Ikhsan

UJIAN AKHIR SEMESTER
BIG DATA 2021/2022

Nomor 1. Klasifikasi Data

Unduh data dari https://www.kaggle.com/fedesoriano/heart-failure-prediction.

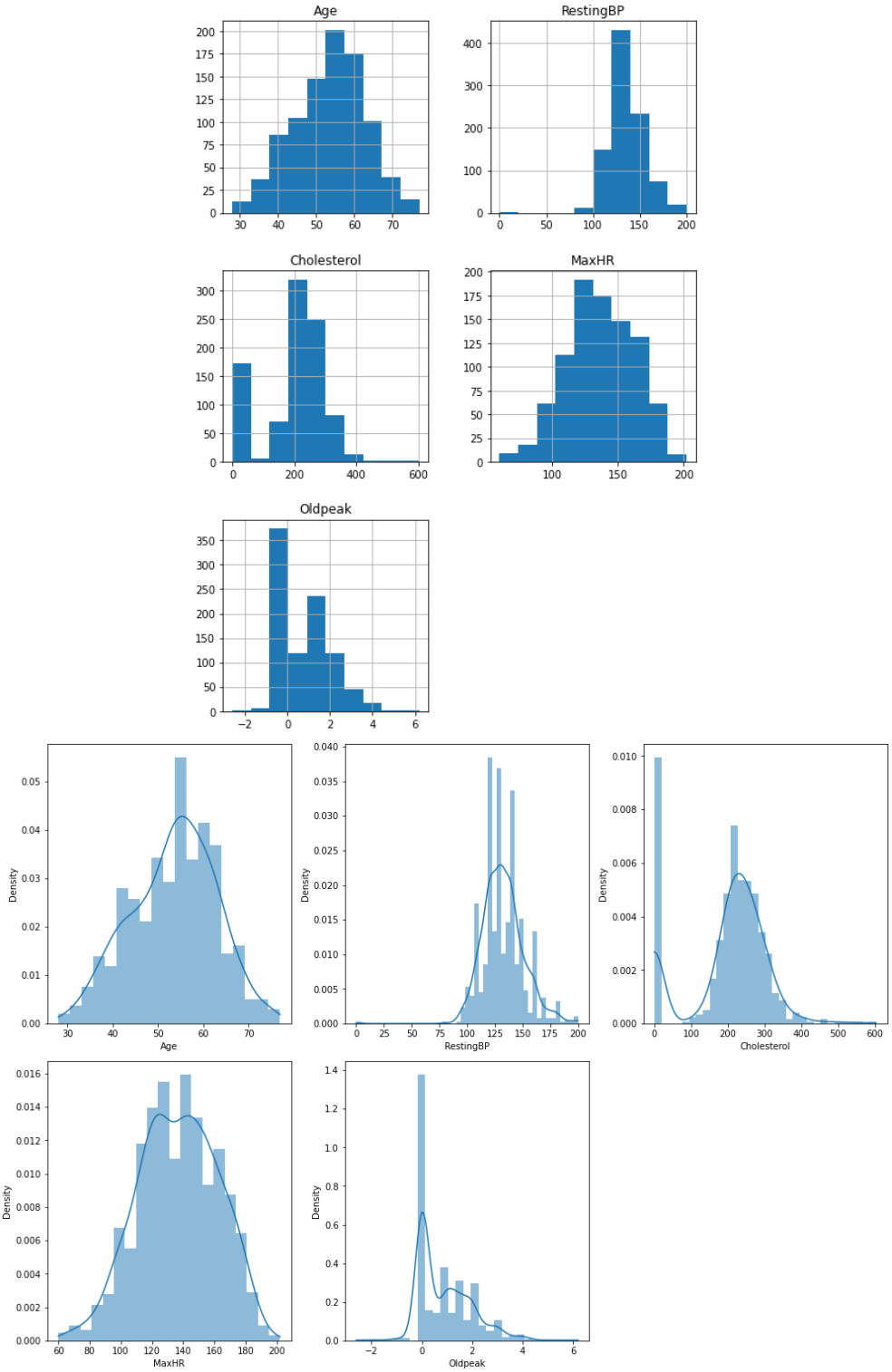
Informasi data:

Jumlah data	918
Jumlah kolom	12
Nilai kosong	Tidak ada

Keterangan kolom:

Nama Kolom	Tipe Data	Keterangan
Age	int64	Umur.
Sex	Object	Gender.
ChestPainType	Object	Tipe rasa sakit pada dada.
RestingBP	int64	Resting Blood Pressure: Tekanan darah saat kondisi tuhuh beristirahat.
Cholesterol	int64	Tingkat kolesterol
FastingBS	int64	Fasting Blood Pressure (Percepatan Tekanan Darah)
RestingECG	Object	Resting Electrocardiogram: Test untuk mengukur aktifitas elektrik di jantung.
MaxHR	int64	Max HeartRate (Detak Jantung Maksimal)
ExerciseAngina	Object	Angina: Suatu penyakit pada dada saat berolahraga, ketika stress atau kegiatan yang membuat jantung bekerja lebih keras.
Oldpeak	float64	Perbandingan kondisi jantung ketika berolahraga dan ketika beistirahat
ST_Slope	Object	Kondisi grafik detak jantung
HeartDisease	int64	Penyakit jantung

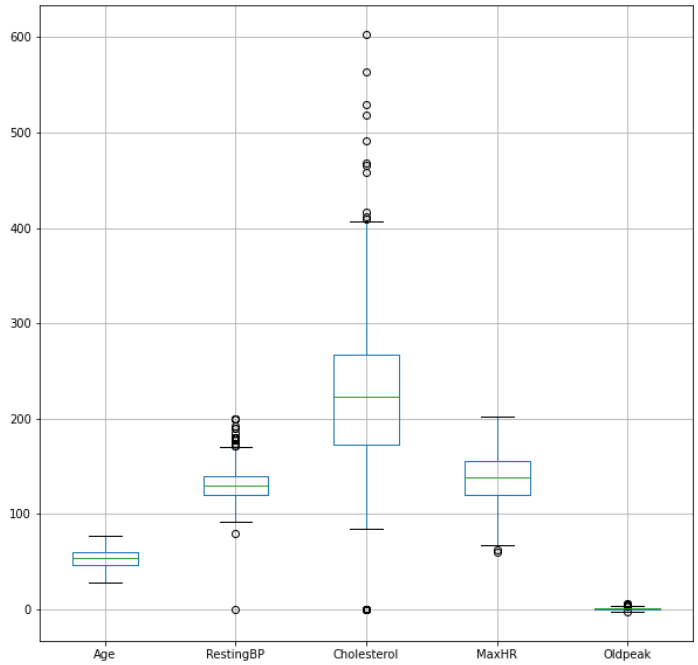
a. Buatlah EDA menggunakan teknik visualisasi data. Kemudian jelaskan hasil dari EDA tersebut.
   - Histogram dan Displot
     Histogram dan Distplot digunakan untuk mengetahui bentuk grafik data sehingga dapat diketahui jenis distribusi data. Histogram digunakan untuk kolom dengan tipe data numerik, hasilnya sebagai berikut:



Dari hasil visualisasi grafik diatas dapat ditarik kesimpulan bahwa 5 kolom data dengan tipe numerik yaitu Age, RestingBP, Cholesterol, HaxHR, dan Oldpeak berjenis distribusi normal karena memiliki visualisasi seperti lonceng (data disekitar median merupakan mean).

- Boxplot

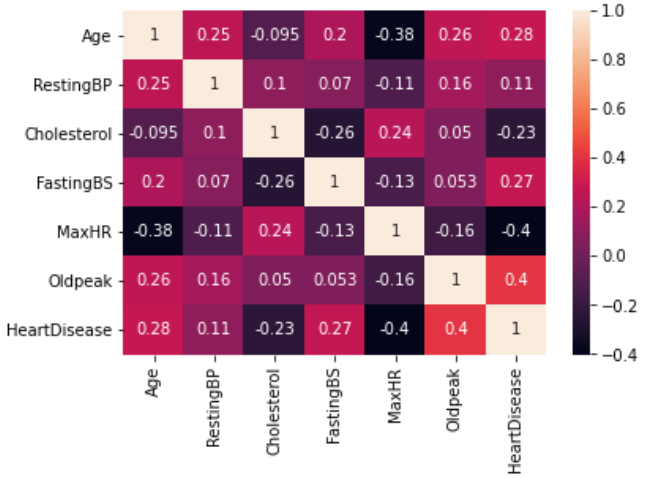
  Boxplot digunakan untuk melihat distribusi data seperti minimum, kuartil 1, median, kuartil 3, maksimum serta outlier pada data, seperti histogram boxplot juga digunakan pada data numerik, berikut ada visualisasinya:



Dari grafik diatas dapat disimpulkan kolom RestingBP, Cholesterol memiliki banyak outlier dan nilai minimum 0, kolom MaxHR memliki sedikit outlier disekitar Q1, kolom Age tidak memiliki outlier jadi data terkonsenterasi disekitar median, dan kolom oldpeak mempunyai beberapa outlier dan merupakan kolom dengan rata-rata nilai data terkecil.

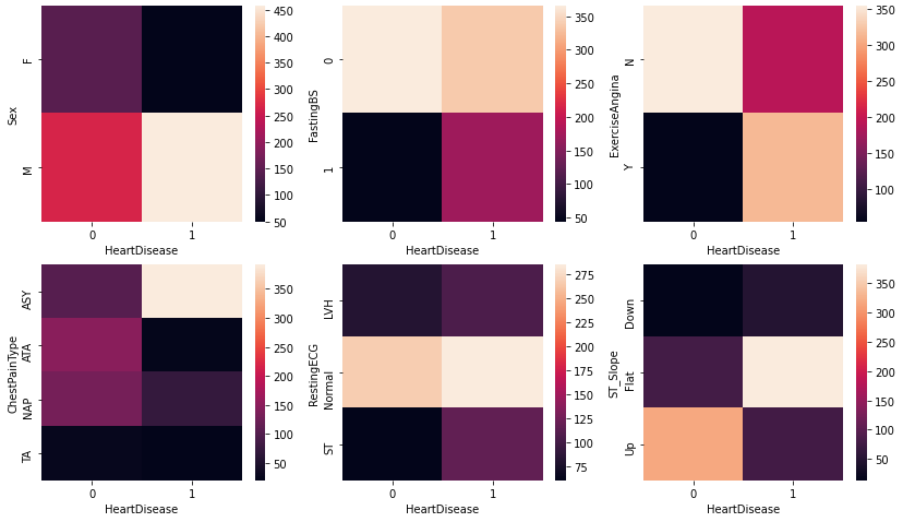
- Heatmap

Heatmap digunakan untuk memvisualisasikan hubungan / korelasi antar kolom, sehingga dapat diketahui apakah kolom cenderung bersifat negatif, positif atau tidak memiliki hubungan dengan kolom lainnya. Berikut adalah hasilnya:



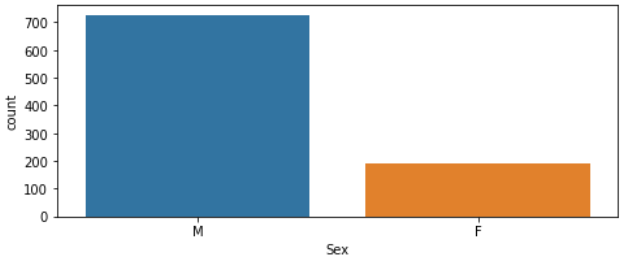
Karena data yang digunakan merupakan gabungan dari numerik dan kategorikal maka metode yang digunakan adalah pearson. Pada grafik diatas kita berfokus pada faktor/kolom yang mempengahui terhadap variabel terikat yaitu HeartDisease. Kolom yang mempengaruhi adalah kolom Oldpeak yaitu sebesar 0.4 dengan arah positif jadi semakin tinggi Oldpeak maka semakin tinggi resiko HeartDisease, meskipun begitu nilai 0.4 masih tergolong berkolerasi lemah. Kolom MaxHR menjadi kolom yang paling berpengaruh negatif dengan nilai -0.4, dimana semakin tinggi MaxHR maka semakin kecil resiko HeartDisease.

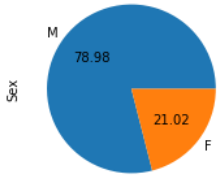
Untuk tipe data kategorikal setiap komposisi *value* dapat dijabarkan kembali untuk melihat nilai yang paling mempengaruhi heart disease:

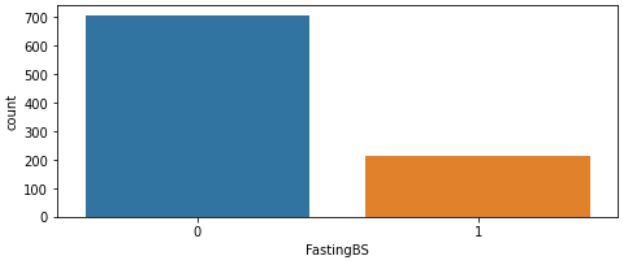


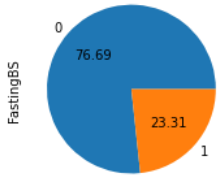
- Barplot Nilai dan frekuensi untuk tipe data kategorikal

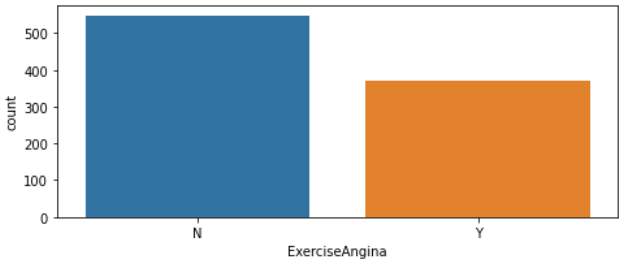
  Barplot dapat digunakan untuk memvisualisasikan nilai dari data kategorikal, sehingga kita dapat mengetahui komposisi nilai dan frekuensinya. Berikut adalah visualisasinya:

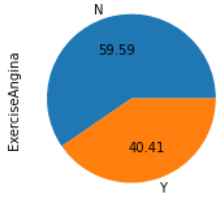

count
Sex
M
F

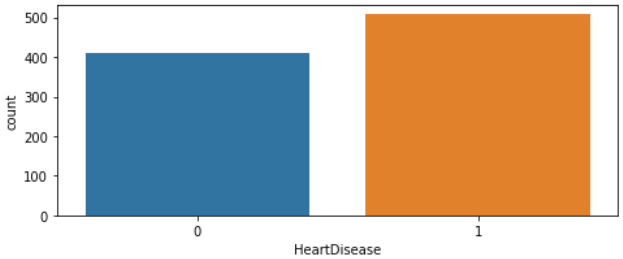

Sex
M
78.98
21.02
F

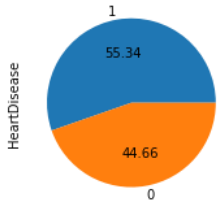

count
FastingBS
0
1

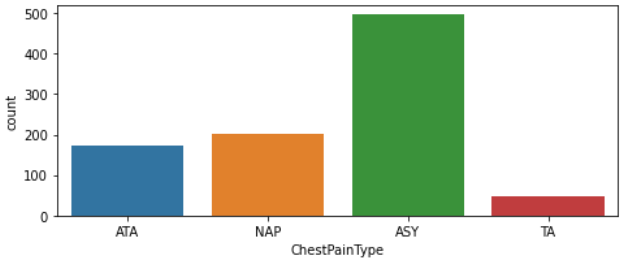

FastingBS
0
76.69
23.31
1

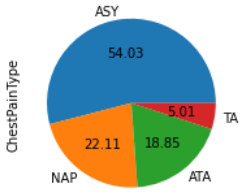

count
ExerciseAngina
N
Y

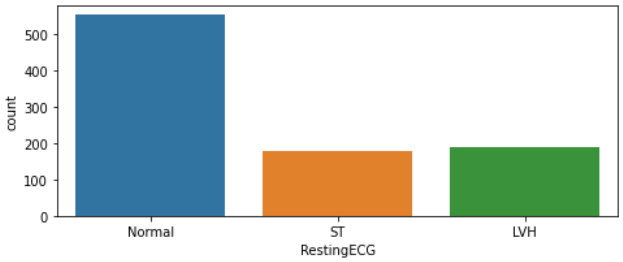

ExerciseAngina
N
59.59
40.41
Y

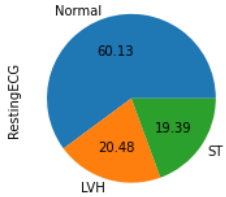

count
HeartDisease
0
1

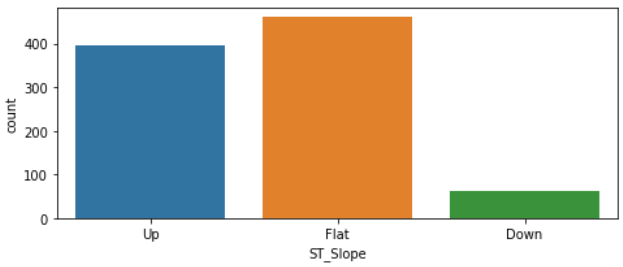

HeartDisease
1
55.34
44.66
0

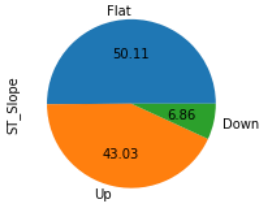

count
ChestPainType
ATA
NAP
ASY
TA


ChestPainType
ASY
54.03
5.01
TA
22.11
18.85
NAP
ATA


count
RestingECG
Normal
ST
LVH


RestingECG
Normal
60.13
19.39
ST
20.48
LVH


count
ST_Slope
Up
Flat
Down


ST_Slope
Flat
50.11
6.86
Down
43.03
Up

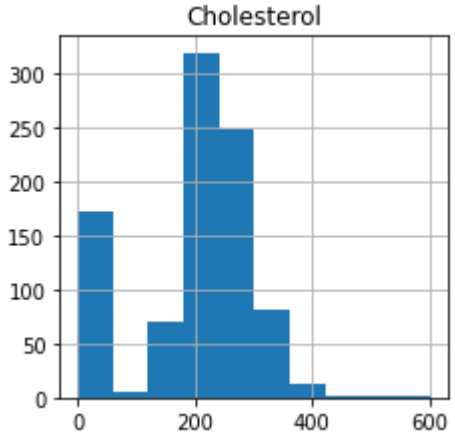
Melalui grafik diatas dapat diketahui nilai dan frekuensinya untuk masing masing kolom, sehingga kita sudah memiliki gambaran mengenai data yang diolah.

b. Apa pra-proses yg cocok dilakukan untuk dataset tersebut.

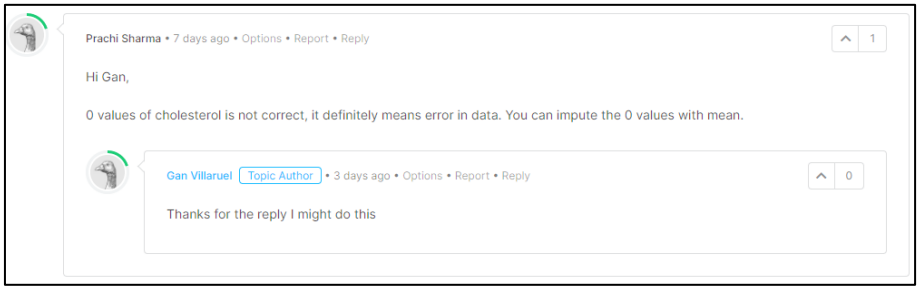
- Melakukan penyesuasian tipe data

  Penyesuasian tipe data dilakukan untuk kolom FastingBS dan HeartDisease yang nilainya dapat diganti dari value 0 dan 1 menjadi boolean False dan True, serta kolom ExerciseAngina dari value 'N' dan 'Y' menjadi boolean False dan True.

- Memperbaiki kolom Cholesterol



Berdasarkan grafik hitogram diatas terjadi keganjilan yaitu banyak data kolesterol dengan nilai 0 hal tersebut menjadi tidak wajar karena memang manusia jarang sekali memiliki kolesterol bernilai nol. Berdasarkan kolom komentar dimana dataset ini diambil (kaggle.com) dikonfirmasi terjadi kesalahan data:



Data Cholesterol dengan nilai 0 dapat diganti dengan nilai mean, akan tetapi karena kolom Cholesterol memiliki outlier nilai 0 diganti menjadi nilai median.

- Normalisasi data numerik

  Normalisasi dilakukan agar data memiliki rentang yang sama, normalisasi dilakukan dengan menggunakan fungsi *min-max scaler* dari *package* Sklearn.

c. Pilih dua metode pembagian data. Kemudian jelaskan alasan menggunakan metode tersebut.

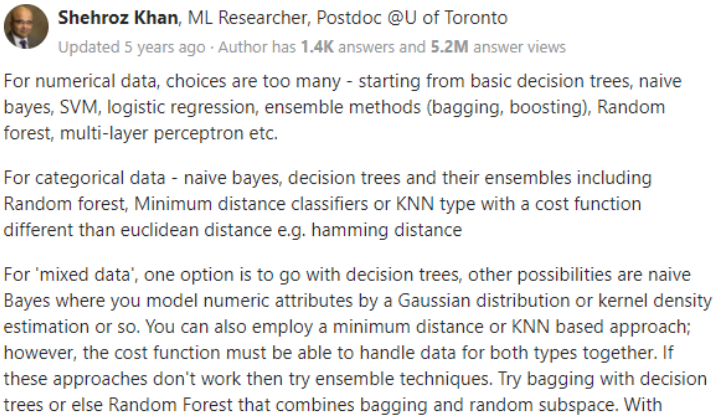
- Split validasi

  Membagi data menjadi 2 bagian yaitu train dan test, split validasi dipilih karena merupakan metode pembagian data yang paling sering digunakan. Karena jumlah data yang digunakan sebanyak 918 data metode split validasi dapat

menjadi pilihan, karena jika jumlah data yang digunakan berjumlah sedikit split validasi akan menghasilkan akurasi yang rendah.

- 10-fold cross validation
  Membagi menjadi beberapa 10 bagian dan secara bergantian 1 bagian menjadi train dan 9 lainnya menjadi test, cross-validation dipilih karena menghasilkan model dengan akurasi terbaik karena dilakukannya iterasi. Jumlah data yang digunakan saat ini juga tidak terlalu bersar sehingga cross-validation tidak menghabiskan waktu yang cukup lama.

d. Pilih dua metode klasifikasi data. Kemudian jelaskan alasan menggunakan metode tersebut.



Shehroz Khan, ML Researcher, Postdoc @U of Toronto
Updated 5 years ago · Author has **1.4K** answers and **5.2M** answer views

For numerical data, choices are too many - starting from basic decision trees, naive bayes, SVM, logistic regression, ensemble methods (bagging, boosting), Random forest, multi-layer perceptron etc.

For categorical data - naive bayes, decision trees and their ensembles including Random forest, Minimum distance classifiers or KNN type with a cost function different than euclidean distance e.g. hamming distance

For 'mixed data', one option is to go with decision trees, other possibilities are naive Bayes where you model numeric attributes by a Gaussian distribution or kernel density estimation or so. You can also employ a minimum distance or KNN based approach; however, the cost function must be able to handle data for both types together. If these approaches don't work then try ensemble techniques. Try bagging with decision trees or else Random Forest that combines bagging and random subspace. With

Dataset yang saat ini digunakan terdiri dari campuran data numerik dan kategorikal, berdasarkan diskusi publik ada yang menyarankan untuk menggunakan algoritma decision tree, naïve bayes, KNN dan ensemble yang contohnya random forest. Dari saran tersebut kemudian dilakukan perbandingan dengan hasil sebagai berikut:

Split Validasi

Pengukuran	Random Forest	CART	Naïve Bayes	KNN
Akurasi	**0,848**	0,735	**0,822**	0,817
Presisi	**0,871**	0,808	**0,859**	0,829
Recall	**0,877**	0,732	0,841	**0,877**

10-fold Cross Validation

Pengukuran	Random Forest	CART	Naïve Bayes	KNN
Akurasi	**0,868**	0,786	0,831	**0,846**
Presisi	**0,858**	0,795	**0,853**	0,843
Recall	**0,878**	0,791	0,828	**0,885**

Berdasarkan tabel diatas jadi dipilih dua metode klasifikasi data dengan performa terbaik:

- Random Forest, random forest menjadi pilihan yang paling menjanjikan dengan nilai akurasi, presisi, recall paling tinggi. Random forest menggunakan konsep decision tree dengan ensemble, yang berbeda dari decision tree biasa karena data akan dibagi secara random menjadi beberapa subset untuk membentuk beberapa tree dan dilakukan voting untuk menentukan kelas.
- KNN dapat menjadi pilihan karena kedua karena dengan cross validasi KNN dapat menghasilkan akurasi dan recall tertinggi melebihi Naïve Bayes. KNN akan mencari jarak antar data dan memilih sejumlah k sample terdekat dan dilakukan voting untuk menentukan kelas.

e. Hitung nilai akurasi, presisi, recall.

Klasifikasi Data	Pembagian data	Confusion Matrix		
		Akurasi	Presisi	Recall
Random Forest	Split Validasi	0,848	0,871	0,877
	10-Fold Cross Validation	0,868	0,858	0,878
K-Nearest Neighbor	Split Validation	0,817	0,829	0,877
	10-Fold Cross Validation	0,846	0,843	0,885

Dari tabel hasil diatas Random Forest dengan menggunakan 10-Fold Cross Validation merupakan model klasifikasi yang paling bagus karena menghasilkan nilai akurasi, presisi dan recall terbesar.

Nomor 2. Klastering Data

Unduh data dari https://www.kaggle.com/uciml/iris.

Informasi data:

Jumlah data	150
Jumlah kolom	6
Nilai kosong	Tidak Ada

Keterangan kolom:

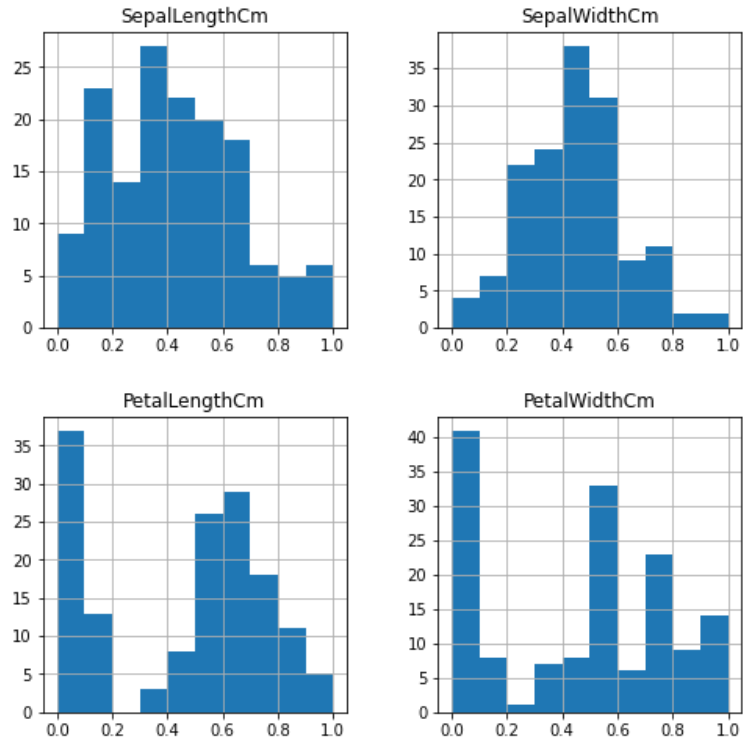
Nama Kolom	Tipe Data	Keterangan
Id	int64	Urutan baris
SepalLengthCm	Float64	Panjang sepal (daun bunga) pada bunga iris
SepalWidthCm	Float64	Lebar sepal (daun bunga) pada bunga iris
PetalLengthCm	Float64	Panjang petal (kelopak bunga) pada bunga iris
PetalWidthCm	Float64	Lebar petal (kelopak bunga) pada bunga iris
Species	Object	Jenis spesies bunga iris [setosa, versicolor, dan virginica]

a. Apa pra-proses yg cocok dilakukan untuk dataset tersebut.

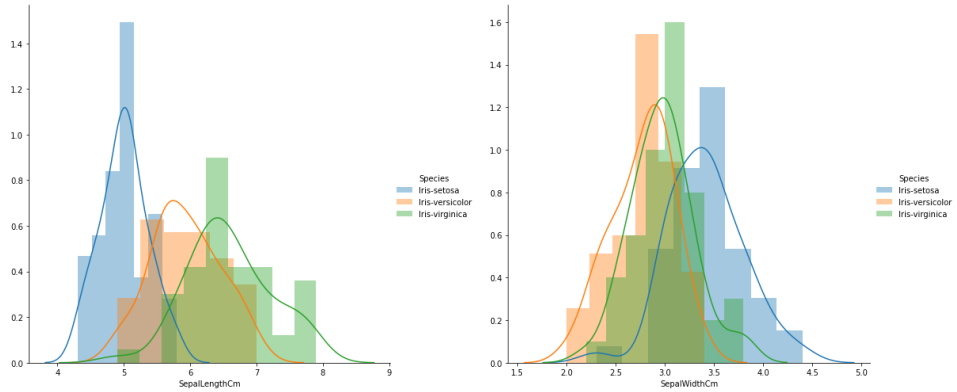
- Menghapus kolom yang tidak digunakan
  Dalam dataset yang digunakan kolom Id tidak digunakan karena hanya berisi iformasi urutan baris, untuk menyederhanakan dataset kolom testebut perlu dihapus.
- Penyesuaian tipe data
  Penyesuaian dilakukan untuk kolom Spesies dengan mengubah tipe data *object* menjadi *categorical*, kemudian tiga nilai penyusun kolom tersebut [setosa, versicolor, dan virginica] akan diganti menjadi berbentuk angka [0, 1, 2].
  Kemudian *dataframe* diubah menjadi *list*, karena fungsi klasterisasi dari *package* Pyclustering yang digunakan menerima input berupa list. Oleh karena itu perlu dilakukan perubahan nilai kolom kategorikal menjadi berbentuk angka.
- Normalisasi
  Pada dataset yang digunakan semua variabel bebasnya bertipe data numerik yang akan dilakukan normalisasi agar data memiliki rentang yang sama, normalisasi dilakukan dengan menggunakan fungsi *min-max scaler* dari *package* Sklearn.
- Visualisasi data
  Visualisasi data dilakukan untuk mendapatkan gambaran / informasi yang lebih detail pada dataset yang digunakan, berikut beberapa visualisasi data yang dilakukan:

- Histogram dan Distplot

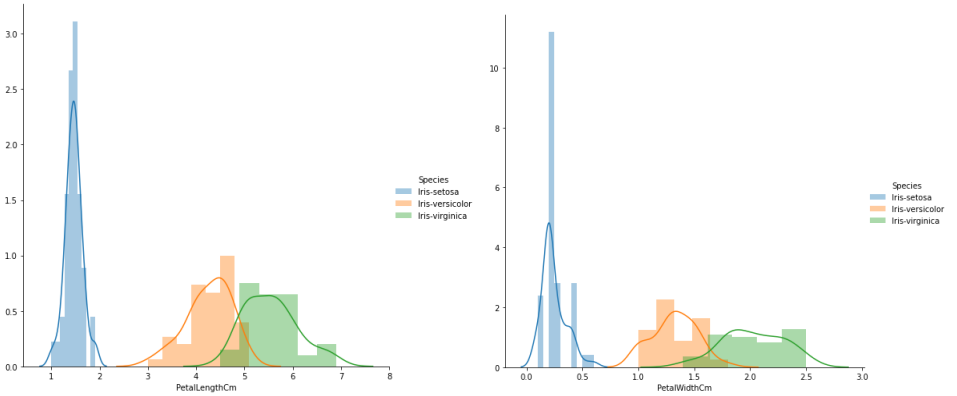
Histogram:



Distplot sepal:

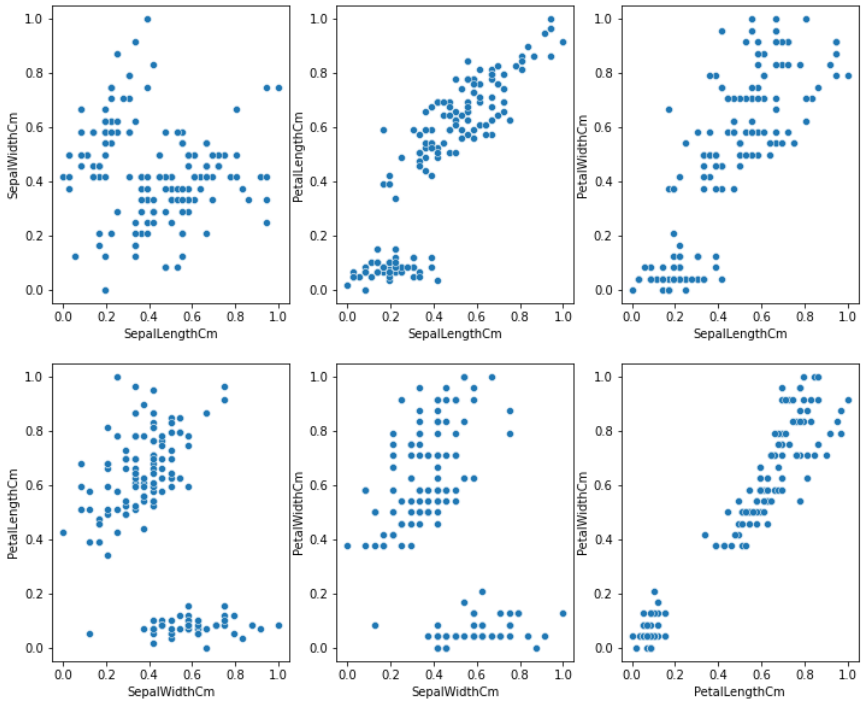


Displot petal:



Dari hasil visualisasi histogram dan distplot data cenderung menghasilkan distribusi normal, berdasarkan distplot iris setosa memiliki petal yang lebih kecil dari spesies lainnya dan tiap speies menghasilkan grafik bell-curve.

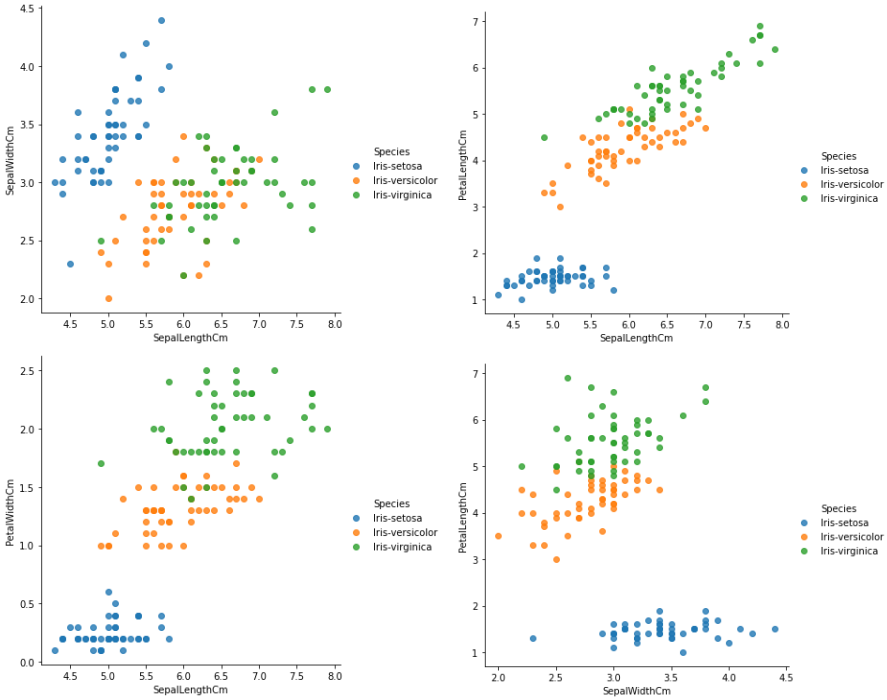
- Scatterplot

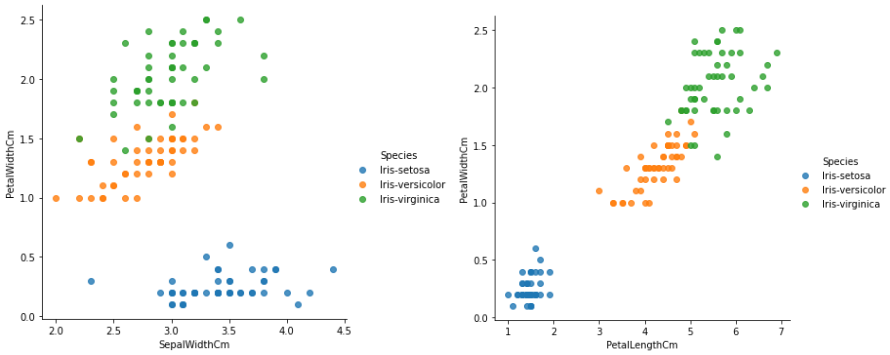


Berdasarkan visualiasasi scatterplot diatas pada umumnya grafik bernilai positif (naik ke arah kanan) jadi semakin besar nilai x maka semakin besar pula nilai y, seperti pada scatterplot terakhir yaitu perbandingan Petal length dengan Petal width. Ada pula grafik dengan nilai tersebar seperti scaterplot pertama yaitu perbandingan Sepal length dan sepal width.

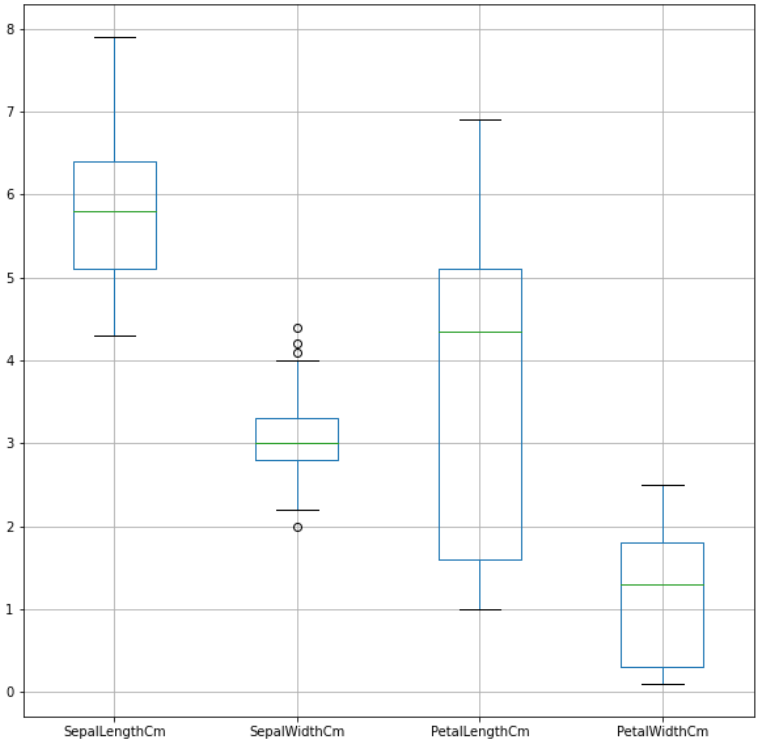
Scatterplot tersebut belum dilakukan klasterisasi tetapi pada grafik tersebut terdapat jarak / gap pada kumpulan nilai yang akan membentuk suatu klaster tersendiri yang terpisah dari kluster lain.

Jika ditambah dengan informasi mengenai jenis spesies grafik akan menjadi sebagai berikut:





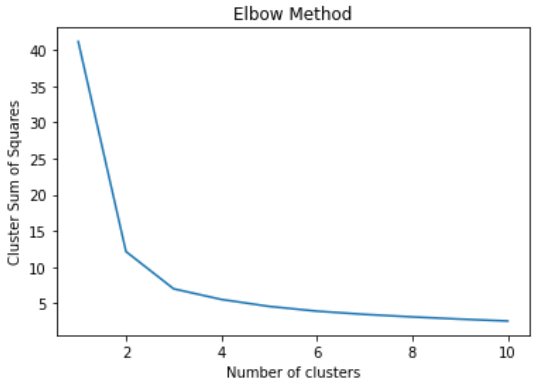
- Barplot



Dari visualisasi barplot diatas hanya kolom Sepal width yang memiliki beberapa outlier, Sepal length memiliki median yang paling besar nilainya sementara Petal width yang terkecil.

- Elbow test

Elbow test digunakan untuk menentukan banyaknya jumlah klaster yang akan ditentukan, berikut ini adalah visualisasinya:



Untuk menentukan jumlah klaster yang optimal dari grafik elbow, sesuai namanya yaitu elbow atau siku, kita dapat memilih titik yang dimana grafik

mulai bergerak menjadi linear, dari grafik diatas dapat diambil keputusan jumlah klaster optimal yang akan dipakai adalah sebanyak 3 klaster.

b. Pilih dua metode pembagian data. Kemudian jelaskan alasan menggunakan metode tersebut.
   - Split validasi
     Membagi data menjadi 2 bagian yaitu train dan test, split validasi dipilih karena merupakan metode pembagian data yang paling sering digunakan. Karena data yang digunakan hanya 150 data metode split validasi yang digunakan mungkin menghasilkan akurasi yang tidak terlalu bagus tetapi masih dapat menjadi pilihan.
   - 10-fold cross validation
     Membagi menjadi beberapa 10 bagian dan secara bergantian 1 bagian menjadi train dan 9 lainnya menjadi test, cross-validation dipilih karena menghasilkan model dengan akurasi terbaik karena dilakukannya iterasi sangat cocok untuk data uang tidak terlalu banyak. Jumlah data yang digunakan saat ini juga tidak terlalu bersar sehingga cross-validation tidak menghabiskan waktu yang cukup lama.

c. Pilih dua metode menghitung jarak antar data. Kemudian jelaskan alasan menggunakan metode tersebut.

   Pemilihan perhitungan jarak data menjadi langkah yang penting dalam klastering, karena akan menentukan bagaimana memperoleh jarak dari dua titik (x, y) yang dikalkulasikan dan akan berpengaruh pada bentuk klaster yang dibuat.

   Berdasarkan paper yang berjudul "K-means with Three different Distance Metrics" yang ditulis oleh Singh, dkk dari jurnal *International Journal of Computer Applications* pada tahun 2013, menghasilkan kesimpulan bahwa pada algoritma K-means dengan menggunakan Euclidean distance memberikan hasil terbaik sedangkan K-means dengan Manhattan distance memberikan hasil terburuk. Dari keputusan paper tersebut dipilihlah dua distance metric yaitu:
   - Euclidean distance

     $$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}$$

     Euclidean distance menhitung jarak terdekat diantara dua titik, euclidean distance menjadi distance metric yang umum digunakan.
   - Minkowski distance

     $$Dist_{XY} = \left(\sum_{K=1}^{d} |X_{ik} - X_{jk}|^{\frac{1}{P}} \right)^{P}$$

     Minkowski distance merupakan generalisasi antara euclidean distance dan manhattan distance

   Distance metrik Euclidean dan Minkowsi selain dipilih karena memberikan hasil yang lebih baik dari Manhattan (berdasarkan jurnal), kedua distance metrik itu juga cocok untuk dataset dengan kolom yang bertipe numerik.

d. Pilih dua metode klasifikasi data. Kemudian jelaskan alasan menggunakan metode tersebut.
   - K-means
     K-means adalah algorima partisi yaitu membagi dataset menjadi beberapa k-grup, k-means akan berusaha mengurangi total squared error. K-means adalah

algoritma klastering yang paling umum digunakan k-means berkerja dengan baik untuk data numerikal dan buruk digunakan untuk tipe data kategorikal.

- K-medoids
  K-medloid bekerja seperti k-means tetapi k-medoids berusaha mengurangi jumlah perbedaan diantara titik untuk menjadi klaster.

  Berdasarkan jurnal "Performance Analysis Of K-Means And K-Medoids Clustering Algorithms For A Randomly Generated Data Set" yang ditulis oleh T. Velmurugan dan Dr. T. Santhanam padat tahun 2008 dari paper *International Conference on Systemics, Cybernetics and Informatics* mendapatkan kesimpulan bahwa kedoids bekerja lebih baik dari k-means jika dataset mengandung outlier dan noise karena kmedoid tidak dipengaruhi oleh hal tersebut. Selain itu k-medoid bekerja dengan efektif untuk data dengan jumlah yang sedikit dan menghasilkan performa yang buruk untuk data yang besar.

e. Hitung nilai SSE dan Centroid.

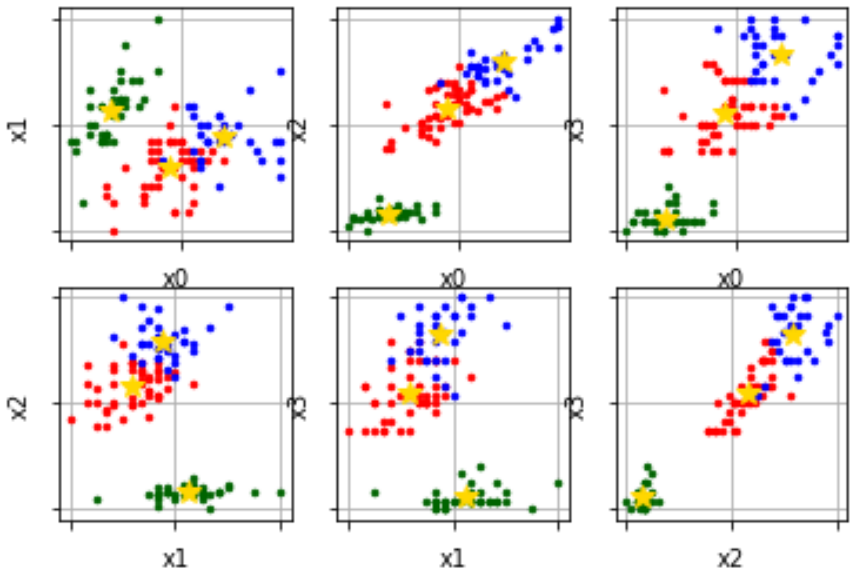
Tabel SSE

K-means		
	Euclidean Distance	Minkowski Distance
Split Validation	23.516181193991816	23.516181193991812
K-Fold Cross Validation	22.761609571259115	22.918995999449184
K-medoids		
	Euclidean Distance	Minkowski Distance
Split Validation	23.812317031831327	23.812317031831324
K-Fold Cross Validation	22.83761584807452	22.837615848074535

Tabel Centroid

K-means		
	Euclidean Distance	Minkowski Distance
Split Validation	[0.184 0.564 0.080 0.054] [0.694 0.441 0.798 0.828] [0.442 0.296 0.578 0.551]	[0.442 0.296 0.578 0.551] [0.694 0.441 0.798 0.828] [0.184 0.564 0.080 0.054]
K-Fold Cross Validation	[0.196 0.591 0.079 0.060] [0.664 0.442 0.735 0.734] [0.397 0.270 0.543 0.505]	[0.196 0.591 0.079 0.060] [0.678 0.443 0.773 0.778] [0.424 0.291 0.550 0.513]
K-medoids		
	Euclidean Distance	Minkowski Distance
Split Validation	[0.194 0.583 0.085 0.042] [0.611 0.417 0.712 0.792] [0.417 0.292 0.492 0.458]	[0.194 0.583 0.085 0.042] [0.611 0.417 0.712 0.792] [0.417 0.292 0.492 0.458]
K-Fold Cross Validation	[0.194 0.583 0.085 0.042] [0.361 0.292 0.542 0.500] [0.667 0.417 0.678 0.667]	[0.194 0.583 0.085 0.042] [0.667 0.417 0.678 0.667] [0.361 0.292 0.542 0.500]

Visualisasi Klaster



Dari hasil SSE dan centroid yang sudah didapat sse yang dihasil dari setiap kombinasi metode pembagian data, algoritma dan penentuan jarak data tidak mengalami perbedaan yang besar, dengan menggunakan split validasi SSE diangka 23 dan dengan k-cross validation turun menjadi 22. SSE terkecil deperoleh dari algoritma kmeans dengan euclidean distance serta menggunakan cross-validation yang menghasilkan SSE sebesar 22.762.

Nomor 3. Menggunakan dataset dari tugas 3 “Praproses Data "Model Prediksi Deret Waktu Titik Panas dengan Memperhatikan Faktor Iklim".

Dimana percobaan yang dilakukan dari sisi datatset.

- Dataset pertama, var terikat dari FIRMS NASA. Var bebas dari BMKG. (Sesuai dengan dataset tugas 3)

Jumlah data	240
Jumlah kolom	8
Nilai kosong	Tidak ada

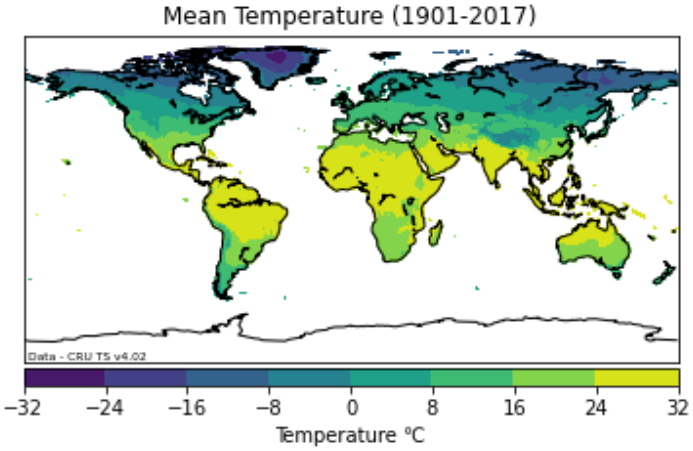
Nama Kolom	Tipe Data
Bulan	Datetime
Suhu	float64
Kelambapan	float64
Curah_Hujan	float64
Radiasi_Matahari	float64
Kecepatan_Angin	float64
Hotspot	int64

- Dataset kedua, var terikat dari FIRMS NASA. Var bebas dari CRU TS v4.05 (https://catalogue.ceda.ac.uk/uuid/c26a65020a5e4b80b20018f148556681)

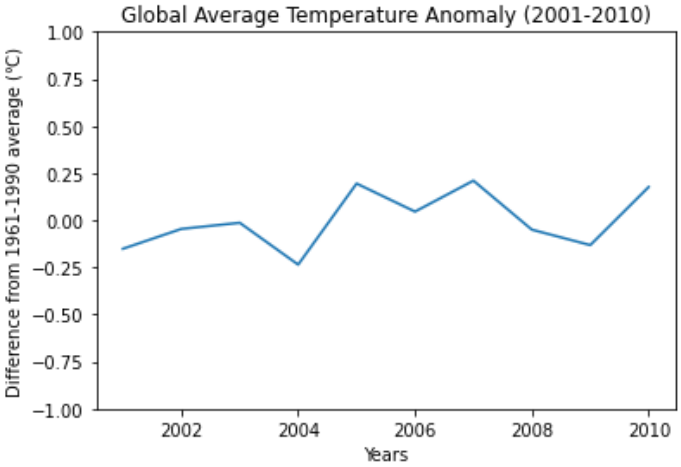
Nomor 3. Prediksi Deep Learning

a. Apa pra-proses yg cocok dilakukan untuk dataset tersebut.
   - Membuat dataset
     Berikut ini ada proses pembuatan dataset:

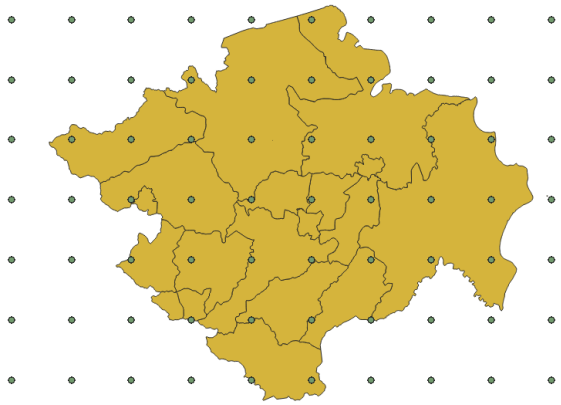
- Mempelajari data dari sumbernya
  Visualisasi kolom temperatur pada dataset:



  Visualisasi time series kolom tempetarur:



- Menggabungkan semua file .nc untuk diambil variabel bebasnya yaitu ['cld', 'dtr', 'frs', 'pet', 'pre', 'tmn', 'tmp', 'tmx', 'vap', 'wet']. Dalam setiap file berisi 120 (12 bulan x 10 tahun) x 360 (latitude) x 720 (latitude) = 31.104.000 jumlah data untuk 10 tahun, karena data yang digunakan berjumlah 20 tahun yaitu dari 2001 ke 2020 berarti data yang digunakan menjadi 62.208.000 jumlah data, data tersebut menjadi sangat banyak karena memproyeksikan seluruh permukaan bumi.
- Melakukan pemotongan data dengan memilih data dengan latitude dan longitude untuk indonesia saja, latitude indonesia tersebar dari 5.6701 sampai -10.934 dan longitude indonesia tersebar dari 140.9419 sampai 95.1345. setelah dilakukan pemotongan data, data bekurang hingga 98% menjadi 728.640 data saja.
- Data kemudian dilakukan clipping dengan menggunakan data shp Sumatra Selatan menggunakan aplikasi QGIS sebagai berikut:



Dari gambar diatas dapat diketahui bahwa latitude dan longitude yang ada di dataset bertambah 0.25 disetiap titik sehingga jaraknya sama.

- Dataset yang sudah diklip terdapat duplikasi pada tanggal, hal tersebut terjadi karena ada beberapa titik (long, lat) di provinsi Sumatra Selatam, untuk mengatasinya data dengan tanggal yang sama akan disatukan dan dikalkulasi dengan menggunakan mean, sehingga jumlah data berkurang menjadi 240 data yaitu jumlah bulan dalam 20 tahun.
- Kemudian dataset tersebut ditambahkan dengan kolom hotspot yang merupakan variabel terikat dari FIRMS NASA.
- Berikut adalah hasil akhir dataset yang sudah dilakukan clipping:

Jumlah data	240
Jumlah kolom	12
Nilai kosong	Tidak ada

Nama Kolom	Tipe Data	Keterangan
Date	Datetime	Tanggal
Cld	float64	Cloud cover
Dtr	float64	Diurnal temperature range
Frs	float64	Frost day frequency
Pet	float64	Potential Evapo-transpiration
Pre	float64	Precipitation
Tmn	float64	Monthly average daily minimum temperatur
Tmp	float64	Daily mean temperature
Tmx	float64	Monthly average daily maximum temperatur
Vap	float64	Vapour pressure
Wet	float64	Wet day frequency
Hotspot	int64	Titik panas

Pada dataset diatas kolom latitude dan longitude dibuang karena sudah tidak digunakan lagi, kedepannya data akan diproses berdasarkan time series.

- Normalisasi

Kedua dataset dilakukan normalisasi agar data memiliki rentang yang sama, normalisasi dilakukan dengan menggunakan min-max scaler yaitu dengan rentang -1 hingga 1.

b. Pilih satu metode pembagian data. Kemudian jelaskan alasan menggunakan metode tersebut.

Metode pembagian data yang digunakan adalah Split Validation akan tetapi saat melatih model lstm dimasukan parameter input berupa cross_validation dengan k sebanyak 2.

c. Tentukan nilai arsitektur NN. Kemudian jelaskan alasan mengapa arsitektur NN dibuat seperti itu.

Konfigurasi Hyper parameter:

Hyper parameter	Nilai
Neuron	[8, 16]
Activation function	['sigmoid', 'tanh', 'relu', 'selu', 'elu', 'softplus']
Optimizer	['adam', 'adamax', 'rmsprop', 'sgd']
Dropout rate	[0.1]
Epochs	[500, 1000, 1500, 2000]
Batch_size	[8, 16, 32, 64],

Hyper-parameter yang digunakan menghasilkan 768 kombinasi hyper-parameter. Model akan dibuat dengan satu layer LSTM saja, karena hanya dengan satu layer LSTM saja dapat memberikan hasil yang cukup baik, jika menggunakan banyak layer untuk menjalankan 768 kombinasi hyper-parameter membutuhkan waktu yang sangat lama.

d. Minimal jumlah kombinasi dari arsitektur NN akan menghasilkan 480 model prediksi. Kemudian pilih salah satu dari 480 model tersebut menggunakan metode hyperparameter tuning grid serach.

(Akan dibahas di poin f)

e. Metode Deep Learning menggunakann LSTM-RNN.
Deep Learning menggunakan LSTM-RNN aplikasikan dengan bantuan *package* Tensorflow.Keras.

f. Hitung nilai RMSE dan waktu komputasi.

Dataset 1: BMKG.

- Metode pembagian data Split validation dan LSTM dengan Cross validation k = 2, Persentasi data latih 80% dan data uji 20%
- Arsitektur NN
  - Hyper parameter

Hyper parameter	Nilai
Neuron	[8, 16]
Activation function	['sigmoid', 'tanh', 'relu', 'selu', 'elu', 'softplus']
Optimizer	['adam', 'adamax', 'rmsprop', 'sgd']
Dropout rate	[0.1]
Epochs	[500, 1000, 1500, 2000]
Batch_size	[8, 16, 32, 64],

  - Model LSTM satu layer
- Nilai RMSE dan Waktu komputasi (Saat memilih model terbaik dari 480 model)
  - RMSE terkecil

Best parameters: -0.336187 using {'activation': 'relu', 'batch_size': 32, 'dropout_rate': 0.1, 'epochs': 1000, 'neurons': 8, 'optimizer': 'rmsprop', 'verbose': 0}

- Waktu kompulasi

  01:34:18.01

Dataset 2 : CEDA.

- Metode pembagian data Split validation dan LSTM dengan Cross validation k = 2, Persentasi data latih 80% dan data uji 20%
- Aristektur NN
  - Hyper parameter

Hyper parameter	Nilai
Neuron	[8, 16]
Activation function	['sigmoid', 'tanh', 'relu', 'selu', 'elu', 'softplus']
Optimizer	['adam', 'adamax', 'rmsprop', 'sgd']
Dropout rate	[0.1]
Epochs	[500, 1000, 1500, 2000]
Batch_size	[8, 16, 32, 64],

  - Model LSTM satu layer
- Nilai RMSE dan Waktu komputasi (Saat memilih model terbaik dari 480 model)
  - RMSE terkecil

    Best parameters: -0.275854 using {'activation': 'tanh', 'batch_size': 64, 'dropout_rate': 0.1, 'epochs': 1500, 'neurons': 8, 'optimizer': 'sgd', 'verbose': 0}
  - Waktu kompulasi

    01:31:23.53

Berdasarkan hasil percobaan diambil keputusan sebagai berikut:

- Percobaan dengan dataset BMKG memperoleh nilai negated RMSE terbaik yaitu sebesar -0.336187 dengan arsitektur {'activation': 'relu', 'batch_size': 32, 'dropout_rate': 0.1, 'epochs': 1000, 'neurons': 8, 'optimizer': 'rmsprop', 'verbose': 0}.
- Percobaan dengan dataset CEDA memperoleh nilai negated RMSE terbaik yaitu sebesar -0.275854 dengan arsitektur {'activation': 'tanh', 'batch_size': 64, 'dropout_rate': 0.1, 'epochs': 1500, 'neurons': 8, 'optimizer': 'sgd', 'verbose': 0}
- Perbedaan dataset mempengaruhi arsitektur terbaik yang dihasilkan.
- Dataset CEDA lebih baik dari dataset BMKG karena memberikan hasil RMSE yang lebih baik yaitu sebesar -0.275854.
- Dataset Ceda memerlukan waktu selama 01:31:23.53 yang memiliki waktu yang lebih singkat dari dataset Y yang memerlakukan waktu selama 01:34:10.01, karena datasetnya sama-sama memiliki 240 data maka waktu yang diperlukan tidak berbeda jauh.