Random Forest Decision Model and Its Applications

Workshop

Albert C. Yang, M.D., Ph.D.

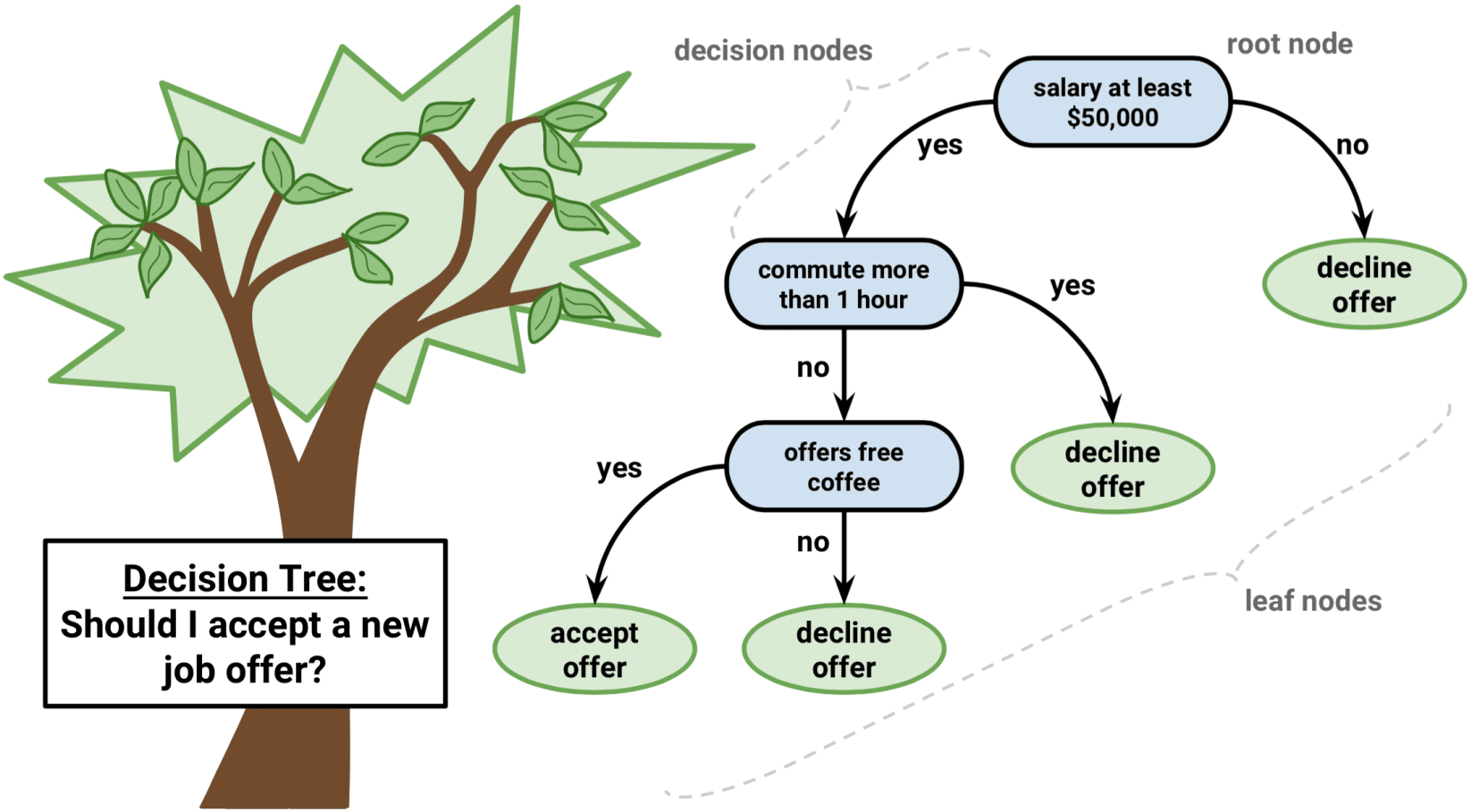
Institutes of Brain Science/Digital Medicine Center

National Yang-Ming University

Apr 16, 2020

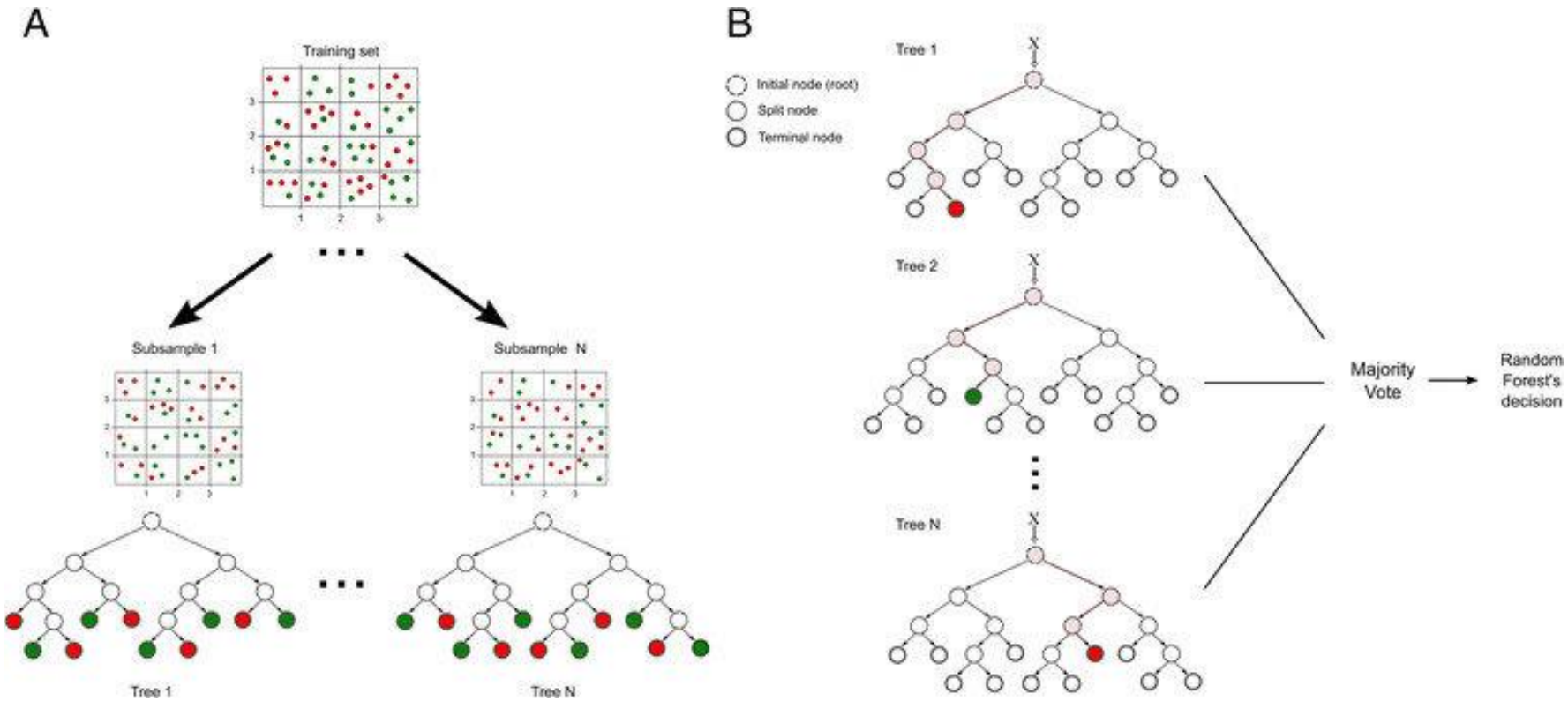
accyang@gmail.com

Classification and Regression Tree



https://www.datacamp.com/community/tutorials/decision-trees-R

Random Forest Decision Tree



https://www.researchgate.net/figure/Random-forest-model-Example-of-training-and-classification-processes-using-random_fig5_280533599

Chronic Kidney Disease Dataset

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Chronic_Kidney_Disease Data Set

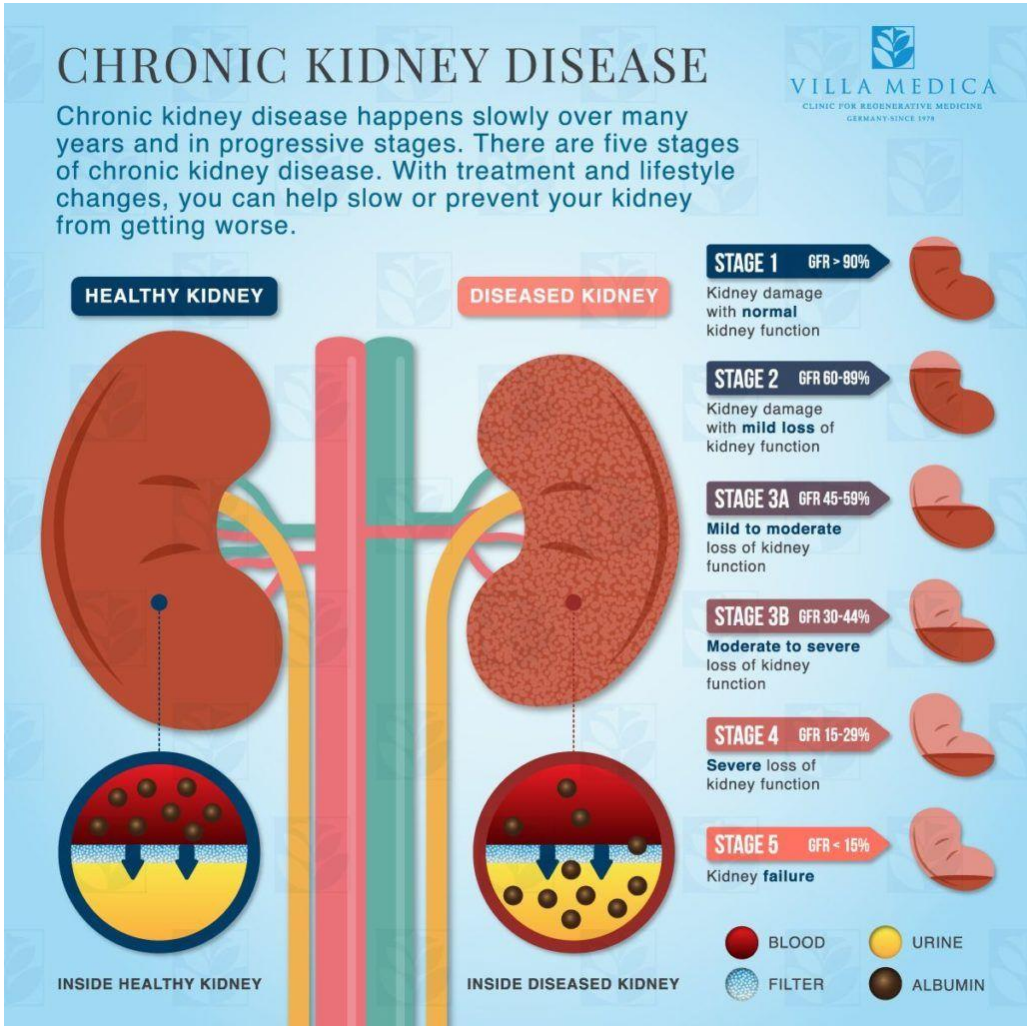
Download: Data Folder, Data Set Description

Abstract: This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period.

Data Set Characteristics:	Multivariate	**Number of Instances:**	400	**Area:**	N/A
Attribute Characteristics:	Real	**Number of Attributes:**	25	**Date Donated**	2015-07-03
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	161585

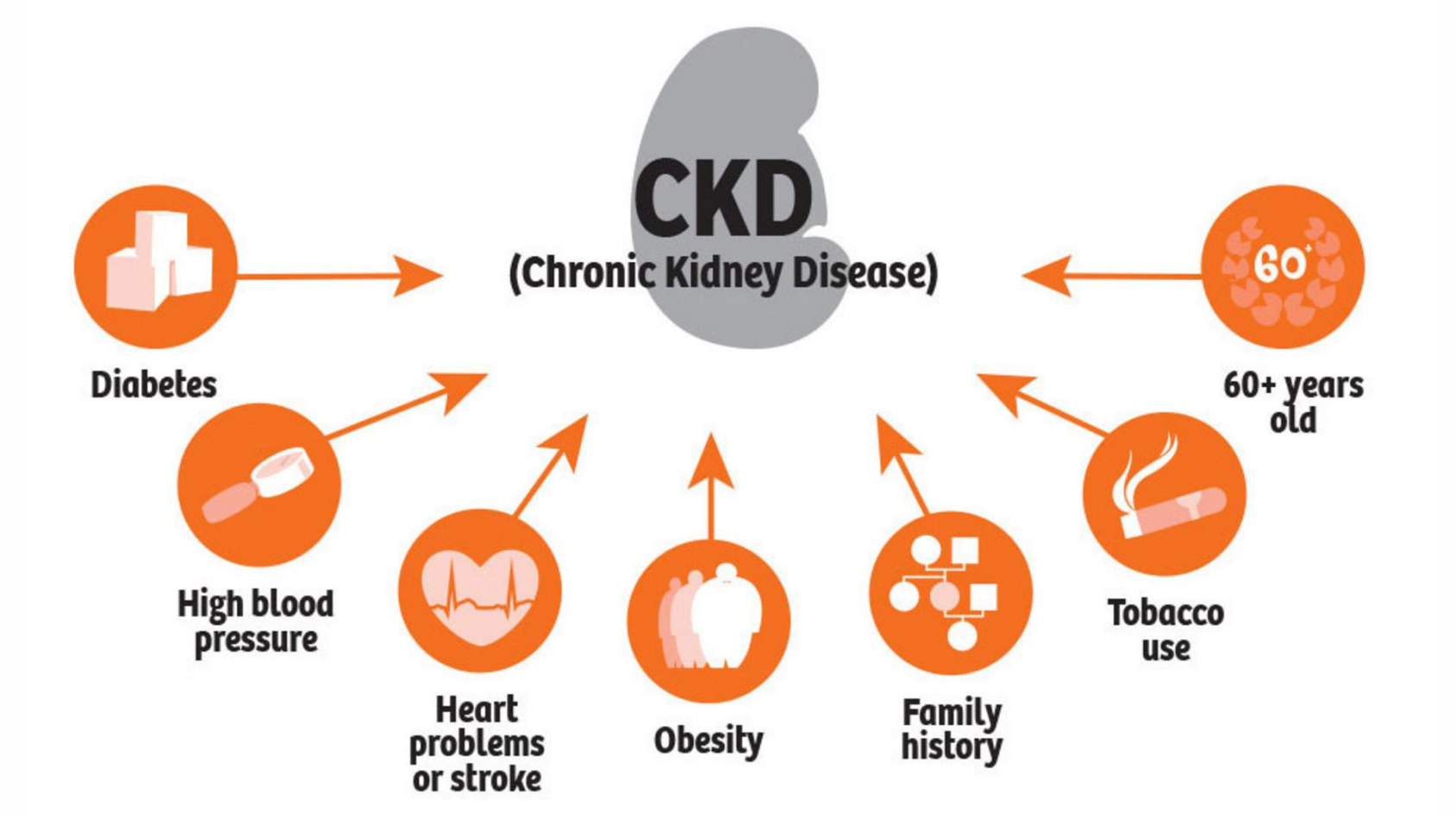
https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

What is Chronic Kidney Disease?



https://www.pinterest.com/pin/401524123025874100/

What is Chronic Kidney Disease?



https://www.siemens-healthineers.com/nl-be/news/chronic-kidney-disease.html

Chronic Kidney Disease Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm
2	0	48	80	1.02	1	0		normal	notpresent	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes
3	1	7	50	1.02	4	0		normal	notpresent	notpresent		18	0.8			11.3	38	6000		no	no
4	2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8			9.6	31	7500		no	yes
5	3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no
6	4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no
7	5	60	90	1.015	3	0			notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes
8	6	68	70	1.01	0	0		normal	notpresent	notpresent	100	54	24	104	4	12.4	36			no	no
9	7	24		1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1			12.4	44	6900	5	no	yes
10	8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9			10.8	33	9600	4	yes	yes
11	9	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes
12	10	50	60	1.01	2	4		abnormal	present	notpresent	490	55	4			9.4	28			yes	yes
13	11	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes
14	12	68	70	1.015	3	1		normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes
15	13	68	70						notpresent	notpresent	98	86	4.6	135	3.4	9.8				yes	yes
16	14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes
17	15	40	80	1.015	3	0		normal	notpresent	notpresent	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no
18	16	47	70	1.015	2	0		normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6				no	no
19	17	47	80						notpresent	notpresent	114	87	5.2	139	3.7	12.1				yes	no
20	18	60	100	1.025	0	3		normal	notpresent	notpresent	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes
21	19	62	60	1.015	1	0		abnormal	present	notpresent	100	31	1.6			10.3	30	5300	3.7	yes	no
22	20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes
23	21	60	90						notpresent	notpresent		180	76	4.5		10.9	32	6200	3.6	yes	yes
24	22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no
25	23	21	70	1.01	0	0		normal	notpresent	notpresent										no	no
26	24	42	100	1.015	4	0	normal	abnormal	notpresent	present		50	1.4	129	4	11.1	39	8300	4.6	yes	no
27	25	61	60	1.025	0	0		normal	notpresent	notpresent	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes
28	26	75	80	1.015	0	0		normal	notpresent	notpresent	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes
29	27	69	70	1.01	3	4	normal	abnormal	notpresent	notpresent	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes

Chronic Kidney Disease Dataset

Data Set Information:

We use the following representation to collect the dataset
age - age
bp - blood pressure
sg - specific gravity
al - albumin
su - sugar
rbc - red blood cells
pc - pus cell
pcc - pus cell clumps
ba - bacteria
bgr - blood glucose random
bu - blood urea
sc - serum creatinine
sod - sodium
pot - potassium
hemo - hemoglobin
pcv - packed cell volume
wc - white blood cell count
rc - red blood cell count
htn - hypertension
dm - diabetes mellitus
cad - coronary artery disease
appet - appetite
pe - pedal edema
ane - anemia
class - class

N = 368 (CKD: 228; non-CKD: 140)

	A	B	C	D	E	F	G	H	I	J
1	age	bu	sc	htn	dm	cad	appet	pe	ane	classification
2	48	36	1.2	yes	yes	no	good	no	no	ckd
3	7	18	0.8	no	no	no	good	no	no	ckd
4	62	53	1.8	no	yes	no	poor	no	yes	ckd
5	48	56	3.8	yes	no	no	poor	yes	yes	ckd
6	51	26	1.4	no	no	no	good	no	no	ckd
7	60	25	1.1	yes	yes	no	good	yes	no	ckd
8	68	54	24	no	no	no	good	no	no	ckd
9	24	31	1.1	no	yes	no	good	yes	no	ckd
10	52	60	1.9	yes	yes	no	good	no	yes	ckd
11	53	107	7.2	yes	yes	no	poor	no	yes	ckd
12	50	55	4	yes	yes	no	good	no	yes	ckd
13	63	60	2.7	yes	yes	no	poor	yes	no	ckd
14	68	72	2.1	yes	yes	yes	poor	yes	no	ckd
15	68	86	4.6	yes	yes	yes	poor	yes	no	ckd
16	68	90	4.1	yes	yes	yes	poor	yes	no	ckd
17	40	162	9.6	yes	no	no	good	no	yes	ckd
18	47	46	2.2	no	no	no	good	no	no	ckd
19	47	87	5.2	yes	no	no	poor	no	no	ckd
20	60	27	1.3	yes	yes	yes	good	no	no	ckd
21	62	31	1.6	yes	no	yes	good	no	no	ckd
22	61	148	3.9	yes	yes	yes	poor	yes	yes	ckd
23	60	180	76	yes	yes	yes	good	no	no	ckd
24	48	163	7.7	yes	no	no	good	no	yes	ckd
25	42	50	1.4	yes	no	no	poor	no	no	ckd
26	61	75	1.9	yes	yes	no	good	no	yes	ckd
27	75	45	2.4	yes	yes	no	poor	no	no	ckd
28	69	87	2.7	yes	yes	yes	good	yes	no	ckd

Read Data

- data = readtable('ckd_subset.csv');

368x11 table

	1 id	2 age	3 bu	4 sc	5 htn	6 dm	7 cad	8 appet	9 pe	10 ane	11 classification
1	0	48	36	1.2000	'yes'	'yes'	'no'	'good'	'no'	'no'	'ckd'
2	1	7	18	0.8000	'no'	'no'	'no'	'good'	'no'	'no'	'ckd'
3	2	62	53	1.8000	'no'	'yes'	'no'	'poor'	'no'	'yes'	'ckd'
4	3	48	56	3.8000	'yes'	'no'	'no'	'poor'	'yes'	'yes'	'ckd'
5	4	51	26	1.4000	'no'	'no'	'no'	'good'	'no'	'no'	'ckd'
6	5	60	25	1.1000	'yes'	'yes'	'no'	'good'	'yes'	'no'	'ckd'
7	6	68	54	24	'no'	'no'	'no'	'good'	'no'	'no'	'ckd'
8	7	24	31	1.1000	'no'	'yes'	'no'	'good'	'yes'	'no'	'ckd'
9	8	52	60	1.9000	'yes'	'yes'	'no'	'good'	'no'	'yes'	'ckd'
10	9	53	107	7.2000	'yes'	'yes'	'no'	'poor'	'no'	'yes'	'ckd'
11	10	50	55	4	'yes'	'yes'	'no'	'good'	'no'	'yes'	'ckd'
12	11	63	60	2.7000	'yes'	'yes'	'no'	'poor'	'yes'	'no'	'ckd'
13	12	68	72	2.1000	'yes'	'yes'	'yes'	'poor'	'yes'	'no'	'ckd'
14	13	68	86	4.6000	'yes'	'yes'	'yes'	'poor'	'yes'	'no'	'ckd'
15	14	68	90	4.1000	'yes'	'yes'	'yes'	'poor'	'yes'	'no'	'ckd'
16	15	40	162	9.6000	'yes'	'no'	'no'	'good'	'no'	'yes'	'ckd'
17	16	47	46	2.2000	'no'	'no'	'no'	'good'	'no'	'no'	'ckd'
18	17	47	87	5.2000	'yes'	'no'	'no'	'poor'	'no'	'no'	'ckd'

Fit Binary Decision Tree for Multiclass Classification

- tree = fitctree(data,'classification')

```
>> tree = fitctree(data,'classification')

tree =

  ClassificationTree
           PredictorNames: {'age'  'bu'  'sc'  'htn'  'dm'  'cad'  'appet'  'pe'  'ane'}
             ResponseName: 'classification'
    CategoricalPredictors: [4 5 6 7 8 9]
               ClassNames: {'ckd'  'notckd'}
           ScoreTransform: 'none'
          NumObservations: 368


  Properties, Methods
```

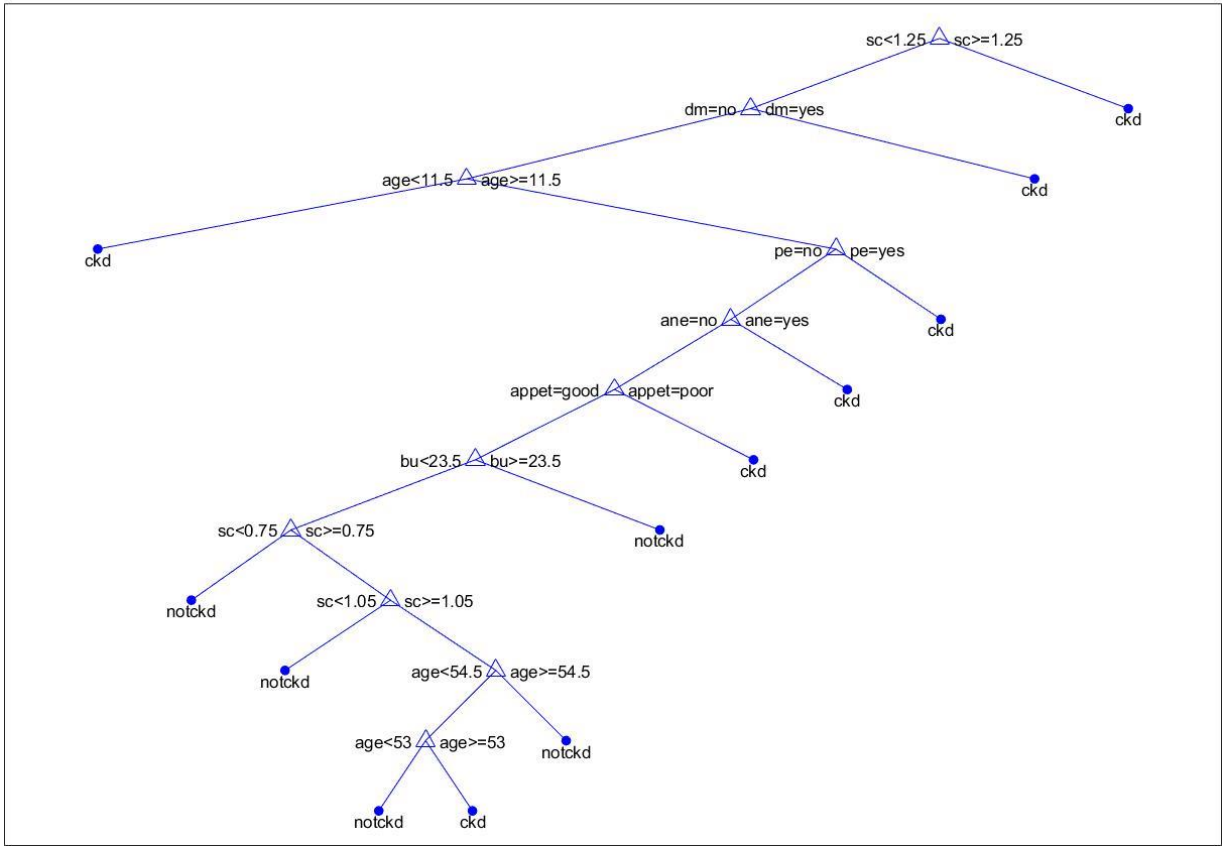
Analyze Decision Tree

```
>> view(tree)

Decision tree for classification
 1  if sc<1.25 then node 2 elseif sc>=1.25 then node 3 else ckd
 2  if dm=no then node 4 elseif dm=yes then node 5 else notckd
 3  class = ckd
 4  if age<11.5 then node 6 elseif age>=11.5 then node 7 else notckd
 5  class = ckd
 6  class = ckd
 7  if pe=no then node 8 elseif pe=yes then node 9 else notckd
 8  if ane=no then node 10 elseif ane=yes then node 11 else notckd
 9  class = ckd
10  if appet=good then node 12 elseif appet=poor then node 13 else notckd
11  class = ckd
12  if bu<23.5 then node 14 elseif bu>=23.5 then node 15 else notckd
13  class = ckd
14  if sc<0.75 then node 16 elseif sc>=0.75 then node 17 else notckd
15  class = notckd
16  class = notckd
17  if sc<1.05 then node 18 elseif sc>=1.05 then node 19 else notckd
18  class = notckd
19  if age<54.5 then node 20 elseif age>=54.5 then node 21 else notckd
20  if age<53 then node 22 elseif age>=53 then node 23 else notckd
21  class = notckd
22  class = notckd
23  class = ckd
```

Analyze Decision Tree

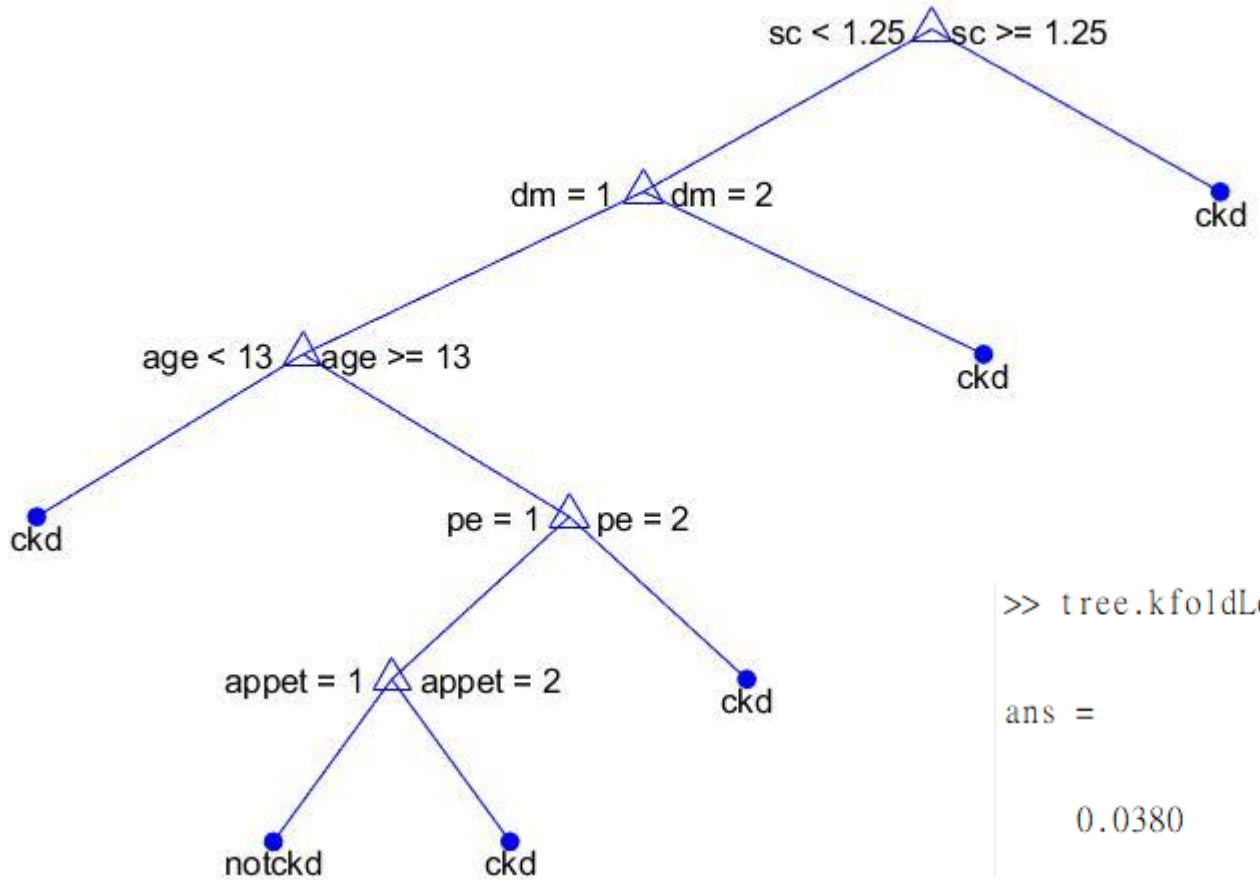
- view(tree,'Mode','graph')



Cross Validation

- tree = fitctree(data,'classification','MaxNumSplits',5,'CrossVal','on’)
- view(tree.Trained{1},'Mode','graph')

Analyze Decision Tree




```
>> tree.kfoldLoss

ans =

    0.0380
```

Train Random Forest Decision Tree

- rftree = fitcensemble(data,'classification')

```
rftree =

  classreg.learning.classif.ClassificationEnsemble
          PredictorNames: {'age'  'bu'  'sc'  'htn'  'dm'  'cad'  'appet'  'pe'  'ane'}
            ResponseName: 'classification'
   CategoricalPredictors: [4 5 6 7 8 9]
              ClassNames: {'ckd'  'notckd'}
          ScoreTransform: 'none'
         NumObservations: 368
              NumTrained: 100
                  Method: 'LogitBoost'
            LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training cycles.'
                 FitInfo: [100×1 double]
       FitInfoDescription: {2×1 cell}


  Properties, Methods
```

Analyze Random Forest Decision Tree

- view(rftree.Trained{1}.CompactRegressionLearner,'Mode','graph')