Data Mining for agriculture Workshop

CGIAR | Platform for Big Data in Agriculture

CIAT 50 1967-2017

Date
Nairobi, Kenia

Hugo Andres Dorado B.

h.a.dorado@cgiar.org

Agenda

- A brief overview to R and the basic functions and graphics
- Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).
- Training machine learning models.
- Interpreting machine learning models outputs.
- Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)

Agenda

- **A brief overview to R and the basic functions and graphics**
- Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).
- Training machine learning models.
- Interpreting machine learning models outputs.
- Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)

Getting start in R

- It’s free
- Less easy than an interphase but easier than a complex program language.
- Versatile.
- Big community. (R-bloggers, stackflow,…)
- Produce nice graphics.

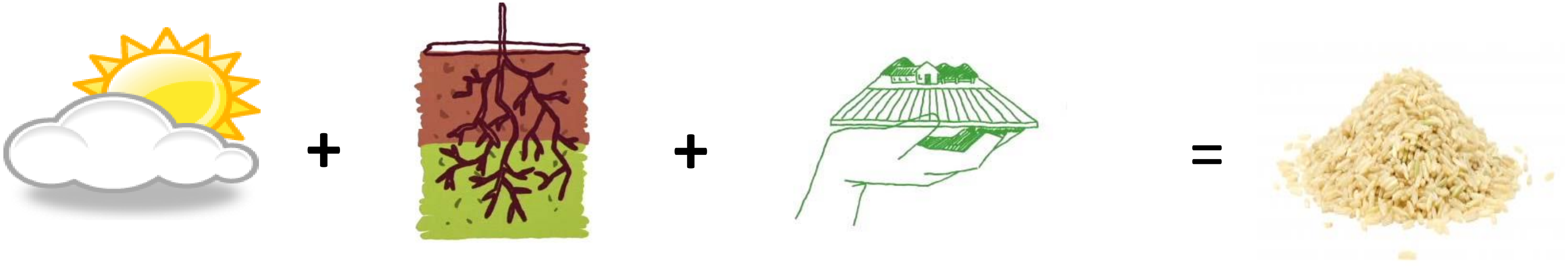
Practice in R

- Help
- Install packages
- Read datasets
- Objects.
- Mathematical operation
- Summary function.
- Basic graphics.
- ggplot graphics

Agenda

- A brief overview to R and the basic functions and graphics
- **Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).**
- Training machine learning models.
- Interpreting machine learning models outputs.
- Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)

Getting data



Wheatear **Soil** **Agronomic management** **Productivity**

Uncontrollable factors

Controllable factors

Getting data

Wheatear:

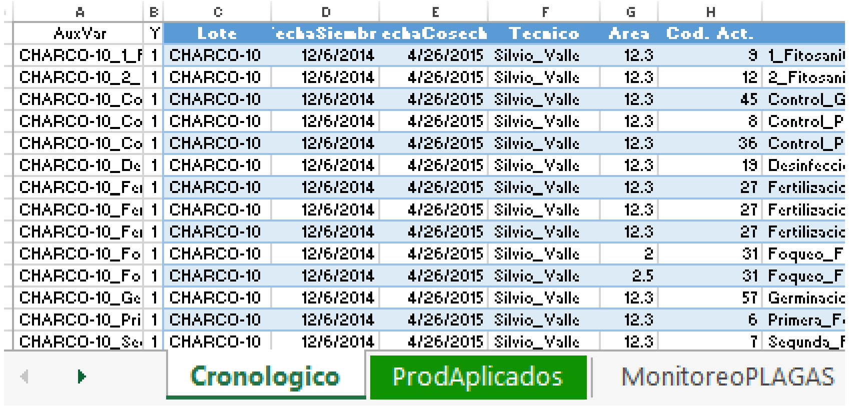
- Station (Airports, meteorological institutes, farmers)
- aWhere, http://www.awhere.com/, https://aqueous-fjord-58270.herokuapp.com/.
- http://www.worldclim.org/
- *www.cru.uea.ac.uk/data*

Soil:

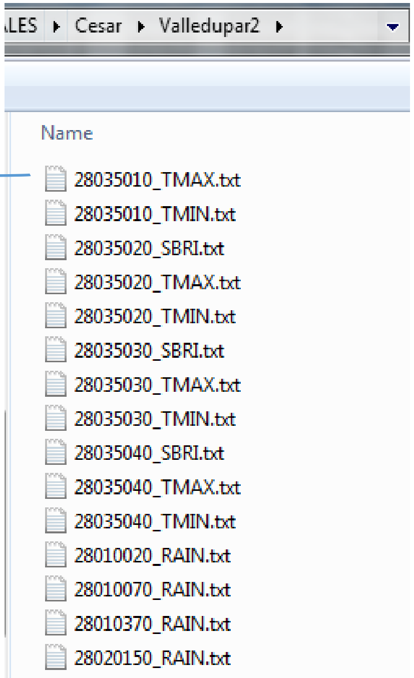
- Soil analysis.
- Soil mapping (Another projects).
- RASTA (https://cgspace.cgiar.org/handle/10568/69682).
- SoildGrid, (https://www.soilgrids.org/, ftp://ftp.soilgrids.org/data/recent/).

Processing data challenges

Many variables formats: numeric, date, text,…

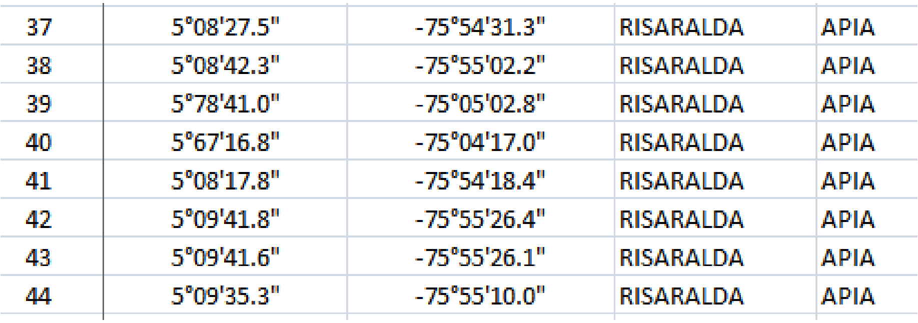


A	B	C	D	E	F	G	H	
AuxVar	Y	Lote	'echaSiembr	echaCosech	Tecnico	Area	Cod. Act.	
CHARCO-10_1_F	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	9	1_Fitosani
CHARCO-10_2_	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	12	2_Fitosani
CHARCO-10_Co	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	45	Control_G
CHARCO-10_Co	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	8	Control_P
CHARCO-10_Co	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	36	Control_P
CHARCO-10_De	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	19	Desinfecci
CHARCO-10_Fe	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	27	Fertilizacic
CHARCO-10_Fe	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	27	Fertilizacic
CHARCO-10_Fe	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	27	Fertilizacic
CHARCO-10_Fo	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	2	31	Foqueo_F
CHARCO-10_Fo	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	2.5	31	Foqueo_F
CHARCO-10_Ge	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	57	Germinacic
CHARCO-10_Pri	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	6	Primera_F
CHARCO-10_Se	1	CHARCO-10	12/6/2014	4/26/2015	Silvio_Valle	12.3	7	Segunda_F



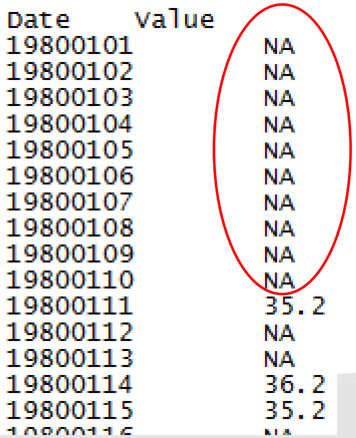
28035010_TMAX.txt
28035010_TMIN.txt
28035020_SBRI.txt
28035020_TMAX.txt
28035020_TMIN.txt
28035030_SBRI.txt
28035030_TMAX.txt
28035030_TMIN.txt
28035040_SBRI.txt
28035040_TMAX.txt
28035040_TMIN.txt
28010020_RAIN.txt
28010070_RAIN.txt
28010370_RAIN.txt
28020150_RAIN.txt

Miracles coordinates



37	5°08'27.5"	-75°54'31.3"	RISARALDA	APIA
38	5°08'42.3"	-75°55'02.2"	RISARALDA	APIA
39	5°78'41.0"	-75°05'02.8"	RISARALDA	APIA
40	5°67'16.8"	-75°04'17.0"	RISARALDA	APIA
41	5°08'17.8"	-75°54'18.4"	RISARALDA	APIA
42	5°09'41.8"	-75°55'26.4"	RISARALDA	APIA
43	5°09'41.6"	-75°55'26.1"	RISARALDA	APIA
44	5°09'35.3"	-75°55'10.0"	RISARALDA	APIA

Datos desagregados y en otros formatos



```
Date      Value
19800101       NA
19800102       NA
19800103       NA
19800104       NA
19800105       NA
19800106       NA
19800107       NA
19800108       NA
19800109       NA
19800110       NA
19800111       35.2
19800112       NA
19800113       NA
19800114       36.2
19800115       35.2
```

Missing values

Platform for
Big Data
in Agriculture

CGIAR

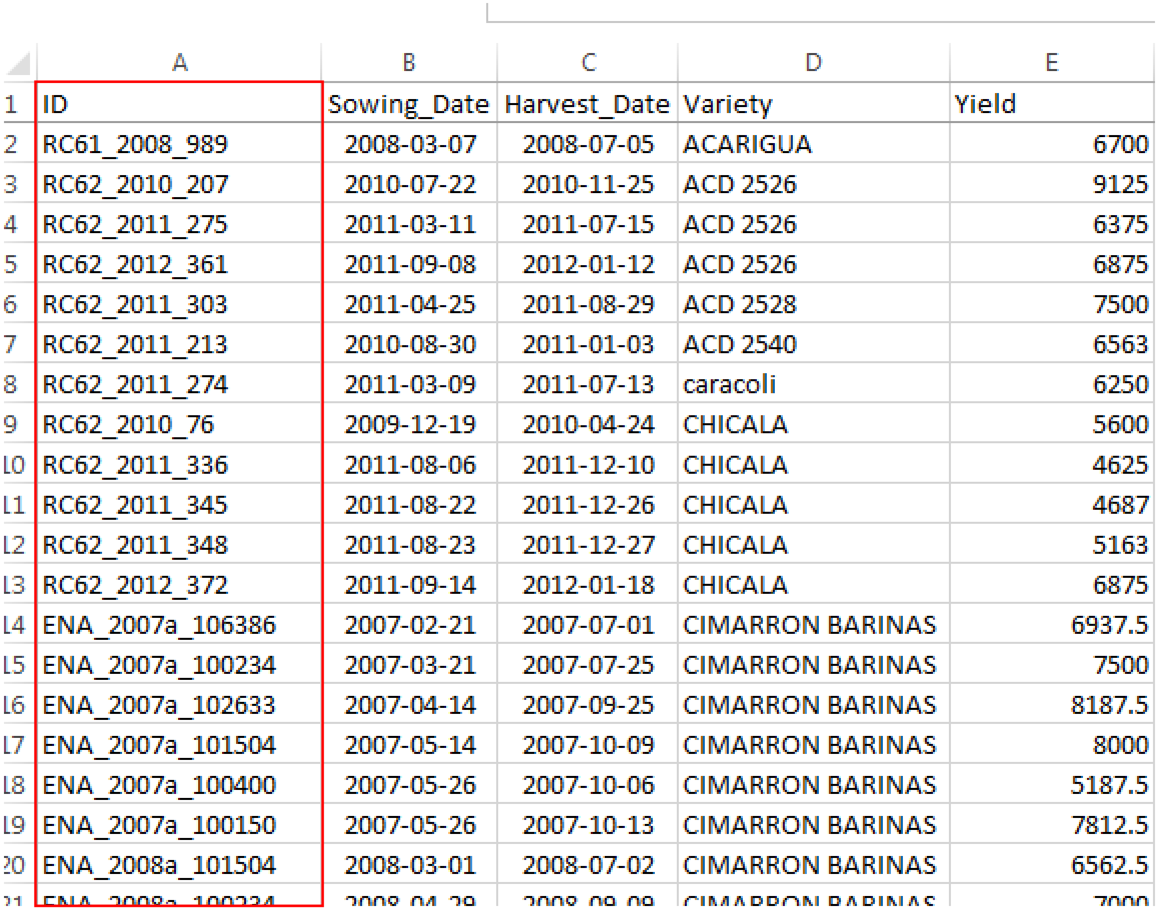
Our vision, a sustainable food future

CIAT 50 1967-2017

Getting started

Dataset structure

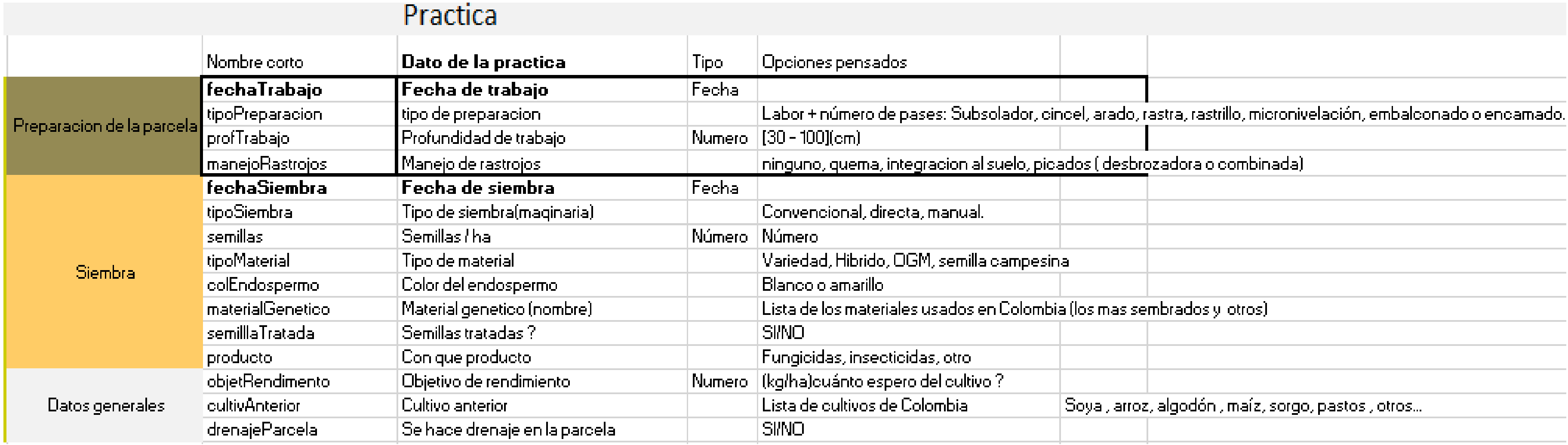
Each row represent a observation and each row represent a variable



	A	B	C	D	E
1	ID	Sowing_Date	Harvest_Date	Variety	Yield
2	RC61_2008_989	2008-03-07	2008-07-05	ACARIGUA	6700
3	RC62_2010_207	2010-07-22	2010-11-25	ACD 2526	9125
4	RC62_2011_275	2011-03-11	2011-07-15	ACD 2526	6375
5	RC62_2012_361	2011-09-08	2012-01-12	ACD 2526	6875
6	RC62_2011_303	2011-04-25	2011-08-29	ACD 2528	7500
7	RC62_2011_213	2010-08-30	2011-01-03	ACD 2540	6563
8	RC62_2011_274	2011-03-09	2011-07-13	caracoli	6250
9	RC62_2010_76	2009-12-19	2010-04-24	CHICALA	5600
10	RC62_2011_336	2011-08-06	2011-12-10	CHICALA	4625
11	RC62_2011_345	2011-08-22	2011-12-26	CHICALA	4687
12	RC62_2011_348	2011-08-23	2011-12-27	CHICALA	5163
13	RC62_2012_372	2011-09-14	2012-01-18	CHICALA	6875
14	ENA_2007a_106386	2007-02-21	2007-07-01	CIMARRON BARINAS	6937.5
15	ENA_2007a_100234	2007-03-21	2007-07-25	CIMARRON BARINAS	7500
16	ENA_2007a_102633	2007-04-14	2007-09-25	CIMARRON BARINAS	8187.5
17	ENA_2007a_101504	2007-05-14	2007-10-09	CIMARRON BARINAS	8000
18	ENA_2007a_100400	2007-05-26	2007-10-06	CIMARRON BARINAS	5187.5
19	ENA_2007a_100150	2007-05-26	2007-10-13	CIMARRON BARINAS	7812.5
20	ENA_2008a_101504	2008-03-01	2008-07-02	CIMARRON BARINAS	6562.5
21	ENA_2008a_100234	2008-04-29	2008-09-09	CIMARRON BARINAS	7000

Be sure to add an ID to the dataset, this is necessary to connect another datasets.

Crear o tener presente un diccionario de datos



Practica

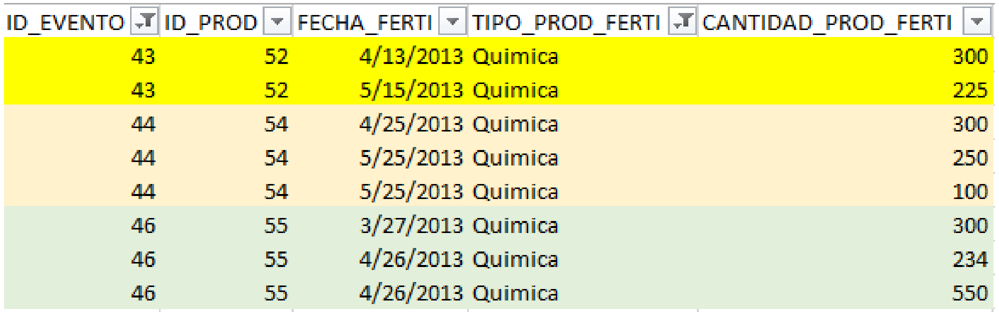
	Nombre corto	Dato de la practica	Tipo	Opciones pensados	
Preparacion de la parcela	**fechaTrabajo**	**Fecha de trabajo**	Fecha		
	tipoPreparacion	tipo de preparacion		Labor + número de pases: Subsolador, cincel, arado, rastra, rastrillo, micronivelación, embalconado o encamado.	
	profTrabajo	Profundidad de trabajo	Numero	[30 - 100](cm)	
	manejoRastrojos	Manejo de rastrojos		ninguno, quema, integracion al suelo, picados (desbrozadora o combinada)	
Siembra	**fechaSiembra**	**Fecha de siembra**	Fecha		
	tipoSiembra	Tipo de siembra(maqinaria)		Convencional, directa, manual.	
	semillas	Semillas / ha	Número	Número	
	tipoMaterial	Tipo de material		Variedad, Hibrido, OGM, semilla campesina	
	colEndospermo	Color del endospermo		Blanco o amarillo	
	materialGenetico	Material genetico (nombre)		Lista de los materiales usados en Colombia (los mas sembrados y otros)	
	semilllaTratada	Semillas tratadas ?		SI/NO	
	producto	Con que producto		Fungicidas, insecticidas, otro	
Datos generales	objetRendimento	Objetivo de rendimiento	Numero	(kg/ha)cuánto espero del cultivo ?	
	cultivAnterior	Cultivo anterior		Lista de cultivos de Colombia	Soya , arroz, algodón , maíz, sorgo, pastos , otros...
	drenajeParcela	Se hace drenaje en la parcela		SI/NO	

At less is suggested to report the next information for each variable:

- A small name.
- The complete name.
- unit of measurement
- Range [Max - Min], posible categories.

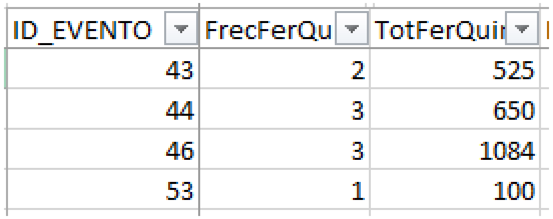
Variables transformation

Repeated rows (fertilizers)



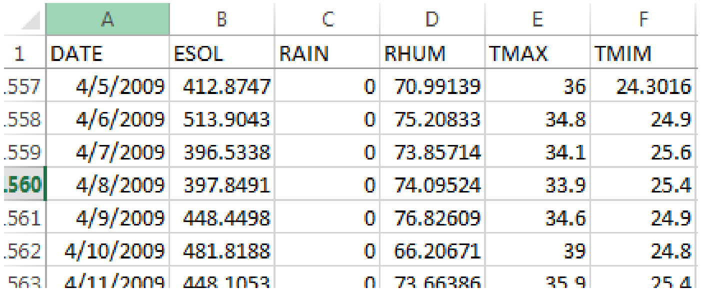
ID_EVENTO	ID_PROD	FECHA_FERTI	TIPO_PROD_FERTI	CANTIDAD_PROD_FERTI
43	52	4/13/2013	Quimica	300
43	52	5/15/2013	Quimica	225
44	54	4/25/2013	Quimica	300
44	54	5/25/2013	Quimica	250
44	54	5/25/2013	Quimica	100
46	55	3/27/2013	Quimica	300
46	55	4/26/2013	Quimica	234
46	55	4/26/2013	Quimica	550

New variables summarized



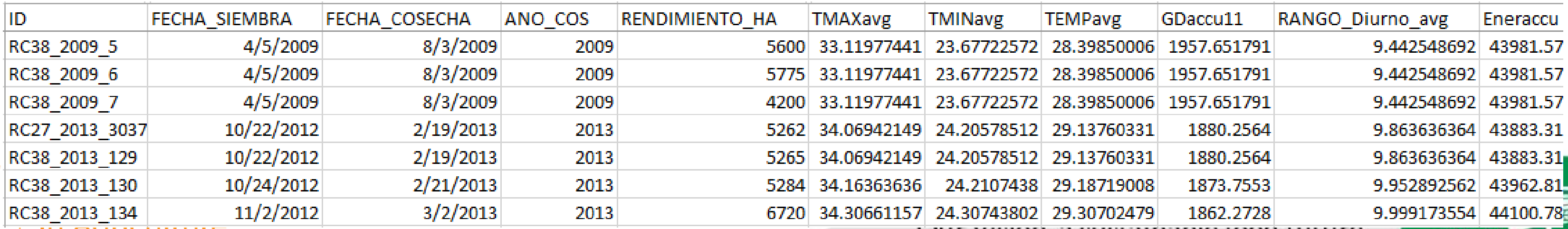
ID_EVENTO	FrecFerQu	TotFerQui
43	2	525
44	3	650
46	3	1084
53	1	100

Daily information



	A	B	C	D	E	F
1	DATE	ESOL	RAIN	RHUM	TMAX	TMIM
.557	4/5/2009	412.8747	0	70.99139	36	24.3016
.558	4/6/2009	513.9043	0	75.20833	34.8	24.9
.559	4/7/2009	396.5338	0	73.85714	34.1	25.6
.560	4/8/2009	397.8491	0	74.09524	33.9	25.4
.561	4/9/2009	448.4498	0	76.82609	34.6	24.9
.562	4/10/2009	481.8188	0	66.20671	39	24.8
563	4/11/2009	448.1053	0	73.66386	35.9	25.4

Weather indicators (accumulated, average, frequency, maximum o minimum)



ID	FECHA_SIEMBRA	FECHA_COSECHA	ANO_COS	RENDIMIENTO_HA	TMAXavg	TMINavg	TEMPavg	GDaccu11	RANGO_Diurno_avg	Eneraccu
RC38_2009_5	4/5/2009	8/3/2009	2009	5600	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC38_2009_6	4/5/2009	8/3/2009	2009	5775	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC38_2009_7	4/5/2009	8/3/2009	2009	4200	33.11977441	23.67722572	28.39850006	1957.651791	9.442548692	43981.57
RC27_2013_3037	10/22/2012	2/19/2013	2013	5262	34.06942149	24.20578512	29.13760331	1880.2564	9.863636364	43883.31
RC38_2013_129	10/22/2012	2/19/2013	2013	5265	34.06942149	24.20578512	29.13760331	1880.2564	9.863636364	43883.31
RC38_2013_130	10/24/2012	2/21/2013	2013	5284	34.16363636	24.2107438	29.18719008	1873.7553	9.952892562	43962.81
RC38_2013_134	11/2/2012	3/2/2013	2013	6720	34.30661157	24.30743802	29.30702479	1862.2728	9.999173554	44100.78

Exercise with summarize, merge and weather indicators.

Use the information contained in the link below, to process fertilizer data.

- https://github.com/hdorado/Workshop_Nairobi

Compute the weather indicators for crop stage, according to the exercise planted in.

- https://github.com/hdorado/Indicadores-climaticos

Data cleaning

Check the coordinates

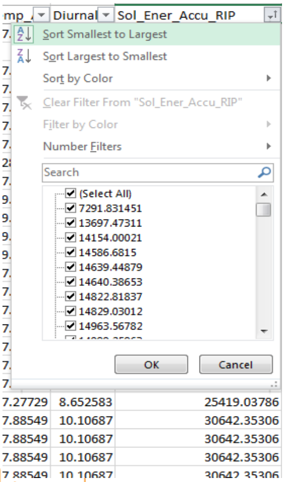
Padula
Arrozal Treinta y Tres
-33.050636, -53.6990

Useful software

Google earth
Quantum gis
Diva gis
Arc gis

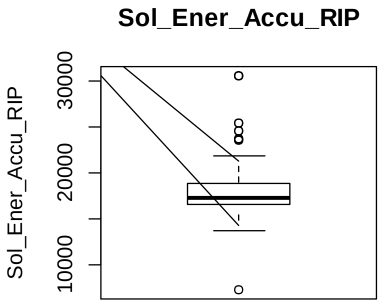
Uppercase or Lowercase

Outliers



☑ (Select All)
☑ 7291.831451
☑ 13697.47311
☑ 14154.00021
☑ 14586.6815
☑ 14639.44879
☑ 14640.38653
☑ 14822.81837
☑ 14829.03012
☑ 14963.56782

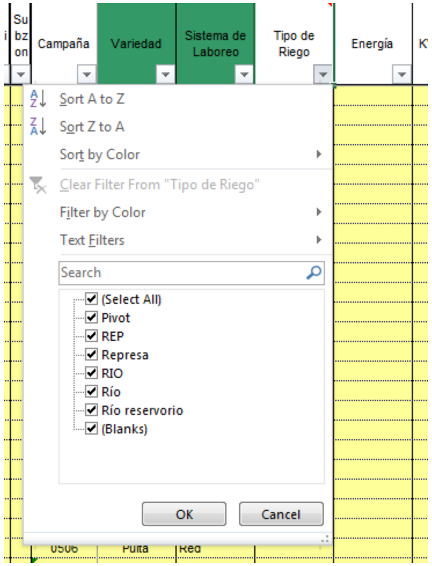
mp_	Diurnal	Sol_Ener_Accu_RIP
7.27729	8.652583	25419.03786
7.88549	10.10687	30642.35306
7.88549	10.10687	30642.35306
7.88549	10.10687	30642.35306
7.88549	10.10687	30642.35306



Boxplot

Library in R

tidyr



☑ (Select All)
☑ Pivot
☑ REP
☑ Represa
☑ RIO
☑ Río
☑ Río reservorio
☑ (Blanks)

Excel filters

Agenda

- A brief overview to R and the basic functions and graphics
- Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).
- **Training machine learning models.**
- Interpreting machine learning models outputs.
- Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)

Variables

Predictors, features

$X_1 \quad X_2 \quad X_3 \ \dots \ X_p \longrightarrow Y$

pH Soc temp Rain,…

Yield, quality, incoming

Supervised vs unsupervised

$X_1 \quad X_2 \quad X_3 \ \dots \ X_p \qquad Y$

Categorical: Classification

Numerical: Regression

Linear regression

Neural network

Supervised

$X_1 \quad X_2 \quad X_3 \ \dots \ X_p \qquad \xcancel{Y}$

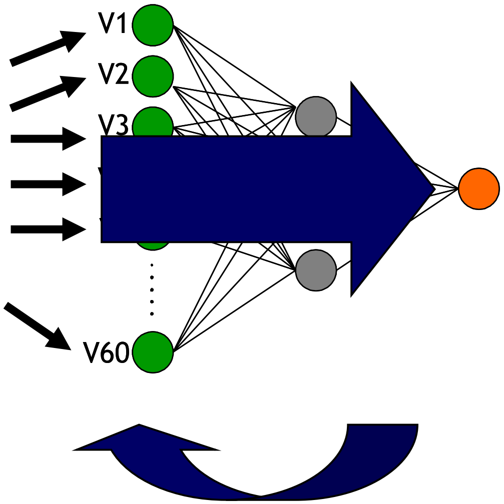
Clustering
Principal component analysis
Factorial analysis
.
.
.

Unsupervised

Neural networks (Multilayers perceptron)

	V1	V2	V3	V4	V5	...	V60	L 1	L 2	L 3	L 4	L 5	...	*Kg/lote*
Obs 1	0.1	18	3	312	0.3	...	89	0	1	0	1	0	...	2.39
Obs 2	0.2	15	4	526	0.1	...	52	1	0	0	0	1	...	30.35
Obs 3	0.6	14	1	489	0.2	...	64	0	1	1	1	1	...	42.25
Obs 4	0.05	19	2	523	0.5	...	13	0	0	0	0	1	...	52.50
Obs 5	0.4	13	3	214	0.6	...	57	1	1	1	1	1	...	
Obs 6	0.8	12	4	265	0.4	...	24	1	1	0	1	0	...	82.25
Obs 7	0.2	15	1	236	0.8	...	26	0	0	1	0	0	...	89.28
Obs 8	0.1	17	3	541	0.1	...	35	0	1	1	1	0	...	125.0
Obs9	0.6	16	2	845	0.3	...	51	0	0	1	1	0	...	142.8
Obs10	0.1	18	1	126	0.1	...	43	1	1	0	0	1	...	150.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Obs3000	0.04	15	3	235	0.6	...	85	1	1	1	1	0	...	180

Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
0.1	0.2	0.6	0.05	0.4	0.8	0.2	0.1	0.6	0.1	...	0.04
18	15	14	19	13	12	15	17	16	18	...	15
3	4	1	2	3	4	1	3	2	1	...	3
312	526	489	523	214	265	236	541	845	126	...	235
0.3	0.1	0.2	0.5	0.6	0.4	0.8	0.1	0.3	0.1	...	0.6
...	...	...	...	...	...	...	...	...	...	...	...
89	52	64	13	57	24	26	35	51	43	...	85



Predicted

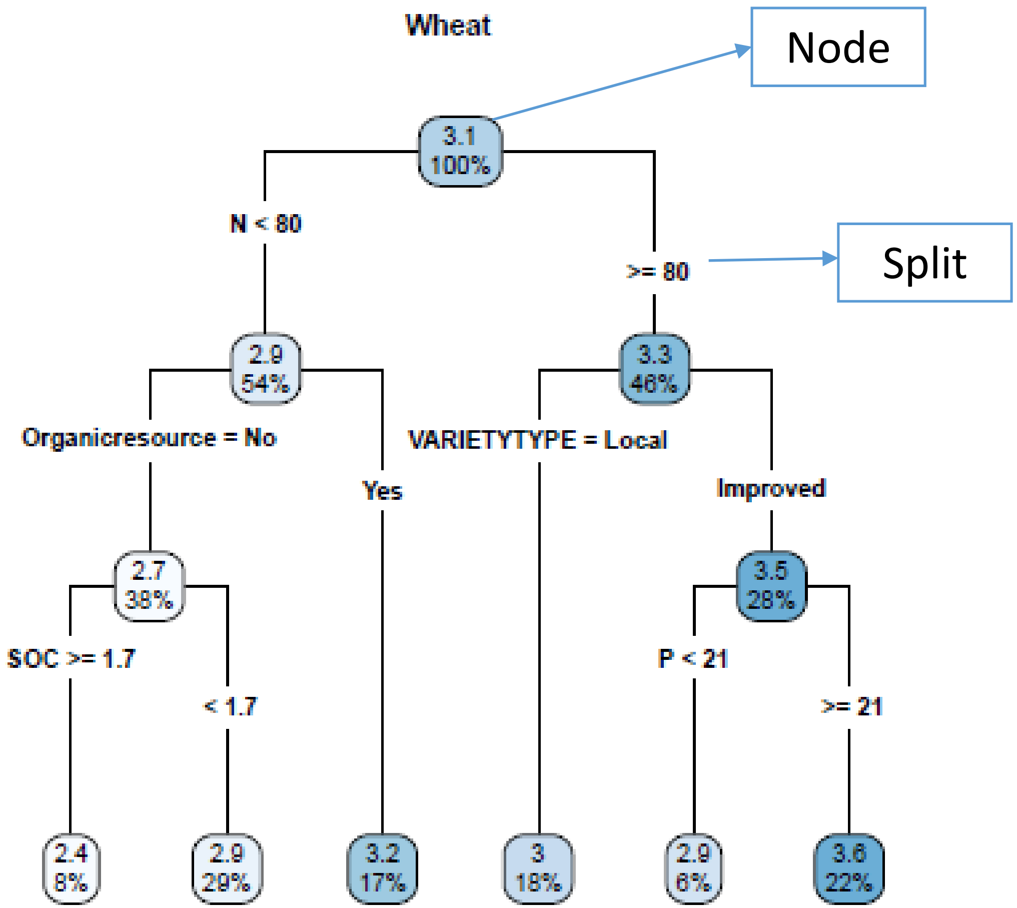
Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
2.07	29.0	53.5	50.5		89.5	99.2	120	172	170	...	188

Observed

Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	...	Obs3000
2.3	30.3	42.5	52.5		82.2	89.2	125	142	150	...	180

CART(Clasification and regression trees)

Index
Gini
information

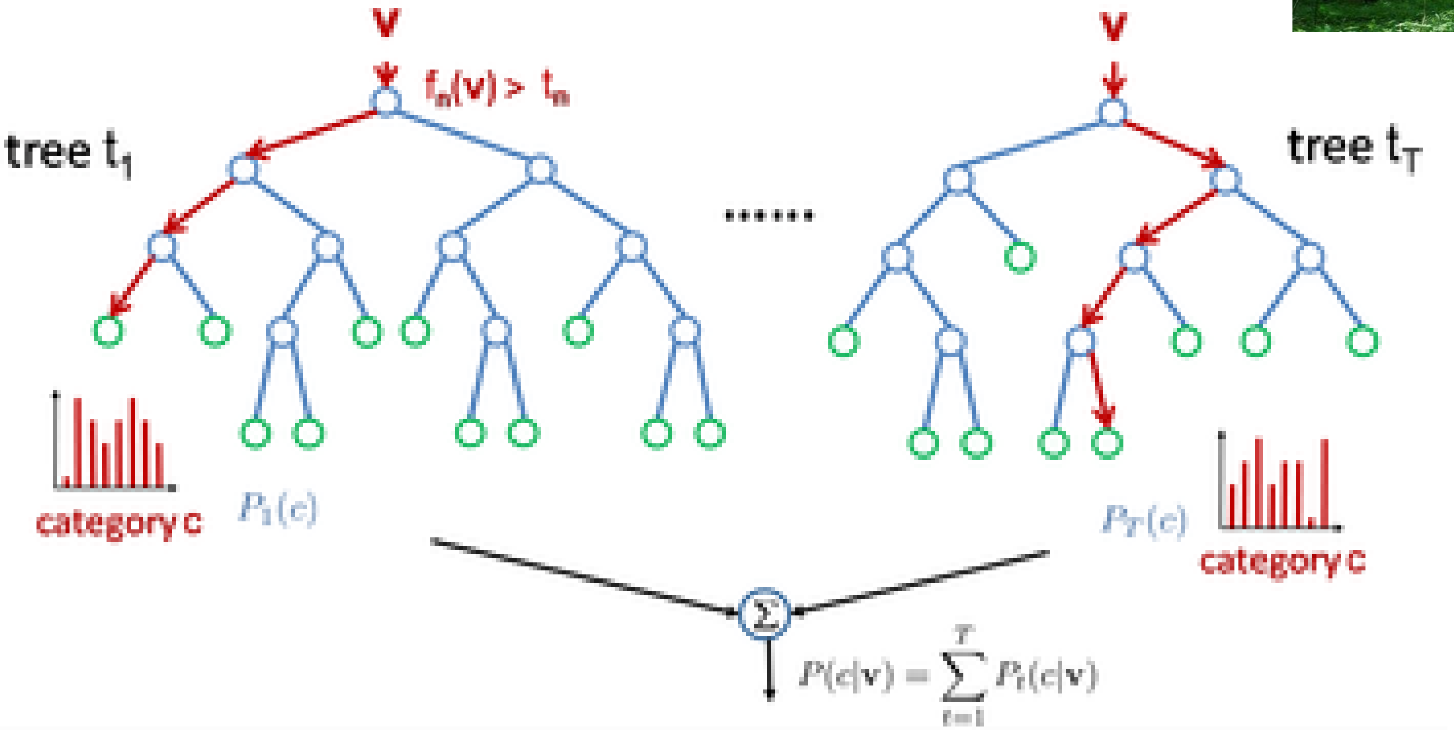


pruning

Random forest

mtry = number of variables

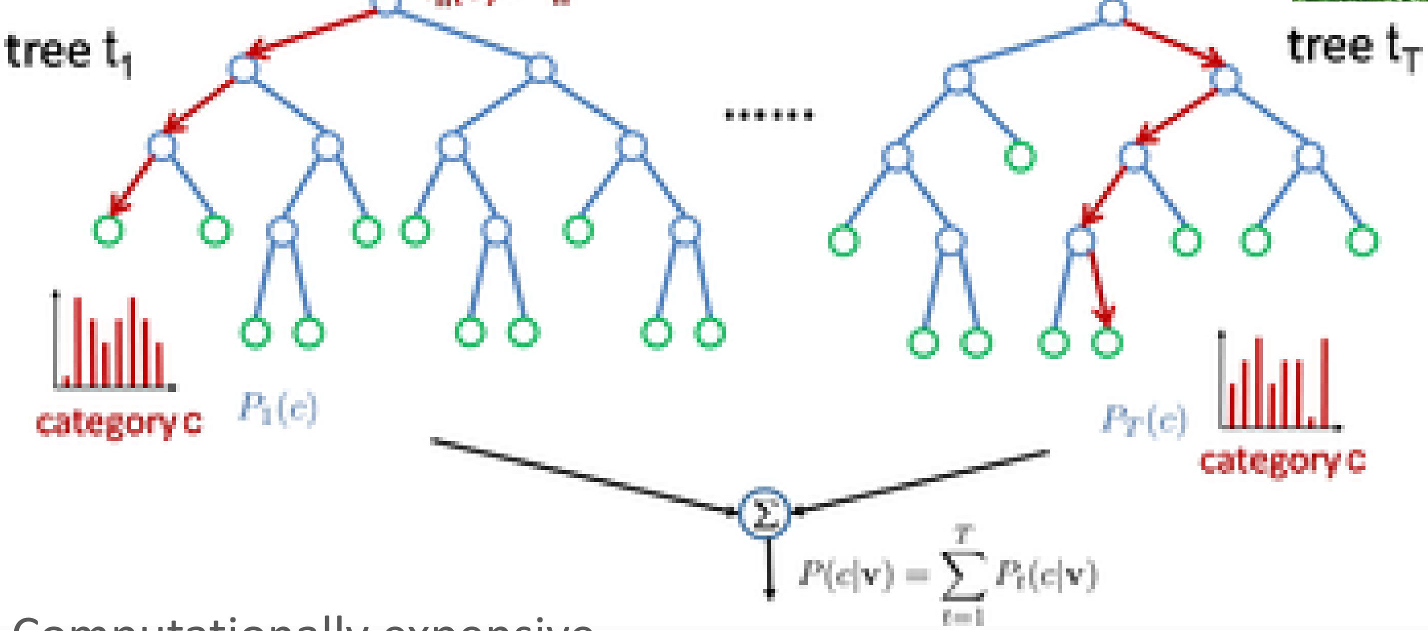
ntrees = number of tress



The split is based in gini coefficient or information index

Conditional forest

mtry = number of variables
ntrees = number of tress



The split is based in permutation tests

Computationally expensive
Reduce the random forest bias

Agenda

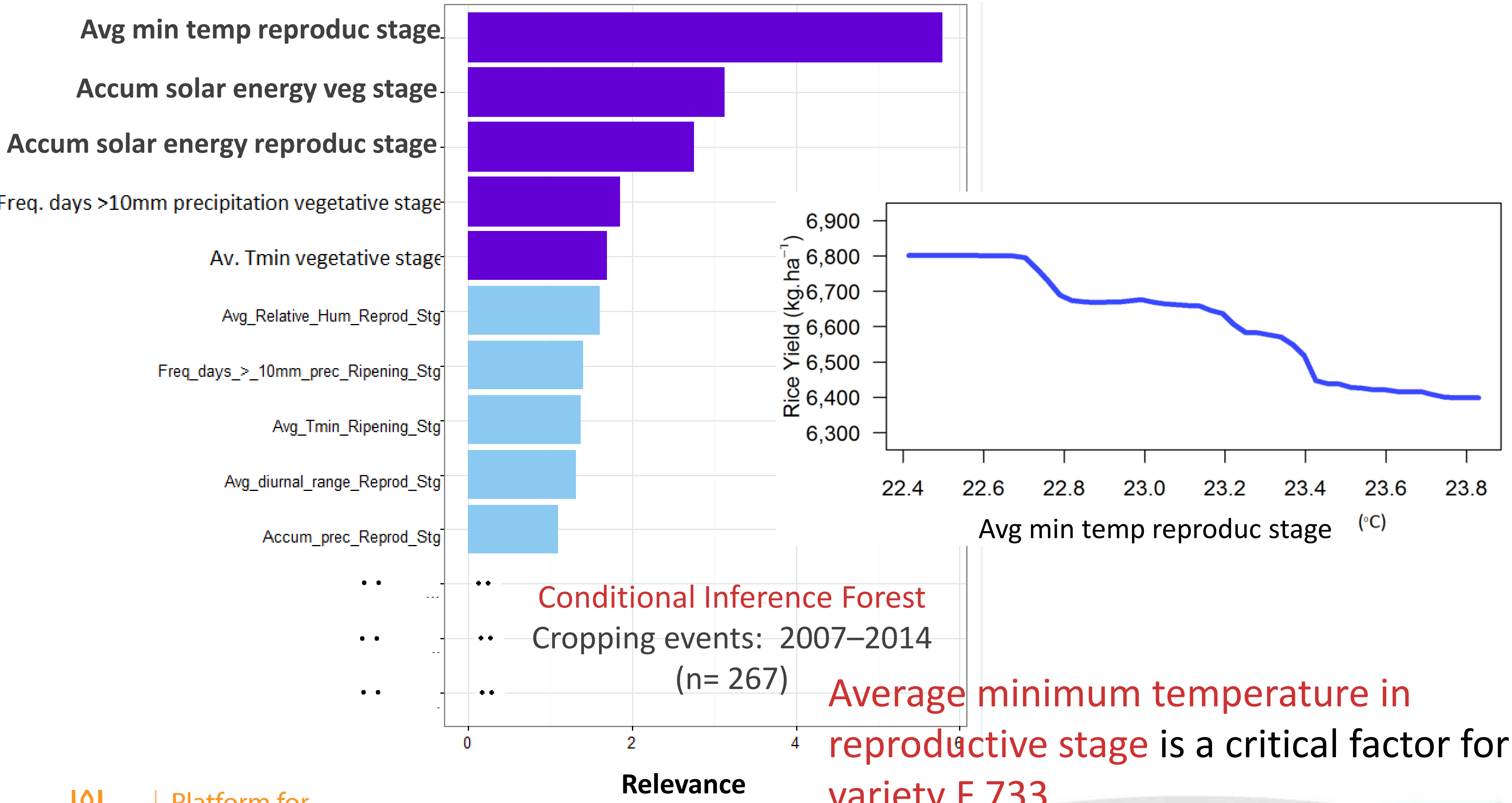
- A brief overview to R and the basic functions and graphics
- Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).
- Training machine learning models.
- **Interpreting machine learning models outputs.**
- Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)

The models outputs

Climate accounts for about 30% to production variability in irrigated rice – Variety F 733

Importance of variables

R² = 30%



Conditional Inference Forest

Cropping events: 2007–2014

(n= 267)

Average minimum temperature in reproductive stage is a critical factor for variety F 733

Agenda

- A brief overview to R and the basic functions and graphics
- Getting and processing data with a big data approach (sources, how to collect and to process weather and soil data, how to organize the data in an analyzable structure).
- Training machine learning models.
- Interpreting machine learning models outputs.
- **Practicing exercise with own data. (would be amazing if you have dataset which you want to analyze)**

CIAT

International Center for Tropical Agriculture
Since 1967 Science to cultivate change

CGIAR

A CGIAR Research Center

Headquarters
Km 17 Recta Cali-Palmira C.P. 763537
P.O. Box 6713, Cali, Colombia
Phone: +57 2 445 0000

✉ ciat@cgiar.org
www.ciat.cgiar.org

f ciat.ecoefficient

@ciat_cgiar

@CIAT_