

Crafting an Exceptional Portfolio: Your Path to Success

Extra Class



Apa saja yang sudah kamu pelajari!

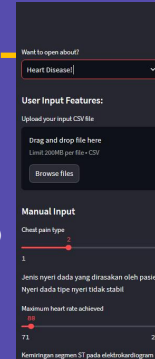
Machine Learning Concept

1. Data Preprocessing dan Cleaning
2. Data Visualization and EDA
3. Modelling (Supervised/Unsupervised)
4. Dimensionality Reduction and Feature Selection
5. Hyperparameter Tuning
6. Project Deployment (Streamlit)

Project Capstone

Heart Disease Dataset

*Kita buat
jadikan
example
portofolio



Welcome to my machine learning dashboard

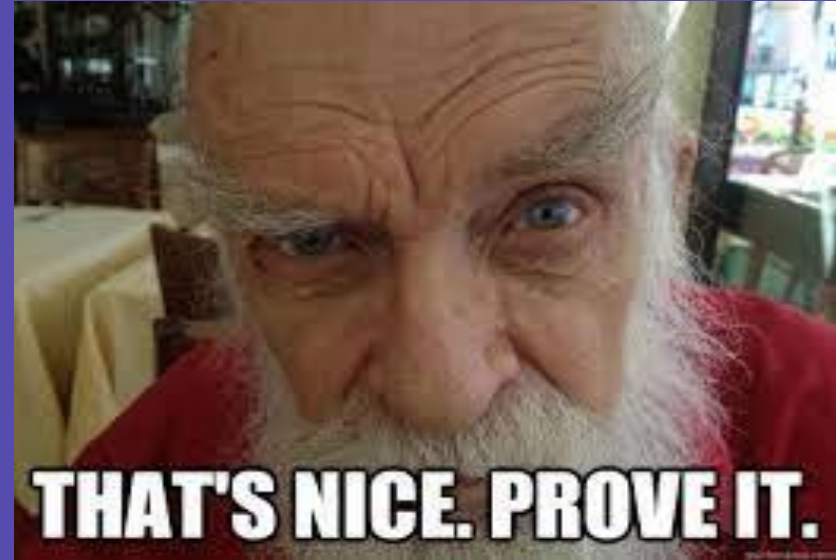
This dashboard created by : @dabellusidu

This app predicts the Heart Disease

Data obtained from the [Heart Disease dataset](#) by UCML.



**Kamu sudah mempelajari
banyak hal, yuk kita
implementasikan!**



Outline

1. What is Portofolio?
2. What is the Content
3. Tips and Trick
4. Best Practices

What is Portofolio?



Perbedaan antara?

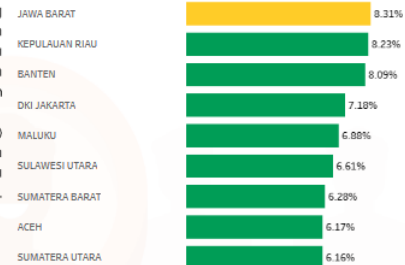


**General
Information**



**Project
Portfolio**

	CV	Portofolio
TUJUAN	Menjelaskan siapa kamu melalui riwayat hidup	Menunjukkan apa yang kamu kerjakan
KONTEN	Latar belakang kamu dan pengalaman terkait pekerjaan yang dilamar	Hasil karya kreatif seperti foto, video, cerita, gambar
FORMAT	Dokumen simpel dan clean sepanjang 1-2 halaman	Video, Website, Buku, Blog
GAYA PENULISAN	Alur jelas , tekankan pada substansi	Kreatif , untuk menunjukkan hasil karyamu

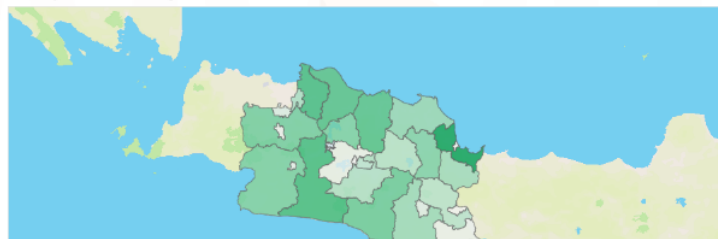


Gambar 1. Persentase Pengangguran di Jawa Barat

Portofolio yang kita miliki dapat dibuat dalam bentuk website, link ataupun demografi yang memudahkan orang lain melihat bagaimana hasil yang kita miliki.

SEBARAN PENCARI KERJA DI JAWA BARAT

Menurut data Dinas Tenaga kerja dan Transmigrasi Provinsi Jawa Barat, rata-rata pencari kerja di Jawa Barat tahun 2022 adalah sebanyak **14.422 orang**. Lima daerah di Jawa Barat dengan pencari kerja terbanyak antara lain **Kabupaten Cirebon** dengan angka pencari kerja terbanyak yaitu sebanyak **43.428 orang**, **Kabupaten Bekasi** sebanyak **30.163 orang**, **Kabupaten Cianjur** sebanyak **30.070 orang**, **Kabupaten Karawang** sebanyak **27.400 orang**, dan **Kabupaten Subang** sebanyak **27.167 orang**.



Kenapa Perlu Portofolio?



1. Membangun dan Mengasah Kemampuan

Belajar mengode, membangun model, meningkatkan akurasi model, dan menerapkan model semuanya merupakan bagian dari alur kerja ilmu data

2. Memamerkan Pengalaman

Semakin bervariasi portofolio yang dimiliki, semakin banyak kamu dapat memamerkan berbagai keterampilan teknis yang dapat dibicarakan

3. Mendemonstrasikan kemampuan yang dimiliki

Dengan memamerkan konsistensi, ketekunan, perhatian terhadap detail, dan kemauan untuk belajar dan terus meningkat.

What is the content?

What is Portofolio?



What is the content?



Buatlah Portfolio dan/atau dashboard berdasarkan project yang diberikan! Gunakan ilmu yang telah diajarkan dan kreativitas kalian dalam menyusun project Heart Disease

*(Halaman Selanjutnya adalah sebagai potong contoh yang dapat dipelajari)

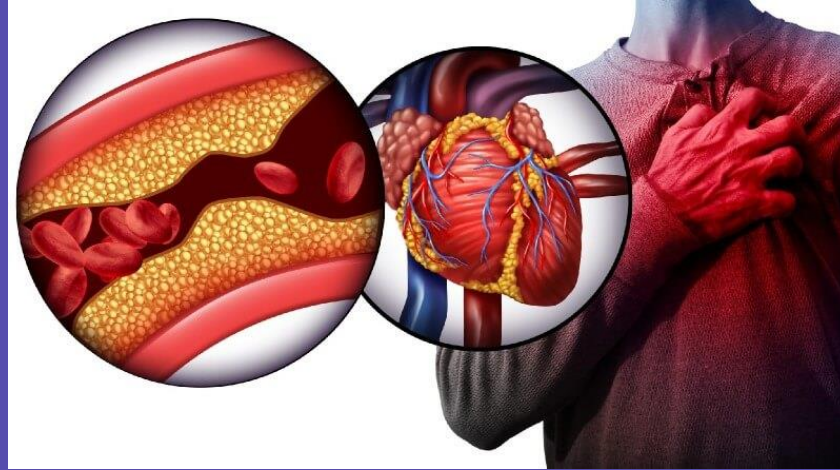
Designing Heart Disease Dashboards:

<<GOAL DASHBOARD>>

<<your name>>

Outline

1. Goals
2. Data Understanding
3. Tools yang digunakan
4. Data Analysis
5. Data Modelling
6. Conclusion
7. My Profile



- Heart Disease Project -

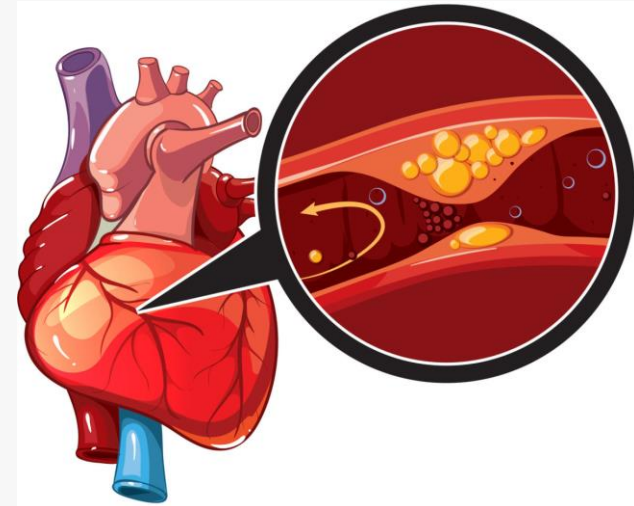
Terdapat 17,9 juta kasus kematian penyakit jantung setiap tahunnya yang didorong oleh adanya peningkatan kasus hipertensi, obesitas, dan gaya hidup yang tidak sehat (<<Latar Belakang>>).

Mari kita (<<POINT YANG INGIN DISELESAIKAN>>)!

Problem Identification

Goal: Menentukan perawatan dini penyakit jantung.

- ☐ **Penyakit Jantung dapat disebabkan oleh berbagai faktor.** Kemudian, faktor-faktor apa saja yang menjadi penyebab dan gejala penyakit jantung?
- ☐
- ☐ **Apabila sudah mengetahui faktor yang memiliki signifikansi yang tinggi.** Kemudian, bagaimana mengetahui apabila seseorang menderita penyakit jantung atau tidak?



Data Understanding

Dataset yang digunakan adalah data Heart Disease yang diunduh dari UCI ML (<https://archive.ics.uci.edu/dataset/45/heart+disease>)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         1025 non-null   int64
1   sex         1025 non-null   int64
2   cp          1025 non-null   int64
3   trestbps    1025 non-null   int64
4   chol        1025 non-null   int64
5   fbs         1025 non-null   int64
6   restecg     1025 non-null   int64
7   thalach     1025 non-null   int64
8   exang       1025 non-null   int64
9   oldpeak     1025 non-null   float64
10  slope       1025 non-null   int64
11  ca          1025 non-null   int64
12  thal        1025 non-null   int64
13  target      1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Tools yang digunakan

Software



Google Colaboratory atau **Google Colab** adalah executable document yang memungkinkan kamu dalam menulis, mengedit, serta membagikan program yang sudah disimpan pada drive maupun yang baru kamu buat.

Bahasa Pemrograman



Python merupakan bahasa pemrograman komputer yang biasa dipakai untuk membangun situs, software/aplikasi, mengotomatiskan tugas dan melakukan analisis data. Bahasa pemrograman ini termasuk bahasa tujuan umum.

Library Python

<<Tambahkan Library Yang Digunakan>>
<<Pandas>> <<Matplotlib>> <<Streamlit>>

Data Quality (1/2)

Melakukan pengecekan karakter, handling missing value, outlier, duplikat dan data imbalance serta skewness

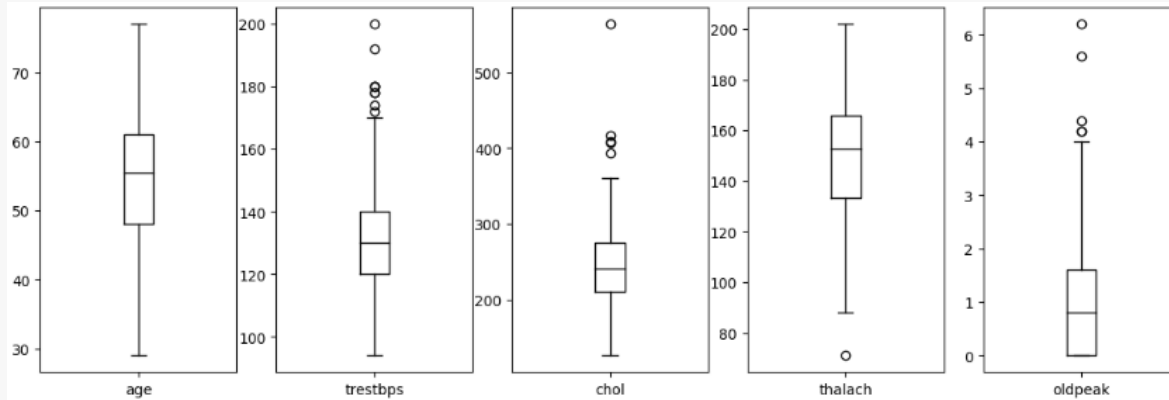
```
Berikut ini merupakan informasi: sex dengan ['Male' 'Female']  
Berikut ini merupakan informasi: cp dengan ['typical angina' 'atypical angina' 'non-anginal pain' 'asymtomatic']  
Berikut ini merupakan informasi: fbs dengan ['No' 'Yes']  
Berikut ini merupakan informasi: restecg dengan ['normal' 'probable or definite left ventricular hypertrophy'  
'ST-T Wave abnormal']  
Berikut ini merupakan informasi: exang dengan ['No' 'Yes']  
Berikut ini merupakan informasi: slope dengan ['upsloping' 'downsloping' 'flat']  
Berikut ini merupakan informasi: ca dengan ['Number of major vessels: 2' 'Number of major vessels: 0'  
'Number of major vessels: 1' 'Number of major vessels: 3' 4]  
Berikut ini merupakan informasi: thal dengan ['reversable defect' 'fixed defect' 'normal' 0]  
Berikut ini merupakan informasi: target dengan ['No disease' 'Disease']
```

Variabel yang di handling

1. **Feature 'CA':** Memiliki 5 nilai dari rentang 0-4, maka dari itu nilai 4 diubah menjadi NaN (karena seharusnya tidak ada)
2. **Feature 'thal':** Memiliki 4 nilai dari rentang 0-3, maka dari itu nilai 0 diubah menjadi NaN (karena seharusnya tidak ada)

Data Quality (2/2)

Melakukan pengecekan karakter, handling missing value, outlier, duplikat dan data imbalance serta skewness

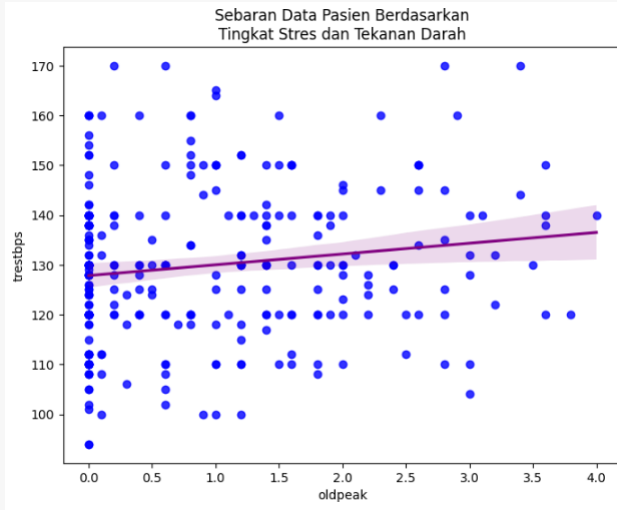


Data yang memiliki outlier dengan menggantikan dengan

1. Nilai Maks: $Q3 + 1.5 \text{ IQR}$
2. Nilai Min: $Q1 - 1.5 \text{ IQR}$

Data Analysis (1/...)

Melakukan pengalihan berdasarkan infografik dari dataset yang dimiliki.



Adanya korelasi Positif antara Tingkat Stress dengan Tekanan Darah dengan ditunjukkan **line ungu**.

Hal ini terjadi karena<<Tambahkan Penjelasan>>

Data Analysis (2/...)

Melakukan pengalihan berdasarkan infografik dari dataset yang dimiliki.

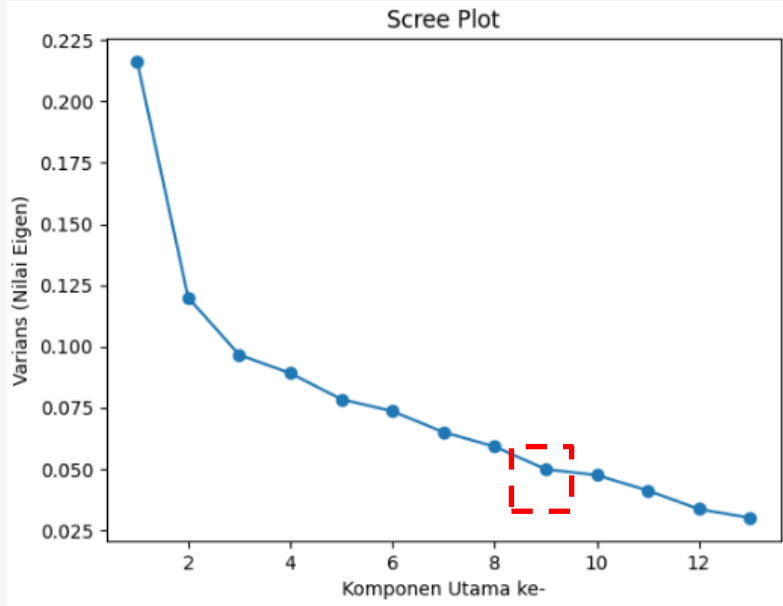
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.064118	-0.058687	0.282669	0.171015	0.106885	-0.110517	-0.411108	0.093718	0.209254	-0.149095	0.392130	0.054752	-0.222416
sex	-0.064118	1.000000	-0.091357	0.007572	-0.123863	0.066692	-0.083290	-0.042981	0.182596	0.157352	-0.060014	0.122489	0.245682	-0.318896
cp	-0.058687	-0.091357	1.000000	0.083242	-0.080369	0.084389	0.080836	0.285605	-0.388610	-0.122582	0.095881	-0.202923	-0.188487	0.416319
stbps	0.282669	0.007572	0.083242	1.000000	0.115138	0.127221	-0.139228	-0.071600	0.001726	0.144438	-0.083024	0.101059	-0.014615	-0.115614
chol	0.171015	-0.123863	-0.080369	0.115138	1.000000	0.013066	-0.142285	-0.020128	0.076547	-0.009534	0.039352	0.124800	0.078868	-0.105627
fbs	0.106885	0.066692	0.084389	0.127221	0.013066	1.000000	-0.077417	-0.023484	0.006080	0.015070	-0.069563	0.150552	-0.042766	-0.027210
stecg	-0.110517	-0.083290	0.080836	-0.139228	-0.142285	-0.077417	1.000000	0.089556	-0.104440	-0.089255	0.111841	-0.126825	0.035452	0.171453
alach	-0.411108	-0.042981	0.285605	-0.071600	-0.020128	-0.023484	0.089556	1.000000	-0.387726	-0.341190	0.376494	-0.296480	-0.134498	0.422559
exang	0.093718	0.182596	-0.388610	0.001726	0.076547	0.006080	-0.104440	-0.387726	1.000000	0.318620	-0.259780	0.154768	0.223241	-0.431599
lpeak	0.209254	0.157352	-0.122582	0.144438	-0.009534	0.015070	-0.089255	-0.341190	0.318620	1.000000	-0.525142	0.245318	0.189228	-0.434108
slope	-0.149095	-0.060014	0.095881	-0.083024	0.039352	-0.069563	0.111841	0.376494	-0.259780	-0.525142	1.000000	-0.067890	-0.088110	0.326473
ca	0.392130	0.122489	-0.202923	0.101059	0.124800	0.150552	-0.126825	-0.296480	0.154768	0.245318	-0.067890	1.000000	0.140048	-0.456989
thal	0.054752	0.245682	-0.188487	-0.014615	0.078868	-0.042766	0.035452	-0.134498	0.223241	0.189228	-0.088110	0.140048	1.000000	-0.370759
arget	-0.222416	-0.318896	0.416319	-0.115614	-0.105627	-0.027210	0.171453	0.422559	-0.431599	-0.434108	0.326473	-0.456989	-0.370759	1.000000

Korelasi tertinggi
terdapat pada **thalach**
0.422 (korelasi Positif
Kuat)

thalach (semakin tinggi detak jantung maksimum yang dicapai pasien selama tes latihan, maka resiko terkena penyakit jantung semakin tinggi)

Data Analysis (../10)

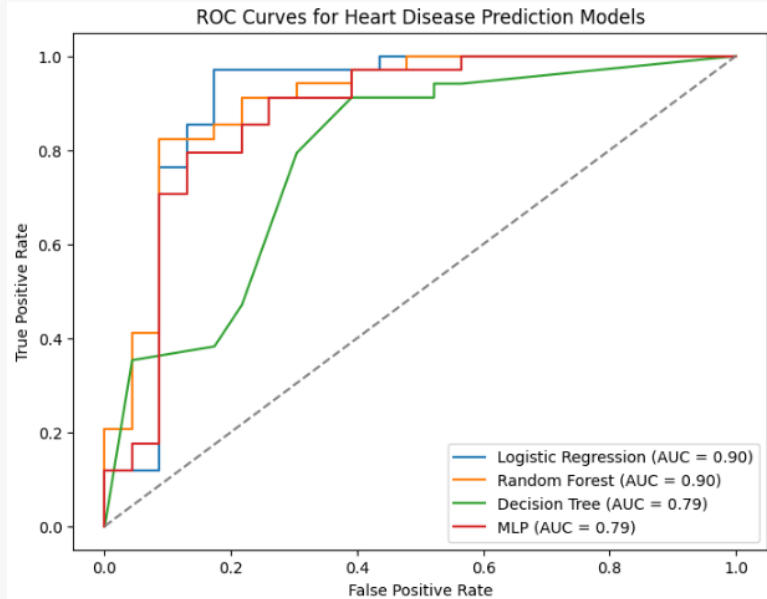
Melakukan pengalihan berdasarkan infografik dari dataset yang dimiliki.



Awalnya kita memiliki 14 variabel, setelah dilakukan analisa dengan Scree Plot didapati bahwa dengan **9 variabel cukup mewakili variansi data.**

Data Modelling (../5)

Melakukan pemodelan data yang telah layak dianalisa.



Jadi dapat disimpulkan, bahwa model yang peformanya lebih bagus ialah model Random Forest dengan ditandai dengan nilai akurasi pada classification report tertinggi, yaitu **sebesar 82%** dan adanya score AUC-ROC Terbesar dibandingkan 3 model lainnya, yaitu **sebesar 90%**.

Conclusion

Untuk menjawab bagaimana menentukan perawatan dini penyakit jantung, kita dapat melakukan:

1. Data Quality dengan mengecek terlebih dahulu data yang kita miliki baik categorical dan numerical
2. Data Analysis untuk mendapatkan overview data
3. Dari hasil proses didapat 9 variabel/features yang terpilih dengan pengaruh signifikan
4. Data Modelling untuk memilih Random Forest sebagai model terbaik dari 4 model yang ditentukan
5.

Membuat Dashboard

×

Want to open about?

Heart Disease!

▼

User Input Features:

Upload your input CSV file

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Manual Input

Chest pain type

124

Jenis nyeri dada yang dirasakan oleh pasien
Nyeri dada tipe nyeri tidak stabil

Maximum heart rate achieved

8071202


Kemiringan segmen ST pada elektrokardiogram

Welcome to my machine learning dashboard

This dashboard created by : [@abelkristanto](#)

This app predicts the Heart Disease

Data obtained from the [Heart Disease dataset](#) by UCIML.



Silahkan akses dengan mengakses
<<Tuliskan link yang mau kamu bagikan>>

My Profile

FOTO SAYA

Working Experience.



PRESENT
Data Scientist in DQLAB
2012 – present



3 years as Data Analyst in Tokopedia
Build Consumer Data Analysis

Connect me.

<Linkedin LINK> <Another Social Media>>

Tips n Tricks

Cek berikut ini



Design dan perpaduan warna yang tepat (keep it simple, content not design)



Jgn masukan semua karya, cukup karya terbaik



Susun berdasarkan kategori



State of originality



relevan



Put some details (objective, user, data source)



Understand business questions



All-in-One

Jgn satu porto untuk semua



File size kecil (kecuali untuk videographer)



Try put your self in other people shoes

Best Practices

Secara umum, dalam mengerjakan project ada beberapa hal-hal penting yang harus dimasukkan antara lain :

- **About Me!** (ceritakan tentang diri kamu!)
- **Project Background** (Project ini membahas tentang apa dan tujuan project ini dilakukan)
- **Tools** (Tools yang digunakan)
- **Opening Insight** (Insight yang digunakan nanti untuk menggambarkan seluk beluk dataset yang ada dalam project ini)
- **Problem Definition** (Masalah-masalah yang harus dipecahkan dalam project ini)
- **Data Understanding** (Membaca data sebagai input untuk pandas dan memahami tipe datanya)
- **Exploratory Data Analysis** (Mengidentifikasi dan menangani anomali, seperti missing values, outlier, data formatting, dan data duplikat)
- **Insights or Analysis Result** (Hasil analisis yang menjadi jawaban dari Problem Definition yang ada)



Check The
requirement



Put github in notebook
form



Finds job Vacancy you want



Put responsibilities
into your portfolio



Write blog or in
medium

Thanks!

