

# **PENERAPAN MODEL TWO CLASS DECISION FOREST DALAM PREDIKSI PENYAKIT PARKINSONS**

Diajukan Untuk Memenuhi Tugas Mata Kuliah Pemodelan dan Simulasi

**Rahmadi Ridwan**

**10116418**



**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNIK DAN ILMU KOMPUTER  
UNIVERSITAS KOMPUTER INDONESIA**

**2019**

# **Model Two-Class Decision Forest dan Penerapan Modulnya pada Microsoft Azure Machine Learning Studio**

## **I. PENGANTAR**

Data mining adalah suatu proses yang digunakan untuk mencari informasi dan knowledge yang berguna, dimana diperoleh dari data-data yang dimiliki. Dari buku Data Mining Technique yang dikarang oleh Berry and Linoff, proses terjadinya data mining dapat dideskripsikan sebagai virtuous cycle. Didasari oleh pengembangan berkelanjutan dari proses bisnis serta didorong oleh penemuan knowledge ditindaklanjuti dengan pengambilan tindakan dari penemuan tersebut.

Salah satu sarana pengimplementasian dalam data mining yang sederhana dan cukup lazim adalah implementasi berbasis model 'Two-Class'. Penerapan model 'Two-Class' adalah model prediktif yang menentukan 'arah' dari sekumpulan data yang bersifat biner. Penerapan model ini dikatakan cukup lazim dan sederhana karena contoh kasus penerapannya seperti untuk menentukan berdasarkan beberapa parameter yang merupakan data pendidikan sekumpulan individu, apakah individu-individu tersebut akan atau sekarang memiliki pendapatan yang tinggi atau khususnya dalam penerapannya dalam bidang bioinformatika, apakah berdasarkan parameter-parameter berisi informasi kondisi kesehatan individu, suatu simulasi machine learning yang menerapkan model two-class ini dapat memprediksikan apakah masing-masing dari sekumpulan individu tersebut menderita penyakit tertentu atau mungkin berisiko menderitanya di masa depan.

Ada banyak penerapan dari model 'Two-Class' (secara harfiah berarti dua kelas) ini, seperti Bayes Point Machine, Averaged Perception dan lain sebagainya. Namun yang dibahas secara khusus pada literatur ini adalah model Two-Class Decision Forest yang diterapkan pada dataset berisi parameter informasi kesehatan individu-individu dan relasinya terhadap penyakit system saraf Parkinson. Two-Class Decision Forest sendiri merupakan sejenis ensemble model yang bersifat cepat dan tersupervisi. Karena khususnya pada kasus ini, target prediksi hanya melibatkan dua hasil akhir (yaitu apakah individu menderita atau tidak penyakit Parkinson), model Decision Forest sangat cocok untuk digunakan.

## **II. SARANA DAN METODE PENERAPAN**

Tools yang digunakan adalah Microsoft Azure Machine Learning Studio dan dataset mengenai penyakit Parkinson dengan data kualitatif berupa data bertipe string yaitu atribut nama dan data bertipe Boolean yaitu atribut status serta atribut-atribut lainnya berisi data kuantitatif dengan rincian sebagai berikut:

name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE,D2 - Two nonlinear dynamical complexity measures






DFA - Signal fractal scaling exponent

spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

Parkinson Dataset 2-Class (Decision Forest) > DP-dataset.csv > dataset

rows  
195








columns  
24

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)
view as					
	phon_R01_S01_1	119.992	157.302	74.997	0.00784
	phon_R01_S01_2	122.4	148.65	113.819	0.00968
	phon_R01_S01_3	116.682	131.111	111.555	0.0105
	phon_R01_S01_4	116.676	137.871	111.366	0.00997
	phon_R01_S01_5	116.014	141.781	110.655	0.01284
	phon_R01_S01_6	120.552	131.162	113.787	0.00968
	phon_R01_S02_1	120.267	137.244	114.82	0.00333
	phon_R01_S02_2	107.332	113.84	104.315	0.0029
	phon_R01_S02_3	95.73	132.068	91.754	0.00551

Parkinson Dataset 2-Class (Decision Forest) > DP-dataset.csv > dataset

rows  
195







columns  
24

MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:F0
						
0.00784	0.00007	0.0037	0.00554	0.01109	0.04374	0.426
0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626
0.0105	0.00009	0.00544	0.00781	0.01633	0.05233	0.482
0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517
0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584
0.00968	0.00008	0.00463	0.0075	0.01388	0.04701	0.456
0.00333	0.00003	0.00155	0.00202	0.00466	0.01608	0.14
0.0029	0.00003	0.00144	0.00182	0.00431	0.01567	0.134
0.00551	0.00006	0.00293	0.00332	0.0088	0.02093	0.191









Parkinson Dataset 2-Class (Decision Forest) > DP-dataset.csv > dataset

rows  
195

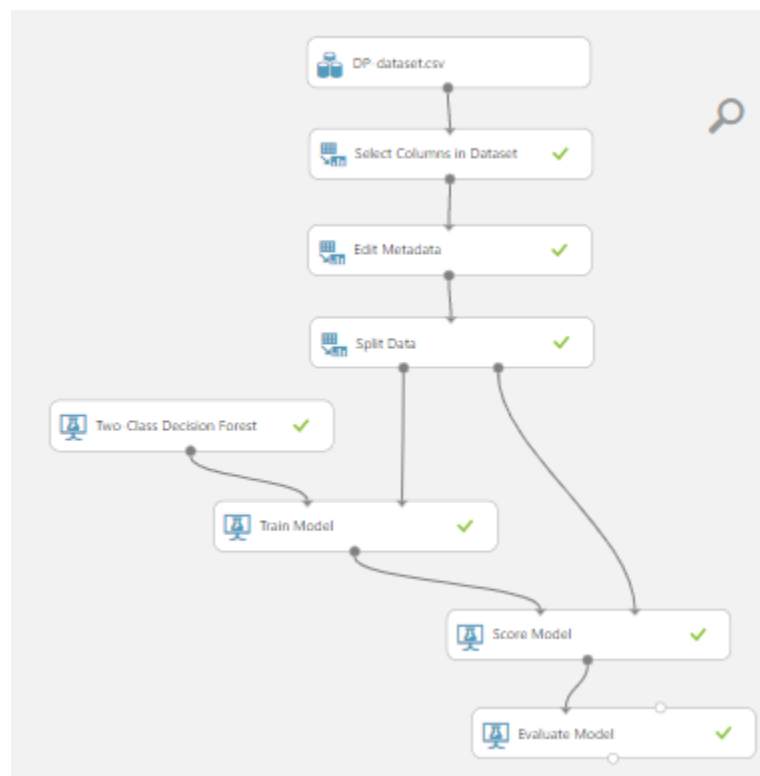
columns  
24

MDVP:Shimmer(dB)	Shimmer:APQ3	Shimmer:APQ5	MDVP:APQ	Shimmer:DDA	NHR
					
0.426	0.02182	0.0313	0.02971	0.06545	0.02211
0.626	0.03134	0.04518	0.04368	0.09403	0.01929
0.482	0.02757	0.03858	0.0359	0.0827	0.01309
0.517	0.02924	0.04005	0.03772	0.08771	0.01353
0.584	0.0349	0.04825	0.04465	0.1047	0.01767
0.456	0.02328	0.03526	0.03243	0.06985	0.01222
0.14	0.00779	0.00937	0.01351	0.02337	0.00607
0.134	0.00829	0.00946	0.01256	0.02487	0.00344
0.191	0.01073	0.01277	0.01717	0.03218	0.0107

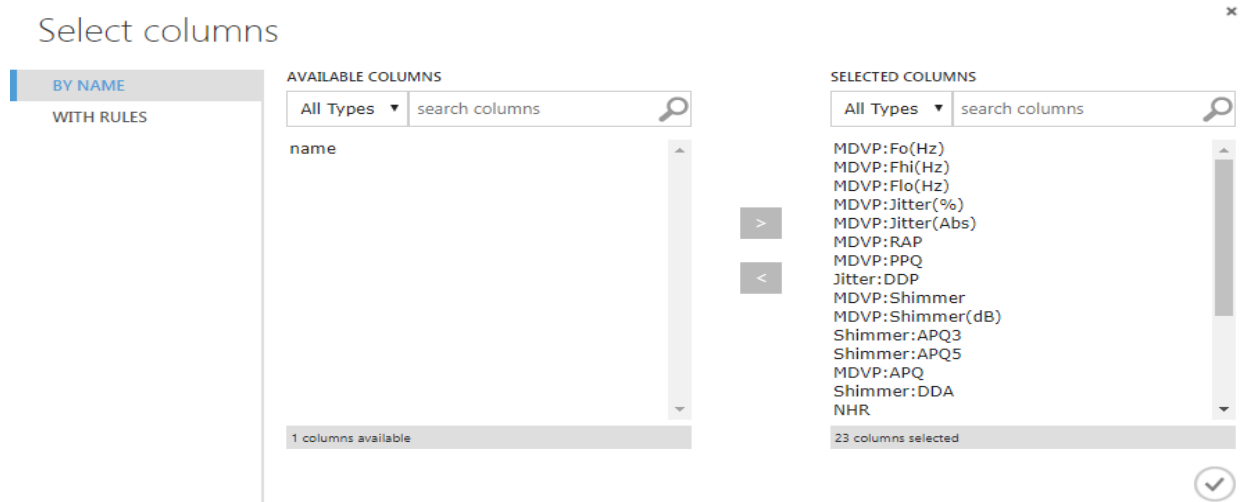
Parkinson Dataset 2-Class (Decision Forest) > DP-dataset.csv > dataset

rows	columns							
195	24							
HNR	status	RPDE	DFA	spread1	spread2	D2	PPE	
								
21.033	1	0.414783	0.815285	-4.813031	0.266482	2.301442	0.284654	
19.085	1	0.458359	0.819521	-4.075192	0.33559	2.486855	0.368674	
20.651	1	0.429895	0.825288	-4.443179	0.311173	2.342259	0.332634	
20.644	1	0.434969	0.819235	-4.117501	0.334147	2.405554	0.368975	
19.649	1	0.417356	0.823484	-3.747787	0.234513	2.33218	0.410335	
21.378	1	0.415564	0.825069	-4.242867	0.299111	2.18756	0.357775	
24.886	1	0.59604	0.764112	-5.634322	0.257682	1.854785	0.211756	
26.892	1	0.63742	0.763262	-6.167603	0.183721	2.064693	0.163755	
21.812	1	0.615551	0.773587	-5.498678	0.327769	2.322511	0.231571	

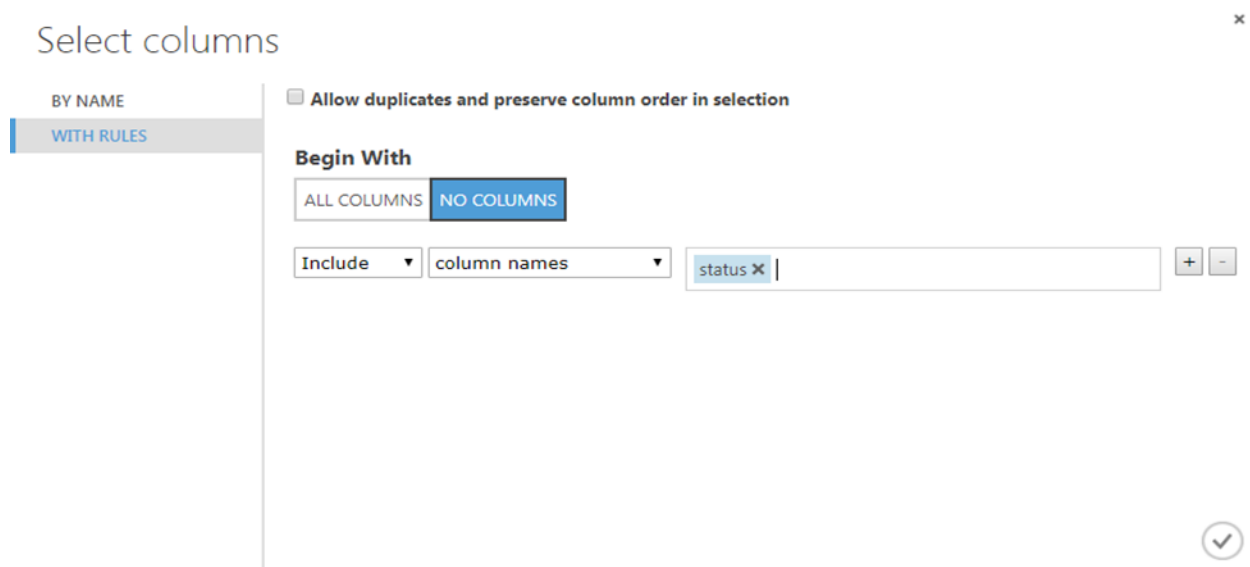
Berikut adalah simulasi yang diterapkan:



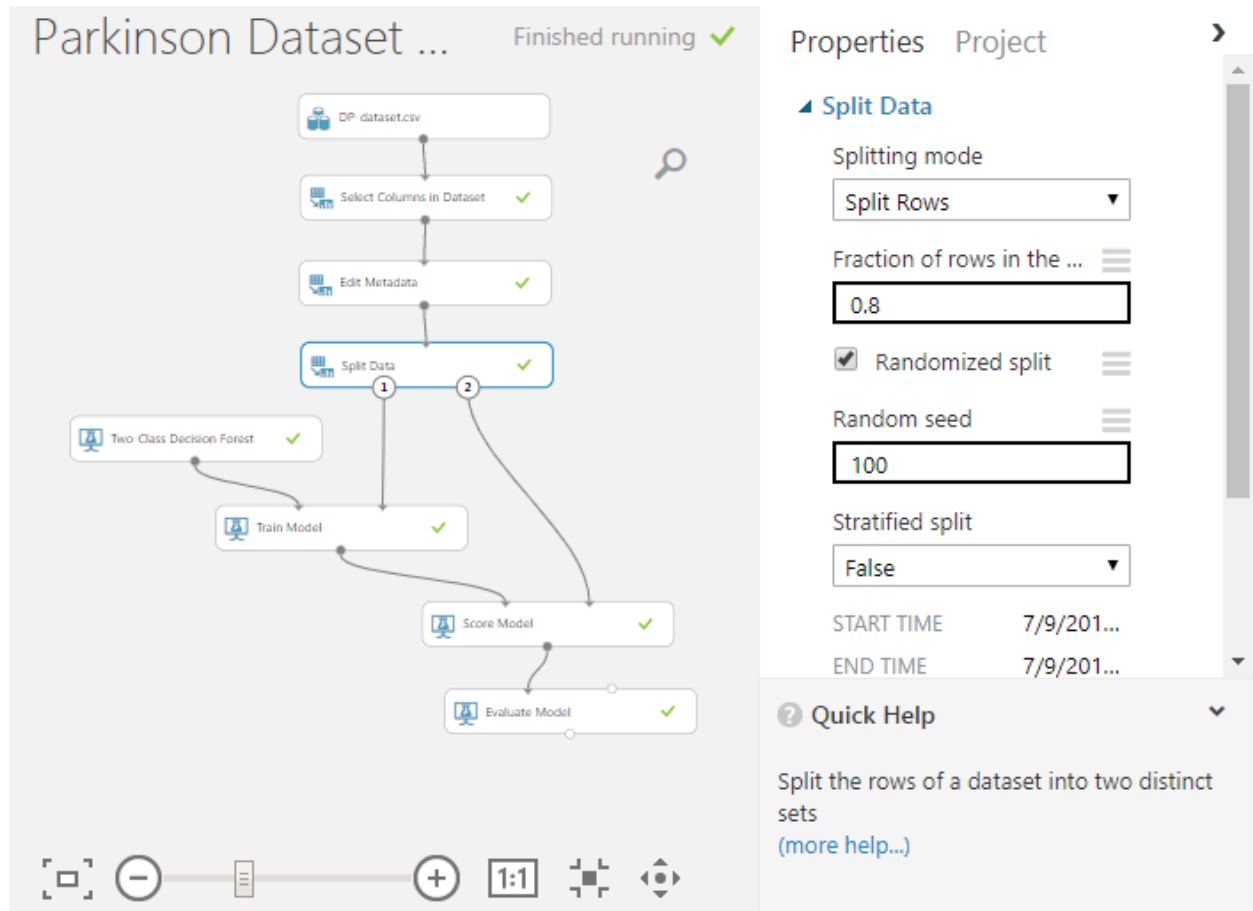
Dataset Parkinson disimpan di dalam dataset DP-dataset.csv yang kemudian akan di filter dengan menyeleksi kolom-kolom yang akan digunakan dalam simulasi. Dalam kasus ini kolom/atribut pertama tidak digunakan karena tidak relevan apabila dimuatkan sebagai parameter (karena nama dan ID individu dalam dunia nyata bukan parameter yang relevan dalam menentukan penyakit yang diderita suatu individu. Semua ini diperoleh menggunakan penerapan modul Select Columns in Dataset → Launch Column Sector → {Pindahkan semua nama kolom-kolom ke bagian Selected Cols kecuali 'name', yang tetap berada di bagian Available Columns si ruas kiri}



Langkah berikutnya dalam persiapan data sebelum diterapkannya model ensemble adalah dengan memprosesnya melalui modul Edit Metadata. Edit Metadata ini digunakan untuk menandai atribut-atribut non kuantitatif yang telah melalui filter tahap sebelumnya yaitu Select Columns in Dataset. Kolom-kolom yang diseleksi untuk ditandai adalah kolom 'status' yang kemudian ditandai dalam bagian Categorical sebagai Make Categorical

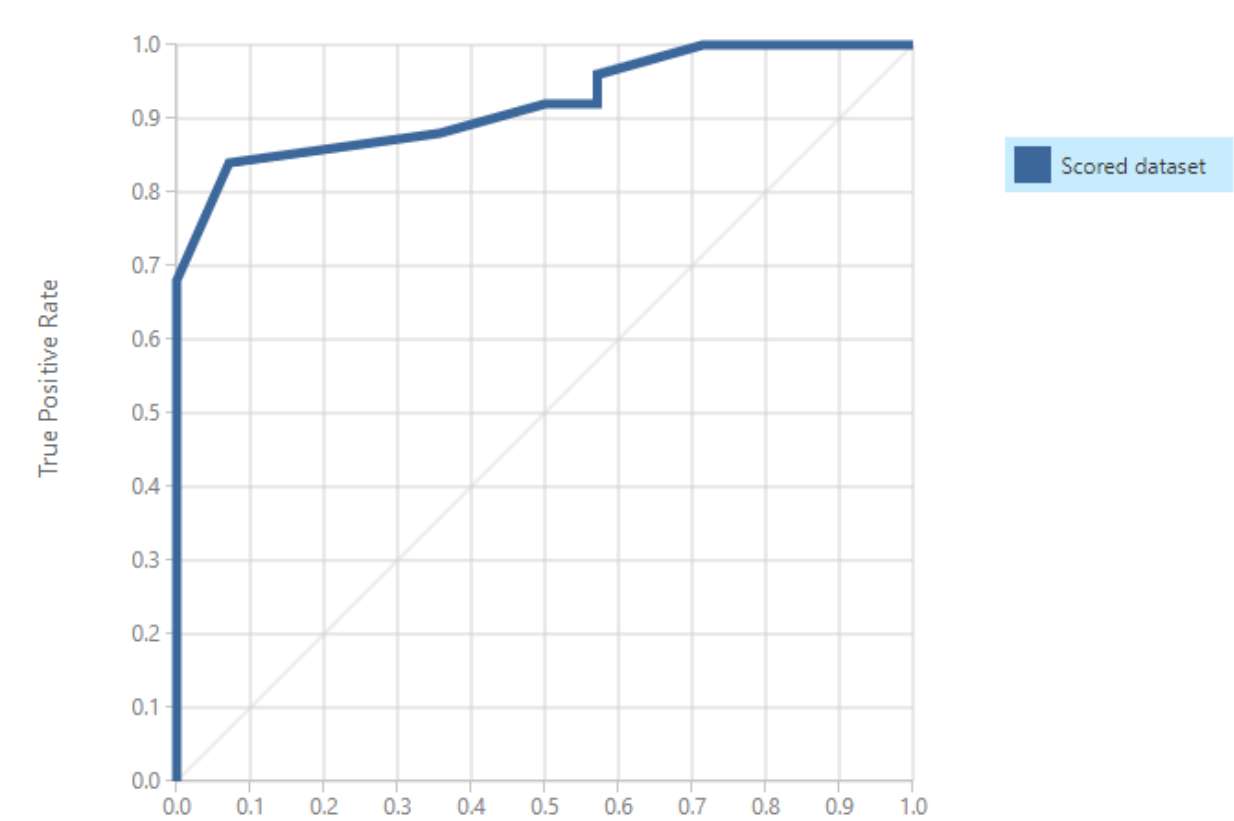


Setelah tahap tersebut telah dilaksanakan, dataset kemudian akan dibelah menggunakan modul Split Data. Pada tahap ini fraksi baris-baris keluaran untuk output pertama ditentukan sebagai 0.8 (maka dari itu pula fraksi baris-baris keluaran untuk output kedua adalah 0.2). Random Seed = 100.



Setelah melalui ketiga tahap persiapan data awal tersebut data kemudian dapat diproses dan disimulasikan dengan konfigurasi yang telah dipaparkan sebelumnya. Keluaran pertama dari modul Split Data dihubungkan dengan masukkan dataset modul Train Model dan modul Two-Class Decision Forest dihubungkan dengan masukkan untrained module dari modul Train Model. Modul Train Model juga dikonfigurasi sedemikian rupa sehingga kolom yang terseleksi pada tahap simulasi modul ini adalah status, hal ini dikarenakan model prediksi yang ingin disimulasikan menggunakan atribut ini sebagai parameter tolak ukur kesuksesan. Setelah itu keluaran dari model Train Model dihubungkan dengan masukkan trained model modul Score Model dan keluaran kedua dari modul Split Data dihubungkan dengan modul Split Data untuk memfasilitasi umpan balik pada model training Decision Forest yang diterapkan. Keluaran dari modul Score Model kemudian dihubungkan dengan modul Evaluate Model dan hasil visualisasi dari keluaran modul inilah yang menjadi hasil olahan dari penerapan simulasi ini yang dapat kemudian dianalisa.

Parkinson Dataset 2-Class (Decision Forest) > Evaluate Model > Evaluation results



Parkinson Dataset 2-Class (Decision Forest) > Evaluate Model > Evaluation results

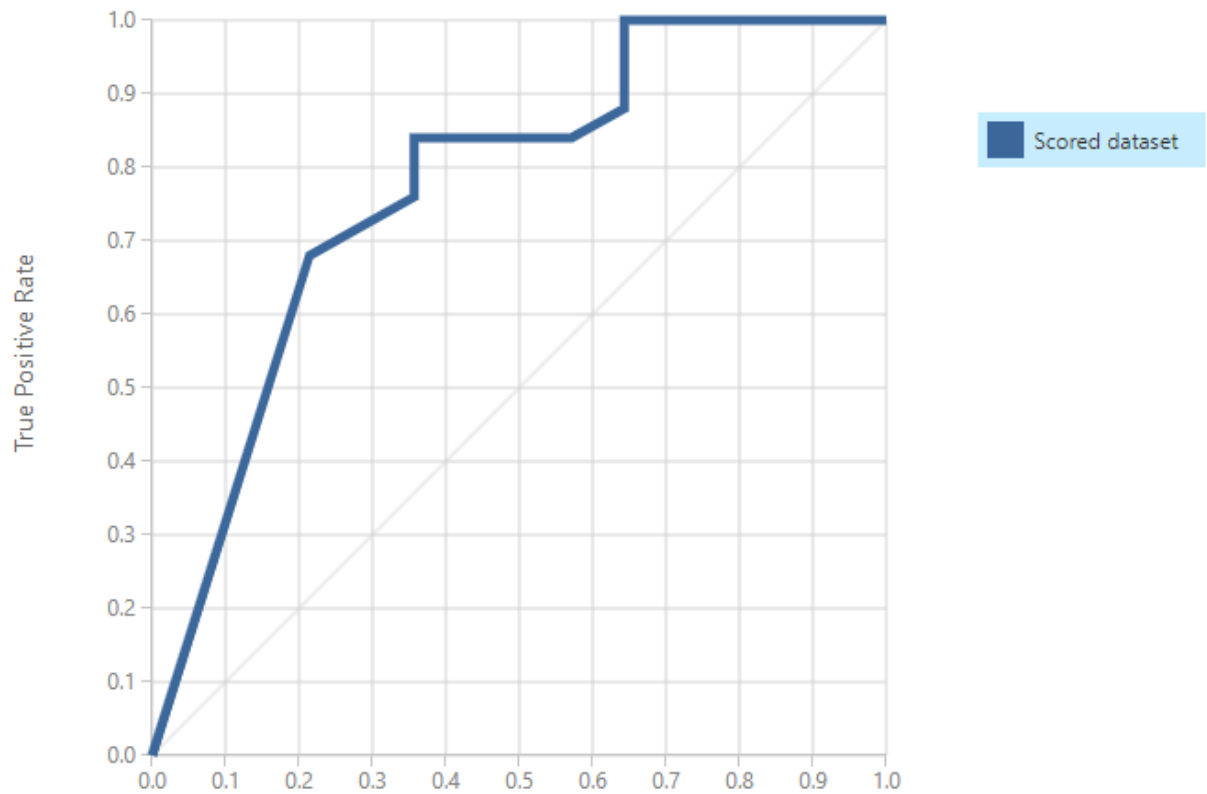
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
23	2	0.769	0.767	0.5	0.920
False Positive	True Negative	Recall	F1 Score		
7	7	0.920	0.836		
Positive Label	Negative Label				
1	0				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	17	0	0.436	0.795	0.810	1.000	0.680	0.636	1.000	0.000
(0.800,0.900]	4	1	0.564	0.872	0.894	0.955	0.840	0.765	0.929	0.054
(0.700,0.800]	1	4	0.692	0.795	0.846	0.815	0.880	0.750	0.643	0.300
(0.600,0.700]	1	2	0.769	0.769	0.836	0.767	0.920	0.778	0.500	0.429
(0.500,0.600]	0	1	0.795	0.744	0.821	0.742	0.920	0.750	0.429	0.494
(0.400,0.500]	0	0	0.795	0.744	0.821	0.742	0.920	0.750	0.429	0.494
(0.300,0.400]	1	0	0.821	0.769	0.842	0.750	0.960	0.857	0.429	0.494
(0.200,0.300]	1	2	0.897	0.744	0.833	0.714	1.000	1.000	0.286	0.634



Berikut adalah hasil simulasi lain menggunakan Decision Jungle (dengan poin-poin yang lain tetap sama seperti yang diterapkan pada simulasi Decision Forest sebelumnya) sebagai pembandingan:

### Parkinson Dataset 2-Class (Decision Forest) > Evaluate Model > Evaluation results



### Parkinson Dataset 2-Class (Decision Forest) > Evaluate Model > Evaluation results

True Positive <b>22</b>	False Negative <b>3</b>	Accuracy <b>0.692</b>	Precision <b>0.710</b>	Threshold <b>0.5</b>	AUC <b>0.701</b>
False Positive <b>9</b>	True Negative <b>5</b>	Recall <b>0.880</b>	F1 Score <b>0.786</b>		
Positive Label <b>1</b>	Negative Label <b>0</b>				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	17	3	0.513	0.718	0.756	0.850	0.680	0.579	0.786	0.000
(0.800,0.900]	4	3	0.692	0.744	0.808	0.778	0.840	0.667	0.571	0.163
(0.700,0.800]	0	2	0.744	0.692	0.778	0.724	0.840	0.600	0.429	0.283
(0.600,0.700]	0	0	0.744	0.692	0.778	0.724	0.840	0.600	0.429	0.283
(0.500,0.600]	2	1	0.821	0.718	0.807	0.719	0.920	0.714	0.357	0.344
(0.400,0.500]	0	0	0.821	0.718	0.807	0.719	0.920	0.714	0.357	0.344
(0.300,0.400]	1	0	0.846	0.744	0.828	0.727	0.960	0.833	0.357	0.344
(0.200,0.300]	1	0	0.872	0.769	0.847	0.735	1.000	1.000	0.357	0.344

## **KESIMPULAN**

Penerapan dari model Two-Class Decision Forest menghasilkan grafik yang lebih menyimpang (bagian sekung cenderung lebih menjauhi sumbu simetri grafik true positive terhadap true negative) ketimbang dengan penerapan model Two-Class Decision Jungle. Dapat dikatakan simulasi yang menerapkan Two-Class Decision Forest lebih mutakhir dalam menghasilkan prediksi lebih akurat ketimbang Two-Class Decision Jungle untuk dataset Parkinson yang digunakan pada penerapan kasus ini.

## **REFERENSI**

- [1] 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007
- [2] Parkinson Data Set Available: <http://archive.ics.uci.edu/ml/datasets/parkinsons> [Accessed 6-9-2019]
- [3] Faisal, M. Reza & Kurniawan Erick. 2019. 'Seri Belajar Data Science Pengenalan Microsoft Azure Machine Learning Studio'. INDC

## **AKUN AZURE**

Email : xurenhuan.ridwan@email.unikom.ac.id  
Password : ciscoenpa55  
Eksperimen 1 : Parkinson Dataset 2-Class (Decision Forest)  
Eksperimen 1 : Parkinson Dataset 2-Class (Decision Jungle)  
Account Type : Personal Account