PASS

PREDICTING AND SECURING SUCCESS

NHS

Health Education North West

Manchester Metropolitan University

PASS

PREDICTING AND SECURING SUCCESS

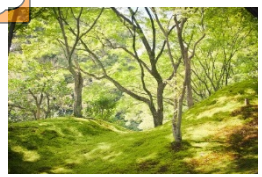
Random forest (rF)

rF overview

rF in PASS

Importance graphs

CI trees



Challenge of H-D data

Many variables

"big data"

Many cases

Different data types

Missing data

Traditional statistics

X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
Qualify	Anonymis	1_Higher	2_Lower	3_Interme	4_SmallEn	5_LowerSi	6_SemiRo	7_Routine	8_NeverW	NSSEC_Pr	Distance f	POLAI	POLA	Starti	HadP	Mode	Mode	Chan	Chan	IsAPE	IsCoU	Cour	HadS	Susp	HadS	Susp	HadS	Susp	Current	MonthsT
1	FALSE	0.069957	0.191898	0.151113	0.10065	0.07687	0.169501	0.146274	0.093737	0.544317	9217.459	4	3	24	0	1	1	0	0	0	1	4	0	0	0	0	0	0		1364
1	FALSE	0.082919	0.231563	0.136989	0.116546	0.076423	0.170615	0.141001	0.043943	0.527778	19539.29	2	3	15	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	TRUE	0.065879	0.166921	0.138846	0.08492	0.077971	0.196942	0.181237	0.087283	0.592813	24841.65	3	2	4	0	1	1	0	0	0	0	0	1	1	0	0	0	0	16	0
1	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0	76	0
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1350
-1	FALSE	0.095735	0.230055	0.13606	0.104439	0.093415	0.1613	0.13403	0.044967	0.516403	24298.63	3	3	8	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
0	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		0
1	FALSE	0.141915	0.2798	0.1468	0.119443	0.058745	0.112115	0.083293	0.05789	0.396552	16886.61	5	5	24	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	FALSE	0.090163	0.236838	0.173885	0.071808	0.075672	0.168894	0.128321	0.05442	0.470288	25521.31	1	3	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE	0.090163	0.236838	0.173885	0.071808	0.075672	0.168894	0.128321	0.05442	0.470288	26670.42	1	2	36	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1362
-1	FALSE	0.050046	0.093605	0.105036	0.100711	0.053445	0.189991	0.205746	0.201421	0.688588	954.8073	4	1	36	0	1	1	0	0	0	0	0	0	0	0	0	0	0	40	0
-1	FALSE	0.241294	0.295814	0.139289	0.107985	0.041505	0.092508	0.050827	0.030777	0.302123	20891.87	5	5	8	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
-1	FALSE	0.151688	0.286492	0.158769	0.121187	0.064815	0.111656	0.081155	0.024237	0.388222	14048.22	2	1	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
0	FALSE													6	0	1	1	0	0	0	0	0	1	1	0	0	1	12		0
1	TRUE	0.091901	0.255217	0.138618	0.074478	0.082328	0.17174	0.128087	0.05763	0.484559	24900.09	4	3	6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1352
0	FALSE	0.0302	0.120232	0.105132	0.06426	0.098887	0.222184	0.255563	0.103542	0.714919	15950.79	1	1	49	0	1	2	1	0	0	0	0	1	1	0	0	0	0	64	0
1	FALSE	0.124794	0.254686	0.186781	0.087964	0.070207	0.146498	0.09487	0.034199	0.413687	15295.39	4	4	24	0	1	1	0	0	0	1	0	0	0	0	0	0	0		1364
-1	FALSE	0.160212	0.265429	0.156852	0.085588	0.065782	0.117418	0.092485	0.056233	0.382799	11222.43	4	4	25	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
1	FALSE													13	0	1	1	0	0	0	0	0	0	0	0	0	0	0	113	0
-1	FALSE	0.15327	0.275414	0.176517	0.144208	0.06383	0.104019	0.054374	0.028369	0.377129	27886.11	2	2	8	0	2	2	0	0	0	0	0	0	0	0	0	0	0	28	0
-1	FALSE	0.04929	0.168745	0.115976	0.059438	0.084372	0.205857	0.194259	0.122064	0.619551	23851.3	2	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE													9	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													9	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													4	0	1	1	0	0	0	0	0	0	0	0	0	0	0	119	0
0	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
-1	FALSE	0.147114	0.273976	0.197393	0.113128	0.065642	0.117318	0.061453	0.023976	0.366325	30791.95	3	3	25	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	TRUE													43	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1333
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357
-1	FALSE	0.182576	0.298424	0.165585	0.07507	0.060859	0.118628	0.0692	0.029657	0.333652	20060.53	2	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
-1	FALSE	0.137383	0.266196	0.172422	0.070834	0.080161	0.144442	0.094026	0.034535	0.403394	25992.61	4	3	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
-1	FALSE	0.137383	0.266196	0.172422	0.070834	0.080161	0.144442	0.094026	0.034535	0.403394	22033.21	1	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
1	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1350
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357
1	FALSE	0.219255	0.306685	0.172648	0.093894	0.042694	0.100187	0.043885	0.020752	0.286608	24866.58	5	5	24	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	FALSE	0.181217	0.340608	0.172619	0.064484	0.056878	0.10582	0.055886	0.022487	0.289581	16228.52	5	4	8	0	2	2	0	0	0	0	0	0	0	0	0	0	0	28	0
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357

What is “random forest”?

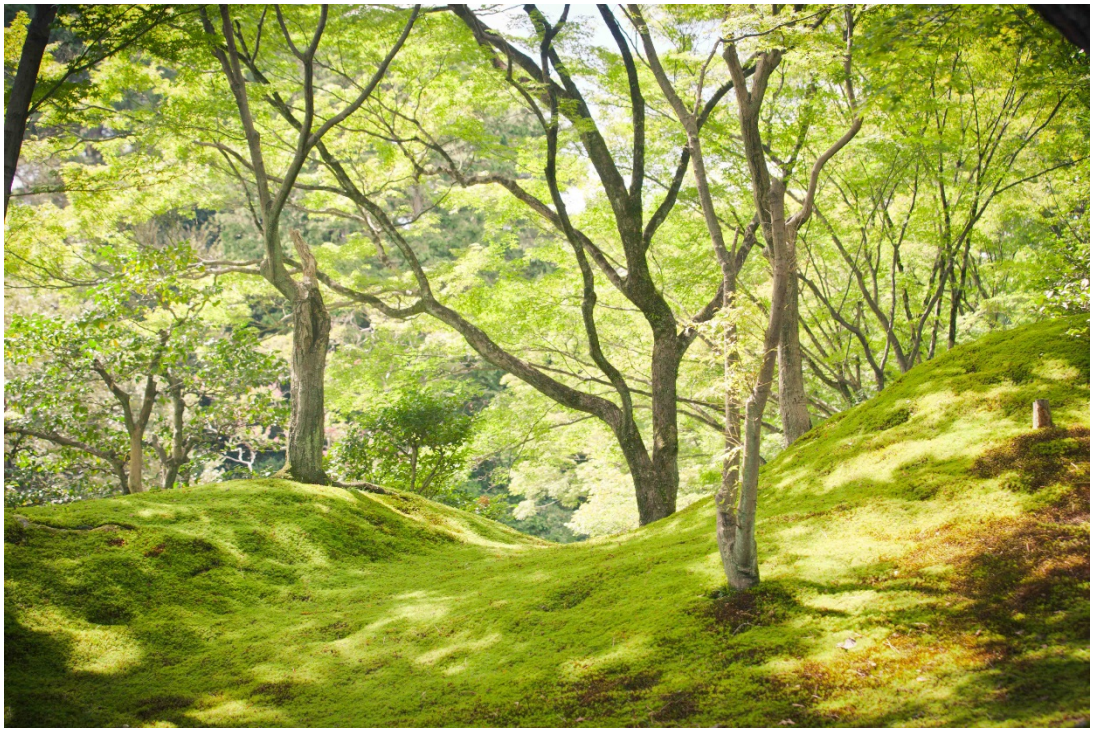
You can only see the forest by looking at the trees

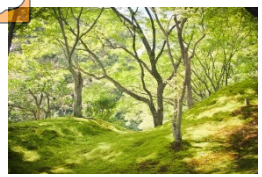
Machine learning

Varied applications

Big data

Outputs

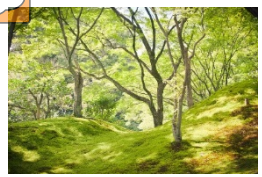




How does random forest work?

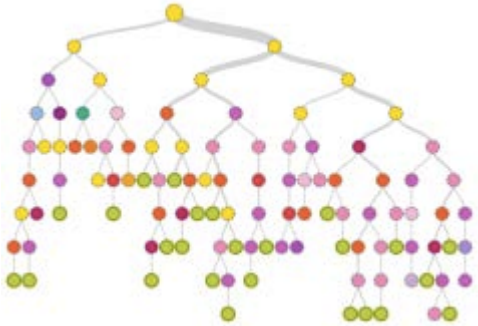
Recursive partitioning

X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
Qualify	Anonymis	1_Higher\	2_Lower\	3_Interme	4_SmallEn	5_LowerSi	6_SemiRo	7_Routine	8_Never\	NSSEC_Pr	Distance f	POLA!	POLA	Start	HadP	Mode	Mode	Chan	Chan	IsAPE	IsCou	Cour	HadS	Susp	HadS	Susp	HadS	Susp	Current	MonthsT
1	FALSE	0.069957	0.191898	0.151113	0.10065	0.07687	0.169501	0.146274	0.093737	0.544317	9217.459	4	3	24	0	1	1	0	0	0	1	4	0	0	0	0	0	0		1364
1	FALSE	0.082919	0.231563	0.136989	0.116546	0.076423	0.170615	0.141001	0.043943	0.527778	19539.29	2	3	15	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	TRUE	0.065879	0.166921	0.138846	0.08492	0.077971	0.196942	0.181237	0.087283	0.592813	24841.65	3	2	4	0	1	1	0	0	0	0	0	1	1	0	0	0	0	16	0
1	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0	76	0
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1350
-1	FALSE	0.095735	0.230055	0.13606	0.104439	0.093415	0.1613	0.13403	0.044967	0.516403	24298.63	3	3	8	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
0	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		0
1	FALSE	0.141915	0.2798	0.1468	0.119443	0.058745	0.112115	0.083293	0.05789	0.396552	16886.61	5	5	24	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	FALSE	0.090163	0.236838	0.173885	0.071808	0.075672	0.168894	0.128321	0.05442	0.470288	25521.31	1	3	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE	0.090163	0.236838	0.173885	0.071808	0.075672	0.168894	0.128321	0.05442	0.470288	26670.42	1	2	36	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1362
-1	FALSE	0.050046	0.093605	0.105036	0.100711	0.053445	0.189991	0.205746	0.201421	0.688588	954.8073	4	1	36	0	1	1	0	0	0	0	0	0	0	0	0	0	0	40	0
-1	FALSE	0.241294	0.295814	0.139289	0.107985	0.041505	0.092508	0.050827	0.030777	0.302123	20891.87	5	5	8	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
-1	FALSE	0.151688	0.286492	0.158769	0.121187	0.064815	0.111656	0.081155	0.024237	0.388222	14048.22	2	1	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
0	FALSE													6	0	1	1	0	0	0	0	0	1	1	0	0	1	12		0
1	TRUE	0.091901	0.255217	0.138618	0.074478	0.082328	0.17174	0.128087	0.05763	0.484559	24900.09	4	3	6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1352
0	FALSE	0.0302	0.120232	0.105132	0.06426	0.098887	0.222184	0.255563	0.103542	0.714919	15950.79	1	1	49	0	1	2	1	0	0	0	1	1	1	0	0	0	0	64	0
1	FALSE	0.124794	0.254686	0.186781	0.087964	0.070207	0.146498	0.09487	0.034199	0.413687	15295.39	4	4	24	0	1	1	0	0	0	1	0	0	0	0	0	0	0		1364
-1	FALSE	0.160212	0.265429	0.156852	0.085588	0.065782	0.117418	0.092485	0.056233	0.382799	11222.43	4	4	25	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
1	FALSE													13	0	1	1	0	0	0	0	0	0	0	0	0	0	0	113	0
-1	FALSE	0.15327	0.275414	0.176517	0.144208	0.06383	0.104019	0.054374	0.028369	0.377129	27886.11	2	2	8	0	2	2	0	0	0	0	0	0	0	0	0	0	0	28	0
-1	FALSE	0.04929	0.168745	0.115976	0.059438	0.084372	0.205857	0.194259	0.122064	0.619551	23851.3	2	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	FALSE													9	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													9	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													4	0	1	1	0	0	0	0	0	0	0	0	0	0	0	119	0
0	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
-1	FALSE	0.147114	0.273976	0.197393	0.113128	0.065642	0.117318	0.061453	0.023976	0.366325	30791.95	3	3	25	0	1	1	0	0	0	0	0	0	0	0	0	0	0	4	0
1	TRUE													43	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1333
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357
-1	FALSE	0.182576	0.298424	0.165585	0.07507	0.060859	0.118628	0.0692	0.029657	0.333652	20060.53	2	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
-1	FALSE	0.137383	0.266196	0.172422	0.070834	0.080161	0.144442	0.094026	0.034535	0.403394	25992.61	4	3	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
-1	FALSE	0.137383	0.266196	0.172422	0.070834	0.080161	0.144442	0.094026	0.034535	0.403394	22033.21	1	2	35	0	1	1	0	0	0	0	0	0	0	0	0	0	0	16	0
1	FALSE													49	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1338
1	FALSE													6	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1337
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1350
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357
1	FALSE	0.219255	0.306685	0.172648	0.093894	0.042694	0.100187	0.043885	0.020752	0.286608	24866.58	5	5	24	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1364
-1	FALSE	0.181217	0.340608	0.172619	0.064484	0.056878	0.10582	0.055886	0.022487	0.289581	16228.52	5	4	8	0	2	2	0	0	0	0	0	0	0	0	0	0	0	28	0
1	FALSE													33	0	1	1	0	0	0	0	0	0	0	0	0	0	0		1357

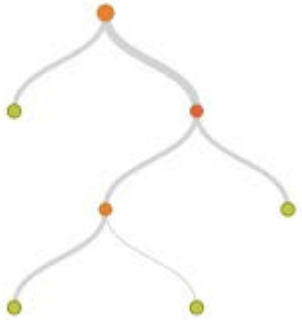


How does random forest work?

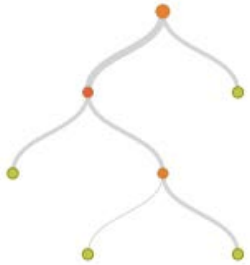
Complicated model



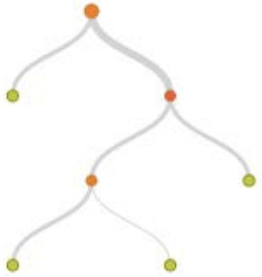
Simple models = “trees”



tree 1

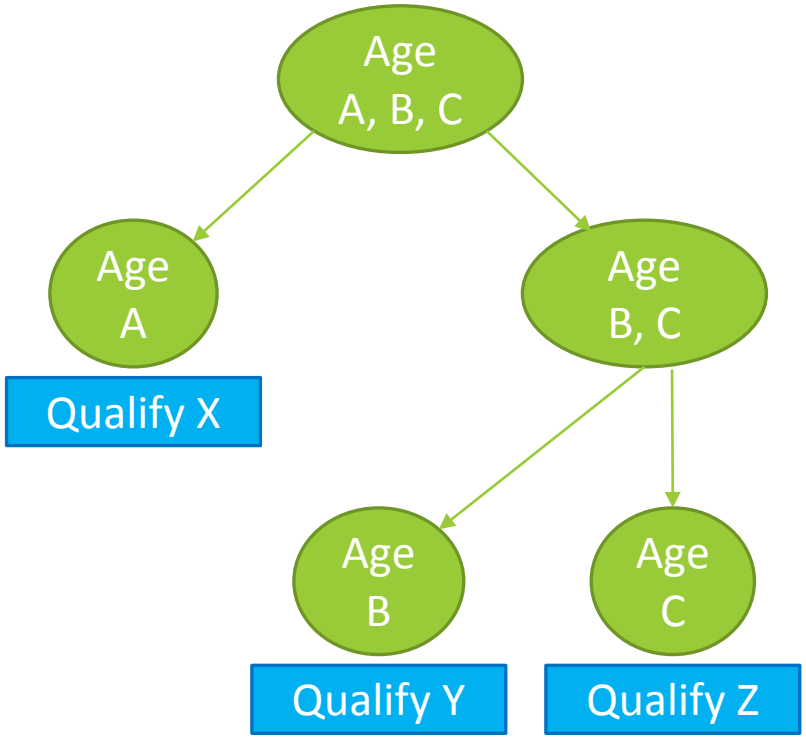


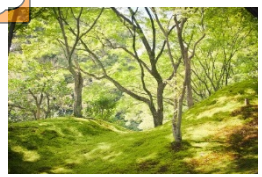
tree 2



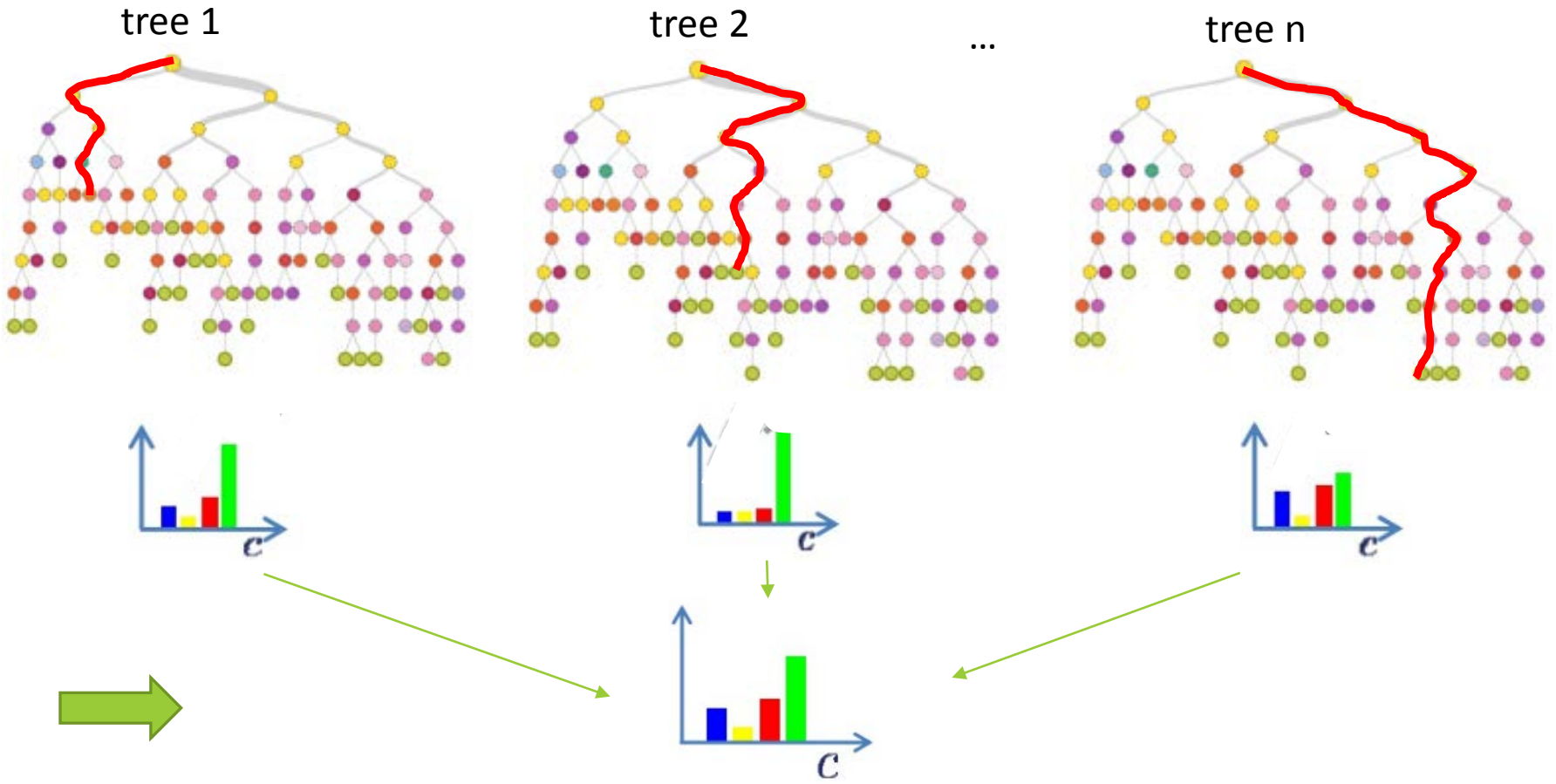
tree 3

How does random forest work?





How does random forest work?



What can we learn?

Identify important variables

Variable importance ranking

Conditional inference trees

PASS

PREDICTING AND SECURING SUCCESS

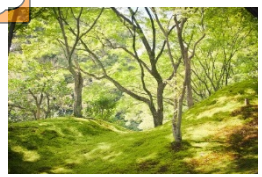
Random forest (rF)

rF overview

rF in PASS

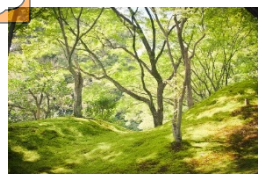
Importance graphs

CI trees



Role of rF in PASS

- Identify important variables
- Rank variable importance
- Identify how important variables differ
- Conditional inference trees



Role of rF in PASS

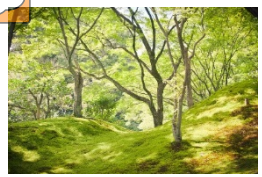
Identify important variables

Distance
Gender
Disabled
Age

YouthHE
AdultHE
Ethnicity

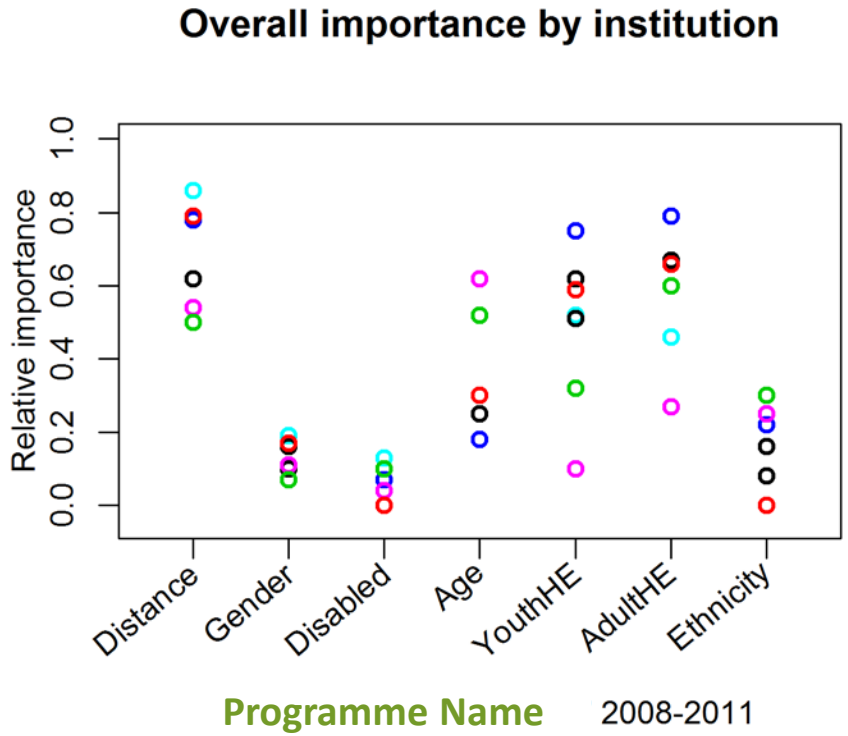
Suspend
Institution
Programme

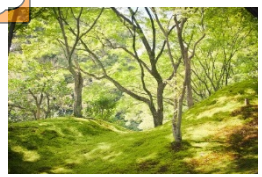
QUALIFY



Role of rF in PASS

Rank variable importance



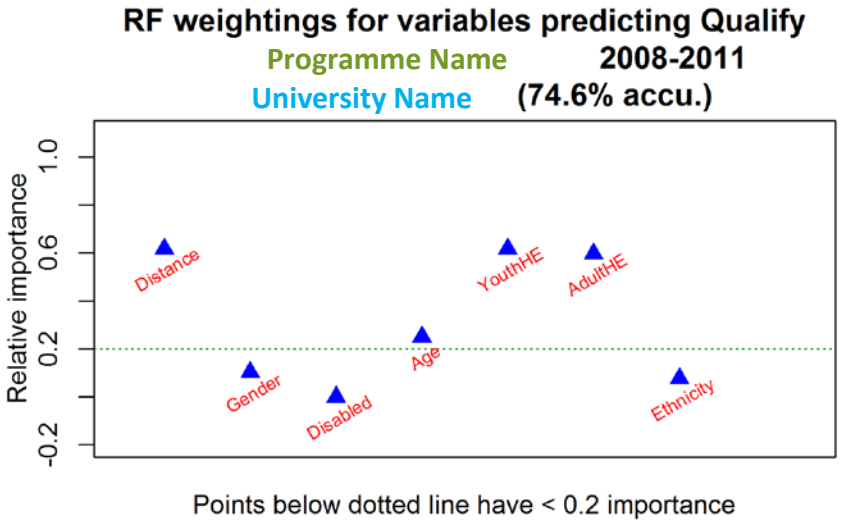


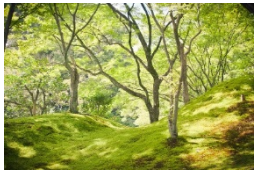
Role of rF in PASS

Identify how important variables differ

by institution

by programme





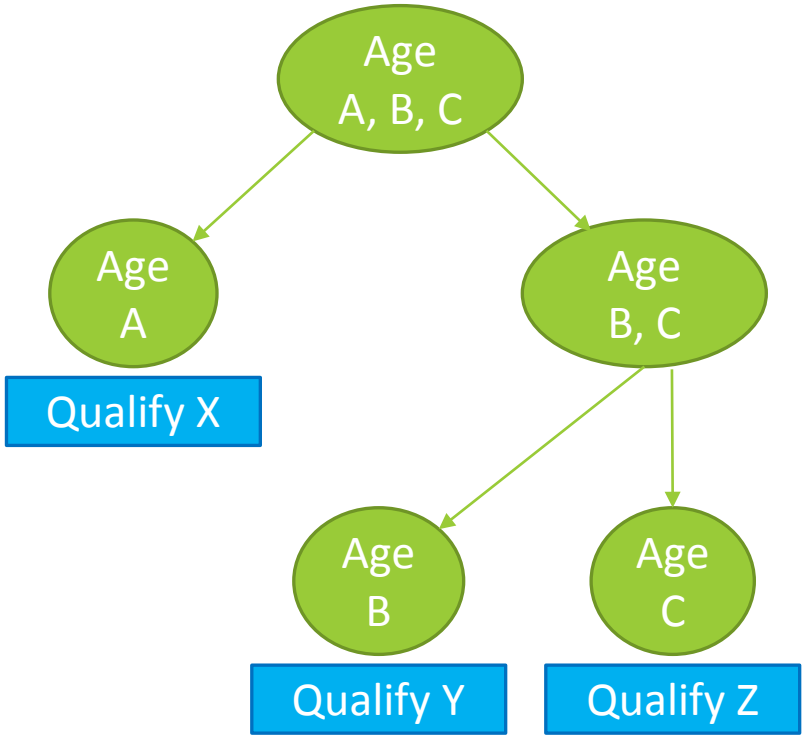
Role of rF in PASS

Conditional inference trees

aids interpretation

action of specific variables

specific student groups



PASS

PREDICTING AND SECURING SUCCESS

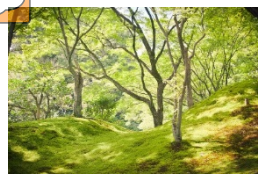
Random forest (rF)

rF overview

rF in PASS

Importance graphs

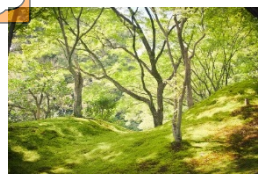
CI trees



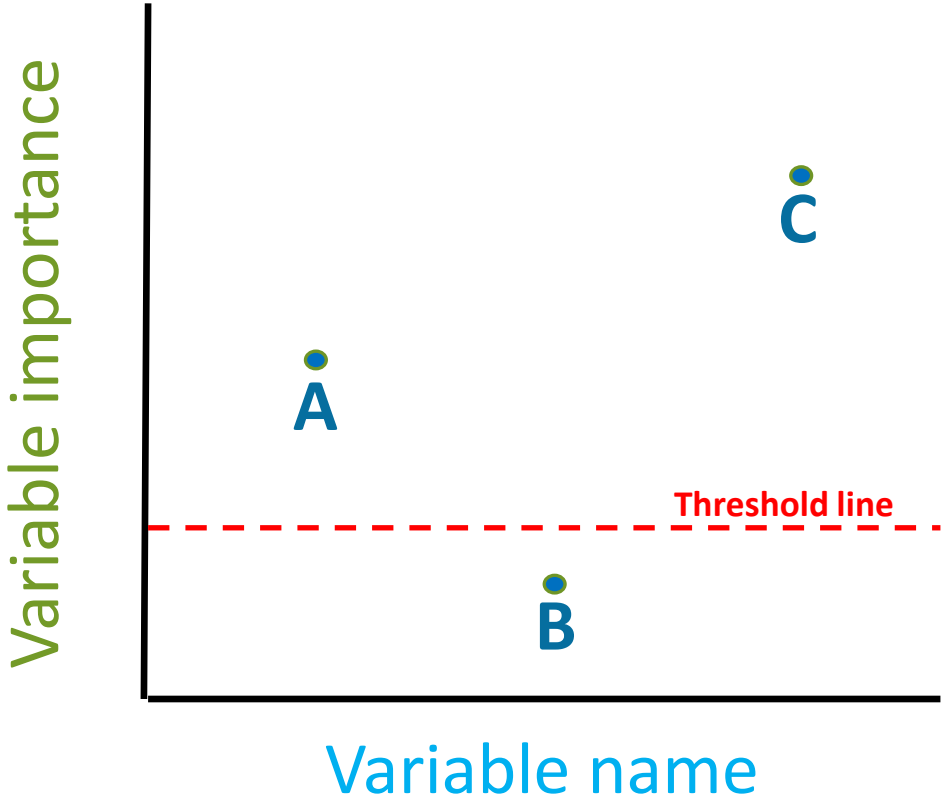
Importance graphs

Rank variable importance

Identify how important variables differ

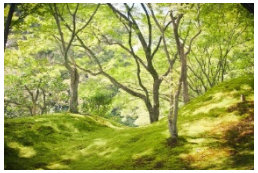


Importance graphs



Variable importance
is a relative measure of the explanatory power of each variable

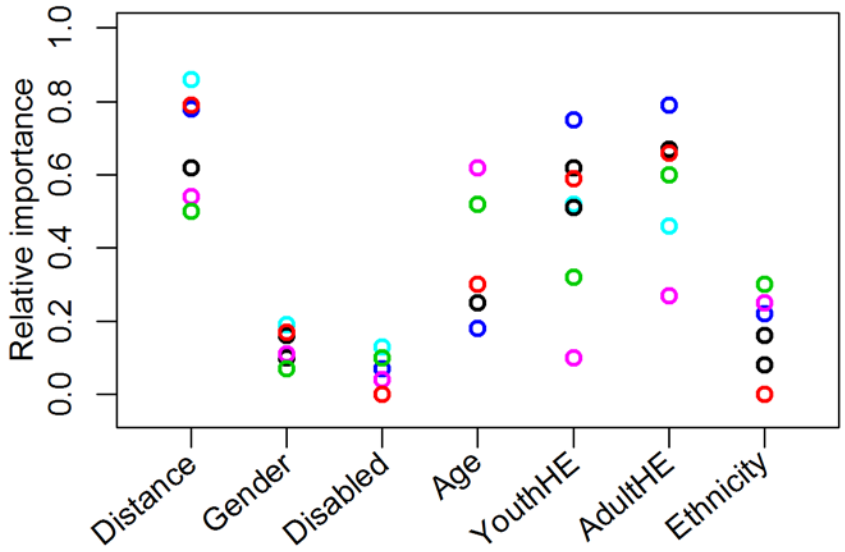
Threshold line
is an arbitrary value to classify variable importance



Importance graphs

Rank variable importance

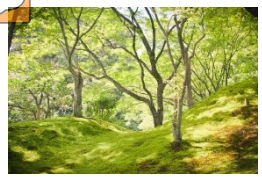
Overall importance by institution



Programme Name 2008-2011

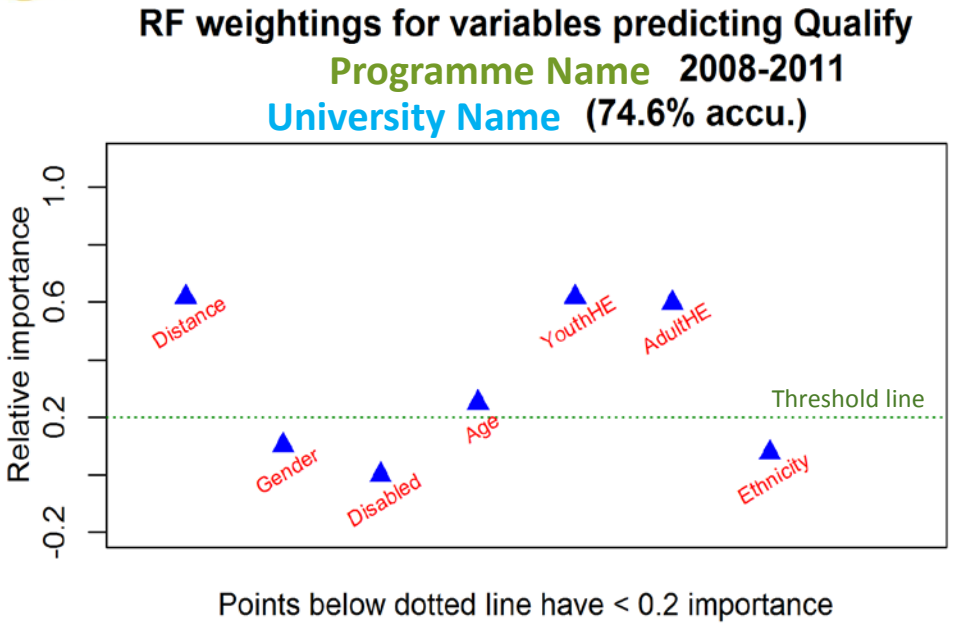
Difference
in importance

Variation
amongst institutions



Importance graphs

Identify how important variables differ



Programme and university pattern of variable importance

PASS

PREDICTING AND SECURING SUCCESS

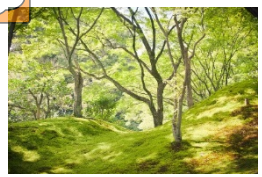
Random forest (rF)

rF overview

rF in PASS

Importance graphs

CI trees



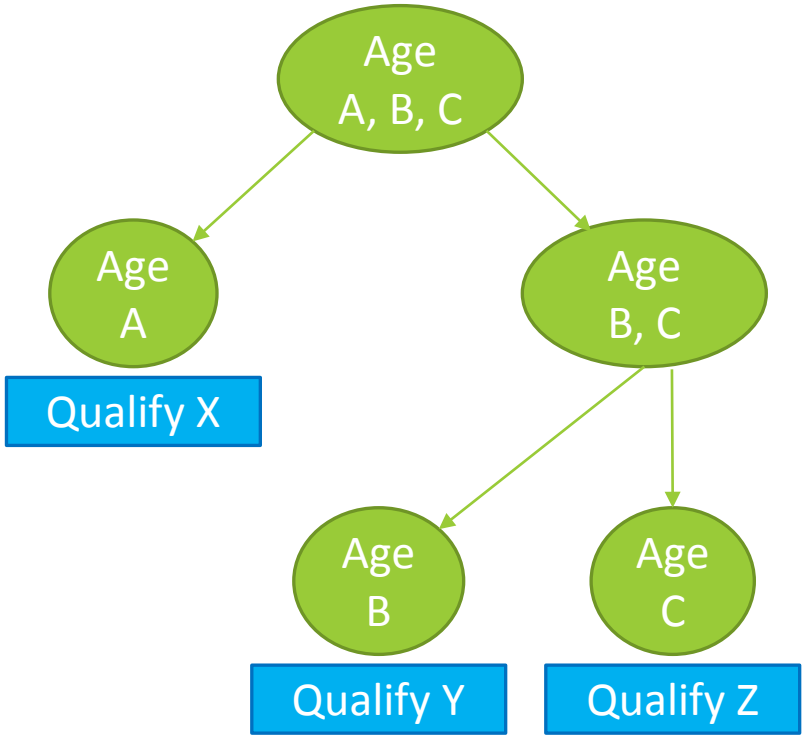
CI trees

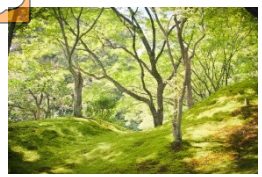
Conditional inference trees

aids interpretation

action of specific variables

specific student groups





CI trees

Conditional inference trees

Programme name

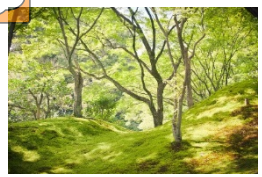
University name

Qualify: I = 0, j = 1, k = unk

Factors affecting Qualify in **Programme name** 2008-2011

University Name

y = (i, j, k); Qualify: i = 0, j = 1, k = unk



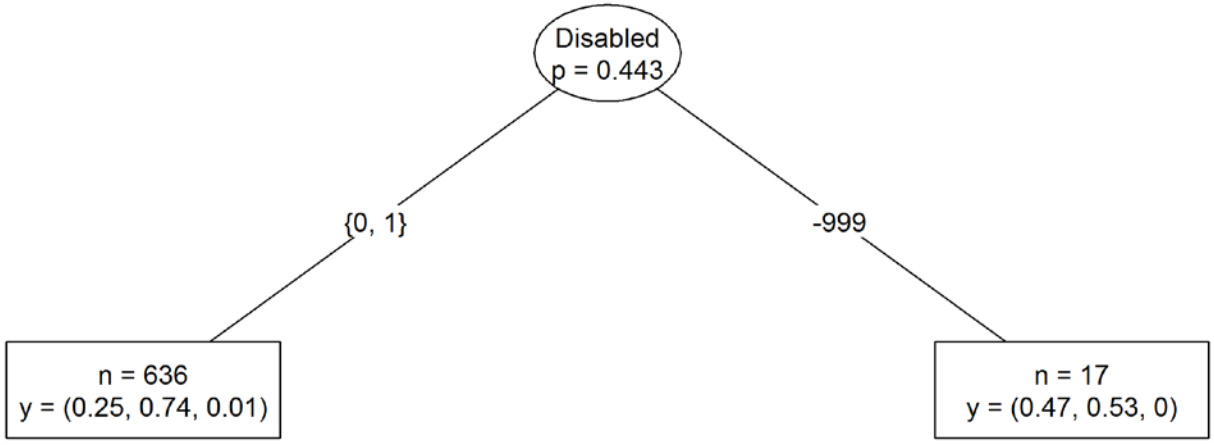
CI trees

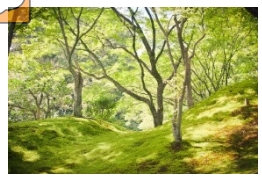
Conditional inference trees

Programme name

University name

y = (i, j, k); Qualify: i = 0, j = 1, k = unk





CI trees

Conditional inference trees

Programme name

University name