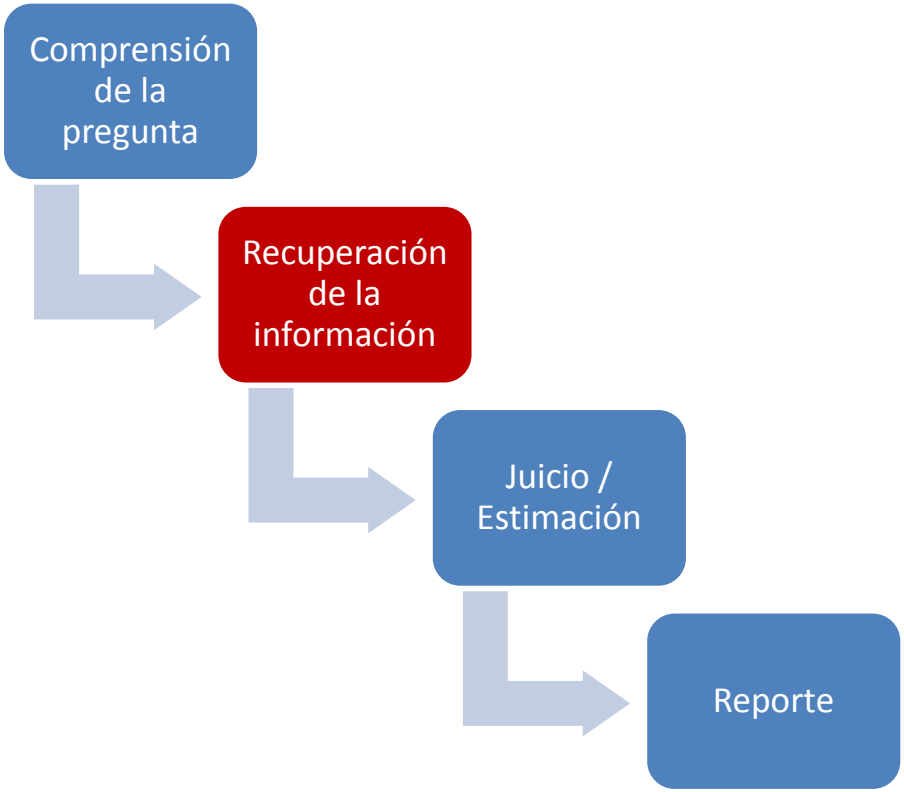
# Estimación electoral con modelos predictivos (*Random Forest*)

Jorge Buendía
Javier Márquez

b&l | buendía&laredo

## Modelo del proceso de respuesta a las encuestas
(Tourangeau y Bradburn)




Asumimos que el individuo usará toda la información relevante para contestar la pregunta de intención de voto.

***No es necesario asumirlo. Se puede verificar empíricamente***

Fuente: Tourangeau y Bradburn (2010) The Psychology of Survey Response.

## Preferencia electoral como variable latente

- Hoy: Estimación de preferencia basada en un solo indicador
- Se pierde información relevante ***que está disponible***
- Información que sabemos está correlacionada con la intención de voto (evaluación economía, partidismo, aprobación)

## Los retos de la estimación electoral

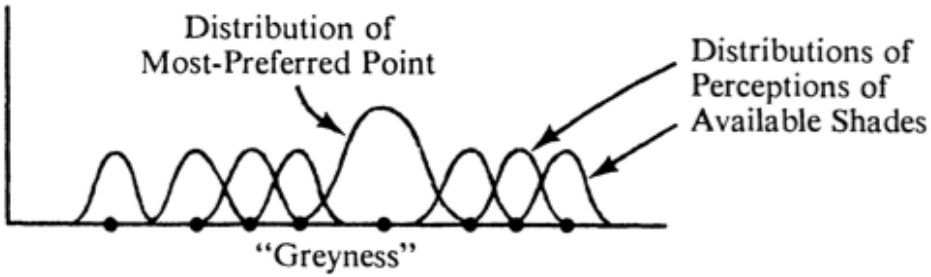
**1) El problema de los *decisos* (Falsos positivos)**

Por ejemplo, sobreestimación del PRI en encuestas preelectorales y de salida

**2) La incertidumbre electoral**

Entrevistados que expresan una preferencia electoral, pero su preferencia tiene un alto grado de incertidumbre.

Las razones pueden ser diversas: *wedge issues*, les gusta el partido pero no el candidato, etc.



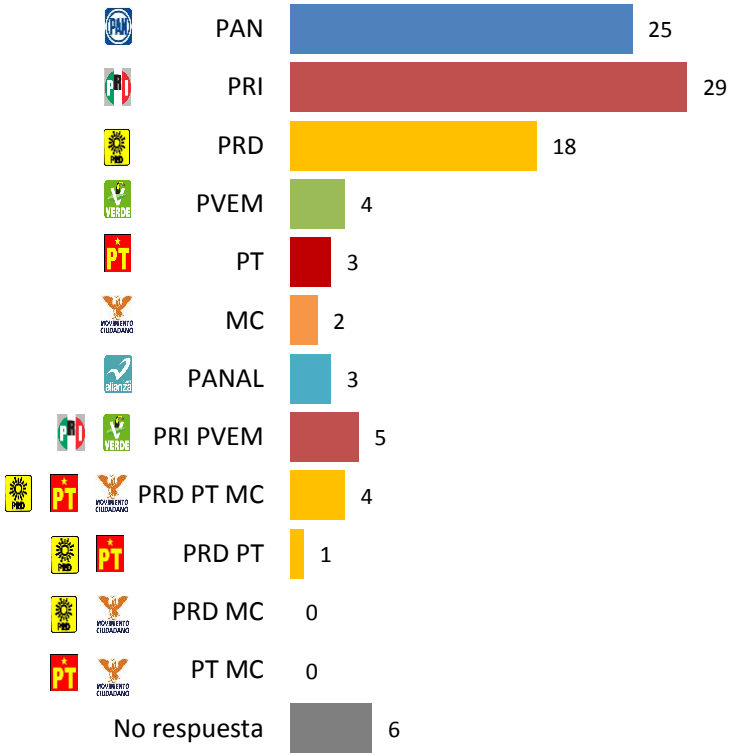

Fuente: Christopher Achen, 1975, *Mass Political Attitudes and the Survey Response*

**3) El problema de los indecisos (Falsos Negativos)**

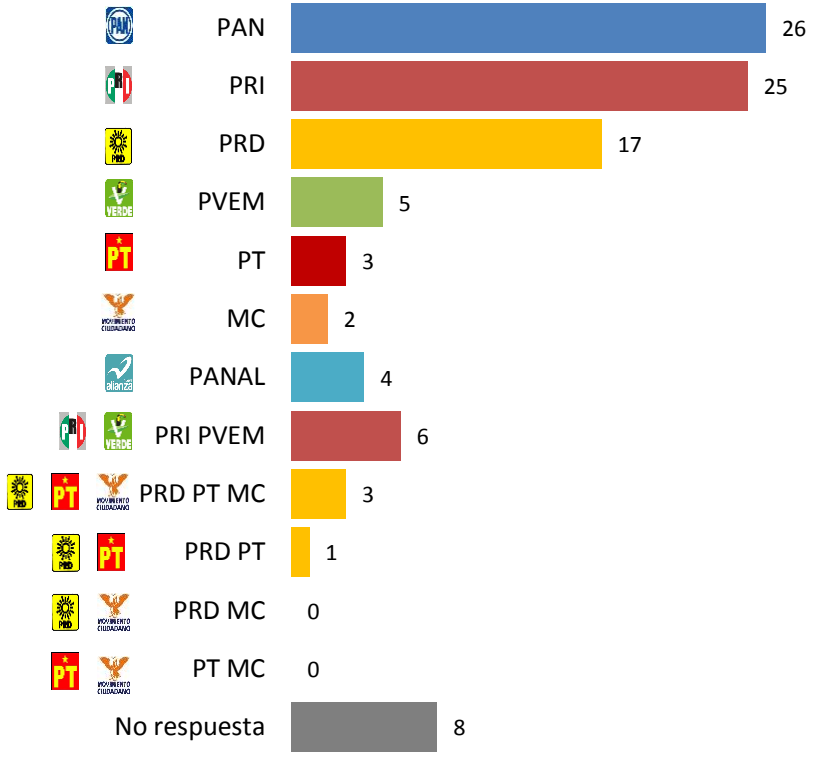
- Son quienes no contestan la pregunta de intención de voto.
- El promedio de las últimas encuestas publicadas es de 25% de no respuesta (*item-nonresponse*) **¿Cómo asignarlos?**
- **La asignación de la no-respuesta (preferencia efectiva/votantes probables) puede ser insatisfactoria**

# Preferencia electoral por decisión de voto 2012

## Entre quienes siempre votan por el mismo partido, decide durante las campañas, o cuando conoce a los candidatos

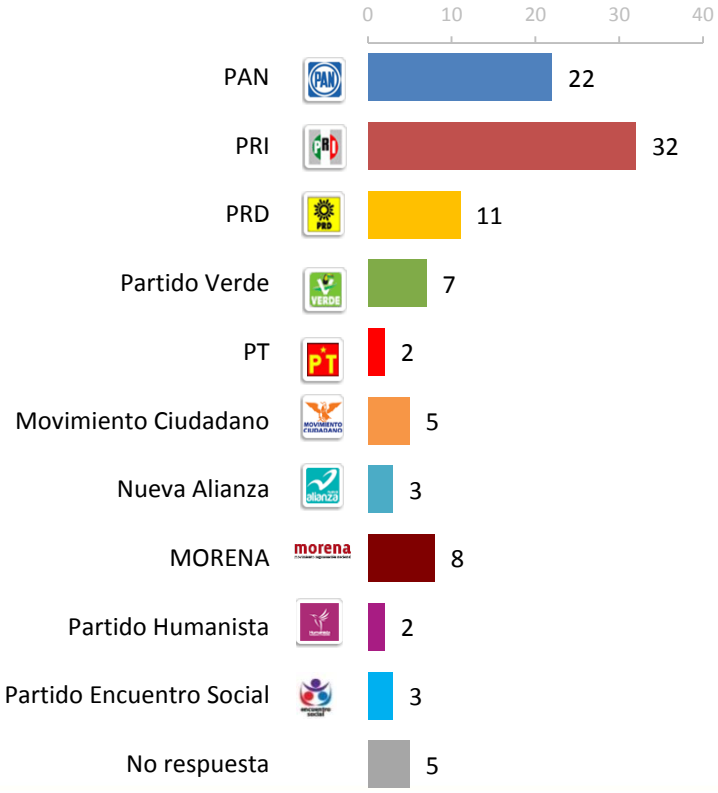



## Entre quienes deciden entre una semana y el día de la elección

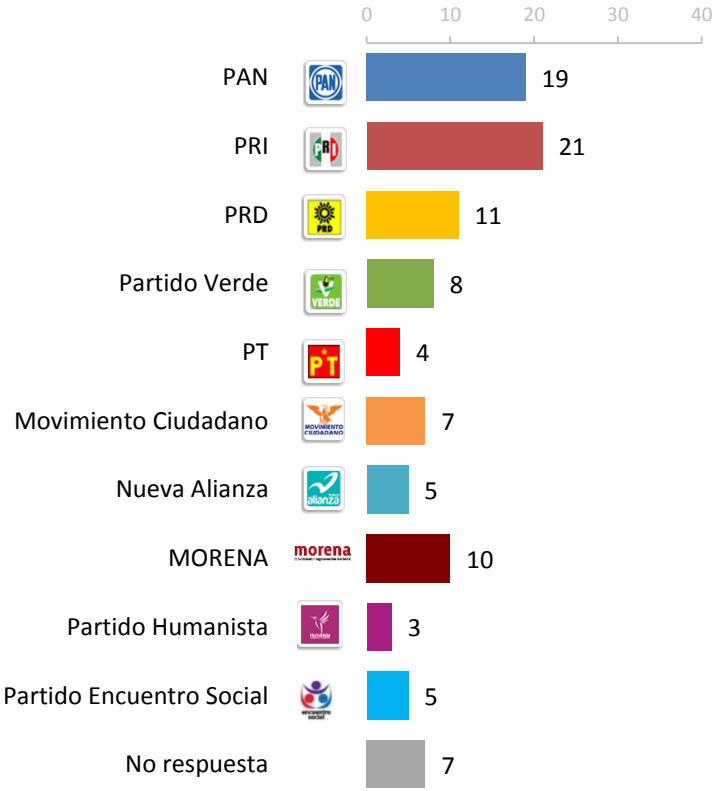


# Preferencia electoral por decisión de voto 2015

## Entre quienes siempre votan por el mismo partido, decide durante las campañas, o cuando conoce a los candidatos




## Entre quienes deciden entre una semana y el día de la elección



## ¿Hay solución?

- Necesitamos un método sólido, plausible y replicable que nos permita enfrentar estos retos.
- Los árboles de regresión y clasificación son uno de estos métodos (CART)
  - No asumen supuestos distribucionales (errores no deben ser normales, exponenciales, etc.)
  - Pueden detectar relaciones no-lineales
  - Pueden detectar interacciones entre variables
  - El ensamblaje de cientos o miles de CARTs (denominados Random Forest) han mostrado tener un desempeño predictivo muy superior a modelos tradicionales (e.g., regresión)

## Random Forest

- Además de su buen desempeño predictivo, generan **medidas de *proximidad*** (similitud) entre cada uno de los individuos encuestados

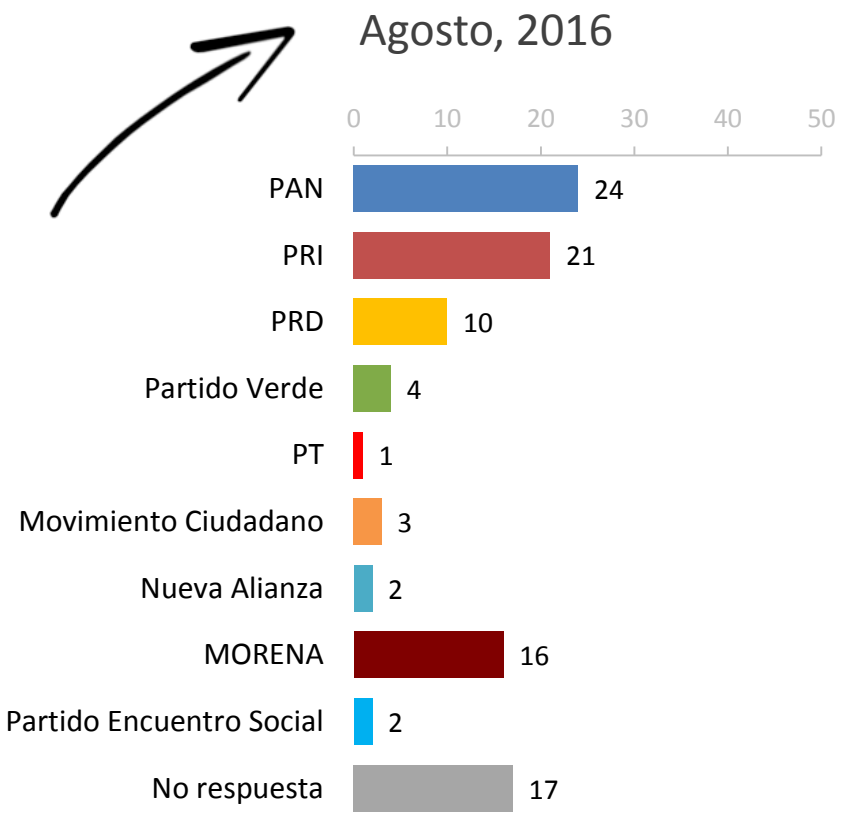
    - Para cada uno de los árboles, si dos observaciones terminan en el mismo nodo, aumenta su proximidad.
    - Al acumular todos los árboles, las observaciones que son "similares" tendrán proximidades cercanas a 1.

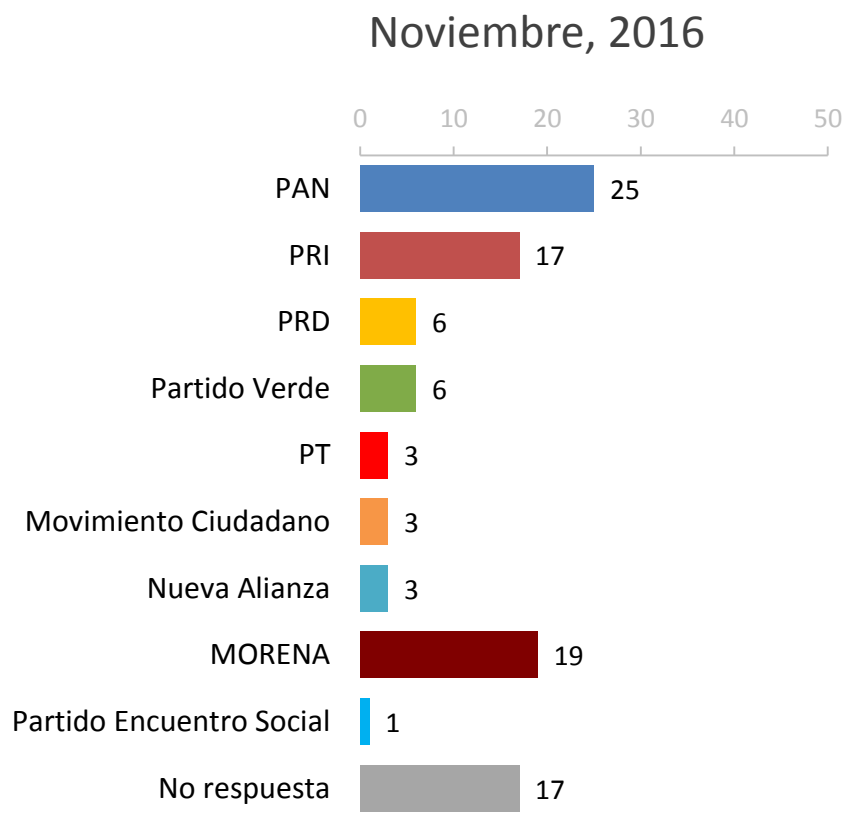
- Detección de valores atípicos

    - ***Outliers*** son casos cuyas proximidades a todos los demás casos en los datos son generalmente pequeñas.
    - Una revisión útil es definir los valores atípicos relativos a su clase. Por lo tanto, un *outlier* en la clase “PAN” es un caso cuyas proximidades a todos los demás casos de clase “PAN” son pequeñas.

# Aplicación:
## Encuesta Nacional Trimestral *Agosto 2016*

Si hoy fueran las elecciones para elegir al próximo Presidente de la República, ¿por cuál partido votaría usted?



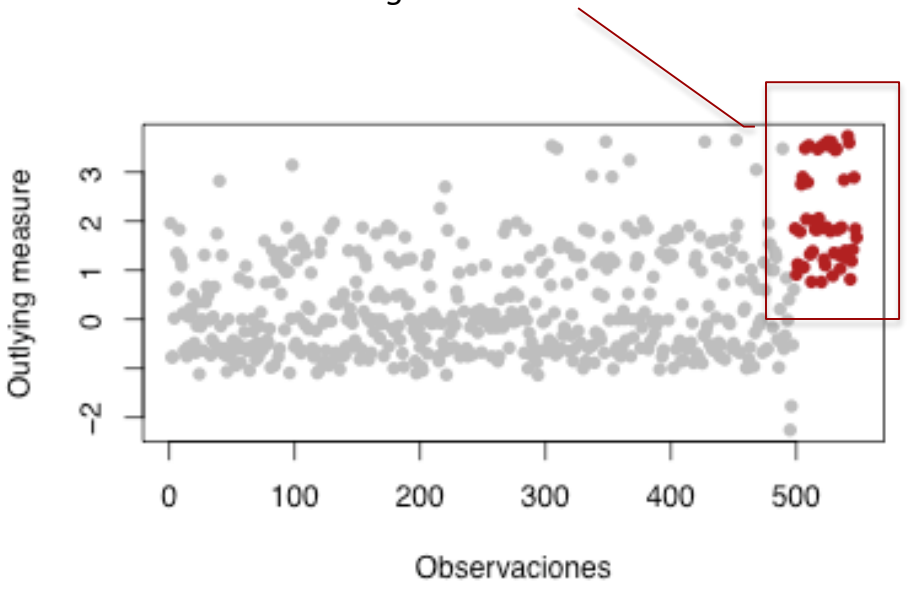



# Aplicación de detección de casos atípicos
# Casos etiquetados incorrectamente (*mislabeled*)

- **El entrenamiento del algoritmo se realizó con 500 casos de la muestra y 21 variables relacionadas con la preferencia electoral**
- 50 casos seleccionados al azar fueron etiquetados incorrectamente y usados durante el entrenamiento.

| | Etiquetas incorrectas | | | | | |
|---|---|---|---|---|---|---|
| Etiqueta correcta | PAN | PRI | PRD | MOR | Otro | NR |
| PAN | **0** | 4 | 2 | 4 | 3 | 5 |
| PRI | 2 | **0** | 1 | 2 | 3 | 0 |
| PRD | 2 | 0 | **0** | 0 | 1 | 1 |
| MOR | 2 | 0 | 2 | **0** | 0 | 3 |
| Otro | 0 | 0 | 3 | 1 | **0** | 0 |
| NR | 0 | 1 | 2 | 4 | 2 | **0** |

*Casos atípicos son detectados por el algoritmo*



## Aplicación de detección de casos atípicos
## Casos etiquetados incorrectamente (*mislabeled*)

Los casos etiquetados incorrectamente son asignados correctamente por el algoritmo:

| | ***Predicción de Random Forest*** | | | | | | |
|---|---|---|---|---|---|---|---|
| ***Etiqueta correcta*** | PAN | PRI | PRD | MOR | Otro | NR | Total |
| PAN | **15** | 0 | 1 | 0 | 0 | 2 | 18 |
| PRI | 0 | **8** | 0 | 0 | 0 | 0 | 8 |
| PRD | 0 | 0 | **3** | 0 | 0 | 1 | 4 |
| MOR | 0 | 0 | 1 | **6** | 0 | 0 | 7 |
| Otro | 1 | 0 | 0 | 0 | **2** | 1 | 4 |
| NR | 1 | 1 | 0 | 0 | 0 | **7** | 9 |
| Total | 17 | 9 | 5 | 6 | 2 | 11 | 50 |

*41 de 50 casos predichos correctamente*

## Resultados de Random Forest
## Predicción para el total de la muestra

Podemos presumir que casos etiquetados incorrectamente de forma natural (inconsistencias, errores, ambivalencias, etc.) también serán asignados de manera "correcta" por el algoritmo.

| | Predicción de Random Forest | | | | | | |
|---|---|---|---|---|---|---|---|
| Observado | PAN | PRI | PRD | MOR | Otro | NR | Total |
| PAN | 218 | ***5*** | ***7*** | ***9*** | ***4*** | ***7*** | 250 |
| PRI | ***10*** | 178 | ***5*** | ***3*** | ***2*** | ***7*** | 205 |
| PRD | ***6*** | ***6*** | 70 | ***7*** | ***2*** | ***7*** | 98 |
| MOR | ***8*** | ***1*** | ***10*** | 127 | ***3*** | ***7*** | 156 |
| Otro | ***13*** | ***6*** | ***17*** | ***9*** | 55 | ***12*** | 112 |
| NR | ***15*** | ***9*** | ***15*** | ***11*** | ***5*** | 119 | 174 |
| Total | 270 | 205 | 124 | 166 | 71 | 159 | 995 |

Casos "consistentes": 77%
PRI observado = PRI predicho
PAN observado < PAN predicho
MORENA observado < MORENA predicho

## Ejemplos (“falsos negativos”):
### Preferencia = No respuesta
### Predicción = PAN

| Preferencia | Nunca votaria por… | Identificación partidista | PAN.vs.PRI | PRI.vs.MOR | PAN.vs.MOR | Opinión PAN | Opinión PRI | Opinión MORENA |
|---|---|---|---|---|---|---|---|---|
| **NR** | PRI | *Indep* | **PAN** | MORENA | **PAN** | **B** | **M** | R |
| **NR** | MOR | *Indep* | **PAN** | MORENA | **PAN** | **B** | R | NC |
| **NR** | NR | *Indep* | **PAN** | MORENA | **PAN** | R | R | R |
| **NR** | NR | *Indep* | **PAN** | PRI | **PAN** | R | R | NC |
| **NR** | NR | *Indep* | **PAN** | MORENA | **PAN** | M | **M** | R |
| **NR** | PRI | *Indep* | **PAN** | MORENA | **PAN** | R | **M** | NC |
| **NR** | PRI | *Indep* | **PAN** | MORENA | **PAN** | **B** | **M** | **B** |
| **NR** | MOR | *Indep* | **PAN** | Ninguno | **PAN** | **B** | R | NC |
| **NR** | PRD | *Indep* | **PAN** | Ninguno | **PAN** | R | **M** | R |
| **NR** | PRI | **Panista** | **PAN** | MORENA | **PAN** | R | **M** | NC |
| **NR** | NR | **Panista** | **PAN** | NS/NC | **PAN** | **B** | **M** | NC |
| **NR** | NR | *Indep* | **PAN** | PRI | **PAN** | R | **M** | NC |
| **NR** | PVEM/NA/ES | *Indep* | **PAN** | NS/NC | **PAN** | R | R | NC |

*B = Muy buena/Buena
R = Regular
M= Mala/Muy mala

## Ejemplos ("falsos positivos"):
## Preferencia = PRI
## Predicción = No PRI

| Predicción | Preferencia | Nunca votaria | Identificación partidista | PAN.vs.PRI | PRI.vs.MOR | PAN.vs.MOR | Opinión PAN | Opinión PRI | Opinión PRD | Opinión MORENA |
|---|---|---|---|---|---|---|---|---|---|---|
| PRD | PRI | PRI | Perredista | PAN | MOR | PAN | B | M | B | NC |
| NR | PRI | PRI | Indep | PAN | MOR | NR | R | R | NC | NC |
| PRD | PRI | PRI | Indep | PAN | MOR | MOR | R | R | NC | R |
| PAN | PRI | NR | Indep | PAN | NR | NR | B | R | NC | NC |
| PAN | PRI | PRD | Panista | PAN | PRI | PAN | B | B | M | NC |
| PAN | PRI | PT.MC | Panista | PAN | PRI | PAN | B | B | NC | NC |
| PAN | PRI | MOR | Priista | PAN | PRI | PAN | B | R | M | M |
| PAN | PRI | PT.MC | Indep | PAN | PRI | PAN | B | R | NC | NC |
| PAN | PRI | NR | Priista | PAN | PRI | MOR | B | R | B | B |
| PRD | PRI | NR | Perredista | Ning | Ning | MOR | R | R | R | R |
| NR | PRI | NR | Indep | Ning | NR | NR | M | M | M | M |
| NR | PRI | NR | Indep | Ning | NR | NR | NC | R | R | R |
| Otro | PRI | PAN | Indep | NR | MOR | PAN | R | M | M | NC |
| NR | PRI | MOR | Priista | NR | NR | NR | R | B | R | R |
| MOR | PRI | PRD | Indep | PRI | MOR | MOR | B | B | NC | NC |

## Ejemplos:
## Preferencia = No respuesta
## Predicción = No respuesta

| Preferencia | Nunca votaria | Identificación partidista | PAN.vs.PRI | PRI.vs.MOR | PAN.vs.MOR | Opinión PAN | Opinión PRI | Opinión PRD | Opinión MORENA |
|---|---|---|---|---|---|---|---|---|---|
| **NR** | PRI | **Indep** | **Ning** | **Ning** | **Ning** | M | M | M | M |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | M | M | M | M |
| **NR** | PAN | **Indep** | **Ning** | **Ning** | **NR** | M | M | M | M |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | R | R | R | R |
| **NR** | NR | **Indep** | **NR** | **NR** | **NR** | M | M | NC | M |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | M | M | M | M |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | R | R | R | R |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | M | M | M | M |
| **NR** | NR | **Indep** | **NR** | **NR** | **NR** | R | M | R | NC |
| **NR** | PRI | **Indep** | PAN | **Ning** | PAN | R | M | R | R |
| **NR** | PRI | **Indep** | **Ning** | PRI | **NR** | M | R | M | M |
| **NR** | PRI | **Indep** | **NR** | MOR | MOR | NC | R | B | R |
| **NR** | NR | **Indep** | **Ning** | **Ning** | **Ning** | M | M | M | M |
| **NR** | PVEM.NA.ES | **Indep** | **NR** | **NR** | **NR** | M | M | M | M |
| **NR** | PRI | **Indep** | PAN | MOR | PAN | M | M | M | M |

## Preferencia electoral

Frecuencias simples *sin* ponderar

Agosto, 2016

| | Observado (Encuesta) | | Predicho (Random Forest) | |
|---|---|---|---|---|
| | Bruta | Efectiva | Bruta | Efectiva |
| PAN | 25.1% | 30.4% | 27.1% | 32.3% |
| PRI | 20.6% | 25.0% | 20.6% | 24.5% |
| PRD | 9.8% | 11.9% | 12.5% | 14.9% |
| MOR | 15.7% | 19.0% | 16.7% | 19.9% |
| Otro | 11.3% | 13.7% | 7.1% | 8.5% |
| NR | 17.5% | - | 16.0% | - |

## ¿Qué tan bien predice elecciones anteriores?

Ejemplo: Encuesta Nuevo León, Junio 2015

| | Observado (Encuesta) | | Predicho (Random Forest) | | |
|---|---|---|---|---|---|
| Candidato/Partido | Bruta | Efectiva | Bruta | Efectiva | Resultado electoral |
| Felipe de Jesús Cantú Rodríguez – PAN | 21.7 | **26.59** | 16.6 | **21.28** | **22.3** |
| Ivonne Álvarez – PRI, PVEM, PANAL, PD | 25.1 | **30.76** | 23.28 | **29.85** | **23.8** |
| Otro | 2.9 | **3.55** | 3.73 | **4.78** | **1.2** |
| Jaime Heliódoro Rodríguez Calderón (Indep) | 31.9 | **39.09** | 34.38 | **44.08** | **48.8** |
| No Respuesta | 18.4 | - | 22.01 | - | - |

## ¿Qué tan bien predice elecciones anteriores?

Ejemplo: Encuesta Querétaro, Junio 2015

| | Observado (Encuesta) | | Predicho (Random Forest) | | |
|---|---|---|---|---|---|
| Candidato/Partido | Bruta | Efectiva | Bruta | Efectiva | Resultado electoral |
| Francisco Domínguez Servién | 34.9 | **46.1** | 36.6 | **49.7** | **48.6** |
| Roberto Loyola Vera | 35.4 | **46.8** | 31.7 | **43.1** | **41** |
| Adolfo Camacho Esquivel | 1.5 | **2** | 1.9 | **2.6** | **2.9** |
| Salvador López Ávila | 0.6 | **0.8** | 0.4 | **0.5** | **1.6** |
| Celia Maya García | 3.3 | **4.4** | 3 | **4.1** | **5.8** |
| NR | 24.4 | - | 26.4 | - | - |