

A single, ripe yellow banana is positioned on the left side of the frame, curving from the top left towards the center. The background is a solid, bright yellow. On the right side, there is a large, semi-transparent white circle containing text.

AgroFair

A machine
learning
algorithm to
predict crown rot
in organic
bananas

Hans-Willem van der Waal



The Problem

AgroFair
RIGHT FROM THE PRODUCER



Crown rot is a major economic problem in organic banana sea transport. It is caused by a complex of pathogenic fungi and bacteria, notably *Lasiodiplodia theobromae*, *Colletotrichum musae* and a complex of *Fusarium spp.* (Alvindia & Natsuaki, 2008).

The cost of crown rot is high, and can amount to 50% of the cargo's value and is at the expense of the producer.





Professionals in the trade know it is related to:

- longer transit times,
- **season**,
- some producers more prone than others,
- **controlled atmosphere transport**



- In conventional bananas: spraying crowns with a systemic fungicide.
- In organic bananas not allowed.
- In the past: Citrex
- Citrex disappeared: was full of quaternary ammonium compounds (QAVs).
- In fact, there is no organic, approved, effective agent available – citric acid?

(van der Waal, 2015; van der Waal & Houtman, 2016)





Does controlled atmosphere really help? Is it better than banavac, which is also a kind of modified atmosphere? *In vitro*, controlled atmosphere was effective.

But: controlled atmosphere containers are more expensive.



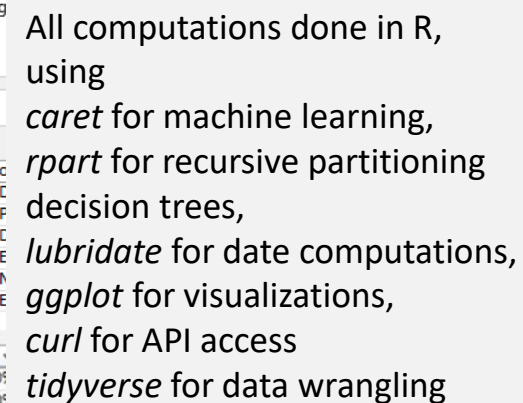
The approach

AgroFair
RIGHT FROM THE PRODUCER



Quality assessment performed on all incoming shipments (individual containers). Take 5 boxes out of 1080, count the number of clusters with medium and severe crown rot and express as percentage of total number of clusters. We have now approx. 25,000 data points since 2016.

A computer monitor displays a complex spreadsheet used for quality assessment. The spreadsheet has multiple columns and rows, with data entered in various colors (green, yellow, orange, purple). The top part of the spreadsheet contains summary statistics, including 'Total A', 'Total B', 'Total C', and 'Total D'. Below this, there are several sections for detailed data entry, including 'Physical inspection' and 'Chemical inspection'. The bottom part of the spreadsheet has a section for 'Final control (signature)' with fields for 'Inspector', 'Date', and 'Signature'. The monitor is positioned in front of the person inspecting the bananas.



llet
71759
86211
86211

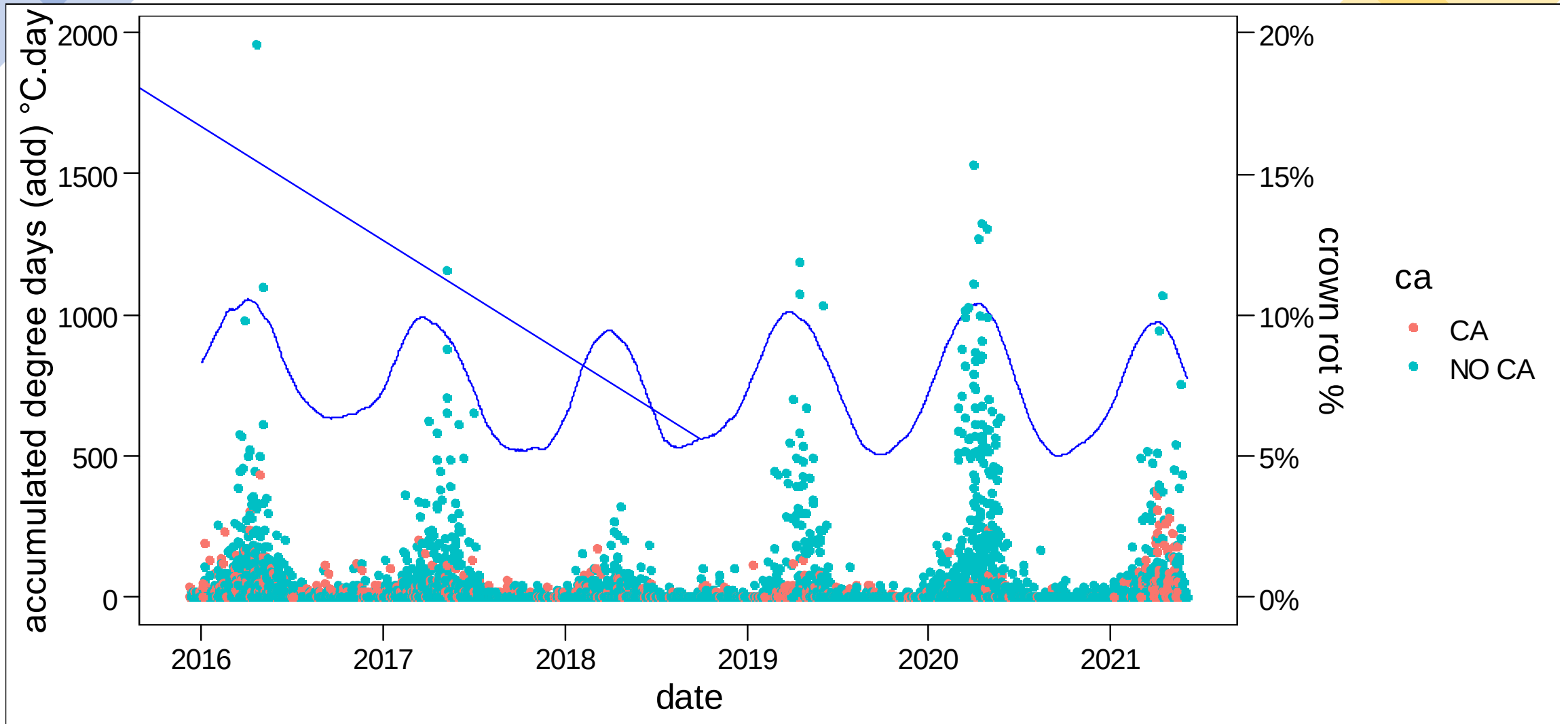
[illegible]

If Crown Rot Severity > 0.02, it is classified as “High”, else “Low”

We want to make a 2-class classification machine learning algorithm to predict Crown Rot Risk *before shipment*.

10

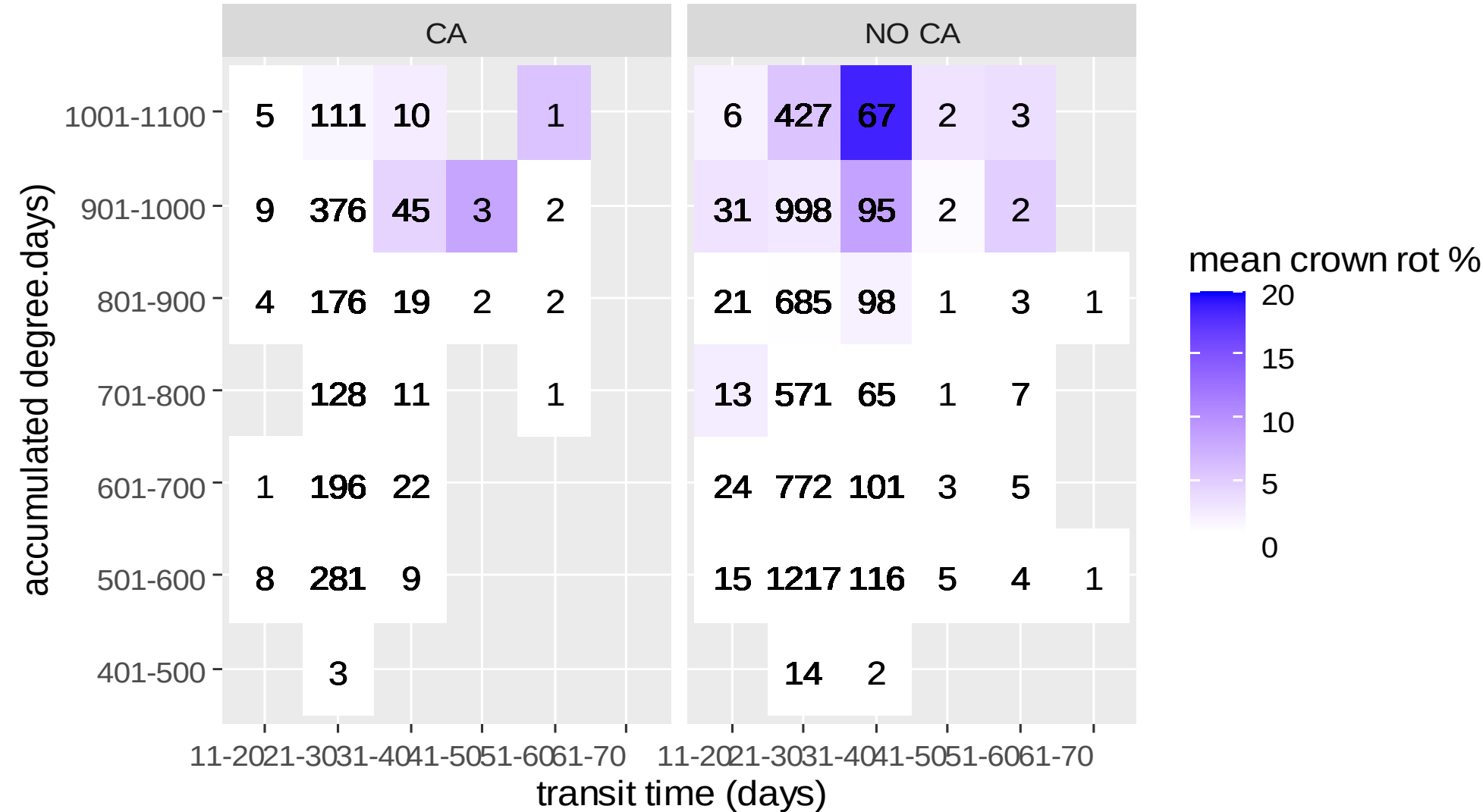
I present a simplified ML predictive model for organic bananas from Peru, based on some 3000 observations.



Accumulated degree days and Crown Rot Risk versus date, to identify a **predictive feature** for season

crown rot versus ADD and transit time

figures indicate number of shipments





The model

Transit time

Controlled atmosphere

Season ->
accumulated
degree days (11
weeks, 13.5°C).

```
#create a function for reading weather from meteostat API:

read_weather <- function(station){
  url <- paste("https://bulk.meteostat.net/hourly/",station,".csv.gz", sep= "")
  tempdir()
  tmp <- tempfile()
  curl_download(url, tmp)
  download.file(url, destfile = paste(station, ".wd", sep=""))
}

#define weather station of choice: Piura with ID 84401
station <- 84401
```

Crown Rot Risk (on arrival)



<https://meteostat.p.rapidapi.com>



Meteostat Developers

Meteostat is an open platform which provides free access to historical weather and climate data.

```
#formula to calculate the cumulative sums based on span and cut-off of 13.5°C
summ_add_PE <- summ_PE %>% filter(date >= (as.Date("2016-01-01")-span*7)) %>%
  filter(date <= (as.Date("2021-06-04"))) %>%
  mutate(grow_temp = (mean_temp - 13.5), cdd = cumsum(grow_temp)) %>% #the grow temperature
  #is the temperature above 13.5°C under which temperature bananas do not grow but die
  mutate(add = cdd - lag(cdd, span*7)) %>% # the accumulated degree.days or temperature sum
  # is calculated for every day from the cumulative grow_temp
  na.exclude() #exclude rows with NaN
```

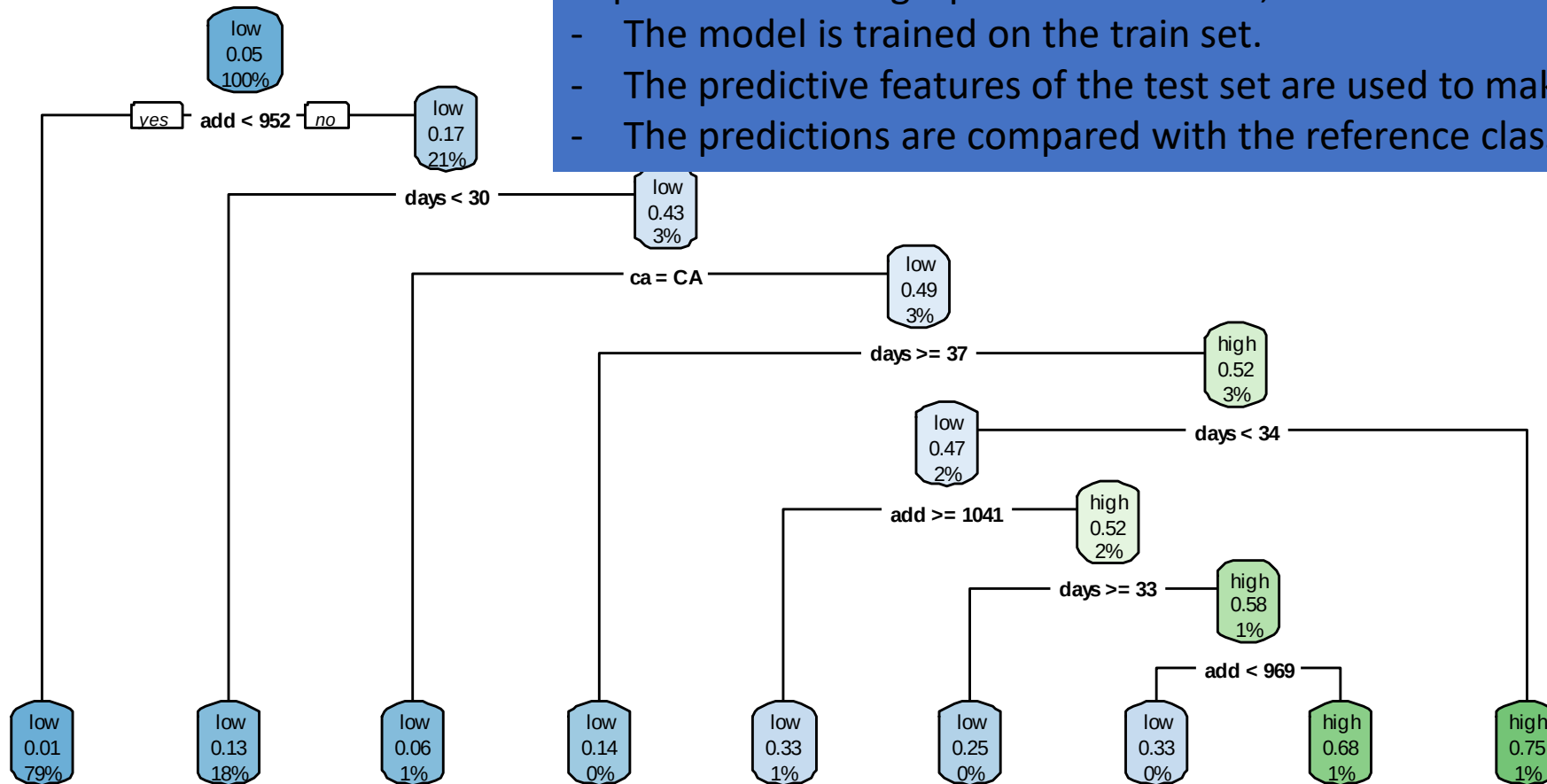
I evaluated different common machine learning algorithms suitable for classification problems included in the *caret* package.

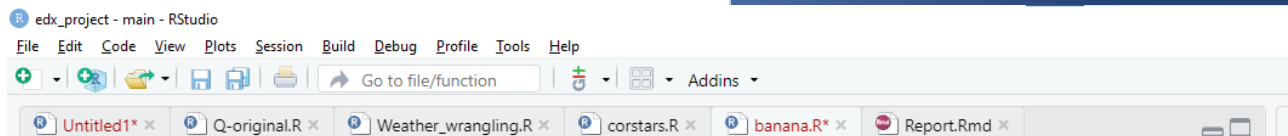
To visualize, here is the recursive partitioning result.

“Recursive partitioning creates a [decision tree](#) that strives to correctly classify members of the population by splitting it into sub-populations based on several dichotomous [independent variables](#).”

Supervised learning: split dataset in two, a train set and a test set.

- The model is trained on the train set.
- The predictive features of the test set are used to make a prediction of class.
- The predictions are compared with the reference class.





The model is good at correctly predicting “Low” CR risk, but bad at predicting “High” CR risk. This is due to unbalanced data: many low observations, few high. *Specificity* is high, but *Sensitivity* is low (is CR risk is positive class). This was resolved by a technique known as *synthetic sampling*. This allows to achieve better sensitivity. Adding more predictive features, such as individual producer, fruit grade at harvest, other climate features, can further improve the model.

```

Reference
prediction low high
low 2958 122
high 17 29

```

```

Accuracy : 0.9555
95% CI : (0.9477, 0.9625)
No Information Rate : 0.9517
P-Value [Acc > NIR] : 0.1689

```

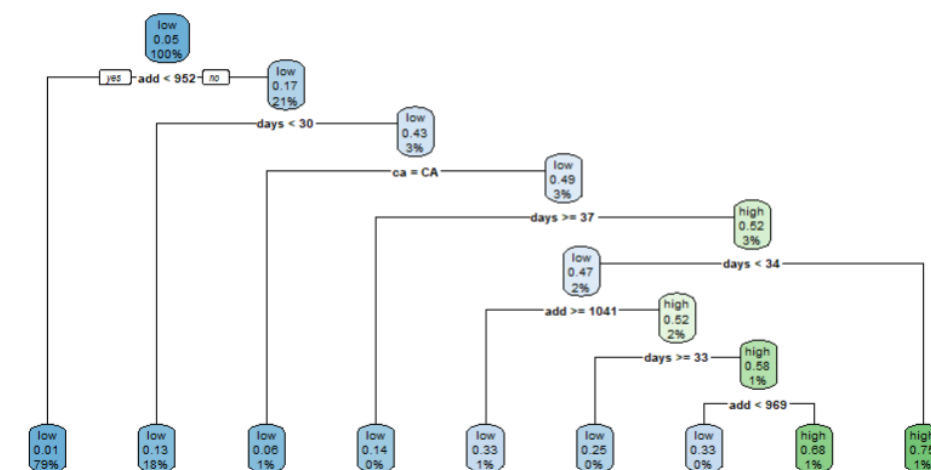
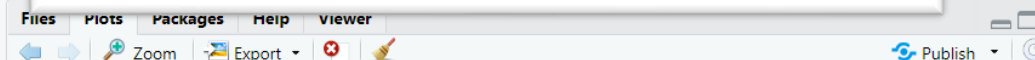
```
Kappa : 0.2781
```

```
McNemar's Test P-Value : <2e-16
```

```

Sensitivity : 0.9943
Specificity : 0.1921
Pos Pred Value : 0.9604

```



278:1 (Top Level) ↕ R Script ↕

Reference		Prediction	
	low	high	
low	2958	122	
high	17	29	

Confusion matrix		
TP	FP	
FN	TN	
Sensitivity/recall	TP/(TP+FN)	0.99
Specificity	TN/(TN+FP)	0.19
Precision	TP/(TP+FP)	0.96


```

Accuracy : 0.9555
95% CI : (0.9477, 0.9625)
No Information Rate : 0.9517
P-Value [Acc > NIR] : 0.1689

Kappa : 0.2781

McNemar's Test P-Value : <2e-16

Sensitivity : 0.9943
Specificity : 0.1921
Pos Pred Value : 0.9604

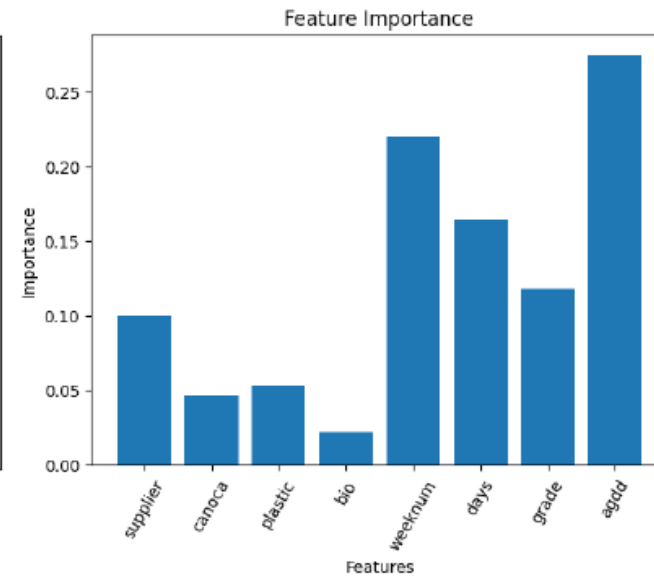
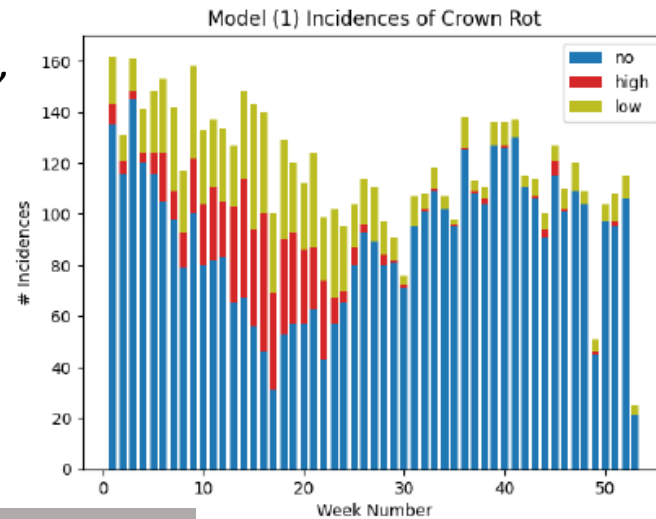
```

We deployed a production model, produced by Dr. Theofrastos Mantadelis, for all our organic bananas, from different countries, based on the same basic idea, but with more predivitve features.

52871	AGRONOR	True	23-2022	45 days	43"	906		BANAVAC	POLYPACK
40749	APPBOSA	False	25-2020	31 days	44"	904	NO CA	no	no
	plastic: BANAVAC	canoca: NO CA		result: no			CA		no
52872	AGRONOR	True	23-2022	45 days	42"	906		BANAVAC	POLYPACK
40749	APPBOSA	False	25-2020	31 days	43"	905	NO CA	no	no
	plastic: BANAVAC	canoca: NO CA		result: low			CA		no
52873	AGRONOR	True	23-2022	45 days	42"	910		BANAVAC	POLYPACK
40749	APPBOSA	False	25-2020	31 days	42"	909	NO CA	high	no
	plastic: BANAVAC	canoca: NO CA		result: no			CA		no
52866	APPBOSA	True	23-2022	24 days	42"	910		BANAVAC	POLYPACK
18101	APPBOSA	True	24-2016	23 days	42"	909	NO CA	high	no
	plastic: BANAVAC	canoca: NO CA		result: no			CA		no
52867	APPBOSA	True	23-2022	19 days	42"	906		BANAVAC	POLYPACK
18101	APPBOSA	True	24-2016	23 days	42"	909	NO CA	high	no
	plastic: BANAVAC	canoca: NO CA		result: no			CA		no
52870	APPBOSA	True	23-2022	24 days	42"	910		BANAVAC	POLYPACK
18101	APPBOSA	True	24-2016	23 days	42"	909	NO CA	high	no
	plastic: BANAVAC	canoca: NO CA		result: no			CA		no
52851	DEFAULTPERU	True	23-2022	25 days	43"	906		BANAVAC	POLYPACK
35123	CEPROBAN	True	27-2019	25 days	43"	904	NO CA	no	no
	plastic: POLYPACK	canoca: CA		result: no			CA		no
52852	DEFAULTPERU	True	23-2022	25 days	43"	906		BANAVAC	POLYPACK
35123	CEPROBAN	True	27-2019	25 days	43"	904	NO CA	no	no
	plastic: POLYPACK	canoca: CA		result: no			CA		no
52903	DEFAULTPERU	True	23-2022	25 days	43"	906		BANAVAC	POLYPACK
35123	CEPROBAN	True	27-2019	25 days	43"	904	NO CA	no	no
	plastic: POLYPACK	canoca: CA		result: no			CA		no
52904	DEFAULTPERU	True	23-2022	25 days	43"	906		BANAVAC	POLYPACK

Model Report - Peru

Production application programmed in *Python*, using a *random forest* algorithm (multiple decision trees).



The performance with more features

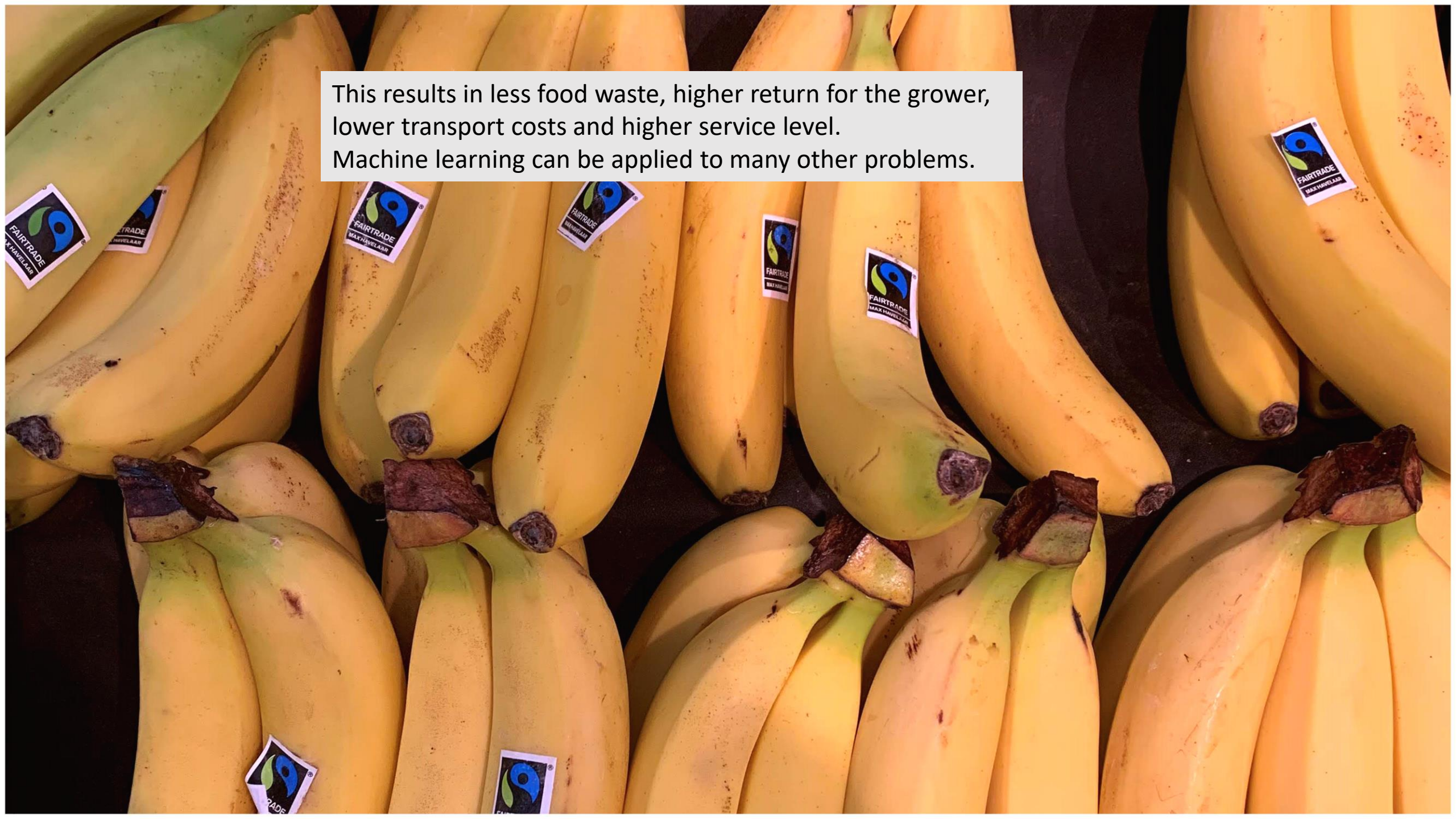
Accuracy: 0.92

	precision	recall	f1-score	support
high	0.85	0.86	0.86	594
low	0.81	0.76	0.79	926
no	0.95	0.96	0.95	4736
accuracy			0.92	6256
macro avg	0.87	0.86	0.87	6256
weighted avg	0.92	0.92	0.92	6256

Label	TP	TN	FP	FN
high	512	5249	90 = 32/low+58/no	82 = 29/low+53/no
low	704	5057	162 = 29/high+133/no	222 = 32/high+190/no
no	4545	1216	243 = 53/high+190/low	191 = 58/high+133/low



The model allows us to choose for controlled atmosphere to reduce the Crown Rot risk, *before shipment*, counting with the expected transit time based on the shipping line schedules.



This results in less food waste, higher return for the grower,
lower transport costs and higher service level.
Machine learning can be applied to many other problems.