




# MACHINE LEARNING : MODELING USING ORANGE

---

*Alvandi Damansyah*

*1103192191*



# Table of Content

1

## Introduction

Familiarize what machine learning is

2

## Machine Learning Model

Machine learning model used for test

5

## Result

Result of test



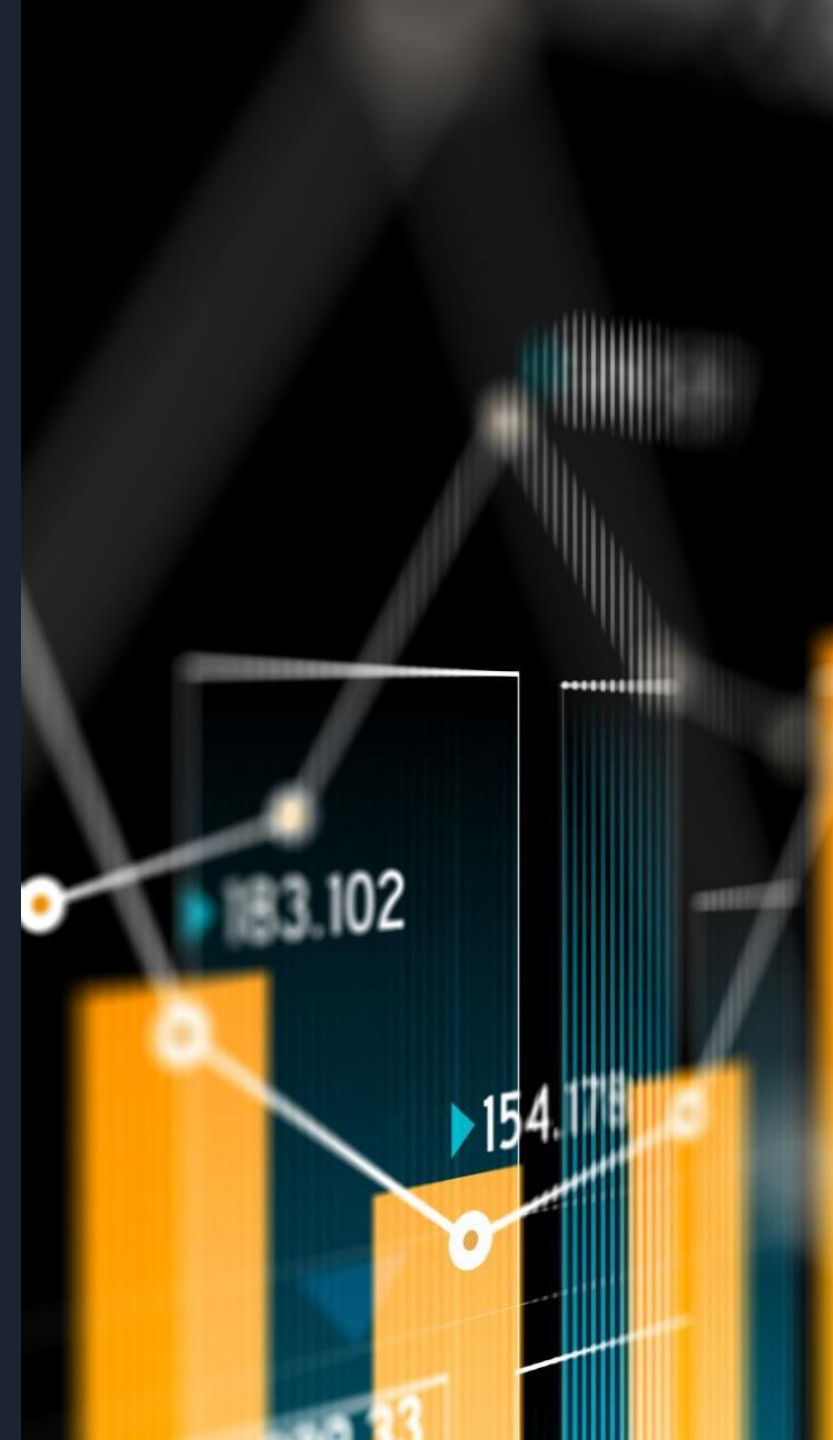
# INTRODUCTION



# ORANGE DATA MINING

---

Orange adalah perangkat lunak sumber terbuka (open source) untuk analisis data dan data mining. Orange dapat membantu pengguna untuk mengimpor, memvisualisasikan, dan menganalisis data, serta membuat model prediksi dan clustering, dengan fitur-fitur yang lengkap.



# MACHINE LEARNING

---

Machine learning adalah suatu metode pengembangan kecerdasan buatan (artificial intelligence/AI) yang memungkinkan sistem komputer untuk mempelajari pola dan memperbaiki kinerjanya dari pengalaman-pengalaman masa lalu, tanpa perlu secara eksplisit diprogram oleh manusia. Machine learning menggunakan algoritma dan model matematika untuk menemukan pola dalam data dan memprediksi hasil di masa depan berdasarkan pola yang ditemukan.



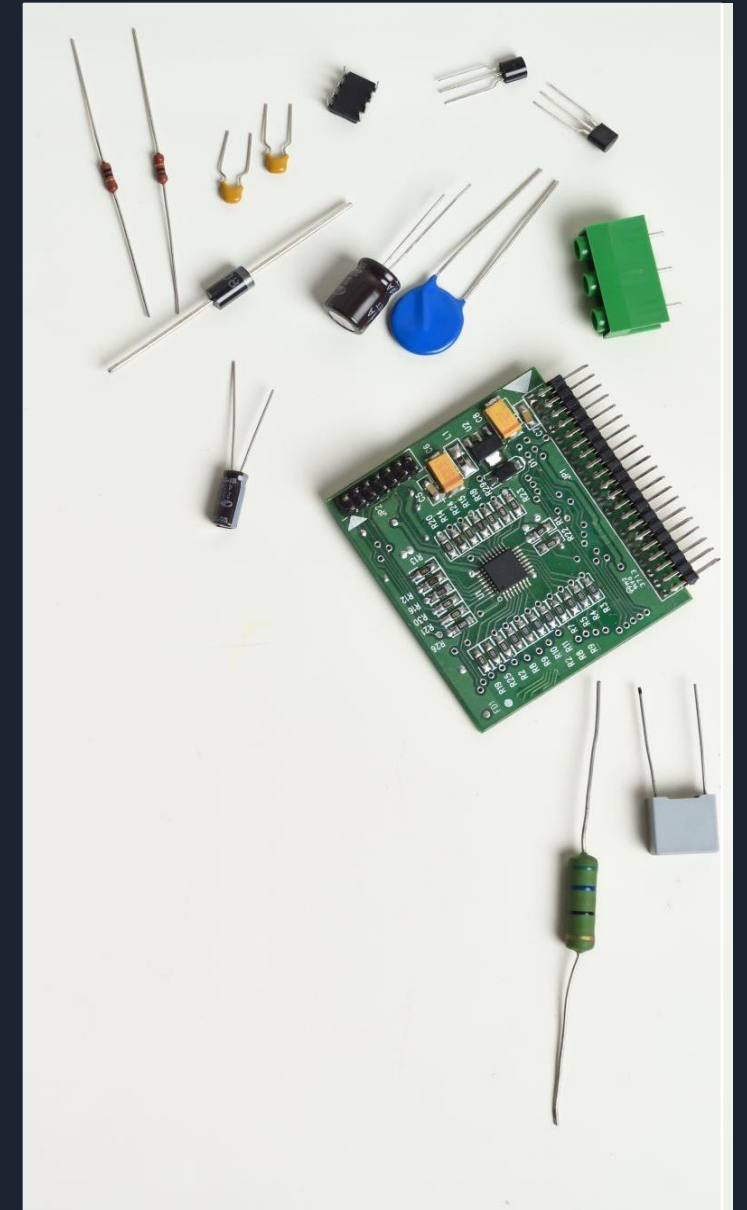
# COMMONLY USED MACHINE LEARNING MODELS

---

Berikut adalah penjelasan singkat mengenai beberapa model machine learning yang umum digunakan:

- Linear Regression: Model machine learning yang digunakan untuk melakukan prediksi nilai berdasarkan hubungan linear antara satu atau beberapa variabel input dengan variabel target.
- Logistic Regression: Model machine learning yang digunakan untuk melakukan klasifikasi pada dua kelas atau lebih dengan menentukan probabilitas kelas tertentu.
- Decision Tree: Model machine learning yang menghasilkan struktur pohon keputusan untuk melakukan klasifikasi atau regresi.
- Random Forest: Model machine learning yang menggunakan banyak pohon keputusan secara acak untuk meningkatkan akurasi prediksi.
- Support Vector Machine (SVM): Model machine learning yang mencari garis pemisah antara dua kelas dengan jarak terbesar untuk melakukan klasifikasi atau regresi.

Model-machine learning yang umum di atas dapat digunakan untuk tugas-tugas machine learning pada berbagai bidang, tergantung pada karakteristik data dan tujuan dari tugas yang ingin dicapai. Pemilihan model-machine learning yang tepat dan akurat dapat membantu dalam meningkatkan akurasi prediksi dan menghasilkan solusi yang efektif.





# RECOMMENDATION OF TOP 3 MACHINE LEARNING MODELS FOR CLASSIFICATION

---

Berikut adalah tiga model machine learning terbaik untuk tugas klasifikasi:

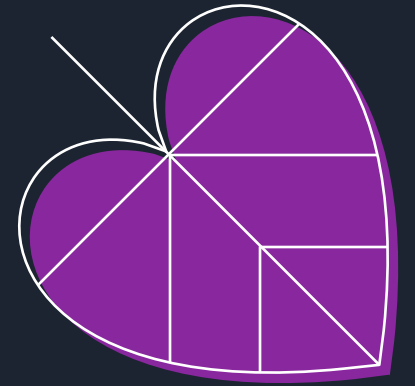
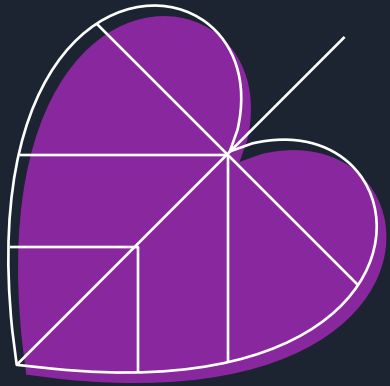
- Random Forest Random Forest adalah model machine learning yang digunakan untuk klasifikasi dan regresi. Model ini menggabungkan banyak pohon keputusan (decision tree) yang dihasilkan secara acak untuk membuat prediksi yang lebih akurat. Model ini juga memiliki kemampuan untuk menangani data yang tidak seimbang dan dapat mengidentifikasi fitur yang penting dalam data.
- Support Vector Machine (SVM) SVM adalah model machine learning yang dapat digunakan untuk klasifikasi dan regresi. SVM mencari garis pemisah yang optimal antara dua kelas dengan memaksimalkan jarak antara kelas tersebut. Model ini dapat mengatasi masalah data yang tidak seimbang dan memiliki kemampuan untuk menangani data dengan jumlah fitur yang sangat besar.
- K-Nearest Neighbors (KNN) KNN adalah model machine learning yang digunakan untuk klasifikasi dan regresi. Model ini bekerja dengan mencari k-nearest neighbors (tetangga terdekat) dari suatu data uji untuk menentukan label dari data tersebut. Model ini mudah diimplementasikan dan cocok untuk data yang memiliki pola yang kompleks.

Pilihan model terbaik untuk tugas klasifikasi dapat bervariasi tergantung pada karakteristik data dan konteks masalah yang dihadapi. Oleh karena itu, sebaiknya dilakukan eksplorasi dan evaluasi model secara lebih komprehensif untuk memastikan pemilihan model yang sesuai.



# MACHINE LEARNING MODELS

---

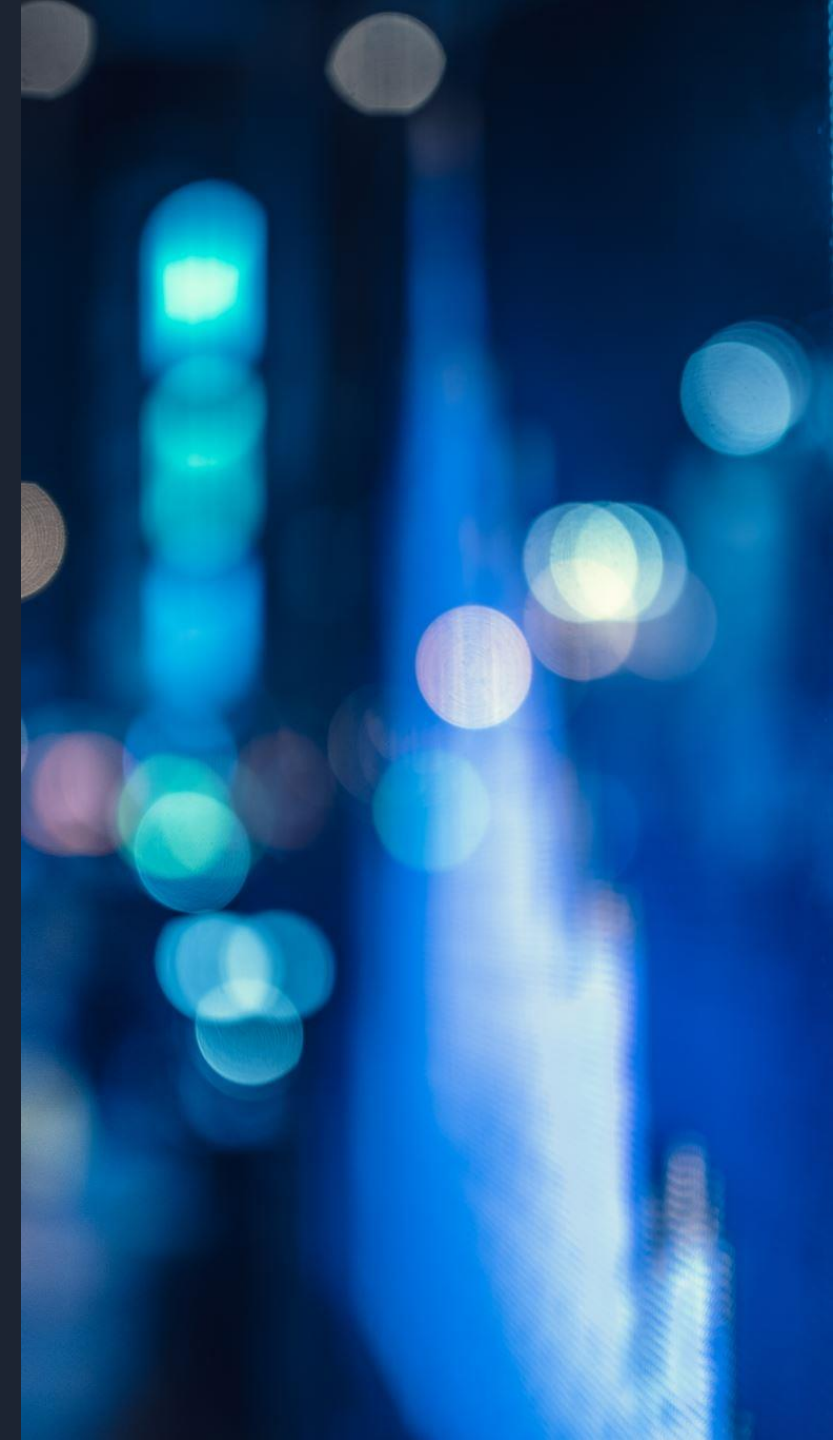




# LINEAR MODEL

---

Linear model adalah suatu pendekatan dalam analisis statistik untuk memodelkan hubungan antara variabel dependen dan satu atau lebih variabel independen dengan asumsi bahwa hubungan tersebut bersifat linear. Model ini menggunakan persamaan garis lurus untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen.



# DECISION TREE MODEL

---

Decision tree model adalah salah satu metode pemodelan prediktif dalam analisis data yang digunakan untuk membuat keputusan berdasarkan serangkaian pertanyaan yang saling terkait. Model ini membentuk pohon keputusan dengan membagi data menjadi kelompok-kelompok yang semakin kecil dan spesifik, dan pada setiap simpul atau cabang pohon keputusan, terdapat pertanyaan yang harus dijawab untuk menentukan cabang mana yang harus diambil.

# RANDOM FOREST MODEL

---

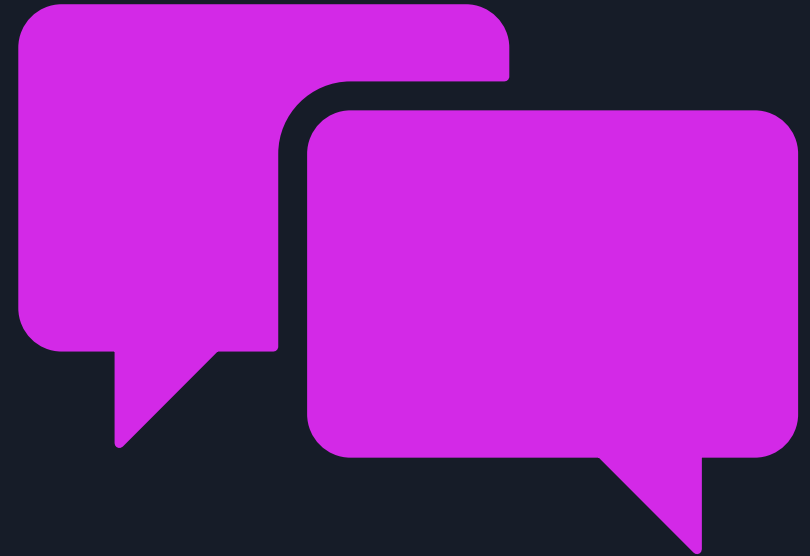
Random forest model adalah suatu algoritma pemodelan prediktif dalam analisis data yang memanfaatkan kumpulan pohon keputusan (decision trees) yang dihasilkan secara acak untuk meningkatkan akurasi prediksi. Model ini bekerja dengan cara membagi data menjadi beberapa bagian, kemudian membangun beberapa pohon keputusan berbeda pada setiap bagian dengan menggunakan subset data yang diambil secara acak.



# CHAT GPT

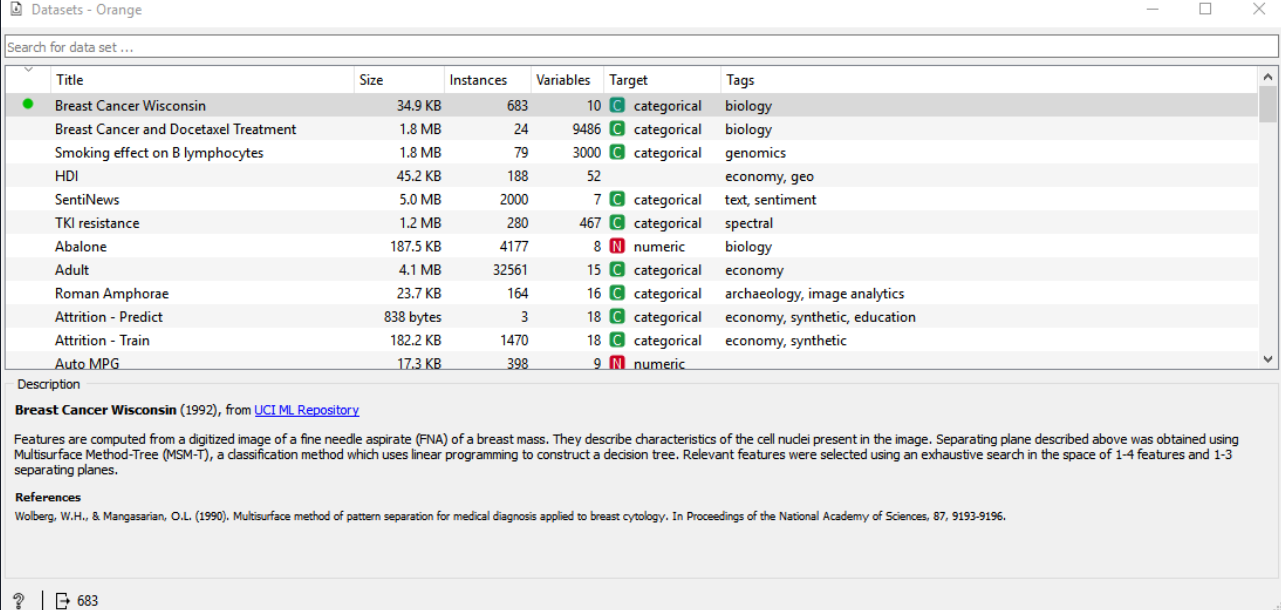
---

Chat GPT adalah model kecerdasan buatan yang dapat memahami bahasa manusia dan menghasilkan teks yang dapat dipahami oleh manusia. Model ini dilatih menggunakan dataset yang sangat besar dan menggunakan teknologi transformer untuk memperbaiki kinerja pemrosesan bahasa alami. Chat GPT dapat digunakan untuk berbagai aplikasi NLP seperti chatbot dan sistem rekomendasi. Dengan kemampuannya, Chat GPT dapat membantu meningkatkan efisiensi, akurasi, dan produktivitas dalam berbagai industri.



# DATASET

Dataset Breast Cancer Wisconsin (Diagnostic) yang dipublikasikan merupakan salah satu dataset yang populer dan sering digunakan dalam penelitian dan tugas machine learning. Dataset ini terdiri dari 569 sampel jaringan payudara yang dikumpulkan dari pasien wanita dengan kanker payudara dan memiliki 30 fitur yang dihitung dari citra digitalisasi jaringan payudara.



Title	Size	Instances	Variables	Target	Tags
Breast Cancer Wisconsin	34.9 KB	683	10	categorical	biology
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3000	categorical	genomics
HDI	45.2 KB	188	52		economy, geo
SentiNews	5.0 MB	2000	7	categorical	text, sentiment
TKI resistance	1.2 MB	280	467	categorical	spectral
Abalone	187.5 KB	4177	8	numeric	biology
Adult	4.1 MB	32561	15	categorical	economy
Roman Amphorae	23.7 KB	164	16	categorical	archaeology, image analytics
Attrition - Predict	838 bytes	3	18	categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	numeric	

Description

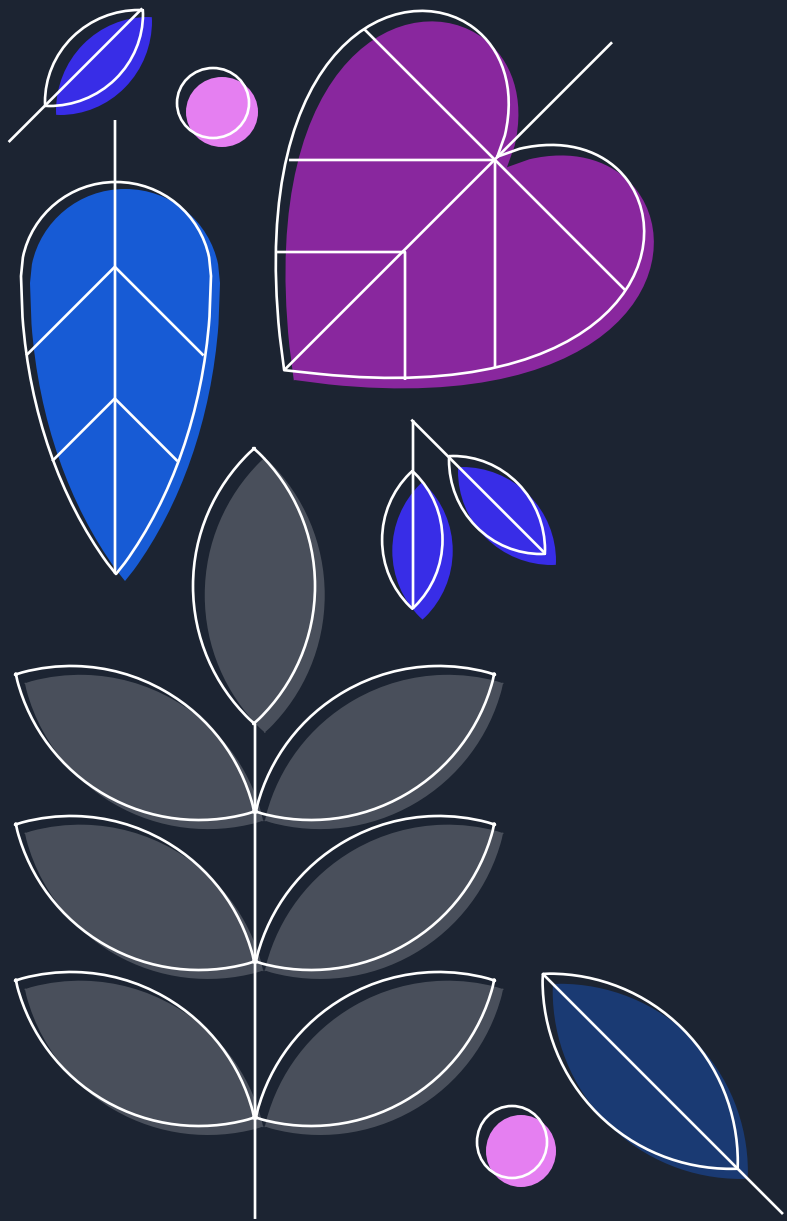
**Breast Cancer Wisconsin** (1992), from [UCI ML Repository](#)

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Separating plane described above was obtained using Multisurface Method-Tree (MSM-T), a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

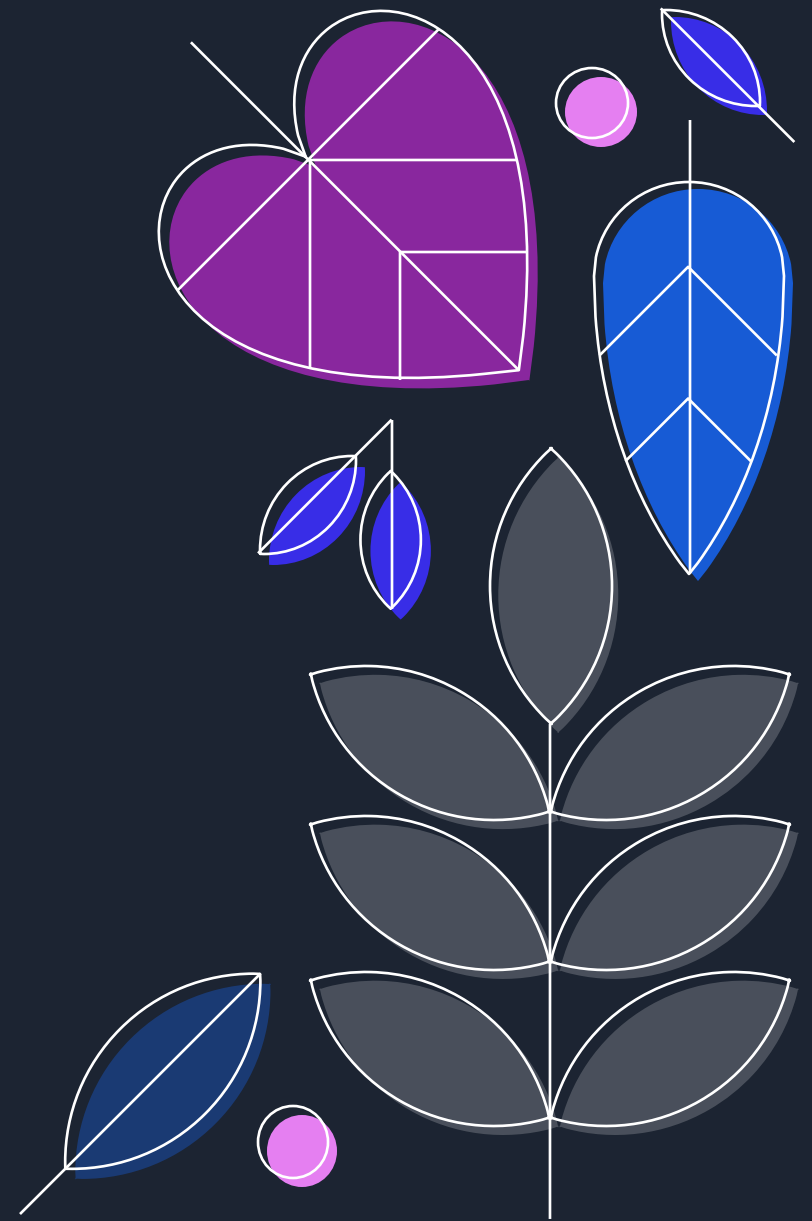
References

Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193-9196.

683



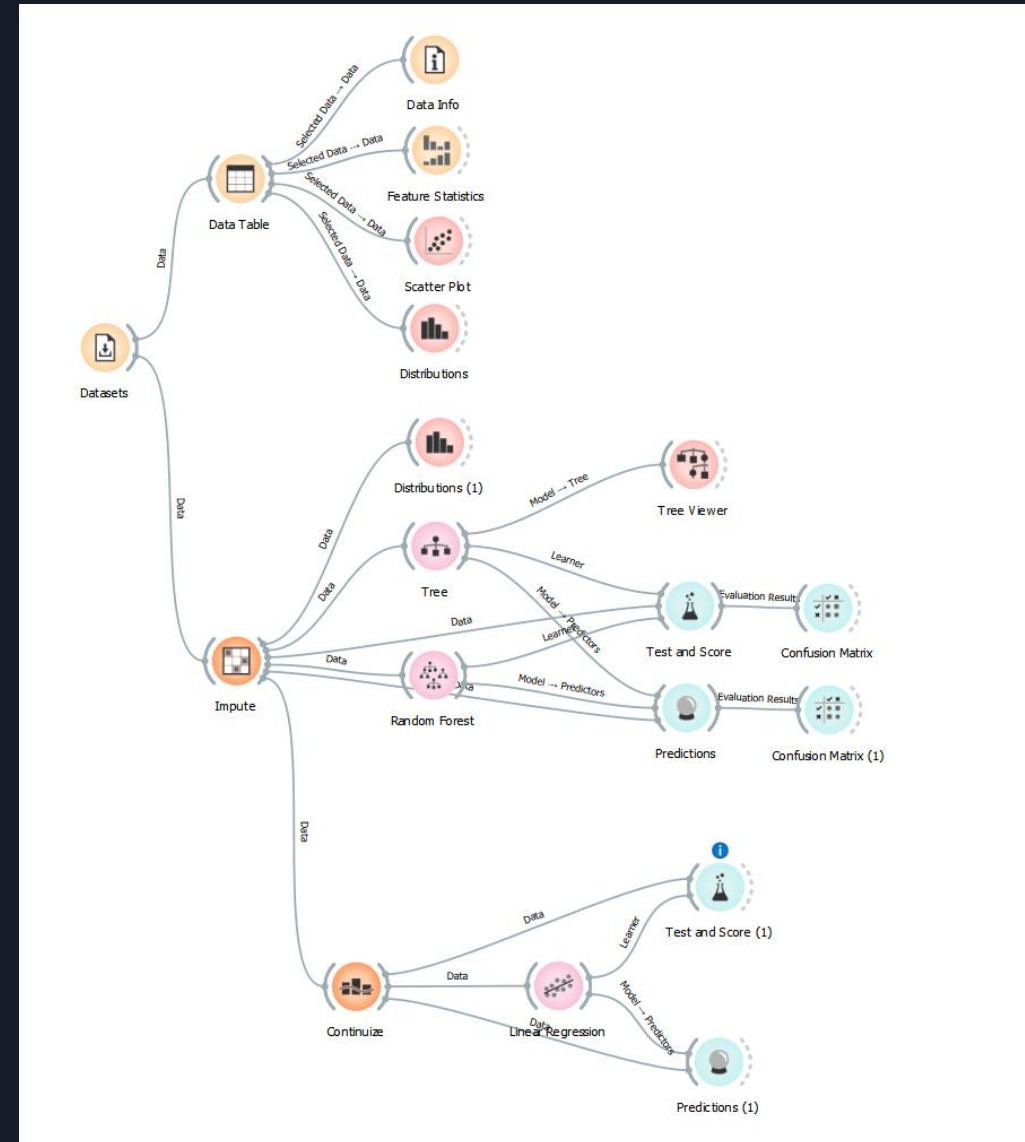
RESULT





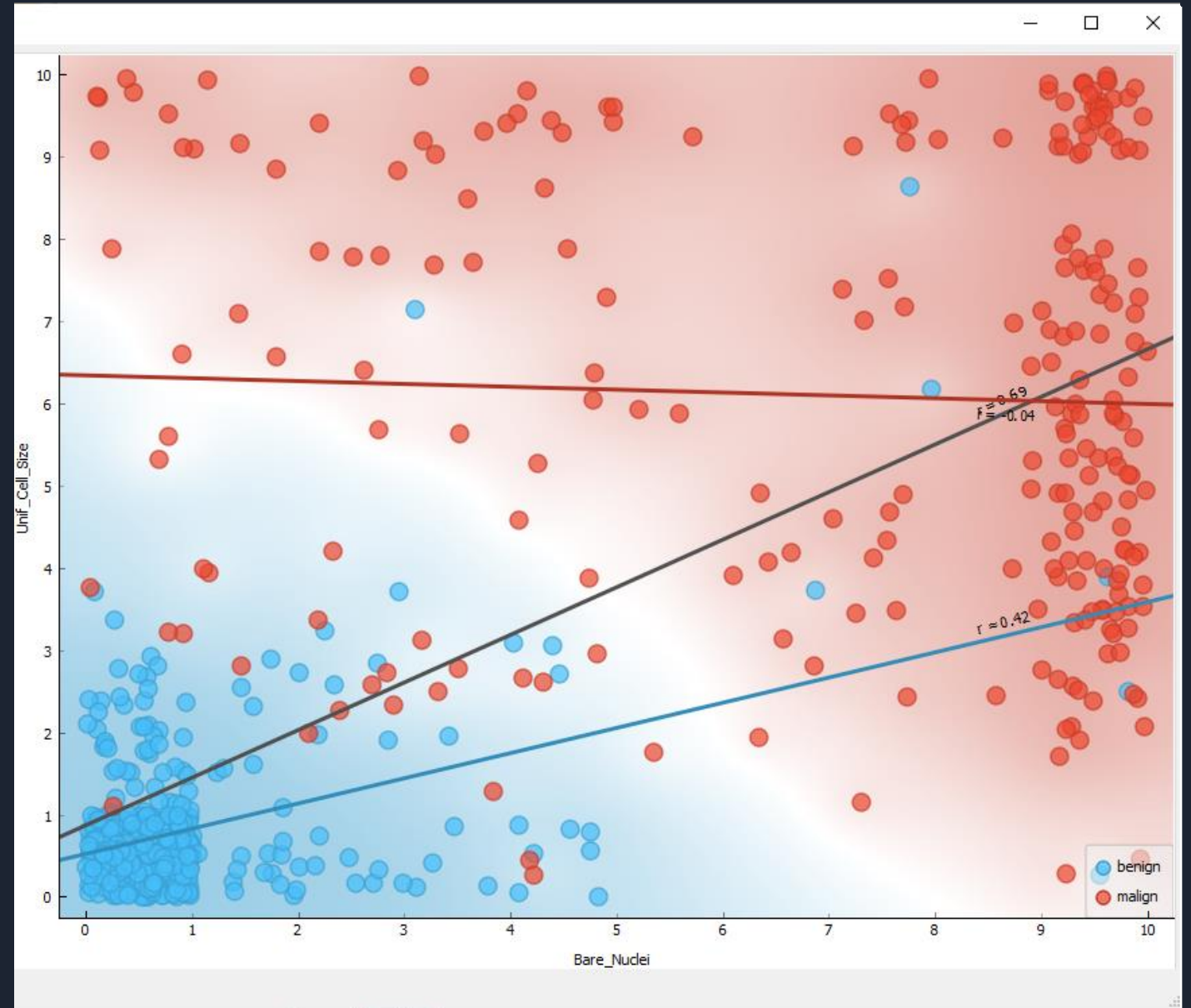
# VISUALISATION (Data Table) SUBDIVISION

Terlihat dari visualisasi program yang dikerjakan. Dataset diolah menjadi data table yang dianalisis dan dilihat data info, dan statistik dari data yang dipakai, yang Selanjutnya divisualisasikan supaya mudah dilihat.



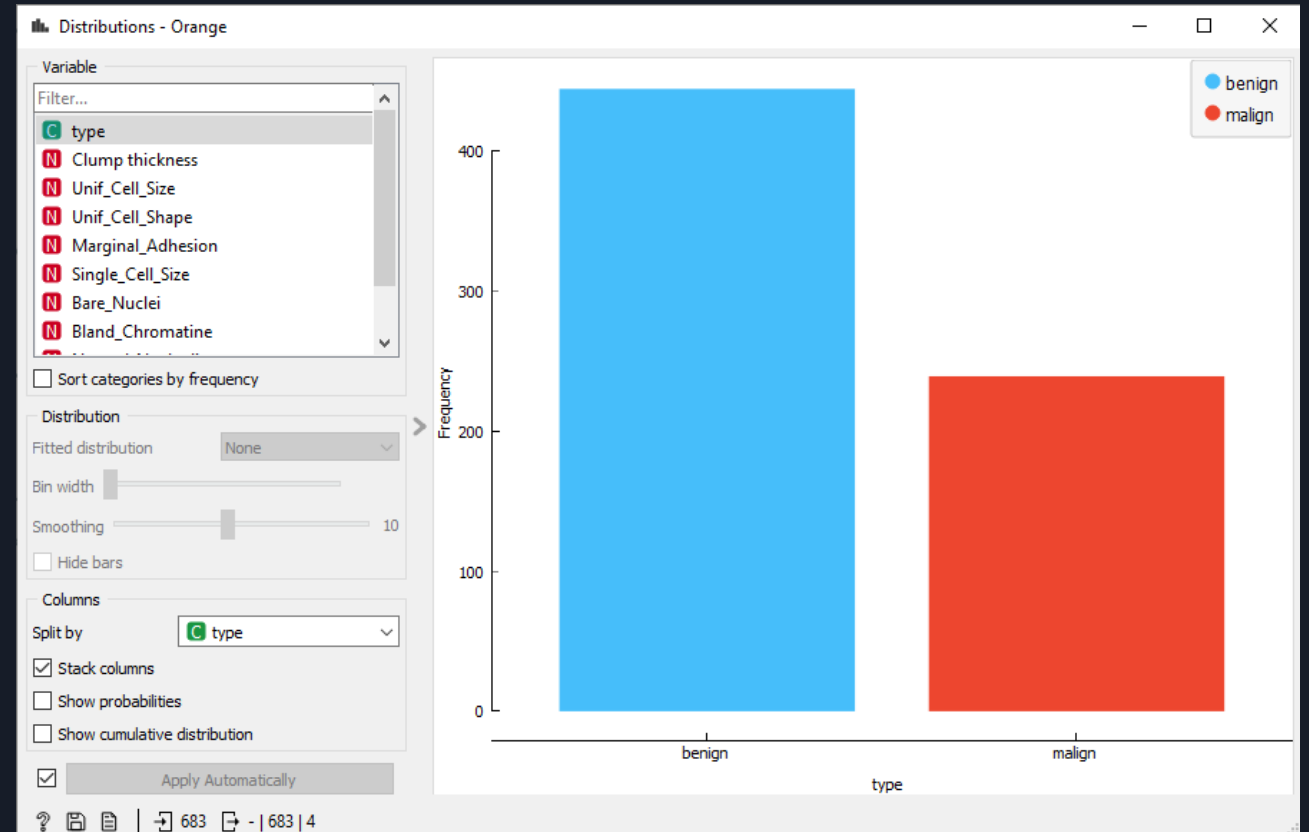
# SCATTERPLOT VISUALISATION

Dalam visualisasi scatterplot dapat terlihat berdasarkan axis y `unif_cell_size` dan axis x `bare_nuclei`, terlihat perbedaan rata-rata ukuran dari kanker benign dan malign.



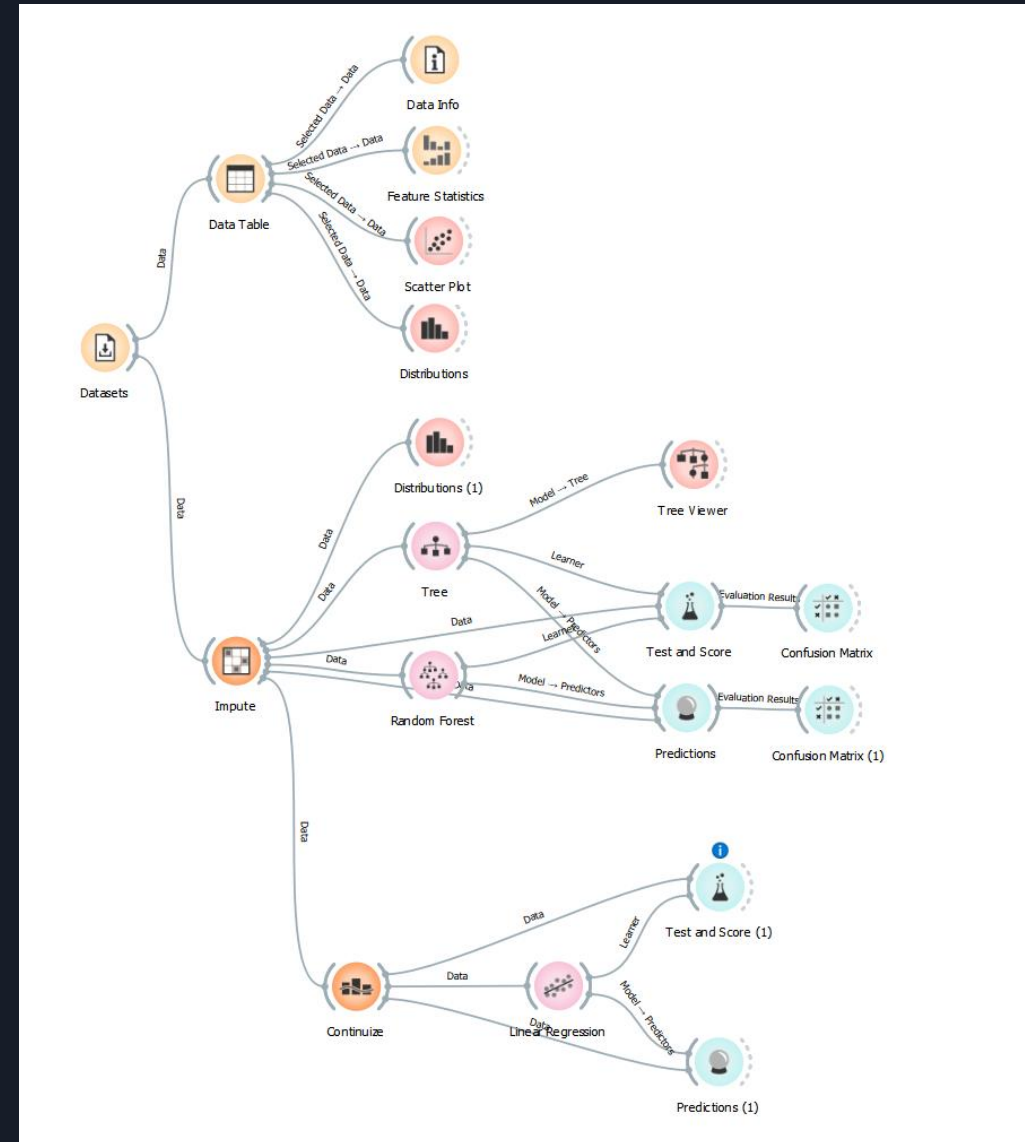
## DISTRIBUTION VISUALISATION

Dalam visualisasi distribusi yang divisualisasikan secara barplot dalam variabel tertentu dapat terlihat perbedaan besar data benign dan malign



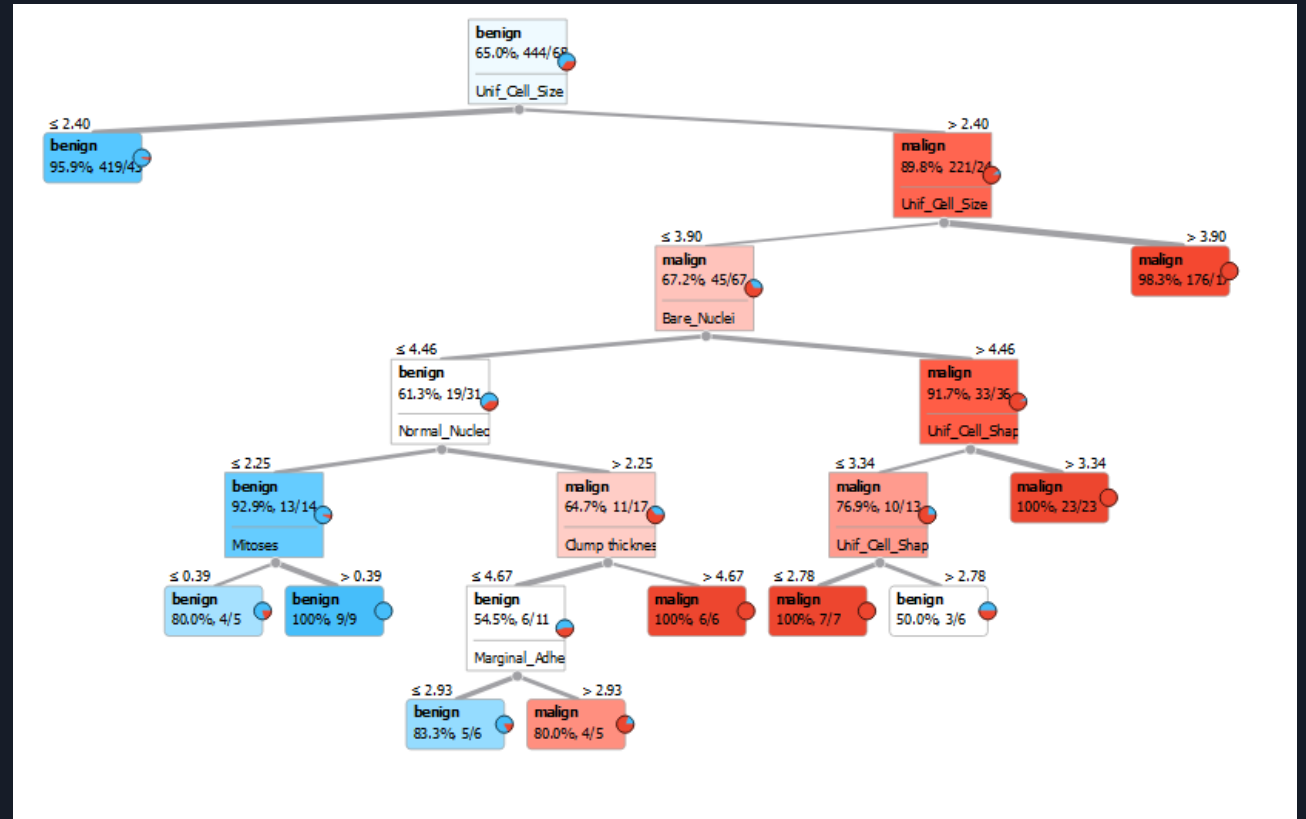
# IMPUTE SUBDIVISION

Dalam cabang impute, menghilangkan data yang buruk dalam dataset supaya model yang digunakan dapat bekerja dengan baik.



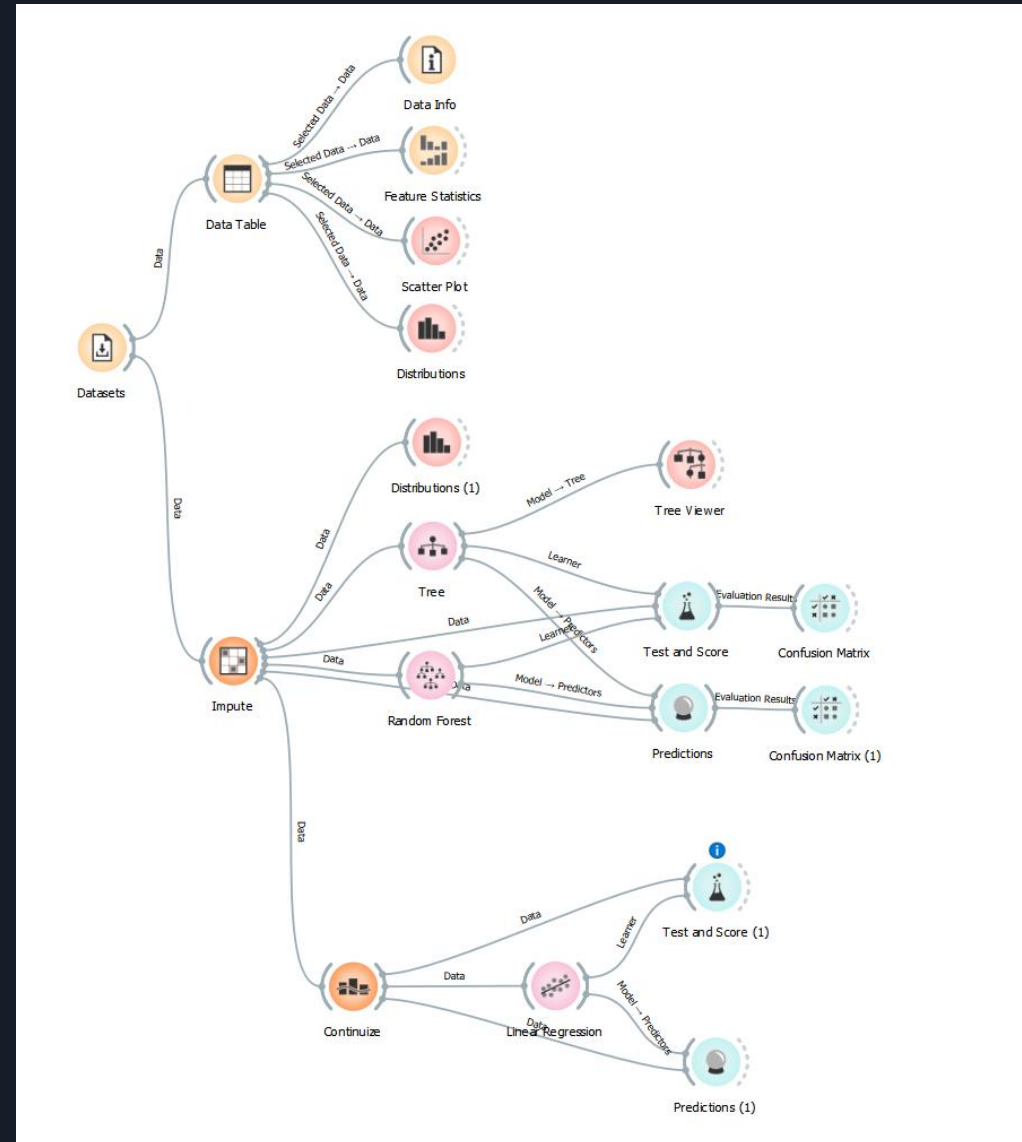
# DECISION TREE VISUALISATION

Dalam visualisasi decision tree dapat terlihat berdasarkan unif\_cell\_size mem perlihatkan diklasifikasi kanker jenis belign atau malign berdasarkan ukuran cell nya



# LINEAR SUBDIVISION

pada model linear menggunakan cabang sendiri dikarenakan linear regression hanya dapat mengolah data kolom data interger. Dikarenakan hal itu merubah kolom type menjadi interger sebelum diolah.





# TEST AND SCORE

---

Terlihat dalam model decision tree dan random forest memiliki nilai F1 tinggi, yang menandakan akurasi model tinggi terhadap data.

Dalam model linear regression terlihat nilai MSE, RMSE, dan MAE rendah, hal ini menandakan nilai error rendah.

Test and Score - Orange

☒ Cross validation  
Number of folds: 5  
☒ Stratified  
☐ Cross validation by feature

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
Tree	0.940	0.941	0.941	0.941	0.941
Random Forest	0.987	0.960	0.961	0.961	0.960

Test and Score (1) - Orange

☒ Cross validation  
Number of folds: 5

Model	MSE	RMSE	MAE	R2
Linear Regression	0.039	0.197	0.129	0.830

# Q&A CHATGPT

---

Terlihat contoh chatGPT menjawab pertanyaan yang diberikan berdasarkan data yang berada dalam databasenya.

A

linear regression error Numeric class variable expected.



Linear regression is a type of supervised machine learning algorithm used for predicting numerical (continuous) values based on input features. In most implementations of linear regression, the target variable (also known as the dependent variable or response variable) is expected to be a continuous numerical variable, and the input features (also known as independent variables or predictors) are also expected to be numeric.



The error message "Numeric class variable expected" typically occurs when you are trying to use linear regression with a target variable that is not numeric, but rather categorical or discrete. Linear regression requires a continuous target variable to estimate the relationship between the input features and the target variable using a linear model.

To resolve this error, you can check the data type of your target variable and make sure it is a numeric (continuous) variable. If it is not numeric, you may need to convert it to a numeric format or choose a different algorithm that is suitable for categorical or discrete variables, such as logistic regression for binary classification or decision trees for multi-class classification. Alternatively, if your target variable represents an ordinal variable with a meaningful order, you could consider using ordinal regression techniques instead of linear regression.