise exports)	Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	Employment in agriculture (Percentage of total employment) (modeled ILO estimate)	Year	Consumption of biomass (Percentage of domestic material consumption)	Consumption of metals (Percentage of domestic material consumption)	Consumption of non-metallic minerals (Percentage of domestic material consumption)
0	Argentina	46.993266	1286060.000	1.612160	0.8690	1.065900	2001	76.94	9.34	13.72
1	Australia	59.318173	4557000.000	6.235404	1.3771	4.754306	2001	43.30	36.84	19.86
2	Brazil	27.329908	2284272.000	4.163496	4.6632	15.324849	2001	71.57	4.89	23.54

**Gambar 5.8** Menggabungkan Dua *Dataset*

Penggabungan kedua *dataset* dilakukan dengan memastikan bahwa kolom ‘*Year*’ memiliki tipe data yang sama di kedua *DataFrame*, lalu menggunakan *inner join* untuk menggabungkan data berdasarkan kolom ‘*Country*’ dan ‘*Year*’. Penggabungan ini menghasilkan satu *dataset* dan antar fitur saling bersatu.

	Missing Values	Percentage Missing
Country	0	0.0
Agricultural land (Percentage of land area)	0	0.0
Agricultural land (sq. km)	0	0.0
Agricultural raw materials exports (Percentage of merchandise exports)	0	0.0
Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	0	0.0
Employment in agriculture (Percentage of total employment) (modeled ILO estimate)	0	0.0
Year	0	0.0
Consumption of biomass (Percentage of domestic material consumption)	0	0.0
Consumption of metals (Percentage of domestic material consumption)	0	0.0
Consumption of non-metallic minerals (Percentage of domestic material consumption)	0	0.0
Energy consumption (Percentage of energy consumption)	0	0.0
Energy intensity per capita (Tonnes of oil equivalent per person)	0	0.0
Energy productivity, GDP per unit of TES (US dollars per tonne of oil equivalent)	0	0.0
Female population (Percentage of population)	0	0.0

**Gambar 5.9** Memperbaiki *Missing Values*

Pada tahap *data cleaning* diperlukan aksi untuk mengetahui fitur mana yang memiliki *missing values*. Maka dari itu, pembersihan data yang dilakukan ialah menghapus fitur dengan persentase nilai hilang di atas 30%, mengambil *instance* dengan kurang dari 30% nilai yang hilang, mengisi semua nilai hilang dengan *forward fill* dan *backward fill* berdasarkan fitur ‘Country’, dan mengisi sisa nilai yang hilang dengan nilai 0.

	Year	Country	Agricultural land (Percentage of land area)	Agricultural land (sq. km)	Agricultural raw materials exports (Percentage of merchandise exports)	Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	Employment in agriculture (Percentage of total employment) (modeled ILO estimate)	Real GDP (US dollars, PPP converted)	Value added (Percentage of gross value added)	Labour tax revenue (Percentage of GDP)	...
0	2001	Argentina	46.993266	1286060.0	1.612160	0.8690	1.065900	921659.19	31.630000	2.54	...
1	2002	Argentina	47.031268	1287100.0	1.550192	0.8453	1.061939	704802.19	31.850000	2.55	...
2	2003	Argentina	47.174981	1291033.0	1.516745	1.1625	1.021523	637257.69	31.336667	3.22	...
3	2004	Argentina	47.318695	1294966.0	1.629307	1.3602	0.969933	581715.31	28.340000	4.09	...
4	2005	Argentina	47.473737	1299209.0	1.362289	1.0349	0.919946	633206.69	28.353333	4.40	...

5 rows x 33 columns

**Gambar 5.10 Data Filtering**

*Data filtering* dilakukan untuk membuang fitur yang tidak relevan dan meningkatkan pemahaman tentang fitur yang akan dipilih dan masih berkaitan dengan konteks utama. Fitur yang dibuang biasanya mempunyai satuan indeks dan fitur yang bermakna ganda tetapi tetap mempunyai hubungan yang sama.

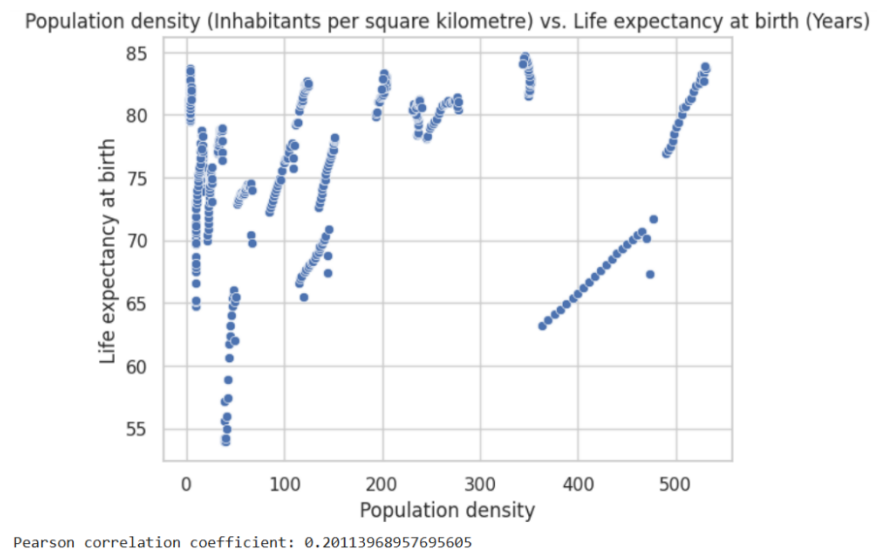
```
# mengatur format tampilan angka untuk tidak menggunakan notasi ilmiah
pd.options.display.float_format = '{:.2f}'.format

# menampilkan statistical five summaries
df_eda.describe().loc[['min', '25%', '50%', '75%', 'max']].T.round(2)
```

	min	25%	50%	75%	max
Year	2001.00	2006.00	2011.50	2017.00	2022.00
Agricultural land (Percentage of land area)	6.37	27.35	47.96	55.78	80.84
Agricultural land (sq. km)	16030.00	173796.84	963410.00	2154940.00	5282176.00
Agricultural raw materials exports (Percentage of merchandise exports)	0.02	0.61	0.95	2.46	7.51
Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	0.12	0.90	2.25	5.05	31.13
Employment in agriculture (Percentage of total employment) (modeled ILO estimate)	0.57	2.26	5.22	16.37	58.85
Real GDP (US dollars, PPP converted)	499047.91	1431436.12	2352502.50	3575444.94	26887390.00
Value added (Percentage of gross value added)	27.88	33.33	33.33	33.33	33.34

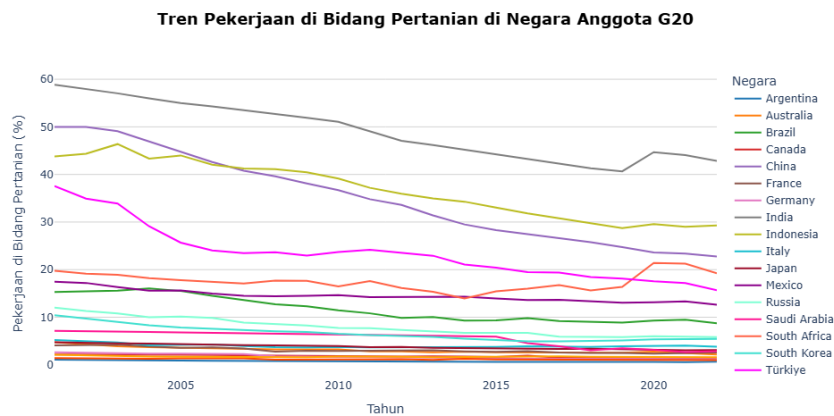
**Gambar 5.11 Statistical Summaries pada EDA**

*Statistical five summaries* memberikan gambaran yang jelas tentang distribusi data. Pengaturan tampilan angka dengan dua desimal membantu untuk menghindari kebingungannya tampilan angka dalam notasi ilmiah, membuat hasil lebih mudah dipahami, terutama bagi pembaca yang tidak terbiasa dengan format tersebut.



**Gambar 5.12** *Bivariate Analysis*

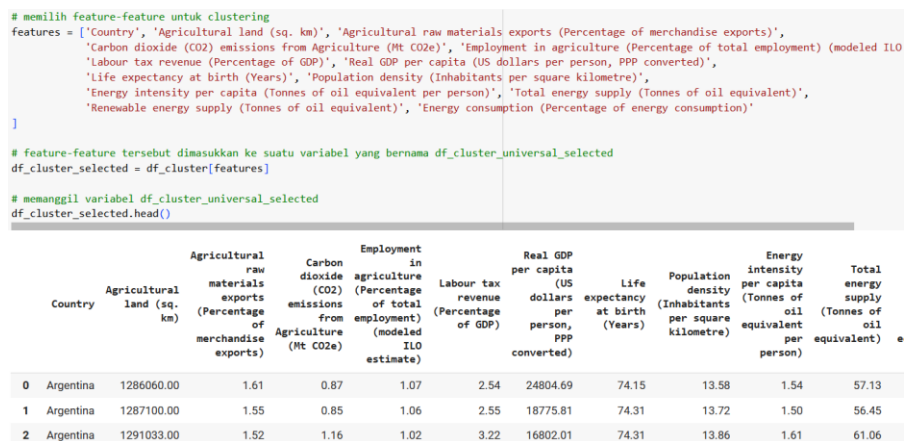
*Bivariate analysis* digunakan untuk mengetahui hubungan antara dua variabel. Beberapa fitur yang dilakukan untuk *bivariate analysis* ini ialah kepadatan penduduk dengan angka harapan hidup, produk domestik bruto dengan pajak tenaga kerja, dan sebagainya. Jika hasil korelasi positif, maka kedua fitur tersebut saling mendukung, dan sebaliknya.



**Gambar 5.13** *Multivariate Analysis*

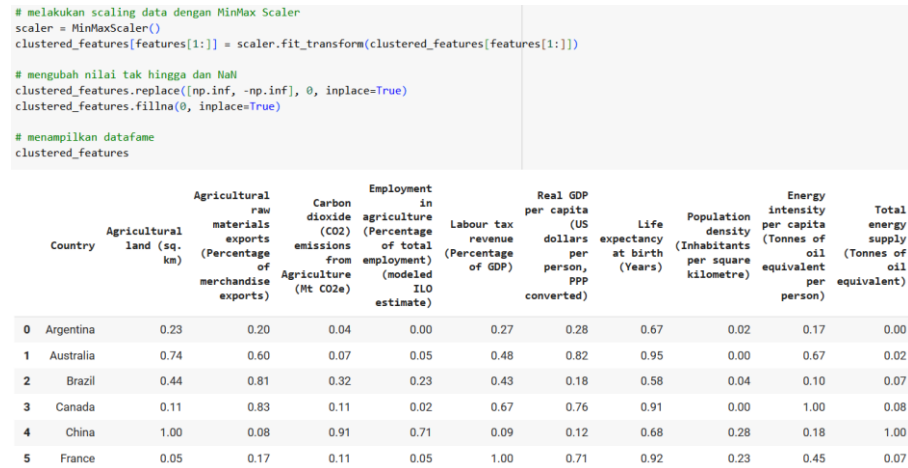
*Multivariate analysis* digunakan untuk mengetahui hubungan antara tiga atau lebih variabel. Beberapa hal yang dilakukan untuk *multivariate analysis* ini ialah tren pekerjaan di sektor pertanian di negara anggota G20, tren pasokan energi terbarukan 5 negara dengan pasokan energi terbarukan tertinggi, dan sebagainya.

#### 5.2.4 Modeling dan Evaluation



**Gambar 5.14** *Feature Selection*

Pemilihan sejumlah fitur yang relevan untuk analisis *clustering* dilakukan dengan berfokus pada data yang berkaitan dengan pertanian, sosio-ekonomi, dan energi.



**Gambar 5.15** *Feature Scaling*

Sebelum melakukan *feature scaling*, diperlukan penghitungan rata-rata dari setiap fitur yang relevan untuk setiap negara dalam *DataFrame* untuk mengelompokkan data berdasarkan kolom ‘*Country*’. Hasilnya adalah *DataFrame* memberikan gambaran umum tentang karakteristik rata-rata setiap negara terkait fitur-fitur yang dipilih.

Data yang ada pada fitur-fitur numerik (selain ‘*Country*’) diubah ke dalam rentang nilai antara 0 dan 1 menggunakan *MinMaxScaler*. Tujuannya adalah untuk menyamakan skala dari setiap fitur sehingga tidak ada fitur yang dominan atau lebih besar nilainya daripada yang lainnya. Ini sangat penting dalam clustering karena algoritma seperti K-Means sensitif terhadap skala fitur yang berbeda.

Setelah proses *scaling*, nilai tak hingga dan hilang diubah menjadi 0. Ini membantu memastikan bahwa data yang digunakan untuk analisis *clustering* tidak terpengaruh oleh nilai tidak valid atau hilang yang bisa menyebabkan *error* dalam pemrosesan.

```

# menentukan feature yang akan digunakan untuk clustering (kecuali 'Country')
features_for_clustering = features[1:]

# membuat dictionary metode pengurangan dimensionalitas
dim_reduction_methods = {
    'PCA': PCA(n_components=2), # mengurangi menjadi 2 komponen
    't-SNE': TSNE(n_components=2, perplexity=1, random_state=42)
}

# membuat dictionary metode clustering
clustering_methods = {
    'KMeans': KMeans(random_state=42),
    'Agglomerative': AgglomerativeClustering()
}

best_dim_reduction = None
best_clustering_method = None
best_n_clusters = None
best_silhouette_score = -1

```

**Gambar 5.16** *Dimensionality Reduction dan Clustering Model Selection*

Dalam pemodelan *clustering* ini menggunakan dua *dictionary* yang terdiri *dimensionality reduction* dan *clustering model*. *Dimensionality reduction* yang dipakai adalah PCA dan t-SNE. Untuk model *clustering* akan menggunakan *K-Means* dan *Agglomerative Clustering*. Tujuan membuat dua *dictionaries* tersebut adalah untuk memilih metode yang menghasilkan *dimensionality reduction* dan *clustering* terbaik berdasarkan iterasi evaluasi *silhouette score* yang menunjukkan kualitas pemisahan antar *cluster*.

```

for dim_method_name, dim_method in dim_reduction_methods.items():
    reduced_data = dim_method.fit_transform(clustered_features[features_for_clustering])

    for cluster_method_name, cluster_method in clustering_methods.items():
        silhouette_scores = []
        print(f"Evaluating {cluster_method_name} clustering with {dim_method_name} dimensionality reduction")
        for n_clusters in range(2, 11): # menguji n_cluster dari 2 hingga 10

            if cluster_method_name in ['KMeans', 'Agglomerative']:
                cluster_method.set_params(n_clusters=n_clusters)

            cluster_labels = cluster_method.fit_predict(reduced_data)
            silhouette_avg = silhouette_score(reduced_data, cluster_labels)
            silhouette_scores.append(silhouette_avg)

            # mencetak iterasi saat ini dengan silhouette score rata-rata
            print(f"{cluster_method_name} Clusters: {n_clusters}, Average Silhouette Score: {silhouette_avg}")

            if silhouette_avg > best_silhouette_score:
                best_silhouette_score = silhouette_avg
                best_dim_reduction = dim_method_name
                best_clustering_method = cluster_method_name
                best_n_clusters = n_clusters

# visualisasi
plt.plot(range(2, 11), silhouette_scores, marker='o', linestyle='--', color='blue')
plt.axvline(x=best_n_clusters, linestyle='--', color='orange', label=f'Model: {best_n_clusters} Clusters')
for i, score in enumerate(silhouette_scores):
    plt.text(range(2, 11)[i], silhouette_scores[i] + 0.002, f'{silhouette_scores[i]:.4f}',
             ha='center', va='bottom', fontsize=12, color='black')
plt.xlabel('Number of Clusters')
plt.ylabel('Silhouette Score')
plt.title(f'Silhouette Score for {cluster_method_name} Clustering ({dim_method_name})')
plt.show()
print()

```

**Gambar 5.17** *Iterasi Model Clustering*

Iterasi model *clustering* ditujukan untuk menemukan kombinasi antara teknik reduksi dimensi dan model *clustering* berdasarkan kinerja *silhouette score*. *Silhouette score* digunakan untuk menilai kualitas pemisahan *cluster*. Pada iterasi ini juga memberikan gambaran terkait performa masing-masing metode *clustering* dengan jumlah cluster yang berbeda.

```
# menyesuaikan model terbaik dengan seluruh dataset
best_dim_method = dim_reduction_methods[best_dim_reduction]
reduced_data = best_dim_method.fit_transform(clustered_features[features_for_clustering])

if best_clustering_method == "KMeans":
    best_cluster_method = KMeans(n_clusters=best_n_clusters, random_state=42)
elif best_clustering_method == "Agglomerative":
    best_cluster_method = AgglomerativeClustering(n_clusters=best_n_clusters)

cluster_labels = best_cluster_method.fit_predict(reduced_data)
clustered_features["Cluster"] = cluster_labels

print(f"Best Dimensionality Reduction: {best_dim_reduction}")
print(f"Best Clustering Method: {best_clustering_method}")
print(f"Best Number of Clusters: {best_n_clusters}")
print(f"Best Silhouette Score: {best_silhouette_score}")

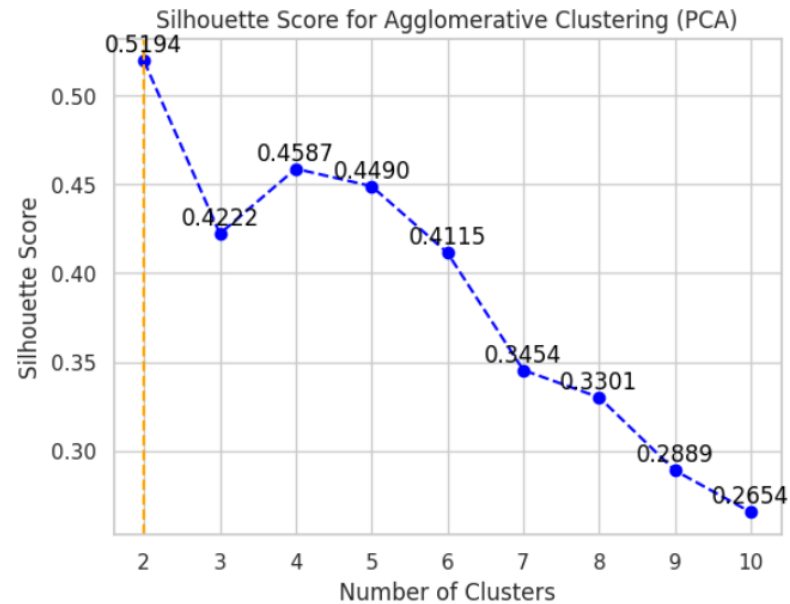
# visualisasi akhir dari metode clustering terbaik
plt.plot(range(2, 11), silhouette_scores, marker='o', linestyle='--', color='blue')
plt.axvline(x=best_n_clusters, linestyle='--', color='orange', label=f'Model: {best_n_clusters} Clusters')
for i, score in enumerate(silhouette_scores):
    plt.text(range(2, 11)[i], silhouette_scores[i] + 0.002, f'{silhouette_scores[i]:.4f}',
             ha='center', va='bottom', fontsize=12, color='black')
plt.xlabel('Number of Clusters')
plt.ylabel('Silhouette Score')
plt.title(f'Best Clustering with Silhouette Score for ({best_dim_reduction}, {best_clustering_method}, n_clusters={best_n_clusters})')
plt.show()
```

**Gambar 5.18** Pemilihan Teknik Reduksi Dimensi dan Model Terbaik

Setelah melakukan iterasi model clustering, maka dilakukan pemilihan teknik reduksi dimensi dan model *clustering* terbaik berdasarkan hasil yang diperoleh sebelumnya. Nilai *silhouette score* yang lebih tinggi menunjukkan *cluster* yang lebih baik.



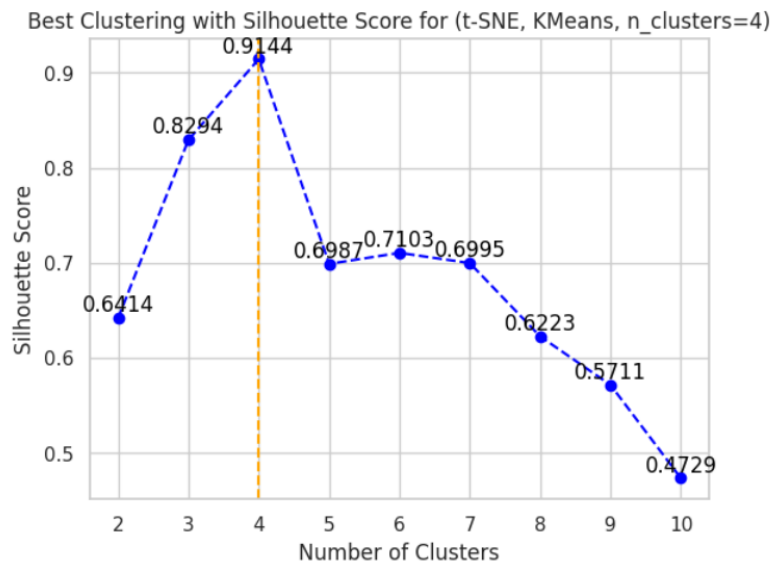
Evaluating Agglomerative clustering with PCA dimensionality reduction  
 Agglomerative Clusters: 2, Average Silhouette Score: 0.519417663029896  
 Agglomerative Clusters: 3, Average Silhouette Score: 0.4221993692503389  
 Agglomerative Clusters: 4, Average Silhouette Score: 0.4587499446266392  
 Agglomerative Clusters: 5, Average Silhouette Score: 0.44895276275833124  
 Agglomerative Clusters: 6, Average Silhouette Score: 0.4114547734074532  
 Agglomerative Clusters: 7, Average Silhouette Score: 0.34542816268266047  
 Agglomerative Clusters: 8, Average Silhouette Score: 0.33006563670386485  
 Agglomerative Clusters: 9, Average Silhouette Score: 0.28888845347621694  
 Agglomerative Clusters: 10, Average Silhouette Score: 0.26537891789192036



**Gambar 5.19** Hasil Iterasi Model *Clustering*

Iterasi kombinasi teknik reduksi dimensi dan metode *clustering* memberikan nilai silhouette score di setiap jumlah *cluster* beserta visualisasi grafiknya. Iterasi ini memberikan hasil untuk masing-masing teknik reduksi dimensi PCA dan t-SNE dengan model masing-masing *clustering K-Means* dan *Agglomerative Clustering*.

Best Dimensionality Reduction: t-SNE  
 Best Clustering Method: KMeans  
 Best Number of Clusters: 4  
 Best Silhouette Score: 0.914484996795654



**Gambar 5.20** Pemilihan Hasil Evaluasi Model Terbaik

Pemilihan hasil evaluasi model terbaik ditemukan pada iterasi sebelumnya menggunakan seluruh *dataset*. Setelah itu, dilakukan pemilihan prediksi *cluster* menggunakan model terbaik dan visualisasi dari *silhouette score* untuk hasil akhir.

Statistik Deskriptif per Cluster:

Cluster	Agricultural land (sq. km)	Agricultural raw materials exports (Percentage of merchandise exports)	Carbon dioxide (CO2) emissions from Agriculture (Mt CO2e)	Employment in agriculture (Percentage of total employment) (modeled ILO estimate)	Labour tax revenue (Percentage of GDP)	Real GDP per capita (US dollars per person, ppp converted)	Life expectancy at birth (Years)	Population density (Inhabitants per square kilometre)	Energy intensity per capita (Tonnes of oil equivalent per person)
0	0.33	0.34	0.28	0.40	0.22	0.26	0.49	0.18	0.24
1	0.54	0.62	0.14	0.03	0.59	0.86	0.88	0.02	0.85
2	0.02	0.13	0.02	0.04	0.78	0.75	0.93	0.57	0.43
3	0.03	0.13	0.05	0.06	0.72	0.67	0.91	0.53	0.42

**Gambar 5.21** Statistik Deskriptif per *Cluster*

Statistik deskriptif per *cluster* menampilkan hasil cluster berdasarkan grup *cluster*-nya. Dengan ini dapat diketahui masing-masing karakteristik tingkat variabel satu sama lain diantara *cluster*.

Cluster 0:

- Argentina
- Brazil
- China
- India
- Indonesia
- Mexico
- Russia
- Saudi Arabia
- South Africa
- Türkiye

Cluster 1:

- Australia
- Canada
- United States

Cluster 2:

- Germany
- Japan

Cluster 3:

- France
- Italy
- South Korea
- United Kingdom

**Gambar 5.22** Pengelompokan Negara berdasarkan *Cluster*

Pengelompokan negara berdasarkan *cluster* menggunakan data yang telah ditetapkan sebelumnya dan kemudian mencetak daftar negara yang termasuk dalam setiap *cluster*. Pengelompokan ini membantu pembaca dalam memahami negara dikelompokkan berdasarkan fitur yang digunakan dalam proses *clustering*.

### 5.3 Pengujian Platform

#### 5.3.1 Perbandingan Algoritma *Clustering*

Adapun perbandingan algoritma *K-Means* dan *Agglomerative Clustering* menggunakan teknik reduksi PCA adalah sebagai berikut.

**Tabel 5.3** Algoritma *Clustering* dengan Teknik Reduksi PCA

K-Means	Agglomerative	Jumlah <i>Cluster</i>
0.5631	0.5194	2
0.4783	0.4221	3
0.5011	0.4587	4
0.4132	0.4489	5

0.3718	0.4114	6
0.3454	0.3454	7
0.2892	0.3300	8
0.2657	0.2888	9
0.2653	0.2653	10

Algoritma terbaik untuk metode PCA adalah *K-Means clustering* dan mencapai *silhouette score* 0,5631 dengan 2 *cluster*. Adapun perbandingan algoritma *K-Means* dan *Agglomerative Clustering* menggunakan teknik reduksi t-SNE adalah sebagai berikut.

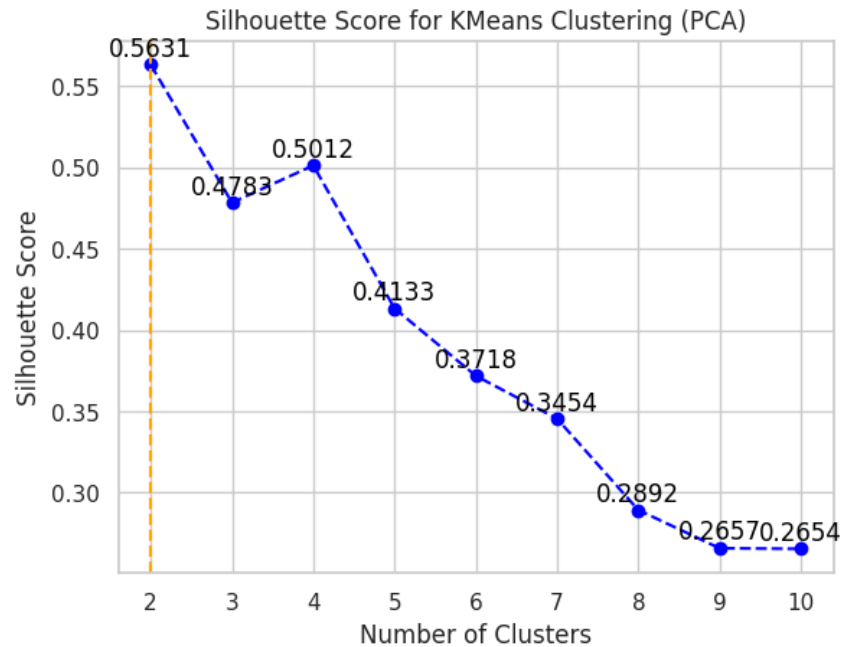
**Tabel 5.4** Algoritma *Clustering* dengan Teknik Reduksi t-SNE

K-Means	Agglomerative	Jumlah <i>Cluster</i>
0.6413	0.6413	2
0.8293	0.8293	3
0.9144	0.9144	4
0.6987	0.6987	5
0.7066	0.7103	6
0.6293	0.6995	7
0.5523	0.6222	8
0.5109	0.5711	9
0.4126	0.4728	10

Algoritma terbaik untuk metode t-SNE adalah *K-Means clustering* dan mencapai *silhouette score* 0,9144 dengan 4 *cluster*.

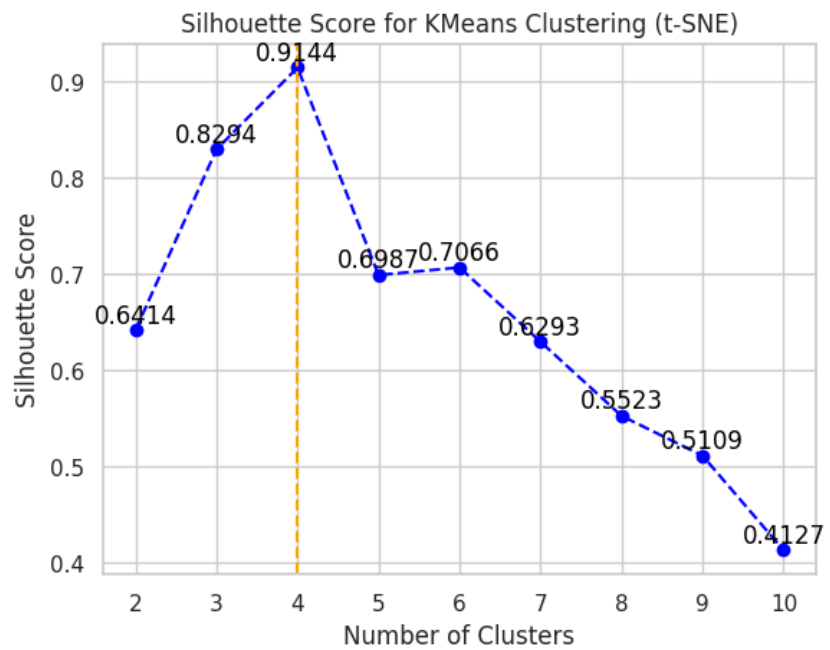
### 5.3.2 Perbandingan Teknik Reduksi Dimensi

Adapun perbandingan teknik reduksi dimensi PCA adalah sebagai berikut.



**Gambar 5.23** Teknik Reduksi Dimensi PCA

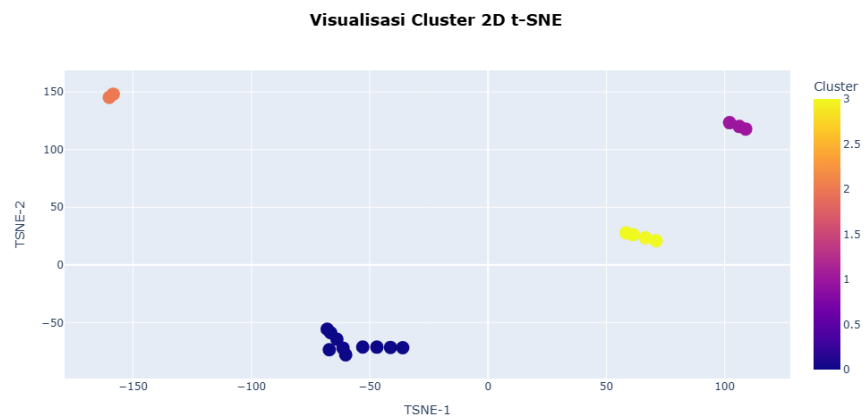
PCA menunjukkan performa *silhouette score* sebesar 0,5631 dengan jumlah *cluster* 2. Tetapi, untuk jumlah berikutnya belum ada *silhouette score* yang di atas 0,5631. Apabila hanya ada 2 *cluster*, maka pengelompokkan kurang bervariasi. Adapun perbandingan teknik reduksi dimensi t-SNE adalah sebagai berikut.



**Gambar 5.24** Teknik Reduksi Dimensi t-SNE

t-SNE menunjukkan performa yang lebih unggul dibandingkan PCA dengan *silhouette score* yang lebih tinggi sebesar 0,9144. Hal ini menunjukkan pemisahan *cluster* yang lebih baik.

### 5.3.3 Visualisasi Hasil *Clustering*



**Gambar 5.25** Visualisasi Hasil *Cluster*

Visualisasi t-SNE 2D ini merupakan hasil *clustering* dengan mengelompokkan negara anggota G20 berdasarkan model *K-Means Clustering* dan menjadi 4 *cluster* berbeda. Hasil evaluasi menunjukkan bahwa *Cluster 0* (Argentina, Brasil, Tiongkok, India, Indonesia, Meksiko, Rusia, Arab Saudi, Afrika Selatan, dan Turki) berisi negara-negara dengan ketergantungan tinggi pada sektor pertanian namun memiliki efisiensi energi rendah.

*Cluster 1* (Australia, Kanada, dan Amerika Serikat) mencakup negara-negara maju dengan sektor pertanian yang relatif lebih kecil, tetapi efisiensi energi dan tingkat kemakmuran yang tinggi. *Cluster 2* (Jerman dan Jepang) dan *Cluster 3* (Perancis, Italia, Korea Selatan, dan Britania Raya) mencerminkan negara-negara dengan karakteristik yang lebih variatif dalam sektor pertanian dan efisiensi energi, dengan *Cluster 3* (Perancis, Italia, Korea Selatan, dan Britania Raya) menunjukkan negara dengan tingkat kehidupan yang lebih tinggi dan infrastruktur energi yang lebih berkembang.

## **BAB VI**

### **PENUTUP**

#### **6.1 Kesimpulan**

Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa sektor pertanian memberikan kontribusi besar terhadap sosio-ekonomi dan lingkungan di negara-negara G20. Proses *clustering* yang diterapkan berhasil mengelompokkan negara-negara G20 berdasarkan karakteristik sektor pertanian, sosio-ekonomi, dan efisiensi energi negara anggota tersebut.

Ditemukan empat *cluster* utama dengan ciri-ciri yang berbeda. *Cluster 0* (Argentina, Brasil, Tiongkok, India, Indonesia, Meksiko, Rusia, Arab Saudi, Afrika Selatan, dan Turki) mencakup negara-negara yang sangat bergantung pada sektor pertanian, namun memiliki efisiensi energi yang rendah serta tingkat emisi karbon yang relatif tinggi. *Cluster 1* (Australia, Kanada, dan Amerika Serikat) terdiri dari negara-negara maju yang sektor pertaniannya lebih kecil, namun memiliki efisiensi energi yang tinggi dan kemakmuran yang tinggi pula.

*Cluster 2* (Jerman dan Jepang) dan *Cluster 3* (Perancis, Italia, Korea Selatan, dan Britania Raya) menggambarkan negara-negara dengan karakteristik yang lebih beragam, di mana *Cluster 3* (Perancis, Italia, Korea Selatan, dan Britania Raya) menunjukkan negara-negara dengan tingkat kehidupan yang lebih baik dan infrastruktur energi yang lebih maju. Hasil analisis ini memperlihatkan adanya hubungan yang jelas antara ketergantungan pada sektor pertanian, efisiensi energi, dan dampaknya terhadap keberlanjutan sosio-ekonomi dan lingkungan. Negara-negara yang lebih bergantung pada pertanian cenderung memiliki emisi karbon yang lebih tinggi dan efisiensi energi yang lebih rendah, sementara negara-negara maju umumnya memiliki efisiensi energi yang lebih tinggi dan emisi yang lebih rendah.



## 6.2 Saran

Berdasarkan hasil analisis yang diperoleh, beberapa saran berikut dapat diberikan untuk membantu negara-negara G20 mengatasi tantangan yang dihadapi dalam sektor pertanian, sosio-ekonomi, dan efisiensi energi.

Pengembangan teknologi ramah lingkungan di negara bergantung pada pertanian. Negara-negara yang sangat bergantung pada sektor pertanian, seperti dalam *Cluster 0* (Argentina, Brasil, Tiongkok, India, Indonesia, Meksiko, Rusia, Arab Saudi, Afrika Selatan, dan Turki) sebaiknya meningkatkan investasi dalam teknologi hijau yang dapat mendukung efisiensi energi dan mengurangi dampak negatif terhadap lingkungan. Penerapan teknologi yang lebih ramah lingkungan dalam sektor pertanian dapat meningkatkan hasil dan mengurangi emisi karbon.

Keberlanjutan sektor pertanian di negara maju mesti dapat lebih difokuskan. Negara-negara maju yang sudah memiliki infrastruktur energi yang baik, seperti dalam *Cluster 1* (Australia, Kanada, dan Amerika Serikat) dan *Cluster 3* (Perancis, Italia, Korea Selatan, dan Britania Raya) perlu memperkuat sektor pertanian mereka dengan mengurangi jejak karbon dan memaksimalkan penggunaan sumber daya yang efisien. Kebijakan yang mendukung praktik pertanian berkelanjutan serta penggunaan energi terbarukan dapat menjadi langkah strategis.

Kerja sama internasional dibutuhkan untuk penyebaran pengetahuan secara masif: Negara-negara G20, baik negara yang maju maupun berkembang, harus bekerja sama untuk berbagi pengetahuan dan teknologi yang dapat membantu negara-negara yang bergantung pada pertanian dalam meningkatkan efisiensi energi mereka. Pembentukan kemitraan internasional untuk berbagi teknologi dan praktik terbaik akan mempercepat adopsi sistem pertanian yang lebih berkelanjutan.

Penguatan kebijakan energi dan pertanian yang terpadu diperlukan terhadap negara-negara G20. Negara anggota G20 perlu mengembangkan kebijakan yang lebih menyeluruh dan saling terkait antara sektor energi dan pertanian. Pendekatan ini akan menciptakan sistem yang lebih harmonis di mana sektor pertanian dapat beroperasi secara lebih efisien dan ramah lingkungan, sementara sektor energi mendukung transisi tersebut.

Dengan memaparkan rekomendasi-rekomendasi ini, negara-negara G20 diharapkan dapat memaksimalkan potensi sektor pertanian, meningkatkan efisiensi energi, dan mengurangi dampak negatif terhadap lingkungan, sehingga berkontribusi pada pencapaian tujuan pembangunan berkelanjutan yang lebih baik.

## DAFTAR PUSTAKA

- [1] A. Rachman, E. Yochanan, A. I. Samanlangi and H. Purnomo, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*, Karawang: Saba Jaya Publisher, 2024.
- [2] A. E. Hill, I. Ornelas and J. E. Taylor, "Agricultural Labor Supply," *Annual Review of Resource Economics*, vol. 13, no. 1, pp. 39-64, July 2021, doi: [10.1146/annurev-resource-101620-080426](https://doi.org/10.1146/annurev-resource-101620-080426).
- [3] Sayifullah and Emmalian, "Pengaruh Tenaga Kerja Sektor Pertanian dan Pengeluaran Pemerintah Sektor Pertanian Terhadap Produk Domestik Bruto Sektor Pertanian di Indonesia," *Jurnal Ekonomi-Qu*, vol. 8, no. 1, pp. 66-81, April 2018, doi: [10.35448/jequ.v8i1.4962](https://doi.org/10.35448/jequ.v8i1.4962).
- [4] I. Setiawan, "Peran Sektor Pertanian Dalam Penyerapan Tenaga Kerja DI Indonesia," *Jurnal Geografi Gea*, vol. 6, no. 1, pp. 1-6, 2006, doi: [10.17509/gea.v6i1.1733](https://doi.org/10.17509/gea.v6i1.1733).
- [5] S. H. Susilowati, "Fenomena Penuaan Petani dan Berkurangnya Tenaga Kerja Muda serta Implikasinya bagi Kebijakan Pembangunan Pertanian," *Forum Penelitian Agro Ekonomi*, vol. 34, no. 1, pp. 35-55, June 2016.
- [6] B. A. Bachtiar, F. S. Haq, M. Janah, N. R. Amalia, J. Novaldi and Budiasih, "Food Crop Agriculture Sector Labor Absorption In Generation Z," in *Seminar Nasional Official Statistics*, Jakarta, 2023, doi: [10.34123/semnasoffstat.v2023i1.1706](https://doi.org/10.34123/semnasoffstat.v2023i1.1706).
- [7] G20, "Overview," G20.org, [Online]. Available: <https://g20.org/about-g20/overview/>. [Accessed 13 February 2025].
- [8] Kementerian Keuangan Republik Indonesia, "G20," [Online]. Available: <https://www.kemenkeu.go.id/g20>. [Accessed 13 February 2025].
- [9] V. Dhar, "Data science and prediction," *Communications of the ACM*, vol. 56, no. 12, pp. 64-73, December 2013, doi: [10.1145/2500499](https://doi.org/10.1145/2500499).
- [10] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Mary Ann Liebert, Inc.*, vol. 1, no. 1, pp. 51-59, February 2013, doi: [10.1089/big.2013.1508](https://doi.org/10.1089/big.2013.1508).
- [11] F. Balali, J. Nouri, A. Nasiri and T. Zhao, *Data Intensive Industrial Asset Management*, Cham: Springer Cham, 2020, doi: [10.1007/978-3-030-35930-0](https://doi.org/10.1007/978-3-030-35930-0).
- [12] M.-S. Chen, J. Han and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-883, December 1996, doi: [10.1109/69.553155](https://doi.org/10.1109/69.553155).
- [13] C. Schröer, F. Kruse and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, no. 1, pp. 526-534, 2021, doi: [10.1016/j.procs.2021.01.199](https://doi.org/10.1016/j.procs.2021.01.199).

- [14] T. D. K., P. B. G. and F. Xiong, "Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques," *Pattern Recognition Letters*, vol. 128, no. 1, pp. 544-550, December 2019, doi: [10.1016/j.patrec.2019.10.029](https://doi.org/10.1016/j.patrec.2019.10.029).
- [15] I. E. Naqa, R. Li and M. J. Murphy, *Machine Learning in Radiation Oncology*, Cham: Springer Cham, 2015, doi: [10.1007/978-3-319-18305-3](https://doi.org/10.1007/978-3-319-18305-3).
- [16] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, September 1999, doi: [10.1145/331499.331504](https://doi.org/10.1145/331499.331504).
- [17] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90-95, November 2013.
- [18] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv*, vol. 1, no. 1, pp. 1-29, September 2011, doi: [10.48550/arXiv.1109.2378](https://doi.org/10.48550/arXiv.1109.2378).
- [19] B. S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, Chichester: John Wiley & Sons, Ltd, 2011, pp. 2354-2360.
- [20] S. V. S. and S. Surendran, "A Review of Various Linear and Non Linear Dimensionality Reduction Techniques," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 3, pp. 2354-2360, 2015.
- [21] F. Anowar, S. Sadaoui and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review*, vol. 40, no. 1, pp. 1-13, February 2021, doi: [10.1016/j.cosrev.2021.100378](https://doi.org/10.1016/j.cosrev.2021.100378).
- [22] A. Bogdanchikov, M. Zhaparov and R. Suliyev, "Python to learn programming," in *ScieTech*, Jakarta, 2013, doi: [10.1088/1742-6596/423/1/012027](https://doi.org/10.1088/1742-6596/423/1/012027).
- [23] M. J. Nelson and A. K. Hoover, "Notes on Using Google Colaboratory in AI Education," in *ITiCSE*, Trondheim, 2020, doi: [10.1145/3341525.3393997](https://doi.org/10.1145/3341525.3393997).

## LAMPIRAN


### Lampiran 1 Kartu Bimbingan Kerja Praktek



UNIVERSITAS WIDYATAMA  
Program Studi Informatika S1

#### KARTU BIMBINGAN KERJA PRAKTEK

Nama Mahasiswa	Fransiscus Kristian Susanto
N I M	40621100012
Alamat Mahasiswa	Komplek Permata Biru Blok H No. 93 RT 004 RW 015 Desa Cinunuk Kecamatan Cileunyi Kabupaten Bandung Telp. : +6281220894779 e-mail : kristian.fransiscus@widyatama.ac.id
Topik / Judul KP	Clustering Tenaga Kerja Sektor Pertanian terhadap Sosio-Ekonomi dan Efisiensi Energi pada Negara Anggota G20
Pembimbing Kampus / NID	Dr. Feli Sulianta, S.T., M.T.

BATAS WAKTU BIMBINGAN	PENGESAHAN PROGRAM STUDI
10 Februari 2025 s/d 21 Juni 2025	

MELANJUTKAN BIMBINGAN	
REKOMENDASI DOSEN PEMBIMBING KAMPUS	PERSETUJUAN PEMBIMBING KAMPUS
<p><i>Untuk melanjutkan bimbingan yang telah melewati batas waktu bimbingan, mahasiswa harus mengembalikan kartu ini ke jurusan sambil membawa foto kopi FRS ( yg mencantumkan kembali KP ) beserta bukti pembayaran registrasi dan KP. Kemudian kartu ini akan diganti dengan kartu bimbingan KP yang baru.</i></p>	

Versi/Revisi : 1/3  
Tanggal : 27/1/04

## Lampiran 2 Kuesioner Kerja Praktek



FAKULTAS TEKNIK  
PROGRAM STUDI INFORMATIKA

Petunjuk Pengisian :

1. Semua pertanyaan dimohon Bapak/Ibu untuk menjawabnya.
2. Jawaban pertanyaan diharapkan seakurat mungkin sehingga hasilnya dapat digunakan sebagai bahan evaluasi dan perbaikan pada program studi yang kami kelola.
3. Dalam hal jawaban sudah tersedia, pilihlah (dengan melingkar) angka yang sesuai menurut Bapak/Ibu.

Nama Mahasiswa : Fransiscus Kristian Susanto  
 Nomor Pokok Mahasiswa : 40621100012  
 Nama Perusahaan : Yayasan Bakti Achmad Zaky  
 Tgl Pelaksanaan Kerja Praktek : 11 September 2024

Bagian I

Pada bagian ini semua pertanyaan mengikut perilaku dan kemampuan mahasiswa **program Studi Teknik Informatika** Universitas Widyatama secara rata-rata selama berlangsungnya kegiatan **Kerja Praktek**.

1. **Penampilan**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik Sekali

2. **Perilaku**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik sekali

3. **Kemampuan Analisis**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik Sekali

4. **Keterampilan Menggunakan Software Aplikasi**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik sekali

5. **Kreativitas/Inovasi**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik Sekali

6. **Kemadirian**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik Sekali

7. **Kerja sama (bekerja dalam tim)**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik sekali

8. **Komunikasi**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik Sekali

9. **Kedisiplinan**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik sekali

10. **Menulis Laporan**

1	2	3	4	5	6	7	8	9	10
Buruk									Baik sekali

**Bagian II**

Pada bagian II menyangkut informasi dan saran saran yang dapat Bapak/Ibu berikan

1. Berapa lama jangka waktu pelaksanaan kerja Praktek yang diharapkan/disediakan oleh perusahaan/instansi ?

Duras pelatihan studi independen program Data Science: Greener Future with Data Driven
Solution mengikuti aturan oleh Kampus Merdeka MSIB, yaitu kurang lebih 4 bulan.

2. Apa yang diharapkan oleh perusahaan/instansi dengan pelaksanaan kegiatan Kerja Praktek ini sebelumnya dan apakah harapan tersebut sudah terpenuhi setelah kegiatan Kerja Praktek ini berlangsung ?

Setelah mengikuti program studi independen diharapkan peserta dapat menyelesaikan seluruh rangkaian proses pembelajaran sesuai dengan kurikulum. Peserta juga diminta membuat produk akhir berupa Data Science product, seperti interactive dashboard, notebook, dan deck.
Selama pembelajaran Ananda Kristian sudah memenuhi kewajibannya selama studi independen.

3. Bagaimana peluang KP di perusahaan ini untuk mahasiswa kami selanjutnya ?

Ananda Kristian memiliki peluang untuk bergabung dengan mitra kami dengan cara dapat mendaftarkan diri melalui Job Connector pada alumni dashboard yang kami sediakan.

4. Saran-saran Bapak/Ibu berkaitan dengan pengembangan pendidikan untuk memenuhi kebutuhan pasar (pengguna)

Data Science adalah salah satu ilmu yang masih terus berkembang. Oleh karena itu, Ananda Kristian diharapkan dapat menerapkan long-life learning untuk mengikuti perkembangan zaman, terkhusus terkait perkembangan data science.

Jakarta, 9 Desember 2024

Pembimbing/Pananggung Jawab Lapangan



**Startup  
Campus**

Fitri Hestikarani

**Catatan:**

1. Kuisisioner yang telah bapak/Ibu isi diharapkan dikirim menggunakan amplop tertutup yang telah kami sediakan (dapat dititipkan melalui mahasiswa).
2. Ucapan terima kasih pada Bapak/Ibu yang telah menerima mahasiswa kami untuk Praktek dan mengisi form kuisisioner ini.