FINAL PROJECT BASIC MACHINE LEARNING

Analisis Klasifikasi Kanker Jinak dan Kanker Ganas Berdasarkan *Breast Cancer Wisconsin Dataset*

TIM B :

Adam Kemal Fajri
Hasyim
Indra Wahyudi
Nugi Gahara Yasa
Tomy Tjandra

DATASET

SUMBER DATA

Data yang digunakan dalam penelitian ini adalah **data sekunder** yang didapatkan dari situs komunitas sains data, **www.kaggle.com**. Data "*Breast Cancer Wisconsin*" terdiri dari **569 pengamatan** terhadap **31 variabel**.

Dataset "*Breast Cancer Wisconsin*" adalah dataset mengenai **kanker payudara** yang didapat melalui perhitungan terhadap **gambar digital** atas **uji Aspirasi Jarum Halus** (FNA) dari **massa payudara**. Data ini menggambarkan mengenai **karakteristik dari inti sel**

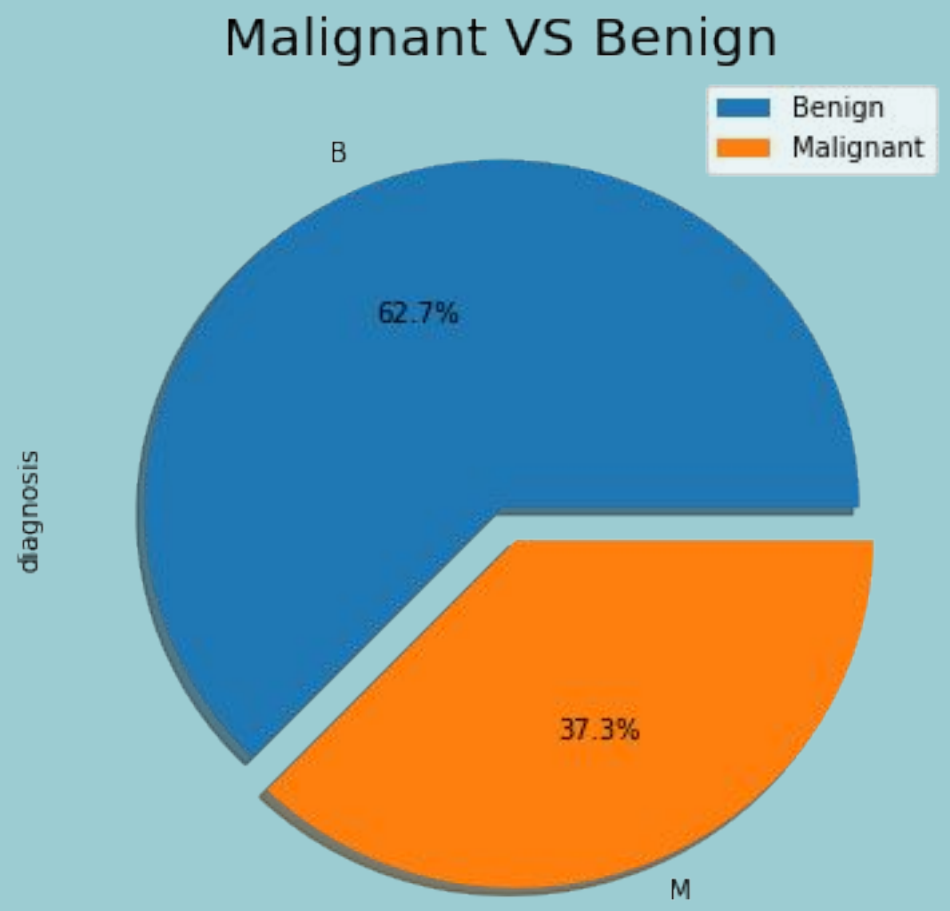
VARIABEL PENELITIAN

Target: Diagnosis (M untuk kanker ganas, B untuk kanker jinak)
Features: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, dan Fractal dimension.

Variabel	Keterangan
Radius	Jarak dari titik pusat ke titik perimeter
Texture	Variansi dari intensitas skala keabuan pada komponen-komponen piksel citra
Perimeter	Garis keliling inti sel
Area	Luas inti sel
Smoothness	Perbedaan antara panjang jari-jari dan rata-rata garis di sekitarnya
Compactness	Kepadatan inti sel
Concavity	Kecekungan kurva pada batas inti sel
Concave Points	Titik kecekungan kurva pada batas inti sel
Symmetry	Selisih garis tegak lurus sumbu utama menuju batas inti sel di kedua arah
Fractal Dimension	Ukuran numerik/dimensi fraktal dari sel
Diagnosis	Diagnosis kanker payudara jinak atau ganas

VISUALISASI DATA

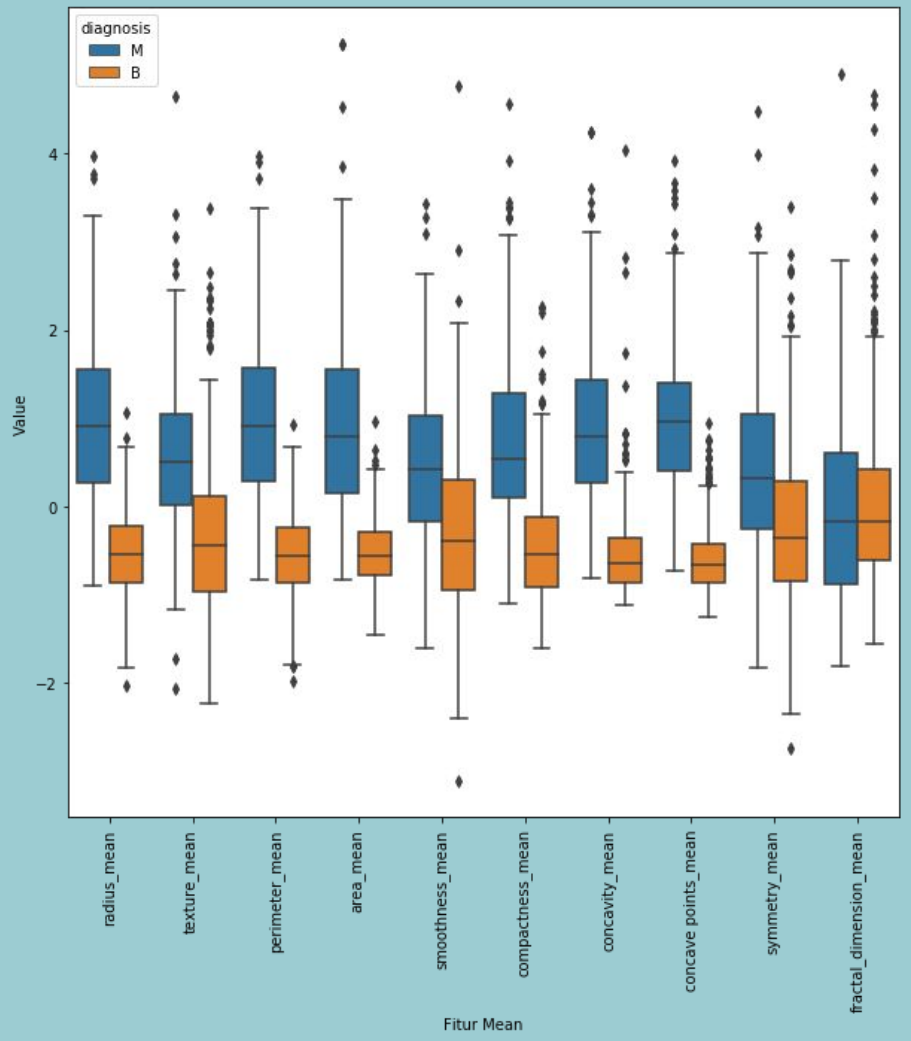
VARIABEL DIAGNOSIS



Diagnosis Kanker Payudara

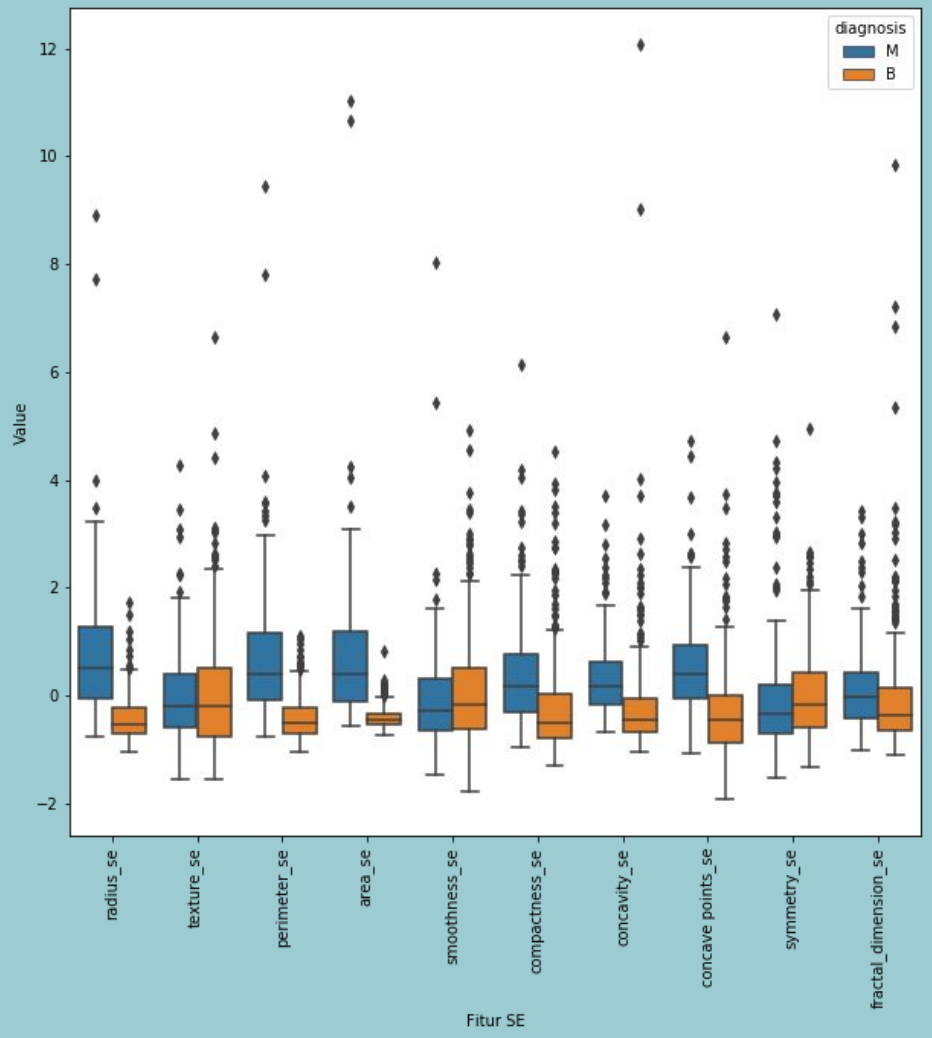
Benign (Jinak) : 357 pengamatan
Malignant (Ganas) : 212 pengamatan

VARIABEL FITUR MEAN



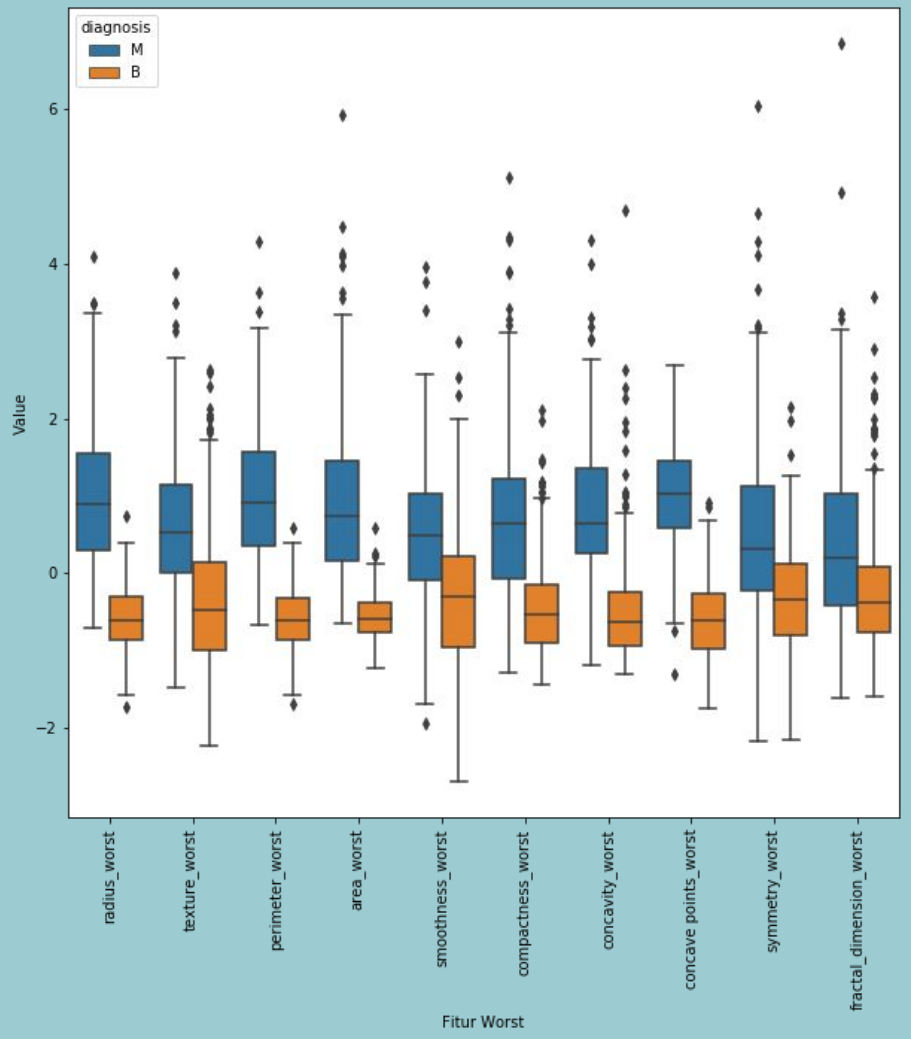
Grafik disamping menggambarkan sebaran atau distribusi dari 10 data pertama (%mean). Dari grafik tersebut, dapat diketahui bahwa fitur-fitur mean pada pasien yang terindikasi kanker ganas memiliki value yang lebih tinggi dibanding pasien yang menderita kanker jinak, misalnya pada ukuran rata-rata perimeter sel seorang pasien kanker ganas memiliki mean yang lebih besar

VARIABEL FITUR STANDAR ERROR



Grafik disamping menggambarkan sebaran atau distribusi dari 10 data kedua (%se). Berdasarkan gambar tersebut, dapat diketahui bahwa 7 fitur standar error pada pasien yang terindikasi kanker ganas memiliki value yang lebih tinggi dibanding pasien yang menderita kanker jinak, sedangkan 3 lainnya yaitu texture, smoothness, dan symmetri adalah sebaliknya

VARIABEL FITUR WORST



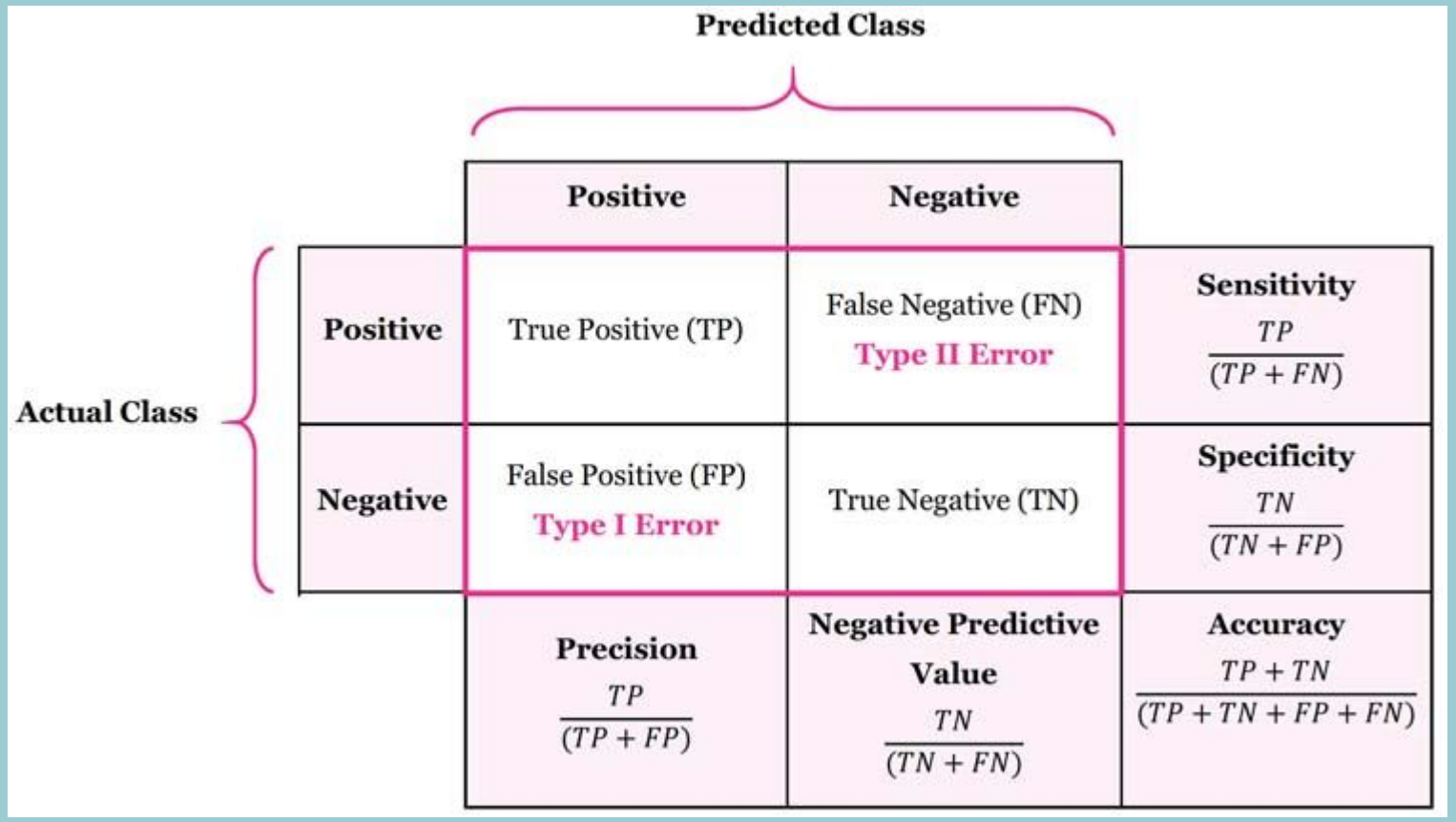
Grafik disamping menggambarkan sebaran atau distribusi dari 10 data ketiga (%worst). Berdasarkan gambar tersebut, dapat diketahui bahwa fitur-fitur worst pada pasien yang terindikasi kanker ganas memiliki value yang lebih tinggi dibanding pasien yang menderita kanker jinak

PERMODELAN 1

METODE KLASIFIKASI

- *Logistic Regression*
- *Support Vector Classifier*
- *Decision Tree Classifier*
- *Random Forest Classifier*

METODE EVALUASI



Actual Class \ Predicted Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Positive: Malignant
Negative: Benign

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

LOGISTIC REGRESSION

Predictor dengan nilai estimate **positif** akan **memperbesar** kemungkinan kanker ganas

	Estimate (Logit)	Odds Ratio
concavity_worst	1.13327	3.10581
symmetry_worst	0.64559	1.90712
compactness_worst	0.57778	1.78208
concave points_worst	0.44589	1.56189
texture_worst	0.38843	1.47466
concavity_mean	0.35144	1.42111
symmetry_mean	0.29657	1.34524
smoothness_worst	0.25622	1.29204
perimeter_worst	0.22815	1.25627
concave points_mean	0.20869	1.23207

Predictor dengan nilai estimate **negatif** akan **memperkecil** kemungkinan kanker ganas

area_worst	0.01025	1.01030
concavity_se	0.01024	1.01029
fractal_dimension_se	-0.01546	0.98466
area_mean	-0.02121	0.97901
compactness_se	-0.06503	0.93704
texture_mean	-0.16285	0.84972
radius_worst	-0.28573	0.75146
radius_mean	-0.65582	0.51902
texture_se	-0.99093	0.37123
intercept	-28.86383	0.00000

SUPPORT VECTOR

PARAMETER GRID SEARCH (12 KOMBINASI)

TIPE KERNEL	linear	poly	RBF	sigmoid
C (REGULARISASI)	0.1	1	10	

TOP 3 PARAMETER BERDASARKAN F1 SCORE

params	split0_test_score	split1_test_score	split2_test_score	mean_test_score
{'C': 10, 'kernel': 'linear'}	0.95327	0.90385	0.96154	0.93955
{'C': 1, 'kernel': 'linear'}	0.95238	0.92453	0.94118	0.93936
{'C': 0.1, 'kernel': 'linear'}	0.95238	0.92308	0.94118	0.93888

DECISION TREE

PARAMETER GRID SEARCH (60 KOMBINASI)

TIPE CRITERION	gini	entropy	
TIPE SPLITTER	best	random	
MAX FEATURES	2	sqrt	log2
MIN SAMPLES LEAF	bilangan ganjil antara 1 sampai 10		

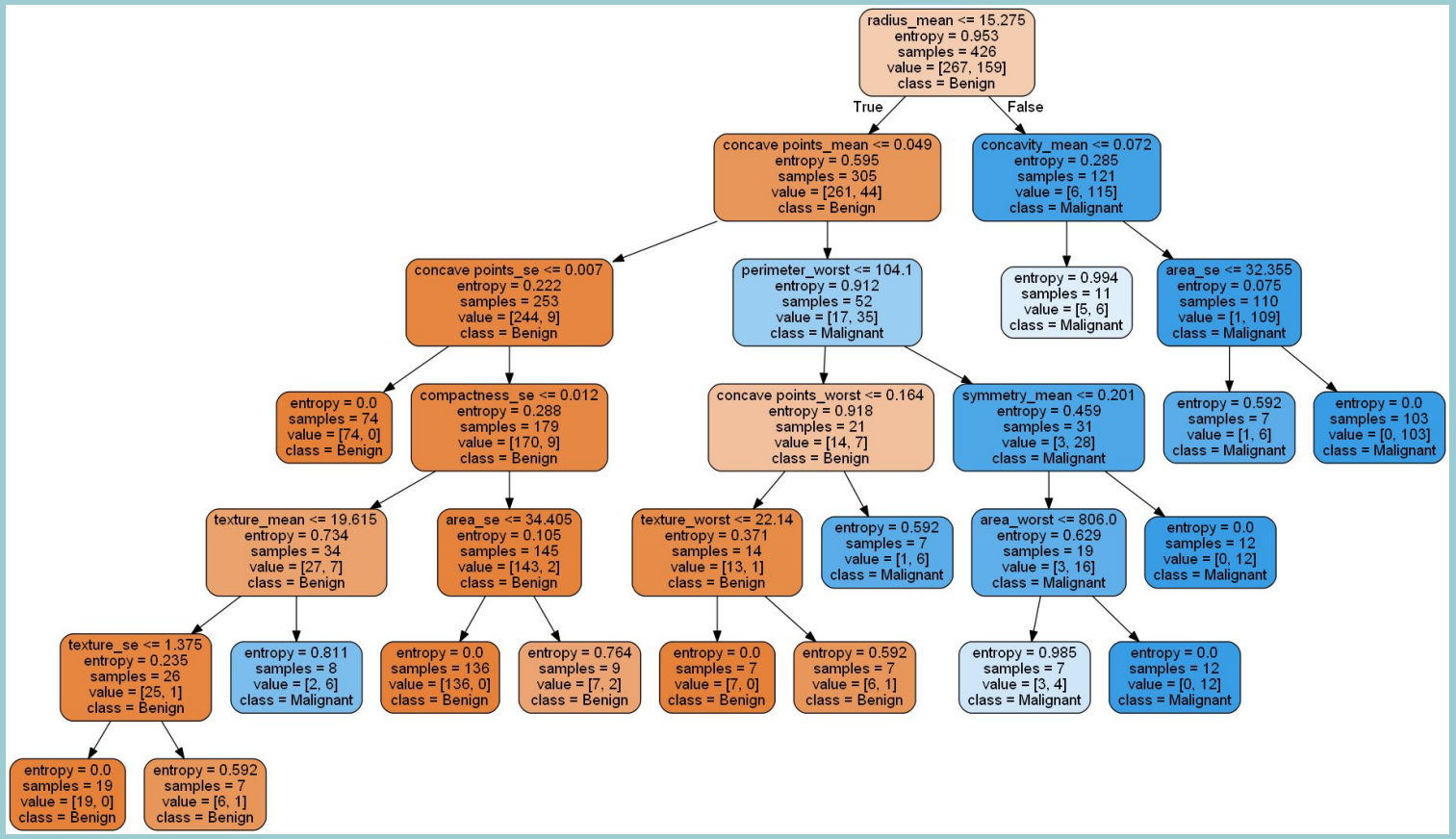
DECISION TREE

TOP 3 PARAMETER BERDASARKAN F1 SCORE

params	split0_test_score	split1_test_score	split2_test_score	mean_test_score
{'criterion': 'entropy', 'max_features': 'log2', 'min_samples_leaf': 7, 'splitter': 'best'}	0.89109	0.91589	0.93204	0.91301
{'criterion': 'entropy', 'max_features': 'log2', 'min_samples_leaf': 1, 'splitter': 'best'}	0.90909	0.91262	0.91071	0.91081
{'criterion': 'gini', 'max_features': 2, 'min_samples_leaf': 1, 'splitter': 'best'}	0.92727	0.93333	0.86726	0.90929

DECISION TREE

VISUALISASI POHON



RANDOM FOREST

PARAMETER GRID SEARCH (18 KOMBINASI)

N ESTIMATOR (TREE)	100	200	300
TIPE CRITERION	gini	entropy	
MAX FEATURES	2	sqrt	log2

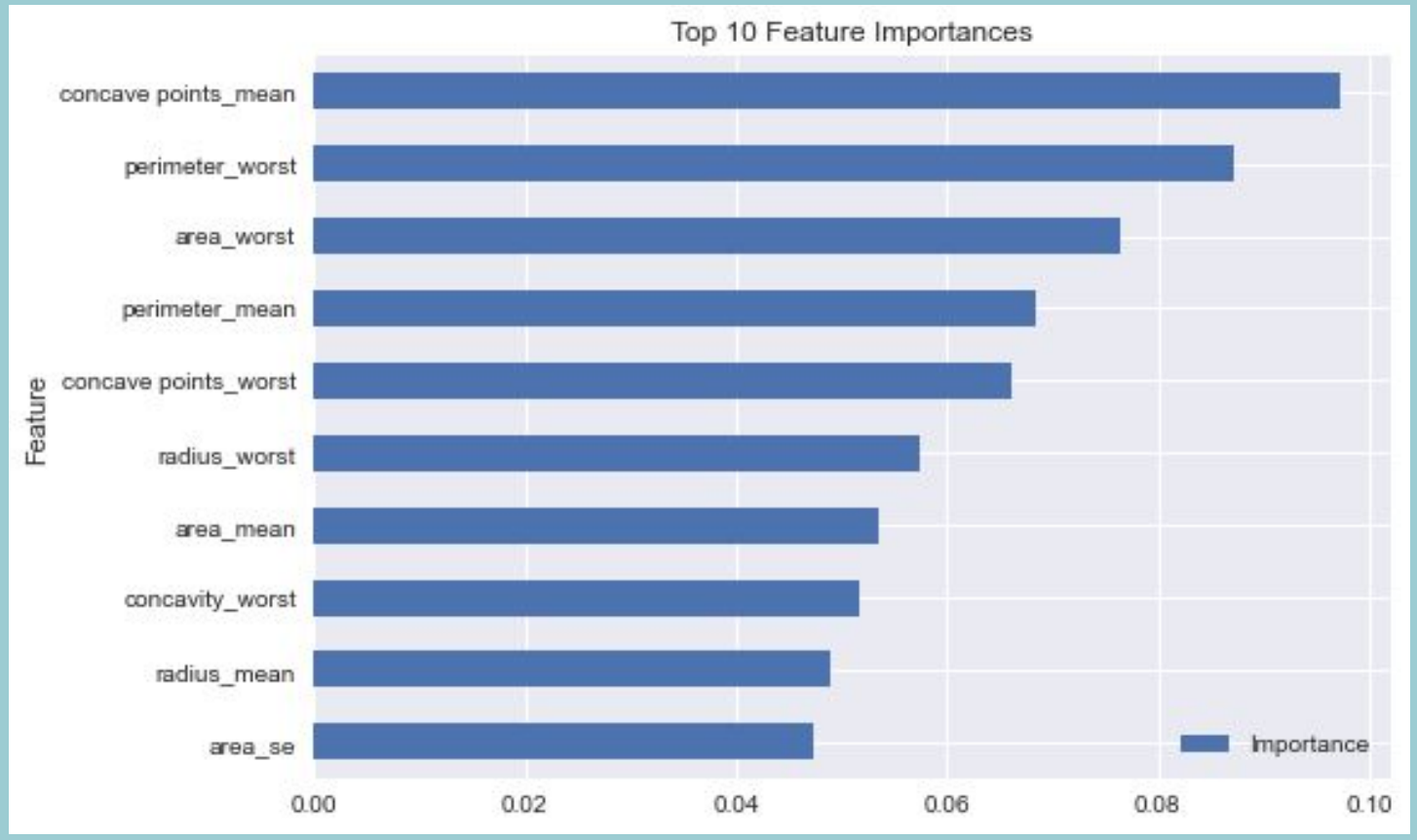
RANDOM FOREST

TOP 3 PARAMETER BERDASARKAN F1 SCORE

params	split0_test_score	split1_test_score	split2_test_score	mean_test_score
{'criterion': 'gini', 'max_features': 2, 'n_estimators': 200}	0.93204	0.93069	0.95327	0.93867
{'criterion': 'entropy', 'max_features': 2, 'n_estimators': 200}	0.93204	0.92157	0.96226	0.93862
{'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 200}	0.92308	0.94231	0.94340	0.93626

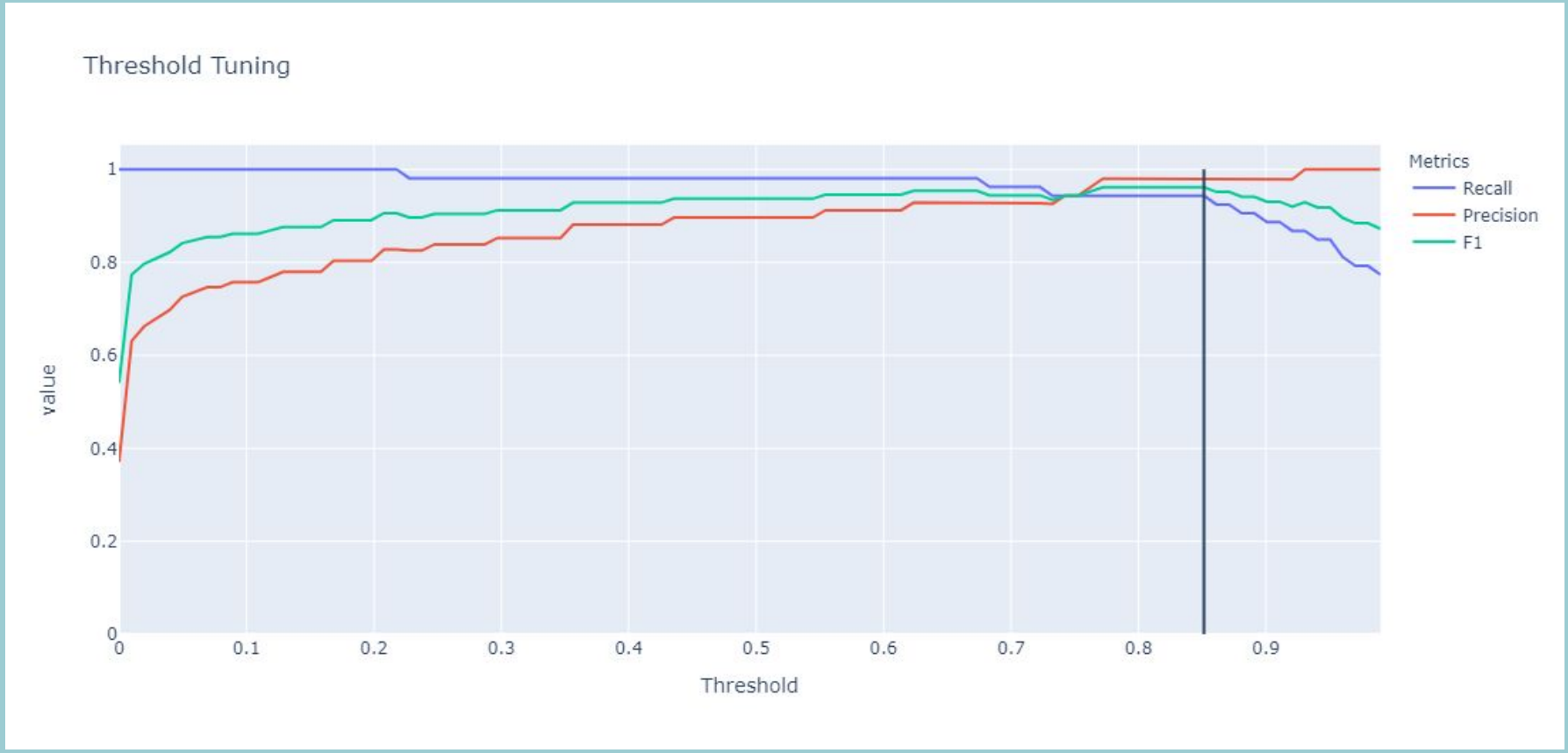
RANDOM FOREST

TOP 10 FEATURES BERDASARKAN IMPORTANCE



THRESHOLD TUNING

`.predict()` secara default menggunakan threshold probability 0.5 untuk mengklasifikasi.



PERMODELAN 2

FEATURE SELECTION

Fitur	F value
Concave Points Worst	964.38
Perimeter Worst	897.94
Radius Worst	860.78
Perimeter Mean	697.23
Area Worst	661.6
Radius Mean	646.98
Area Mean	573.06
Concavity Mean	533.79
Concavity Worst	436.69
Concave Points Mean	8.6176

Metode future selection : Select K Best

K yang digunakan : 10

Score_func : F_classif

MODEL EVALUATION

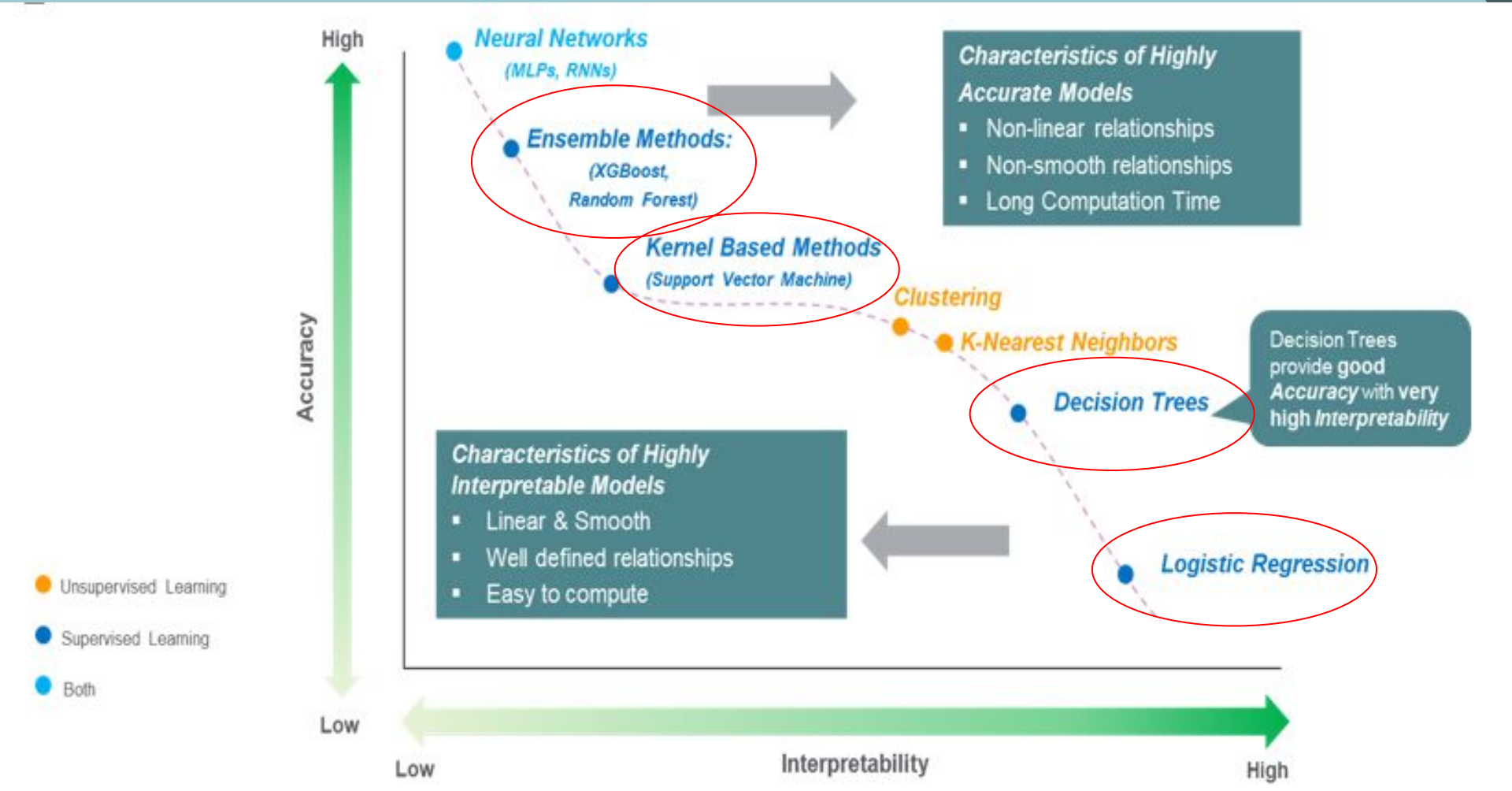
Algoritma Klasifikasi	f1 score
Logistic Regression	**0.98148**
Support Vector Machine	**0.98148**
Gaussian Naive Bayes	0.96226
Random Forest	0.95495
Ada Boost	0.94545
Gradient Boosting	0.94444
Decision Tree	0.9009
Neural Network	0.8965
Multinomial Naive Bayes	0.8888

Test Split : **25 %**

F1 Score Tertinggi : **Logistic Regression & SVM**

KESIMPULAN

MODEL TRADEOFF



MODEL PERFORMANCE 1

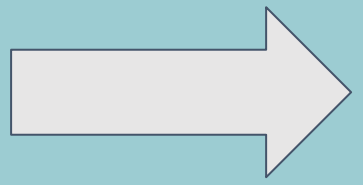
Metrics	Threshold	Recall	Precision	F1
Logistic Regression	0.85000	0.94340	0.98039	0.96154
Support Vector Classifier	0.42000	0.98113	0.94545	0.96296
Decision Tree Classifier	0.50000	0.92453	0.89091	0.90741
Random Forest Classifier	0.46000	0.98113	0.94545	0.96296

MODEL PERFORMANCE 2

Logistic Regression

Penalty : l2

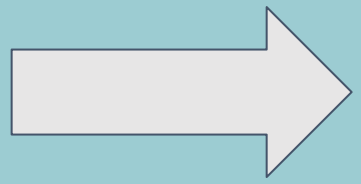
Solver : liblinear



```
Recall Testing : 0.981481481481
Recall Score Training : 0.946666666667
F-1 Score Testing : 0.981481481481
F-1 Score Training : 0.922077922078
```

Support Vector Machine

Kernel : linear



```
Recall Testing : 0.981481481481
Recall Score Training : 0.953020134228
F-1 Score Testing : 0.981481481481
F-1 Score Training : 0.925081433225
```