Introduction to Machine Learning

ANALISIS KELAYAKAN AIR UNTUK DIKONSUMSI

(Water safe to consumption)

oleh

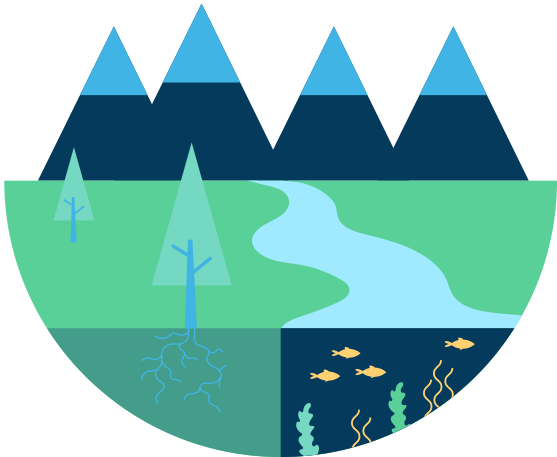
Muhammad Ramadhani - Batch 9

Latar Belakang

Kelangkaan air merupakan masalah sosial dan masalah lingkungan. Air adalah kebutuhan manusia yang berperan penting dalam kehidupan karena kita membutuhkannya untuk pangan, sanitasi, kesehatan, dan energi. Pandemi COVID19 telah menunjukkan pentingnya sanitasi untuk mencegah penyakit. Jelas bahwa dampak dari COVID19 jauh lebih beresiko pada masyarakat miskin perkotaan yang tinggal di daerah kumuh dan tidak memiliki akses air bersih untuk konsumsi. 3 miliar orang di seluruh dunia kekurangan fasilitas dasar cuci tangan di rumah. Oleh karena itu, kelangkaan air diperkirakan dapat menggusur 700 juta orang pada tahun 2030. Di sisi lain, air tidak hanya penting bagi manusia tetapi juga bagi makhluk hidup lainnya di bumi ini. Oleh karena itu, penggunaan dan konsumsi sumber daya yang bertanggung jawab merupakan aspek penting dari kelangkaan air.

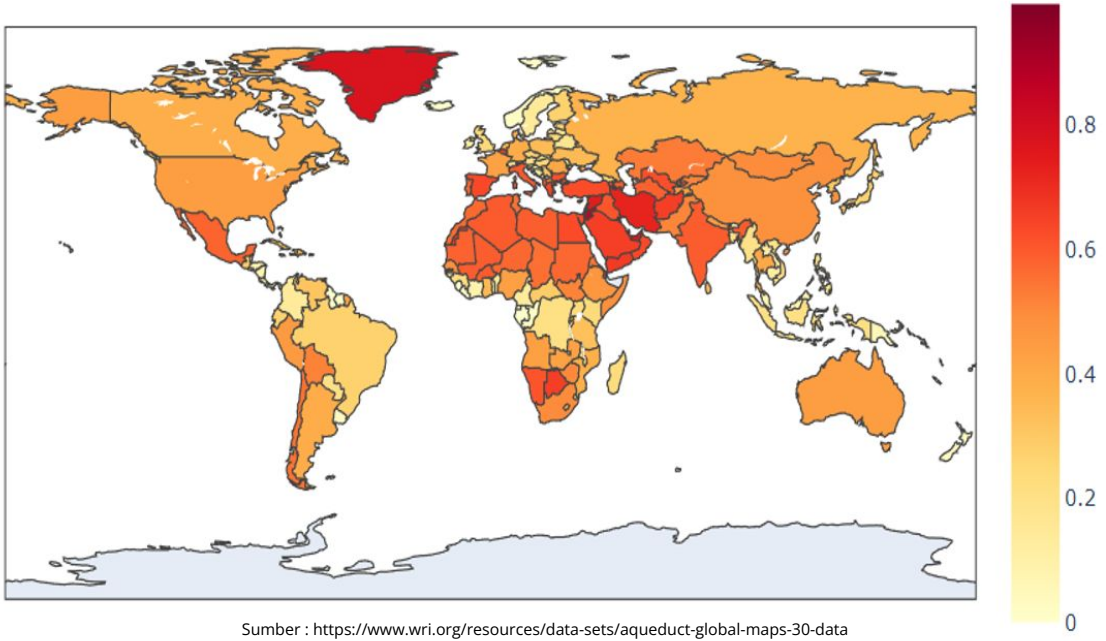
Water Shortage

Meskipun 2/3 bagian bumi diselimuti air, nyatanya hanya sedikit air yang baik untuk kita cerna atau dikonsumsi. Dari keseluruhan jumlah air, sekitar 97 persen merupakan air laut yang tidak bisa kita konsumsi dan sekitar tiga persen merupakan air tawar. Dua dari tiga persen air tawar tersebut merupakan wujud air beku yang berada di kutub utara dan kutub selatan dan hanya 0,62 persen air bersih yang layak konsumsi dari sisa 1 persen air tawar tersebut. Pasokan air bersih akan terus berkurang, karena terganggunya sirkulasi air pembangunan kota yang terbuat dari beton dan aspal yang menghalau air hujan untuk dapat diserap oleh tanah.



Sebaran Kurangnya Air Bersih di Dunia

Seperti yang terlihat pada peta sebaran air di sebelah dimana semakin merah warna peta maka semakin sedikit pula jumlah air bersih di daerah tersebut.



Sumber : https://www.wri.org/resources/data-sets/aqueduct-global-maps-30-data

Sebaran Kurangnya Air Bersih di Dunia

Menurut Perserikatan Bangsa Bangsa (PBB) pada 2019 mencatat bahwa 2,2 miliar orang atau seperempat populasi dunia masih kekurangan air minum yang aman dikonsumsi. Sementara itu, 4,2 miliar orang tidak memiliki layanan sanitasi yang aman dan 3 miliar tidak memiliki fasilitas cuci tangan dasar. Adapun menurut laporan Bappenas, ketersediaan air di sebagian besar wilayah Pulau Jawa dan Bali saat ini sudah tergolong langka hingga kritis. Sementara itu, ketersediaan air di Sumatera Selatan, Nusa Tenggara Barat, dan Sulawesi Selatan diproyeksikan akan menjadi langka atau kritis pada tahun 2045. Kelangkaan air bersih juga berlaku untuk air minum. Menurut RPJMN 2020-2024, hanya 6,87 persen rumah tangga yang memiliki akses air minum aman. Adapun berdasarkanSurvei Sosial Ekonomi Nasional (Susenas) 2020 dari BPS juga menunjukkan ada sebesar 90,21 persenrumah tangga yang memiliki akses air minum layak, meskipun distribusinya tidak merata.

Tabel 1. Standar Kebutuhan Air Departemen Pekerjaan Umum

Keperluan	Konsumsi (Liter/Orang/Hari)
Mandi, cuci, kakus	12,0
Minum	2,0
Cuci pakaian	10,7
Kebersihan rumah	31,4
Taman	11,8
Cuci kendaraan	21,1
Wudhu	16,2
Lain-lain	21,7
Jumlah	126,9

Objective Analysis

Berangkat dari masalah krisis air bersih yang masih dialami sebagian masyarakat. Objektif masalah dari analisis ini adalah memastikan air tersebut layak minum atau tidak. Berdasarkan data historis tujuan analisis ini memberikan prediksi apakah air yang terkandung dapat dikatakan layak minum atau tidak .

Informasi Dataset

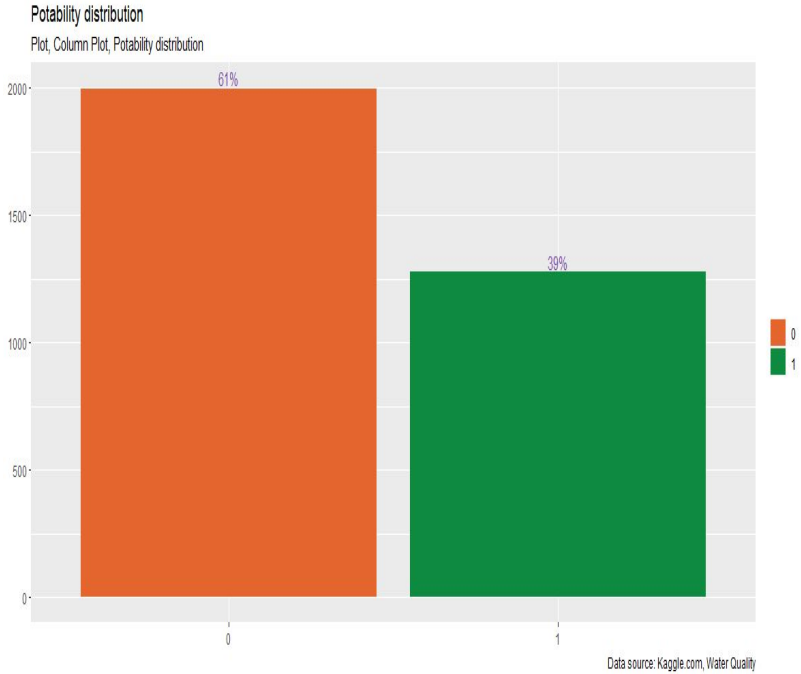
ph	PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status.
Hardness	Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.
Solids	Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc.
Chloramines	Chlorine and chloramine are the major disinfectants used in public water systems.
Sulfates	Sulfates are naturally occurring substances that are found in minerals, soil, and rocks.
Conductivity	Pure water is not a good conductor of electric current rather it's a good insulator. Increase in ions concentration enhances the electrical conductivity of water.
Organic carbon	Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water.
Trihalomethanes	THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated.
Turbidity	The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter.
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

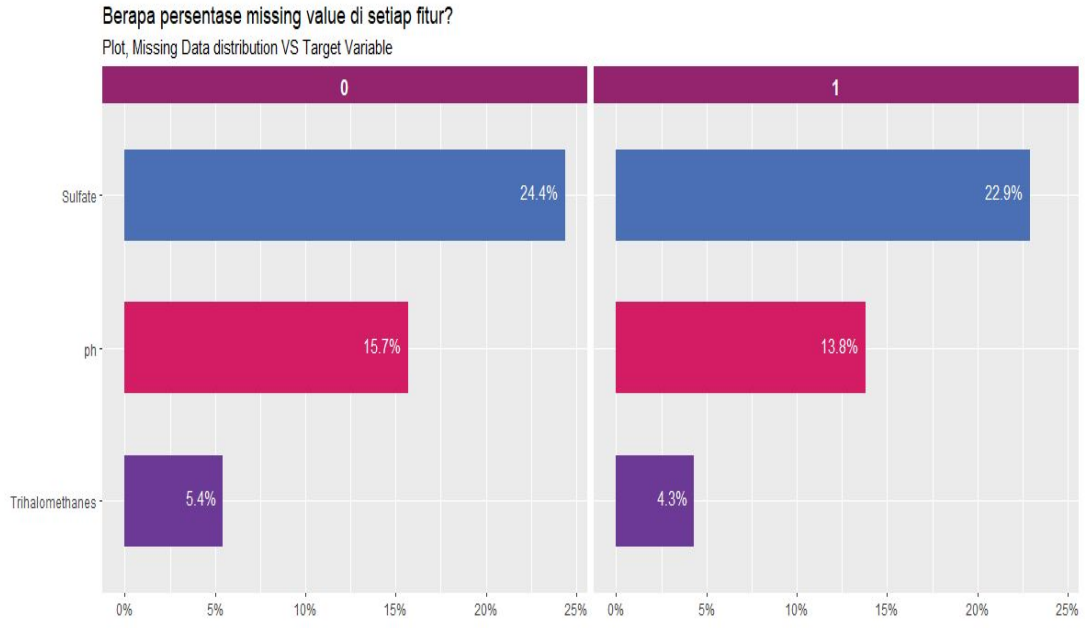
Data Processing

Pre-Processing and Exploratory Data

Cleansing, Exploratory, Scaling, and Clustering

Distribusi Berdasarkan Variabel Target

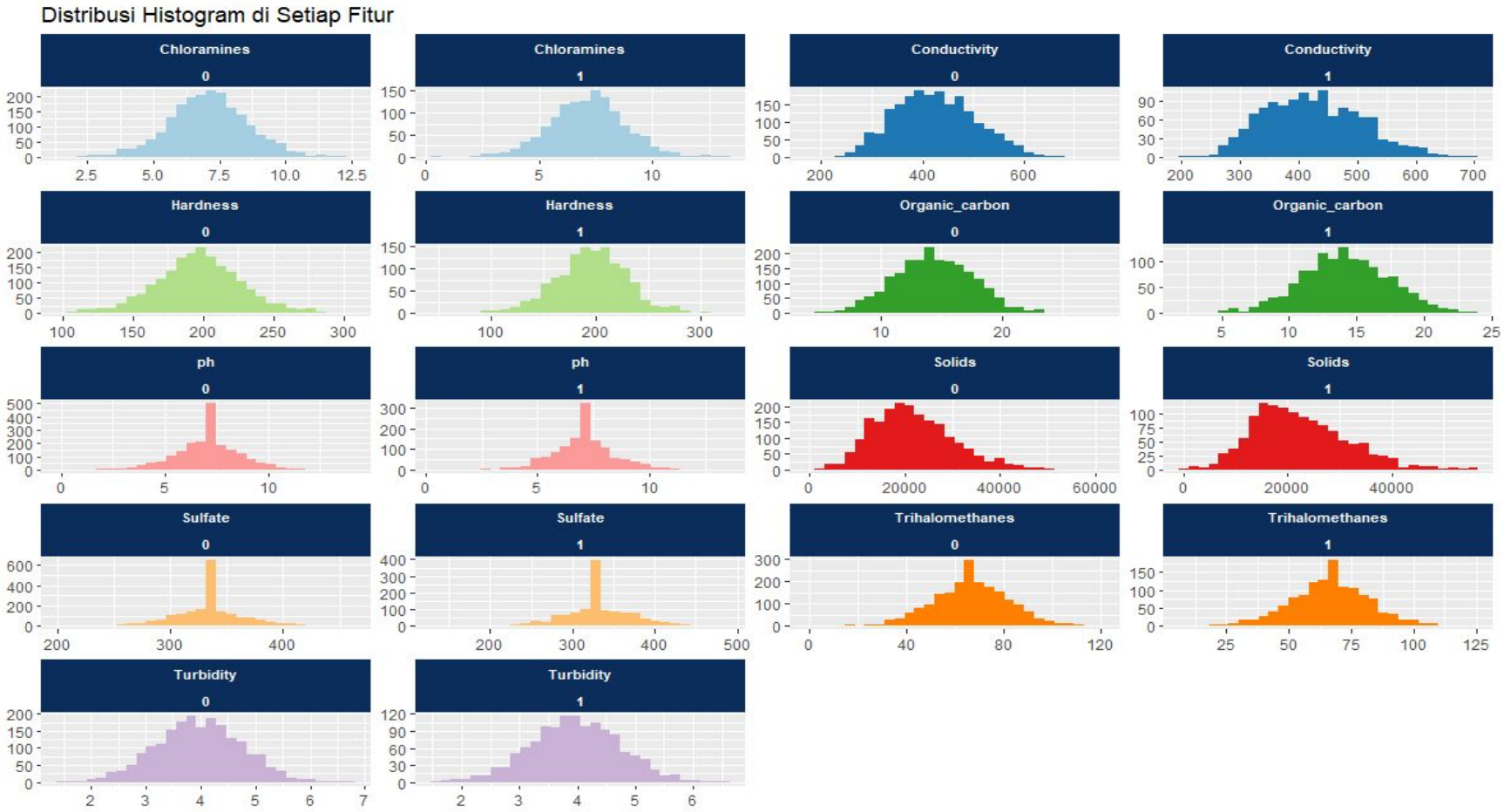




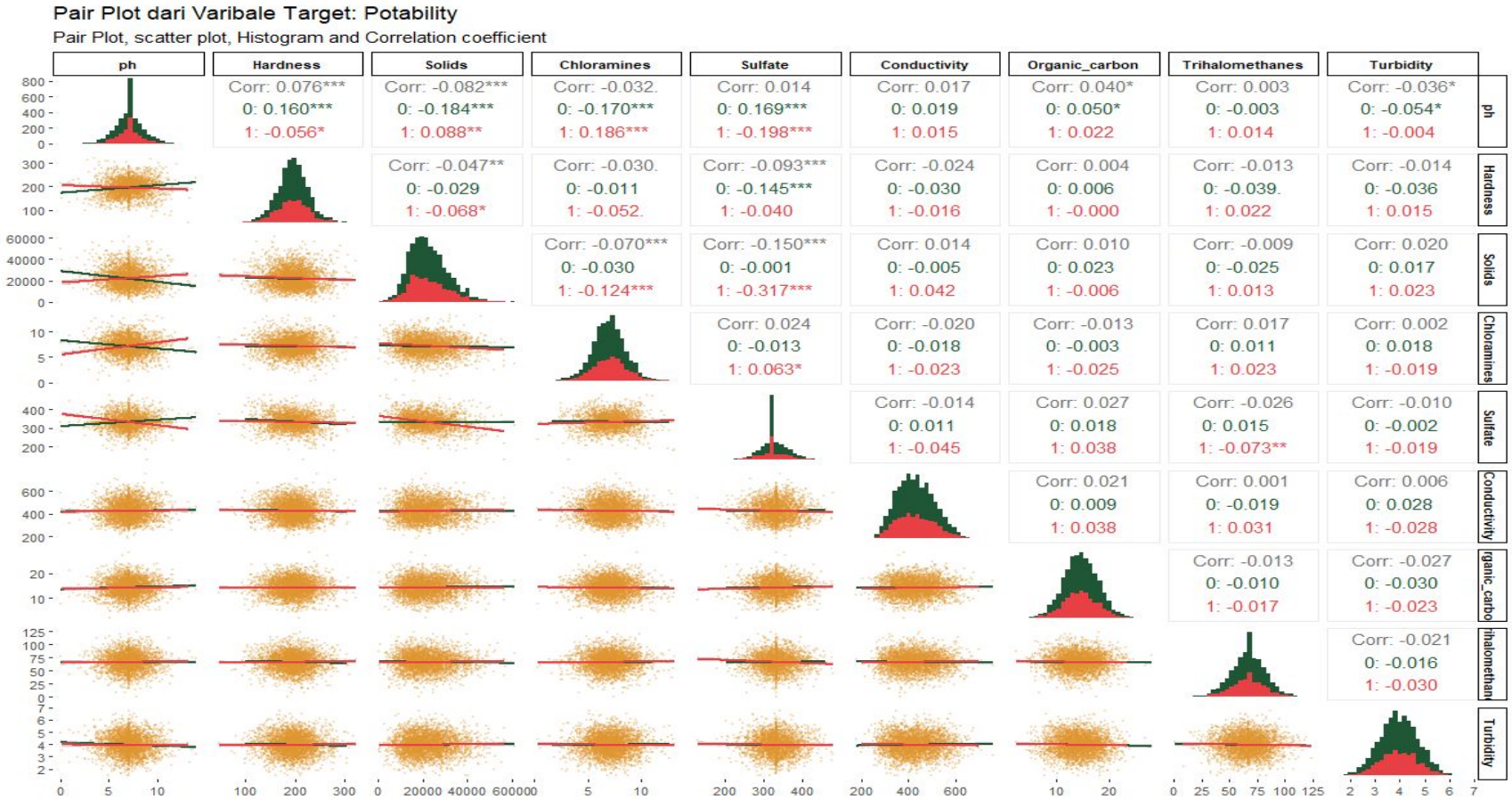
Treatment for Missing Value → Fill with Median

Persebaran Data dengan Box Plot

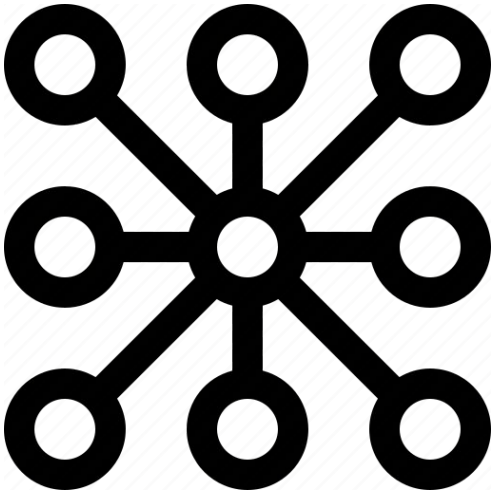
Persebaran Data dengan Histrogram



Correlation Plot



Clustering



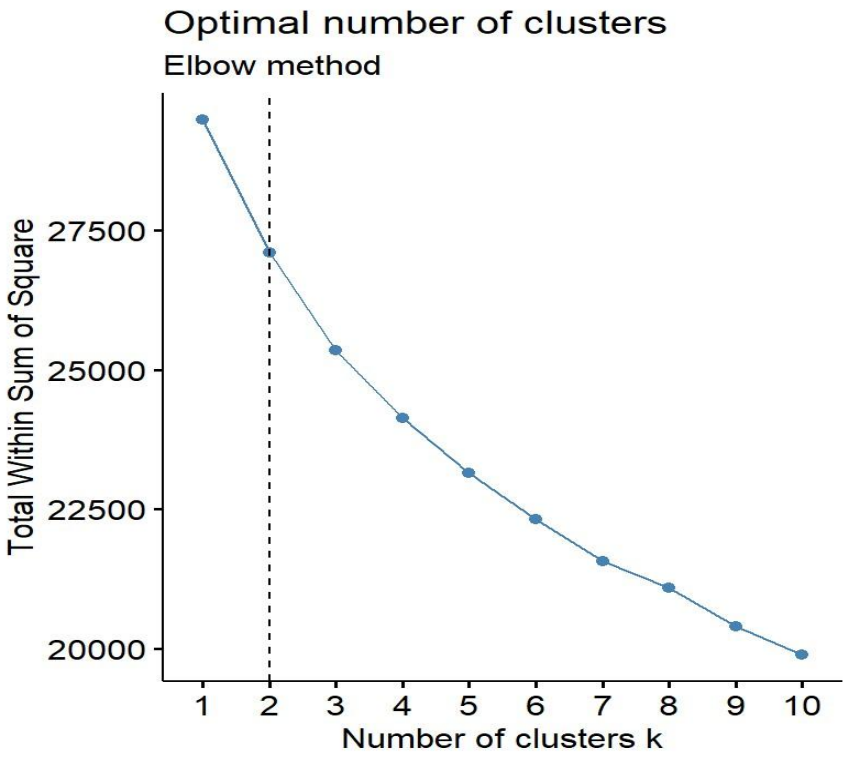
Method dalam Hiearchical Clustering

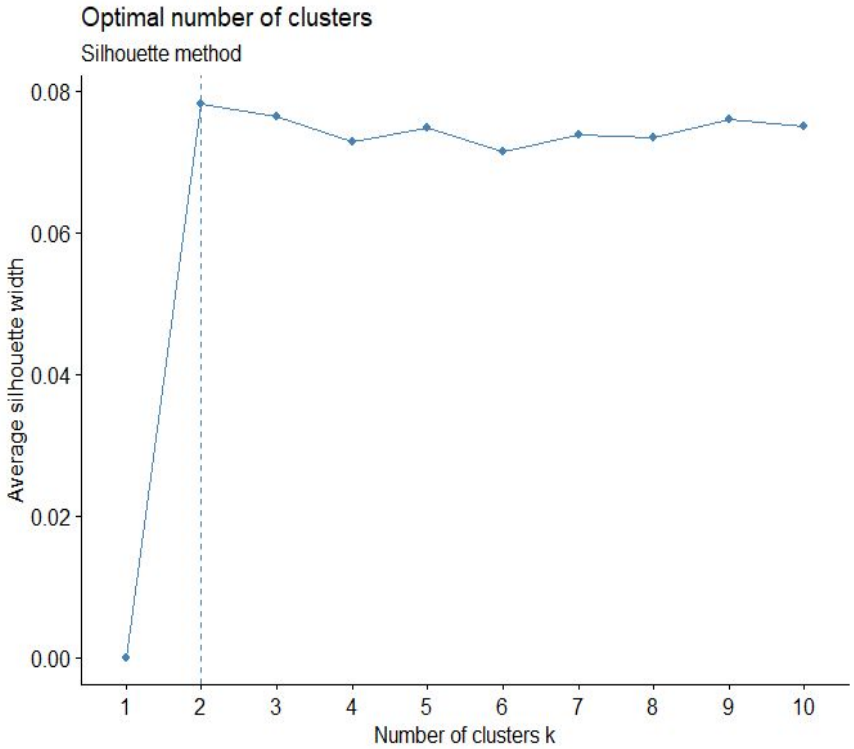
Linkage method yang diuji

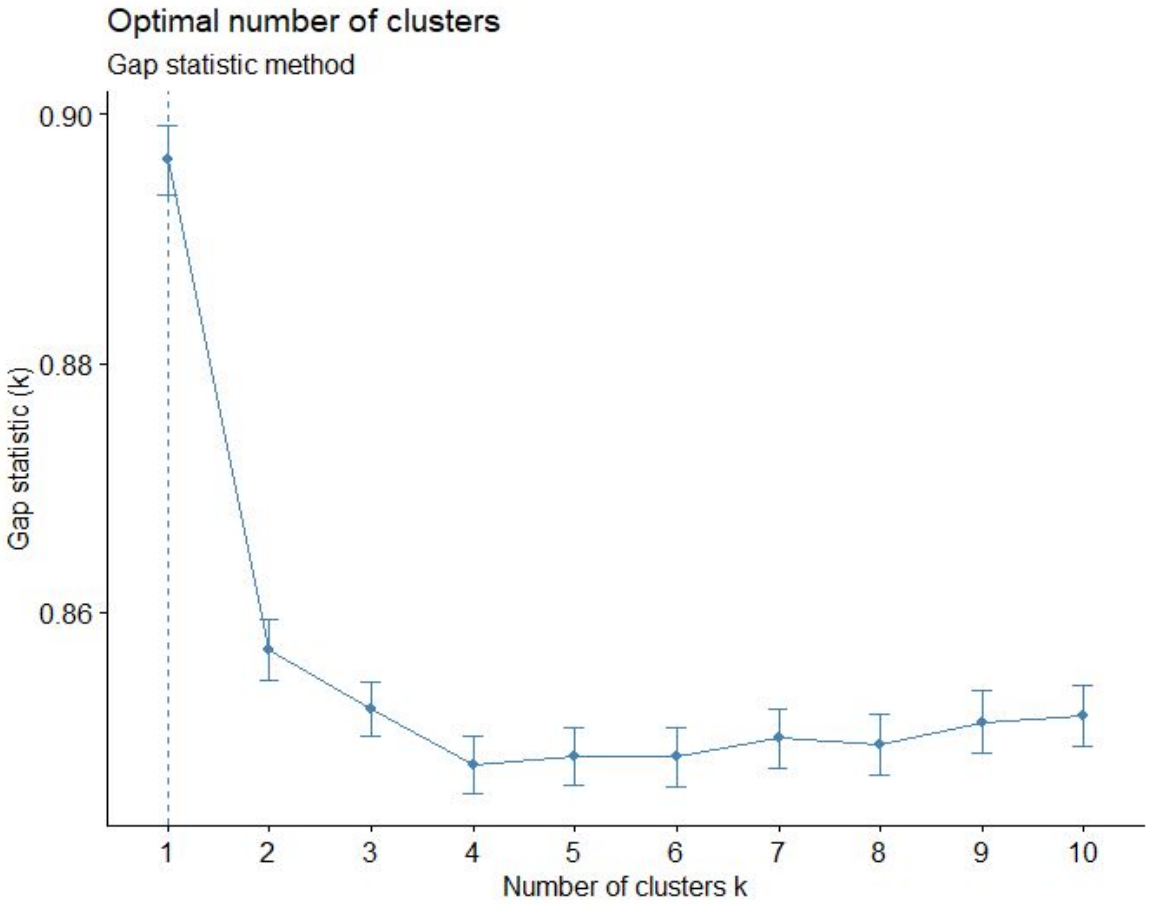
- Average
- Single
- Complete
- Ward

	Result
Average	0.7894
Single	0.6941
Complete	0.8829
Ward	0.9685

Penentuan K Optimal



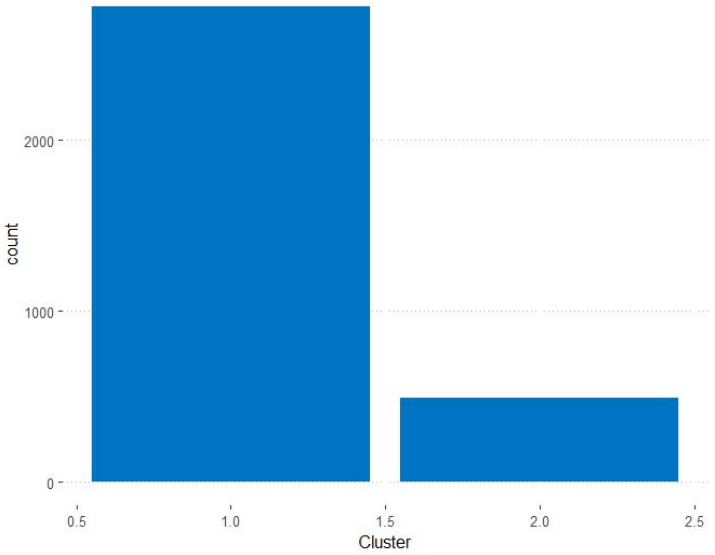


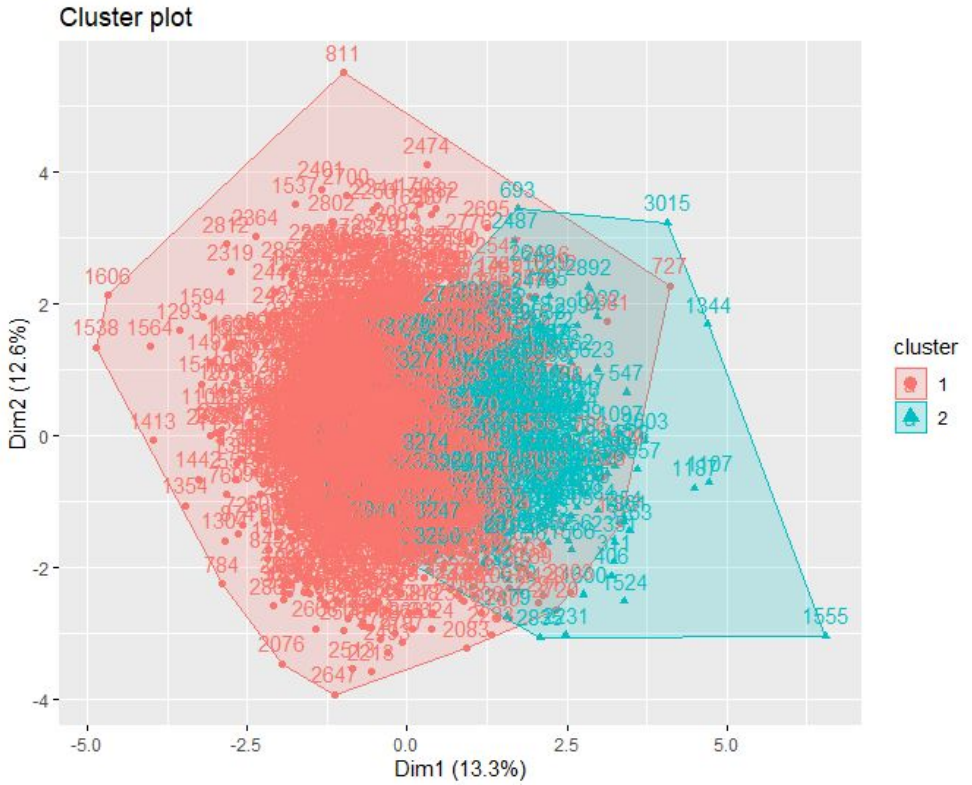


K Optimal

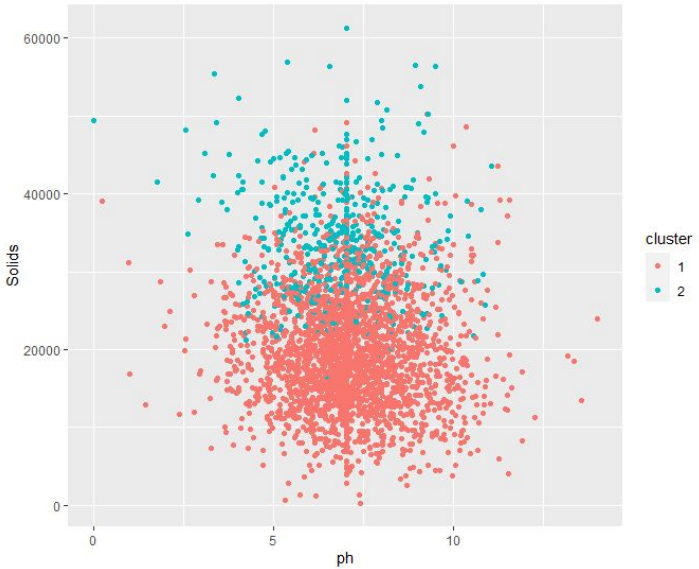
Berdasarkan tiga hasil sebelumnya, kami memutuskan k optimal yang akan digunakan sebesar 2

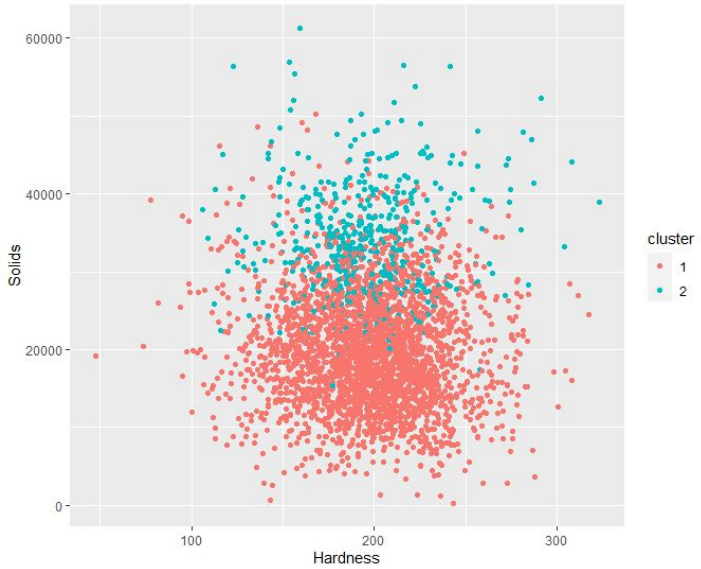
Hasil Hierarchical Clustering





Hierarchical Clustering



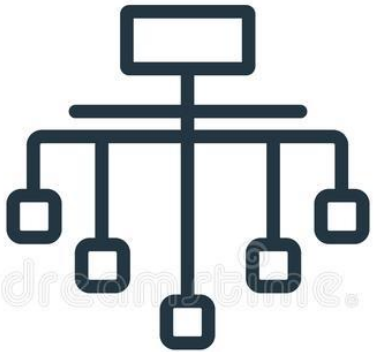


Hierarchical Clustering

Linkage method yang diuji

- Average
- Single
- Complete
- Ward

	Hasil
Average	0.7894
Single	0.6941
Complete	0.8829
Ward	0.9685

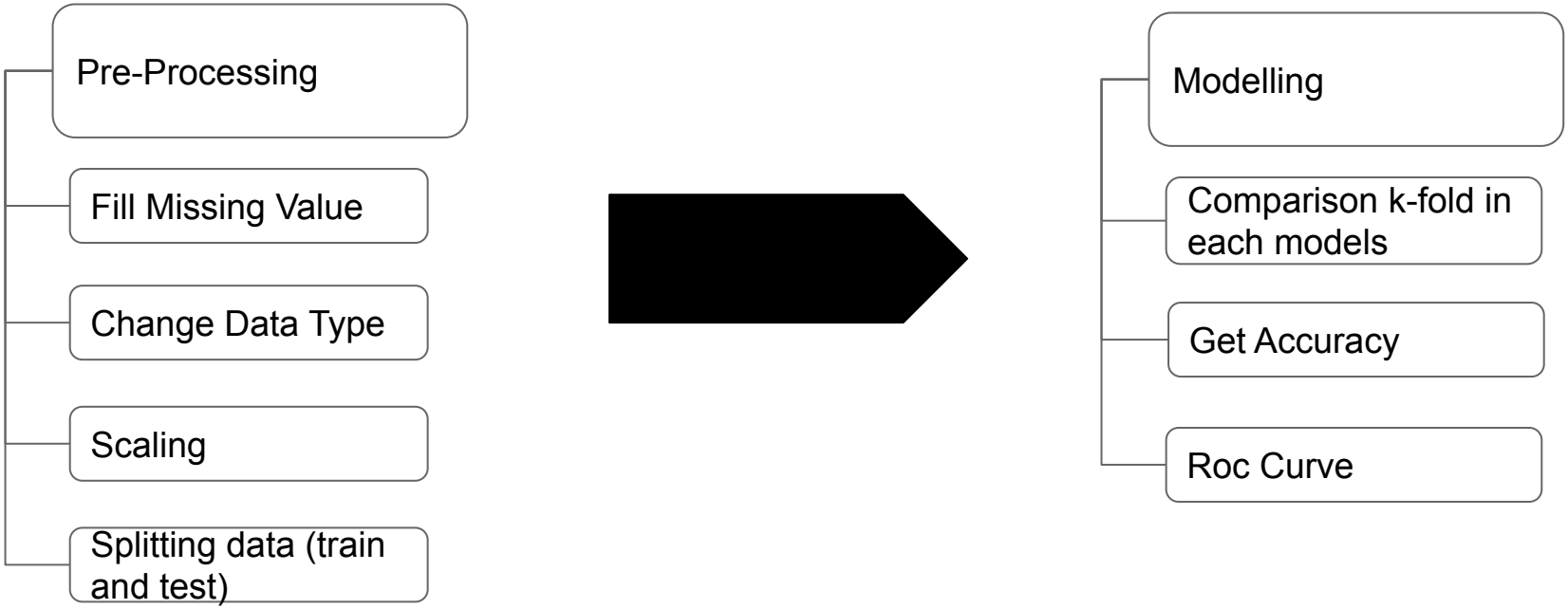


Model Classification

Comparison Accuracy, Cross Validation, dan Fitting Model

Classification

Alur Pengerjaan



Pro-Kontra Model yang Diuji

Logistic Regression

Pro: Memberikan informasi tentang signifikansi statistik dari fitur

Kontra: Butuh asumsi

KNN

Pro: Mudah dimengerti, cepat, dan efisien

Kontra: Butuh jumlah neighbours yang di-*state*

SVM dan Kernel SVM

Pro: Performa lebih bagus karena tidak bias sama outlier dan tidak sensitif terhadap overfit

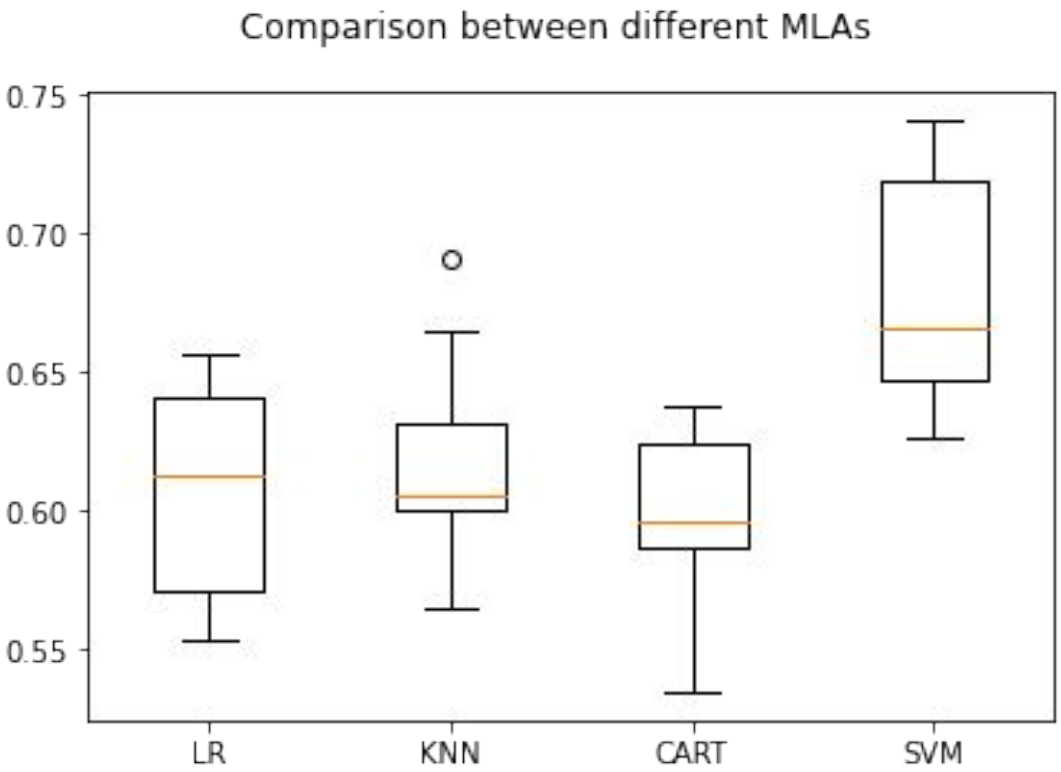
Kontra: Bukan opsi terbaik untuk dataset yang fiturnya besar

Decision Tree dan Random Forest

Pro: Bisa perform dalam linear maupun non linear problem

Kontra: untuk decision tree hasil maksimal butuh data yang lebih besar, lebih mudah mendapatkan hasil yang overfitting

Cross Validation in Multiple Algorithm



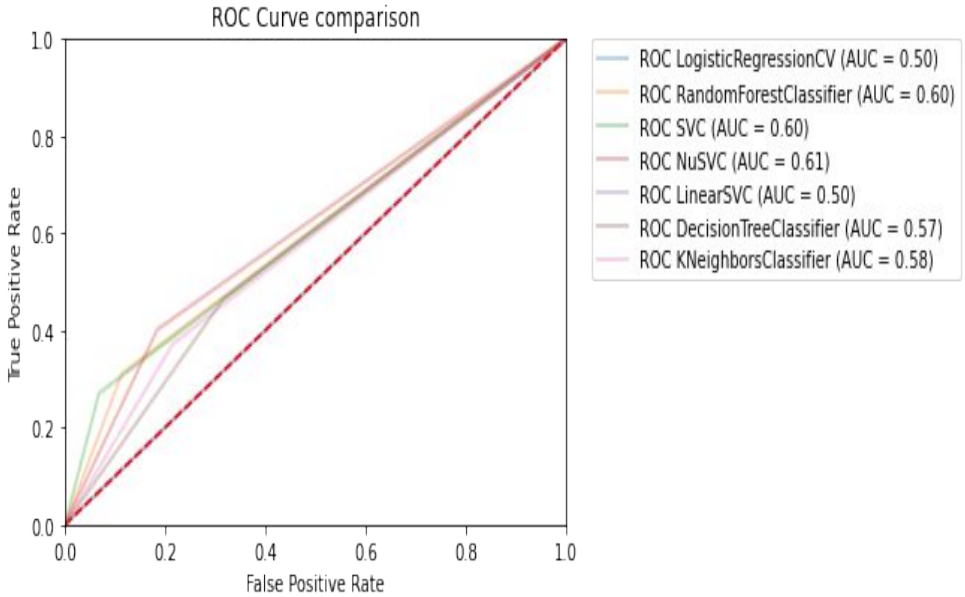
k-folds

Number split = 10

Scoring

Scoring menggunakan accuracy

Train & Test Accuracy



	MLA used	Train Accuracy	Test Accuracy	Precission	Recall	AUC
2	SVC	0.7359	0.6860	0.702128	0.270492	0.601265
1	RandomForestClassifier	1.0000	0.6707	0.618644	0.299180	0.594979
3	NuSVC	0.8645	0.6616	0.563218	0.401639	0.608587
0	LogisticRegressionCV	0.6053	0.6280	0.000000	0.000000	0.500000
4	LinearSVC	0.6053	0.6280	0.000000	0.000000	0.500000
6	KNeighborsClassifier	0.7653	0.6311	0.505556	0.372951	0.578466
5	DecisionTreeClassifier	1.0000	0.5976	0.459677	0.467213	0.570985

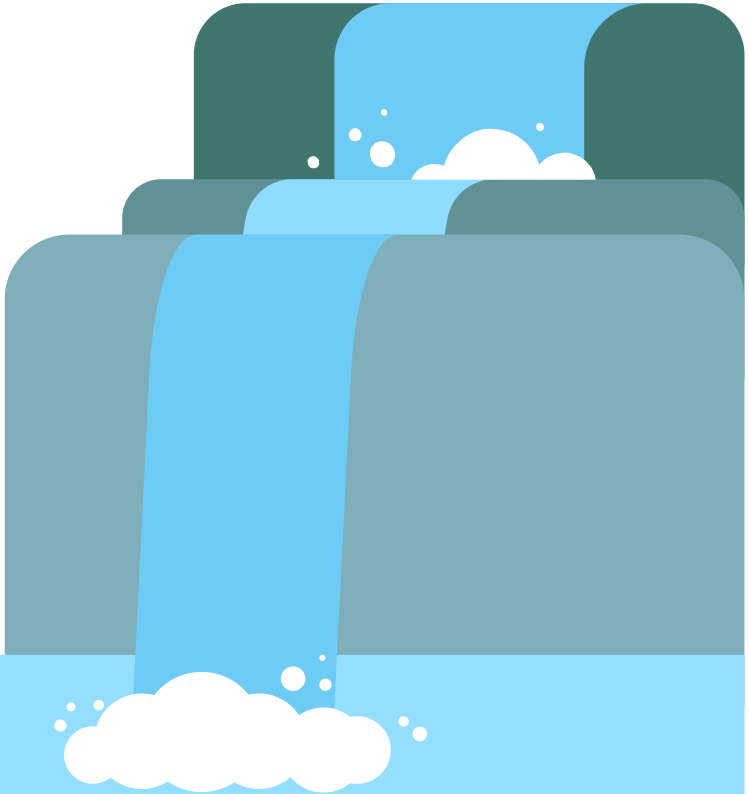
Future Engineering

- Melakukan Cross validation dalam Random forest and tree classifier agar terhindar dari overfitting dan menaikkan accuracy model
- Melakukan HyperParameter Tuning
- Me-*balance*-kan predictors
- Mereduksi fitur-fitur, seperti melakukan pca, features selection, dan research lainnya

Kesimpulan

- Model terbaik dilihat dari akurasinya adalah Support Vector Classifier, Random forest,
- K optimal untuk clustering sebesar 2
- Pemilihan model tergantung dengan objektif masalah, di mana jika ingin memprediksi nilai yang benar, maka model dapat diperhatikan dari sisi hasil akurasinya.

Sekian dan Terima Kasih



Daftar Pustaka

Pai, P. (2021, May 7). *Hierarchical clustering explained | by Prasad Pai*. Towards Data Science. Retrieved December 3, 2022, from https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

Itah, A. Y., & Akpan, C. E. (2005). Potability of drinking water in an oil impacted community in southern Nigeria.

Shekhar, A. (2018, Februari 15). *What Is Feature Engineering for Machine Learning?* Medium. https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a