Facts

1. **Tengo 38 variables (dataset completo)** que se recogen periódicamente.

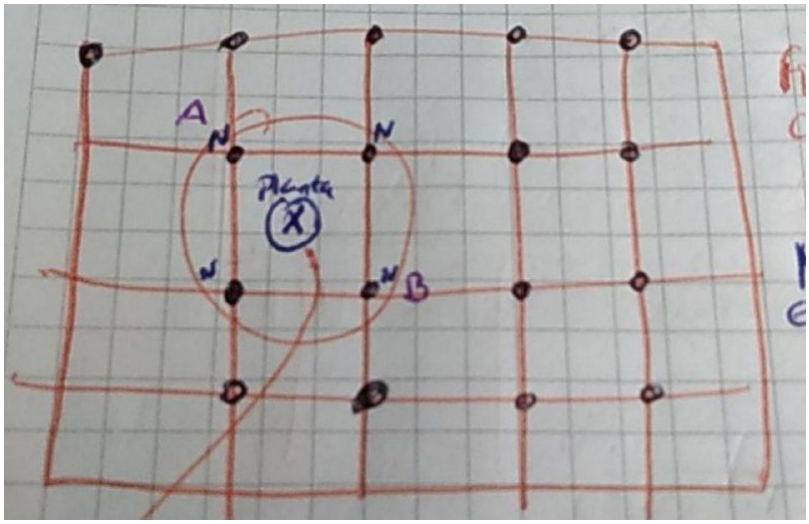
- **Estación meteorológica**
  Luminosidad, Precipitaciones, Dirección del viento (8 variables), Velocidad del viento, Temperatura.
  1 estación por cada finca, cada media hora emiten datos

- **Nodos hídricos**
  Nivel Freático (25 variables)
  Cada lote de las fincas tiene uno o más nodos para medir el NF.
  Lote 2 - 2 nodos
  Lote 16 - 2 nodos
  Lote 18 - 4 nodos
  Lote 21 - 5 nodos
  Lote 24 - 2 nodos
  Lotes 4,5,6,7,8,9,10,13,14,15 - 1 nodo cada uno

  La finca Porvenir tiene 25 nodos hídricos repartidos en sus lotes. Cada hora emiten datos

- **Barcadilla Báscula**
  Peso del racimo (kg)

Éstas variables están distribuidas a lo largo de la finca en relación a sus respectivos lotes



2. EDAD DE LA PLANTA DE BANANO - DATOS

- **Tiempo de vida de la planta de banano (desde que es sembrada hasta que es removida):** 28 semanas, antes de ser cosechada. Cuando cumplen las 28 semanas, se cosecha el racimo y se remueve la planta

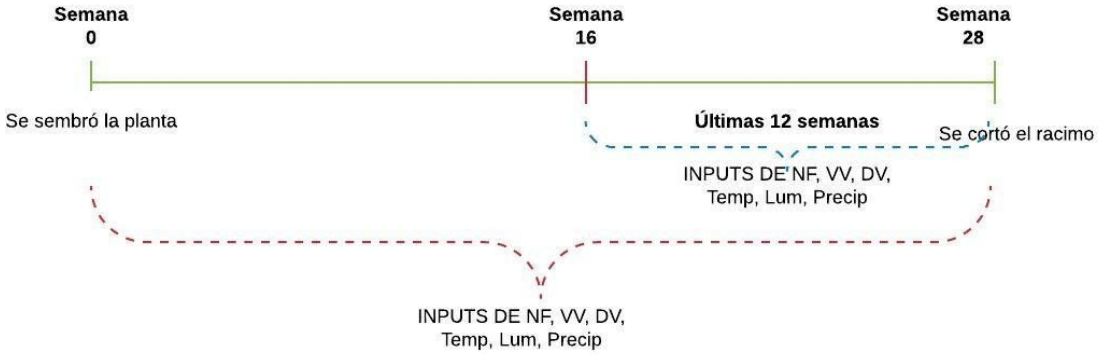
- **Tiempo en que la planta produce el racimo de banano:**

Las últimas 12 semanas de esas 28 semanas de vida, (es decir desde las 16 semanas en adelante), es el tiempo que toma la planta para producir racimos de banano.

Se toma por completo las 12 semanas, por lo que este lapso es el tiempo ideal para recolectar la fruta.

Aunque también se realiza recolección desde las últimas 8,9,10 u 11 semanas de vida de la planta cumplidas, vamos a tomar el número de 12 semanas de edad (las últimas de las 28 semanas) como el tiempo ideal de cortado del racimo

Acorde a la edad de la planta



- La productividad de recolección es semanal.Se toman datos de peso de racimos recolectados por lotes en 1 semana
- Se retrocede 12 semanas a partir de esa semana de datos de peso recolectado y se toman datos de Estacion meteorológica y Nodos hídricos.
- Se retrocede también 28 semanas a partir de esa semana de datos de peso recolectado y se toman datos de Estacion meteorológica y Nodos hídricos

12 Semanas cada hora --- 12 semanas(7 días) * 24 horas = 2016 datos
2 Semanas cada media hora -- 4032 datos
28 Semanas cada hora --- 28 semanas(7 días) * 24 horas = 4074 datos
2 Semanas cada media hora -- 9408 datos

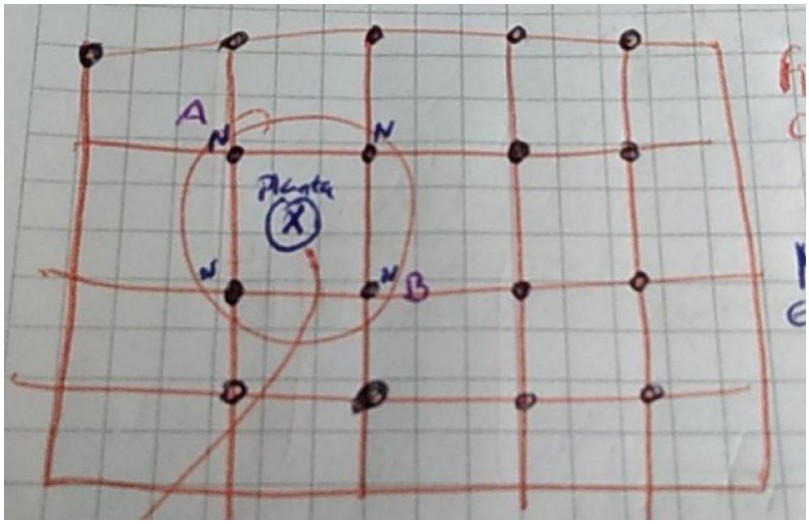
3. **Los features de NIVEL FREÁTICO en el dataset son variables acorde a los lotes de cada finca.**

En cada lote de cada finca hay nodos hídricos que miden el NF para cada lote o porciones de cada lote.
Para sacar un modelo que se pueda aplicar de manera estándar a cada finca, es necesario estandarizar o reducir esta dimensionalidad de features en NF

Tomemos los datos de las 28 semanas que es el tiempo de vida total de la finca

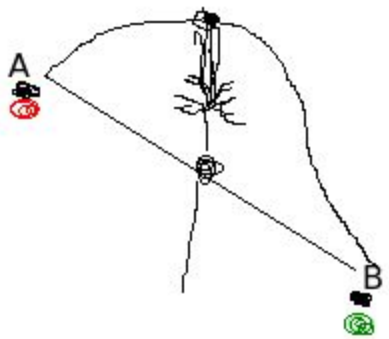
28 Semanas cada hora --- 28 semanas(7 días) * 24 horas = 4074 datos
2 Semanas cada media hora -- 9408 datos

Esto quiere decir que una planta estuvo creciendo durante 4074 mediciones.



- El valor de NF puede ser un promedio geográfico de estos 4 valores de nodos que rodean la planta. Lo mejor que puedo hacer es una combinación lineal de los 4 nodos que rodean a la planta, combinar las mediciones de los 4 nodos para tener el valor del NF
- No conozco el comportamiento de los nodos de los sensores hídricos, podría ser Gaussiano

Tengo un punto A y B
Rojo Nivel Freatico A
Verde Nivel Freatico B

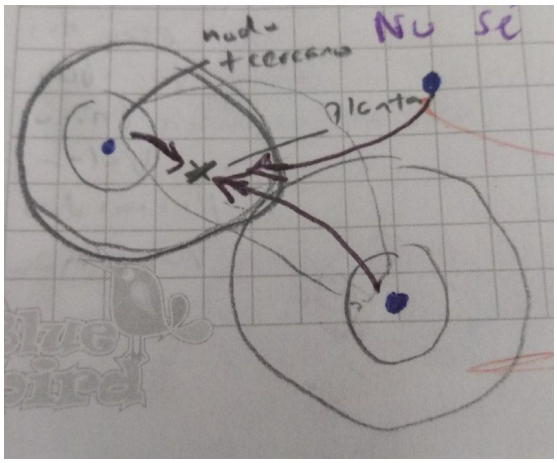


Si hay una planta entre los dos, ¿Cuál es el Nivel freático?
¿O un nodo importa más que otro?.
Si un racimo que se corta, tiene influencia de 15 nodos hídricos. …

No se sabe, entonces el mejor supuesto es una combinación lineal, es decir, espacial

Imaginemos un radar,

- El nodo más cercano tiene mayor influencia que los otros dos.



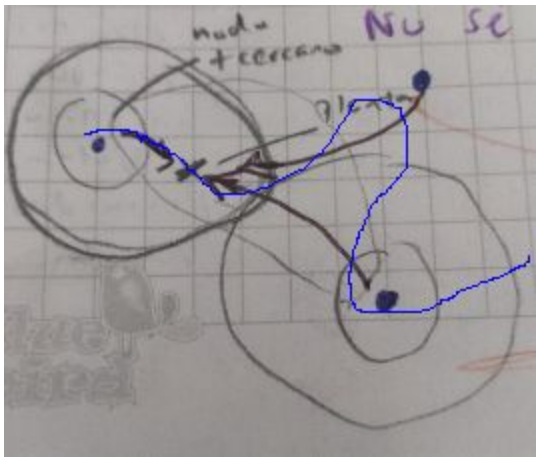
- Si los tres nodos son cercanos, no habría problema, el valor del NF sería el promedio entre ellos

- Pero si quiero ser más exacto, debo tener en cuenta la distancia entre los nodos para promediar, es como un promedio ponderado, donde la distancia más cercanas pesa más que las distancias más lejanas

Hay varias formas de asignar esa distancia o el valor del Nivel Freático:

- Esa planta ¿que valor de NF tiene?
  Solución 1: El valor de NF del nodo más cercano. Esto funciona.

  Pero en caso de que exista mucha disparidad entre ciertos valores de los nodos

  Entonces probablemente el valor a tomar no sea el del nodo más cercano, porque puede variar por debajo de la tierra

El mejor supuesto que puedo hacer es una combinación lineal entre los nodos más cercanos

$\alpha 1A + \alpha 2B$ ¿Que tanto pesan alfa 1 y alfa 2?
Depende de la distancia y no tengo claro eso. Entonces una solución computacionalmente sencilla y sin problemas es el nodo asociado a ese árbol

- **Solución 2**

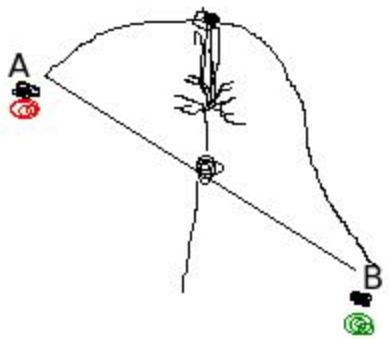
Promedio ponderado de los k nodos más cercanos. ¿Qué impliocca?

- Determinar que nodos son los más cercanos a las plantas.
- El que está más cerquita, lo debería influenciar más que los que están más lejos

Esta solución 2 no gusta mucho porque se tienen muchas variables para jugar:

- ¿Cuantos nodos?
- ¿Cómo hago esa ponderación?
  - 1/distancia
  - 1/distancia al cuadrado
- ¿Por qué ponderé asi? Por que el NF es lineal … ¿lo es? Hay que consultar literatura y expertos

Si mido un NF en rojo y verde A y B, ¿puedo asegurar que en la mitad de la interpolación, es lineal?



Puede que no sea asi
Entonces todos esos problemas o cosas de las cuales debo asegurarme, me los ahorra el nodo más cercano, que siempre es una solución práctica
Si un racimo que se corta, tiene influencia de 15 nodos hídricos. …

4. **LOS DATOS DE ESTACIONES METEOROLÓGICAS SON LO MISMO POR CADA FINCA. ES DECIR LOS MISMOS DATOS PARA TODOS LOS LOTES DE UNA FINCA**

1 estación por finca hace que todas las plantas reciban los mismos valores de las estaciones meteorológicas

- Geográficamente, la información meteorológica me aporta información es de toda la finca.
- Es decir que no puedo diferenciar las condiciones de cada uno de los lotes por los datos de la estación meteorológica
- Los lotes se diferencian por las variables que vienen de los nodos hídricos como el Nivel freático, que es el único que varía por lote.

Entonces, si fuera a realizar una regresión lineal de este estilo:

$y = \alpha X + \beta$ teniendo en las X como mis variables, entonces un nodo hídrico puede hacer la diferencia para todas las plantaciones de la finca. Pero los valores de las estaciones meteorologicas son constantes, todas las plantas recibieron el mismo valor, por lo que hace que no sean relevantes a la hora de hacer una diferencia por lotes.

Serviría tomar en cuenta la info de las estaciones meteorológicas si el análisis se hiciera entre fincas o datos de N fincas, ahí si serviría los datos de las estaciones meteorológicas para diferenciar la productividad

Pero no puedo predecir pretender la productividad entre un lote y otro, con las variables meteorológicas. Sí con los nodos hídricos

- Diferencia entre lotes, es explicada con datos de los nodos hídricos
- Diferencia entre las fincas es explicada con los datos de nodos hídricos + estaciones meteorológicas

Si mi análisis es entre lotes, no me sirve las estaciones meteorológicas y puedo acortar sus variables.

Si analizo solo una finca, LOS DATOS DE las estaciones son constantes y al clasificador no le importa porque el solo cogoe todos esos inputs y los mete en un Beta, digamos una regresión lineal, si no tengo la medida de las estaciones, él la recoge en un beta siempre y cuado hablemos de que esto fue constante en un tiempo t pero no es cierto.

LA PRIMERA VARIABLE DE PARTICIÓN ES LA GEOGRAFÍA y la segunda el TIEMPO

Las estaciones meteorológicas no me aportan nada geográficamente porque tengo solo 1, pero los nodos hídricos si.

- Para un tiempo t dado, las variables de la estación no sirven.
- Pero en el tiempo si sirven.

Entonces, tengo dos particiones, tengo que partir en cuadritos mi finca y pensar esto como un video, en el tiempo

Mi problema es que tengo 4704 datos para un racimo, entonces que voy a hacer para indicar la productividad.

El segundo problema es una serie de tiempo.
Geográficamente ya solucionamos con el nodo más cercano, pa ese valor, y ahí tengo un valor, Pero el problema es que vamos a partirlo en el lote 1-1

Lote 1--1 : Asumimos que los valores son en el nodo más cercano y ese lote 1-1 va a tener una serie de tiempo de 28 semanas con datos horarios donde el NF …

Entonces que hago para decir: Hey racimo, fuiste grande porque esta serie es de esta forma o no.
¿Ese tiempo de 28 semanas cómo lo resumo para un clasificador sencillo?

¿Con el máximo?¿Con el promedio? ¿Con el mínimo?
Todo eso tengo que explorarlo

¿O con la serie completa de los datos?
No puedo perder información. Si tengo la serie completa de los datos, ¿Cómo voy a clasificar esto?

Ahí si ya no tengo 38 variables, sino que cada variable se me convierte en 4704 variables porque si tengo en cuenta toda la serie, a la regresión le voy a tener que meter:

- Promedio del día 1: Digamos 28 semanas por 7
- Promedio del dia 2
- .
- .
- Promedio del día n

Cada variable se me convierte en 4704 variables

Entonces, tengo muchos datos ¿Cómo los resumo?
Entonces, la serie completa de los datos importa? Puede que si.

Si este racimo fue super grande, quiere decir que recibir las primeras 4 semanas es bueno y después recibir un poco más y por ultimo un poco hace que sea más grande …

Entonces tengo mil formas de jugar ahí, y mirar en que parte está influyendo el tamaño de ese racimo

Por eso la recomendación, es que como tengo tantos datos para jugar con las variables y el tiempo, para que me dé futa grande y fruta pequeña, ¿qué influye?
¿Cual de las 38 variables me influye?
¿Dentro de esas variables me influye la temporalidad?
¿Influye que estos valores sean en ciertos tiempos o no?

ENTONCES, si no uso un promedio, se vuelve complejo, puede subir y bajar.

Se puede llegar a pensar:
La etapa de vida más importante del plátano son las primeras 4 semanas,, el NF tiene que mantenerse constante o mantenerse subiendo. Pregunta pa un experto.

Investigar,en artículos, todas esas variables como influyen en el plátano, que expermentos se han hecho:

- Se sientan con el banano y miran las primeras dos semanas ue tanto esta creciendo en promedio ese platano, lo van midiendo y haciendo experimentos de manera que digan si o no …
- O si lo esencial son las primeras 8 semanas de vida, entoinces las otras 20 no importa como estén

Pero todo eso está en la literatura

Entonces, el problema es complejo, pero se puede hacer tan sencillo como quiera, empecemos por coger el promedio de esta baina.
Promedio de las ultimas 28 semanas y esto se me vuelve una sola variable y ya.(El Nivel Freático)

Para este plátano, para las 38 variables, saco el promedio de las últimas 28 semanas y ya.

Si eso me clasifica, quiere decir que el promedio que tuvo durante ese tiempo no importa, pero lo que se sospecha es que importa más que en otras.

5. CLASIFICACIÓN PENSANDO EN CATEGORIAS

Clasifiquemos en gande y pequeña, con las variables promedio de todos oa tener nada más 38 variables y miremos si algo se puede hacer.
Y si ninguna variable no hace nada, entremosle una variable la que haga un poquito más y la parto en periodos digamos mensuales y miro graficamente esos meses como fueron y miro si se puede decir vea:
En el mes m-1 semana 28 el promedio de cada uno de esos meses se afecto asi:
Mes 1 - el promedio fue este
Mes 2 - el promedio fue este
Mes 3 - el promedio fue este

Entonces, EMPEZAR super sencillo que es clasificación binaria, promedio de todas las variables, pero creo que no encontraría ninguna clasificación,
De pronto un modelo sencillo, más el 50 % si pero que me dara el 60% de clasificación bien·

Promedio de mis 38 VARAIBLES, tengo una muestra de cada una.

1 dato de cada variable, pero igual son muchas, son 30, y esto me dá una respuesta y esto lo tengo por cada racimo.

Tengo suficientes datos, pero adicionalmente tengo 4704 datos de cada una de las 30 variables

Problema más sencillo --- CLASIFICACIÓN

- 1 DATO POR variable , empiezo y le tiro Naive Bayes, todo lo que pueda,
- Lo miro graficamente, retiro cada variable y con eso me entretengo 1 mes.
  Y cuando el poder de predicción del mejor MOL que encuentre me de 65% que es lo que se estima, ahí puede hasta ahí llegar mi tesis.

  Para aumentar de un 65% a un 70% o 75% tengo que entrar a mirar la temporalidad, del impacto de las variables en cada semana.

  Habran momentos, le pregunto a expertos, hasta la semana 10, es importante que el NF se mantenga a X nivel ….

Y ahí analizar este tipo de cosas en mis datos, y puede haber variables dummies por ejemplo:

En lugar de coger la temporalidad, si un experto me dice: "Las primeras 10 semanas son importantes de que se mantenga el NF, en este valor de más o menos x,"

Entonces puedo hacer una variable dummy que sea Si Y NO.

Si se mantuvo el nivel freático
No se mantuvo el nivel freático

Y esto me ayuda a clasificar mejor.

- Naive Bayes
- Regresión logística
- SVM
- Neural Networks

Mi problema es que tengo demasiados datos, y no sé que hacer con ellos.
¿Cuál es la primera aproximación?
Necesito las 28 semanas, las resumo en 1 promedio.
Arranquemos con algo. Miremos si eso me permite hacer algo con esas variables

Porque tengo mucha dimensionalidad, 38 variables, son 38 dimensiones en el espacio que deben ser separadas de alguna forma o manera. Solo con eso me dará, pero después le empiezo a hacer preguntas a un experto.

6. YA COMO ANALISTA

Revisar todas las racimos grandes si se le mantuvieron estas semanas

Primeras 10 semanas, se mantuvo el NF?
Si?
No?

Digamos que a todas las frutas grandes, se les mantuvo y ese es un clasificador, que me va a mejorar ese 65% y me va a pasar a otro nivel.
Si quiero seguir mejorando el algoritmo tengo que seguir refinando mis datos.
Este es el trabajo de un data analisis conocer mis datos hasta el fondo,
¿En verdad esto que significa?
¿Alguien ha hecho un experimento de esta variable cimi afecta el crecimiento del plátano?
Buscar en literatura, Si? Entonces le meto variables nuevas.

10 varaibles uyo
10 pablo
10 margi
10 julio
10 john