EVALUASI CLASSIFIER KNN, SVM, DAN DECISION TREE PADA KLASIFIKASI DATASET CARDIOTOCORAPHY

Proyek UTS MK Kecerdasan Buatan 2021

Kelompok 5

Rubyna Hamaswari 1906357925
M. Aidan Daffa 1906300800

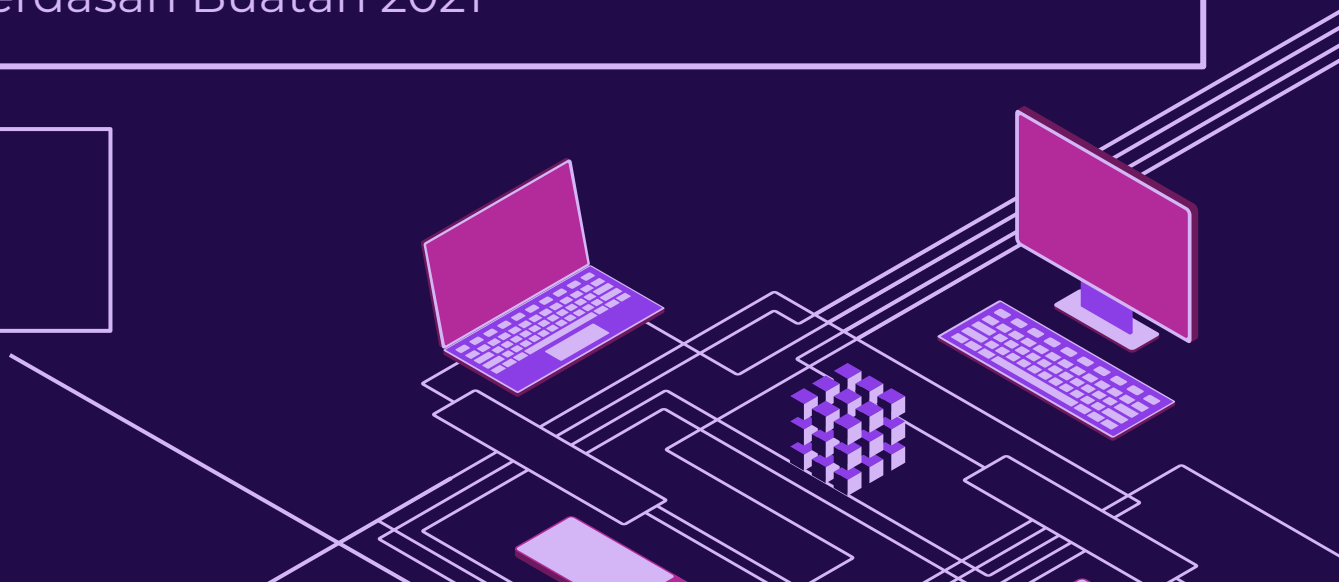


TABLE OF CONTENTS

01

Pembagian Tugas

Muh. Aidan

- Import Library
- Mencari topik & dataset
- Data Preprocessing
- KNN Classification
- Classification Report Function
- Documentation
- Membuat Slide

R. Hamaswari

- Import Library
- Mencari topik & dataset
- Import Dataset
- SVM Classification
- DT Classification
- Documentation
- Membuat Slide

02

Dataset

Penjelasan Dataset

Link dataset

Attribute Characteristics	Real Number
Number of Instances	2126
Number of Attribute	23

Dataset ini berisi 2126 data cardiotocography (CTGs). CTG ini diklasifikasikan oleh tiga ahli kandungan dan label klasifikasi konsensus yang ditetapkan yaitu keadaan janin (N, S, P). N adalah normal, S = suspect, P = pathologic.

Penjelasan Atribut

- LB - FHR baseline (beats per minute)
- AC - # of accelerations per second
- FM - # of fetal movements per second
- UC - # of uterine contractions per second
- DL - # of light decelerations per second
- DS - # of severe decelerations per second
- DP - # of prolongued decelerations per second
- ASTV - percentage of time with abnormal short term variability
- MSTV - mean value of short term variability
- ALTV - percentage of time with abnormal long term variability
- MLTV - mean value of long term variability
- Width - width of FHR histogram
- Min - minimum of FHR histogram
- Max - Maximum of FHR histogram
- Nmax - # of histogram peaks
- Nzeros - # of histogram zeros
- Mode - histogram mode
- Mean - histogram mean
- Median - histogram median
- Variance - histogram variance
- Tendency - histogram tendency
- CLASS - FHR pattern class code (1 to 10)
- NSP - fetal state class code (N=normal; S=suspect; P=pathologic)

03

Classifier yang Digunakan

Classifier

KNN

KNN (K-Nearest Neighbor) adalah salah satu classifier yang mengklasifikasi data berdasarkan jumlah k-data terdekat pada sebuah titik.

SVM

Support Vector Machine merupakan classifier yang mengklasifikasikan data berdasarkan sebuah kernel/hyperplane yang membagi data kedalam beberapa bagian.

Decision Tree

Decision tree membagi variabel menjadi subset berdasarkan tingkatan tertentu secara berkelanjutan hingga seluruh variabel digunakan.

Variasi KNN

1	2	3
N_neighbors = 3	N_neighbors = 5	N_neighbors = 7
Weights = distance	Weights = uniform	Weights = distance
Algorithm = ball_tree	Algorithm = auto	Algorithm = brute

Variasi SVM

1	2	3
kernel = ‘linear’	kernel = ‘rbf’	kernel = ‘poly’

Variasi DT

1	2	3
criterion = 'entropy', max_features = 'log2'	criterion = 'gini'	criterion = 'entropy', max_features = 'log2', max_depth = 20

04

Hasil Pengujian

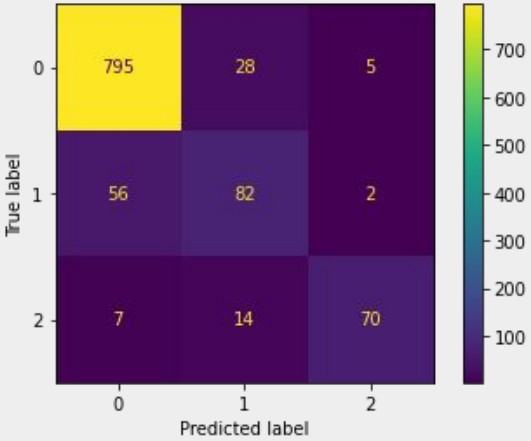
KNN Classifier

		Precision	Recall	F1-Score	Accuracy
Best Parameter	N	0.93	0.96	0.94	0.89
	S	0.66	0.59	0.62	
	P	0.91	0.77	0.83	
Default	N	0.92	0.96	0.94	0.89
	S	0.66	0.55	0.60	
	P	0.89	0.75	0.81	
Random Variation	N	0.92	0.97	0.94	0.89
	S	0.68	0.55	0.61	
	P	0.93	0.77	0.84	

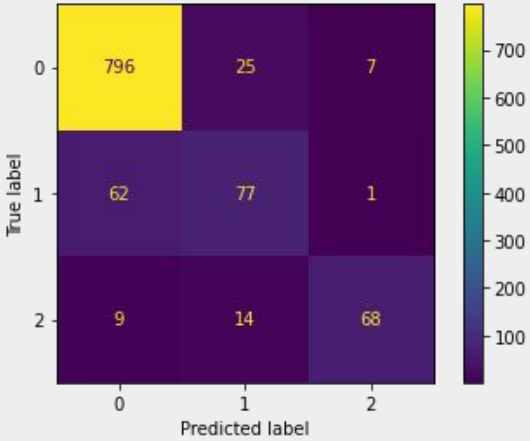
KNN Classifier

Confusion Matrix

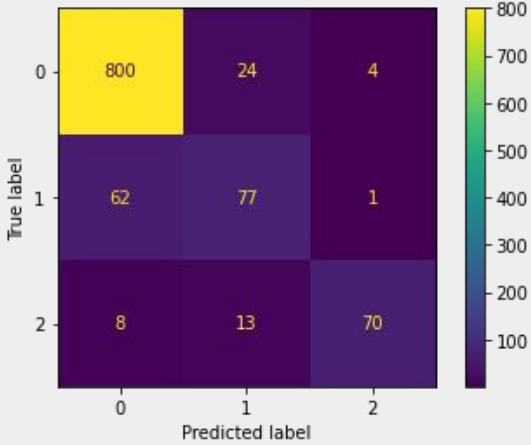
Best Parameter



Default



Random Variaton



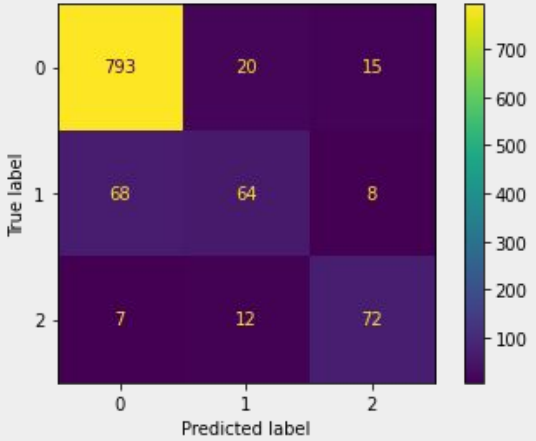
SVM Classifier

		Precision	Recall	F1-Score	Accuracy
Linear Kernel	N	0.91	0.96	0.94	0.88
	S	0.67	0.46	0.54	
	P	0.76	0.79	0.77	
RBF Kernel	N	0.87	0.97	0.92	0.84
	S	0.51	0.33	0.40	
	P	0.97	0.42	0.58	
Poly Kernel	N	0.89	0.97	0.93	0.87
	S	0.65	0.43	0.52	
	P	0.81	0.62	0.70	

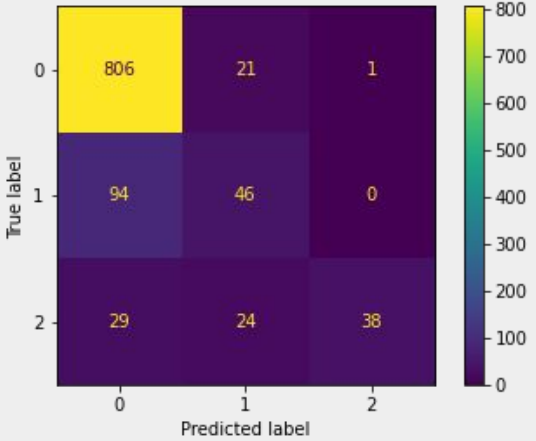
SVM Classifier

Confusion Matrix

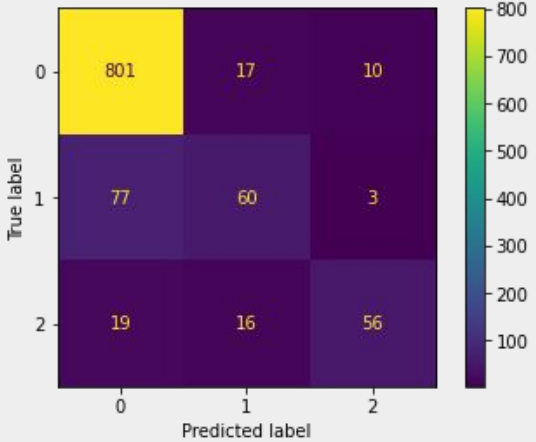
Linear Kernel



RBF Kernel



Poly Kernel



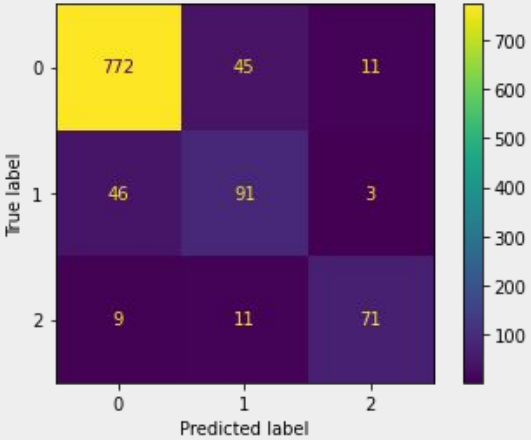
Decision Tree Classifier

		Precision	Recall	F1-Score	Accuracy
Best Parameter	N	0.94	0.97	0.96	0.91
	S	0.75	0.68	0.71	
	P	0.77	0.71	0.74	
Default	N	0.94	0.95	0.95	0.91
	S	0.71	0.71	0.71	
	P	0.94	0.85	0.89	
Random Variation	N	0.92	0.93	0.93	0.88
	S	0.61	0.59	0.60	
	P	0.84	0.81	0.83	

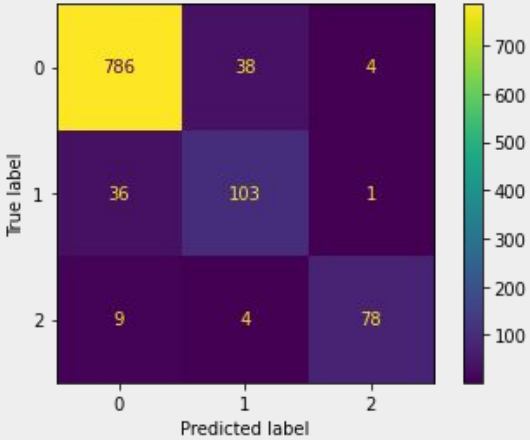
Decision Tree Classifier

Confusion Matrix

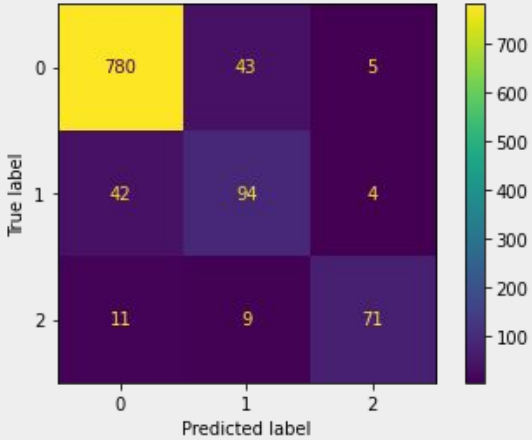
Best Parameter



Default



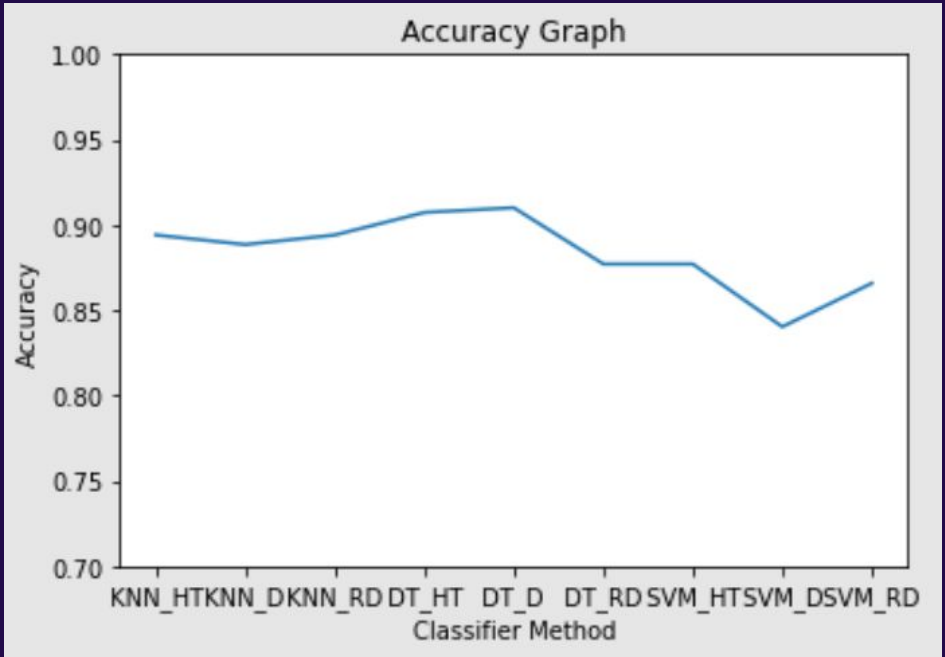
Random Variation



Accuracy Table

No	Variasi	Accuracy
1	KNN Hypertunning	0.8942398489140698
2	KNN Default	0.8885741265344664
3	KNN Variasi lain	0.8942398489140698
4	DT Hypertunning	0.9074598677998111
5	DT Default	0.9102927289896129
6	DT Variasi lain	0.8772426817752597
7	SVM Hypertunning	0.8772426817752597
8	SVM Default	0.8404154863078376
9	SVM Variasi lain	0.8659112370160529

Accuracy Graph



05

Analisis

KNN Classifier

Dari hasil yang didapat pada tabel KNN Classifier maka didapat bahwa hasil terbaik dimiliki oleh percobaan dengan hypertunning grid search. Walaupun perbedaan antara percobaan dengan hypertunning dan percobaan dengan variasi random tidak berjauhan, bahkan accuracy nya sama sampai 10 angka dibelakang koma. Alasannya karena dari 2 percobaan tersebut yang berbeda hanya 2 parameter, yaitu n_neighbor, dan algorithm. Kedua parameter tersebut kurang signifikan di dalam percobaan ini. Hal itu dibuktikan dari percobaan yang telah dilakukan.

DT Classifier

- Pada pengujian menggunakan decision tree classifier, kombinasi antara penggunaan hyperparameter criterion ‘entropy’ dan max_features ‘log2’ memberikan overall accuracy yang paling baik. Hal ini dikarenakan telah dilakukannya hyperparameter tuning menggunakan GridSearch untuk menemukan kombinasi hyperparameter terbaik yang bisa digunakan.
- Penggunaan criterion ‘entropy’ dapat memberikan information gain pada data sehingga model dapat lebih mudah mengidentifikasi feature importance yang terdapat dalam data.
- Penggunaan max_features ‘log2’ berarti pemilihan feature dalam training hanya akan berjumlah log2(23) yang akan mempercepat proses training.
- Penambahan max_depth menyebabkan akurasi berkurang karena penambahan nilai max_depth yang berlebih dapat menyebabkan overfitting

SVM Classifier

- Dari seluruh kernel yang digunakan pada klasifikasi menggunakan SVC, hasil penggunaan kernel linear memiliki overall accuracy yang paling baik. Hal ini menunjukkan bahwa dataset yang digunakan memiliki kecenderungan persebaran yang linear (*linearly separable*).
- Penggunaan kernel radial basis function (RBF) menghasilkan hasil yang kurang baik karena algoritmanya yang mengasumsikan data terdistribusi secara normal sedangkan data yang digunakan tidak terdistribusi secara normal.
- Penggunaan kernel polynomial pada umumnya digunakan untuk pemrosesan gambar karena dapat digunakan pada data dengan derajat yang lebih besar daripada 1. Akan tetapi, penggunaan kernel polynomial kurang cocok dan memakan resource yang lebih banyak jika digunakan pada dataset yang bersangkutan.

Conclusion

Dari percobaan diatas didapat bahwa hasil terbaik didapat dari classifier decision tree. Alasannya karena dataset yang digunakan cocok digunakan untuk classifier DT tersebut. DT adalah classifier yang cocok digunakan pada data yang non-linear, seperti dataset yang digunakan pada percobaan ini. Data yang digunakan juga bukan merupakan data kontinu, data kontinu akan menjadi drawback untuk decision tree classifier.

06

Referensi

Referensi

- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- http://scikit-learn.org/stable/modules/grid_search.html
- https://scikit-learn.org/stable/modules/cross_validation.html
- https://archive.ics.uci.edu/ml/datasets/Cardiotocography