Prediksi Posibilitas Serangan Jantung

Data diambil dari Kaggle dengan nama dataset yaitu Health care: Heart attack possibility.

URL: https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility/data

- Dataset berupa CSV
- Dataset terdiri dari 303 records dengan 14 buah variabel yang diukur.
- Dataset terdiri dari 9 data kategorik (5 di antaranya sudah diubah menjadi data numerik) dan 5 data numerik.
- Dataset memiliki duplicated data sejumlah 1 records
- Dataset memiliki missing value sejumlah 2 records

Stiven Gabriel - 121320054
Dosen Pengampu: Ahmad Suaif, S.Si., M.Si.
Link Github: https://github.com/stivenn13/Tugas-Besar-SDR

01 Business Understanding

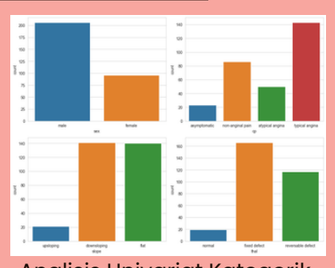
Bagi individu:
- Meningkatkan kesadaran diri tentang risiko serangan jantung.
- Mendorong individu untuk mengambil langkah-langkah preventif untuk mengurangi risiko.
- Membantu individu untuk membuat keputusan yang lebih tepat tentang gaya hidup dan perawatan kesehatan.

Bagi profesional medis:
- Meningkatkan kemampuan untuk mengidentifikasi pasien berisiko tinggi.
- Membantu dalam stratifikasi risiko dan skrining.
- Memandu pengambilan keputusan tentang intervensi dan pengobatan.

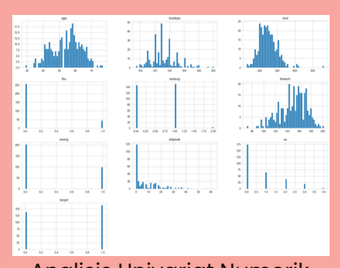
02 Data Understanding

Variabel-variabel pada dataset adalah sebagai berikut:
- age: umur pasien (dalam tahun)
- sex: jenis kelamin pasien
- cp: tipe chest pain
- trestbps: tekanan darah ketika istirahat (dalam mm Hg saat masuk rumah sakit)
- chol: kolesterol serum dalam mg/dl
- fbs: gula darah (trigliserida) saat puasa gula > 120 mg/dl (1 = benar; 0 = salah)
- restecg: hasil elektrokardiografi saat istirahat (0 = normal; 1 = mengalami ST-T wave abnormality; 2 = menunjukkan kemungkinan atau pasti hipertrofi ventrikel kiri berdasarkan kriteria Estes)
- thalach: detak jantung maksimum yang tercapai
- exang: angina diinduksi oleh olahraga (1 = benar; 0 = salah)
- oldpeak: ST depression yang disebabkan oleh olahraga dibandingkan istirahat
- slope: kemiringan puncak exercise ST segment
- ca: jumlah pembuluh besar (0-3) yang diwarnai dengan flourosopy
- thal: hasil thallium stress test
- target: 0 = lebih kecil kemungkinan terkena serangan jantung; 1 = lebih besar kemungkinan terkena serangan jantung

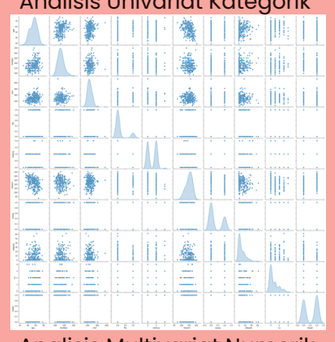
03 Data Analysis



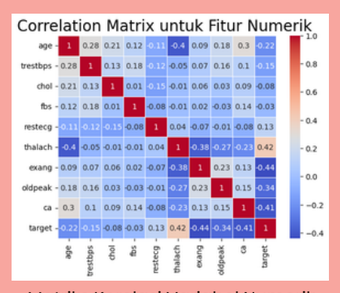
Analisis Univariat Kategorik



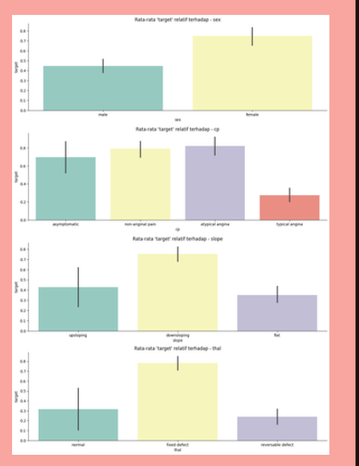
Analisis Univariat Numerik



Analisis Multivariat Numerik



Matriks Korelasi Variabel Numerik



Analisis Multivariat Kategorik

04 Data Preparation

Untuk proses Data Assessing, berikut adalah beberapa pengecekan yang dilakukan:
- Duplicate data (data yang serupa dengan data lainnya)
- Missing value (data atau informasi yang "hilang" atau tidak tersedia)

Pada proses Data Cleaning, secara garis besar, terdapat tiga metode yang dapat digunakan antara lain seperti berikut:
- Dropping (metode yang dilakukan dengan cara menghapus sejumlah baris data)
- Imputation (metode yang dilakukan dengan cara mengganti nilai yang "hilang" atau tidak tersedia dengan nilai tertentu yang bisa berupa median atau mean dari data)

05 Modelling

Dilakukan Modelling menggunakan beberapa ML Model yang berikutnya akan dicari tingkat akurasi dari setiap model

	Model	Accuracy
0	Logistic Regression	88.524590
1	Naive Bayes	81.967213
2	Random Forest	88.524590
3	Extreme Gradient Boost	60.655738
4	Decision Tree	85.245902
5	Support Vector Machine	86.885246

06 Evaluation ML Model

Berdasarkan perolehan tingkat akurasi dan barplot akurasi untuk setiap model diatas menyatakan bahwa model Logistic Regression, Random Forest, dan Support Vector Machine memiliki tingkat akurasi paling tinggi. Maka itu, dilakukan ensemble terhadap ketiga Model tersebut untuk membentuk model baru yang lebih bagus.

```
Accuracy of StackingCVClassifier: 91.80327868852459

              precision    recall  f1-score   support

           0       0.92      0.89      0.91        27
           1       0.91      0.94      0.93        34

    accuracy                           0.92        61
   macro avg       0.92      0.92      0.92        61
weighted avg       0.92      0.92      0.92        61
```

Dapat dilihat bahwa teknik Ensemble ML Model berhasil mendapatkan ML Model baru dengan tingkat akurasi yang lebih tinggi. Setelah itu, akan diperiksa metriks dan melakukan prediksi dari model yang sudah di-Ensemble dan model yang tidak di-Ensemble.

```
Name of the model:  Performa Model Naive Bayes
R^2 of the model: 0.2690631808278867
MSE of the model: 0.4246502900652006
MAE of the model: 0.18032786885245902
......
Name of the model:  Performa Model Extreme Gradient Boost
R^2 of the model: -0.5947712418300652
MSE of the model: 0.62725004818718
MAE of the model: 0.39344262295081966
```

```
Name of the model:  Performa Model Decision Tree
R^2 of the model: 0.4019607843137255
MSE of the model: 0.3841106397986879
MAE of the model: 0.14754098360655737
......
Name of the model:  Performa Model Support Vector Machine
R^2 of the model: 0.6677559912854031
MSE of the model: 0.2862991671569341
MAE of the model: 0.08196721311475409
```

07

Dilakukan prediksi dari model yang sudah di-Ensemble dan model yang tidak di-Ensemble.

	y_true	prediksi_NB	prediksi_XGB	prediksi_DT	prediksi_SCV
226	0	1	1	0	0

Berdasarkan hasil prediksi diatas, ML Models Decision Tree dan Ensembled Model (SCV) memiliki hasil prediksi yang paling dekat dengan nilai y_true.

Referensi:
[1] Pal, Ankita. Logistic regression: A simple primer. Cancer Research, Statistics, and Treatment 4(3):p 551-554, Jul–Sep 2021. | DOI: 10.4103/crst.crst_164_21
[2] AZIZAH, Nur; GOEJANTORO, Rito; SIFRIYANI, Sifriyani. METODE NAIVE BAYES DENGAN PENDEKATAN DISTRIBUSI GAUSS UNTUK KLASIFIKASI PEMINATAN PESERTA DIDIK. Prosiding Seminar Nasional Matematika dan Statistika, [S.l.], v. 1, p. 8-14, may 2019. ISSN 2657-232X. Available at: Berdasarkan hasil prediksi diatas, ML Models Decision Tree dan Ensembled Model (SCV) memiliki hasil prediksi yang paling dekat dengan nilai y_true.. Date accessed: 03 apr. 2024.
[3] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan Random Forest Untuk Mengukur Tingkat Keparahan Penyakit Pada Daun apel," Jurnal Sains dan Seni ITS, vol. 8, no. 2, Jan. 2020. doi:10.12962/j23373520.v8i2.48517
[4] Z. Arif Ali, Z. H. Abduljabbar, H. A. Taher, A. Bibo Sallow, and S. M. Almufti, "Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review", ACAD J NAWROZ UNIV, vol. 12, no. 2, pp. 320–334, May 2023.
[5] Kotsiantis, S.B. Decision trees: a recent overview. Artif Intell Rev 39, 261–283 (2013). Berdasarkan hasil prediksi diatas, ML Models Decision Tree dan Ensembled Model (SCV) memiliki hasil prediksi yang paling dekat dengan nilai y_true.
[6] Awad, M., Khanna, R. (2015). Support Vector Machines for Classification. In: Efficient Learning Machines. Apress, Berkeley, CA. Berdasarkan hasil prediksi diatas, ML Models Decision Tree dan Ensembled Model (SCV) memiliki hasil prediksi yang paling dekat dengan nilai y_true.