**UNIVERSIDAD DE SANTIAGO DE CHILE**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA INFORMÁTICA**

# INFORME LABORATORIO 3 ANÁLISIS DE DATOS

# TIC-TAC-TOE ENDGAME DATA SET

**CARLOS CÁCERES**
**DANY RUBIANO**

Profesor: Felipe Bello
Ayudantes: Bryan Guzmán
Fernanda Lobos

Santiago - Chile
9 de octubre de 2016

# TABLA DE CONTENIDOS

# ÍNDICE DE FIGURAS

# ÍNDICE DE CUADROS

# CAPÍTULO 1. INTRODUCCIÓN

## 1.1 MOTIVACIÓN Y ANTECEDENTES

En estadística la regresión logística es una de las herramientas con mejor capacidad de análisis en problemas donde la variable dependiente puede tomar valores en un conjunto finito como es el caso de estudios clínicos y epidemiológicos (Molinero, 2001). Su uso se impone de manera creciente desde la década de los 80 debido a las facilidades computacionales con que se cuenta desde entonces (Alonso Fernández, s.f.).

En esta oportunidad se realizará un análisis al dataset *tic tac toe* utilizando esta herramienta pues la regresión logística no se limita al área de la salud, sino que a cualquier área en que sea de utilidad la predicción de la pertenecía a una clase considerando el efecto de otras variables. Dado que cada variable del dataset corresponde a una casilla en el tablero del juego, se intentará entonces obtener conocimiento sobre que casillas inciden en el resultado final, que se representa a través de la clase.

## 1.2 OBJETIVOS

Para este laboratorio se tiene como objetivo la obtención de conocimiento del problema abordado mediante el uso de estudios de regresión logística y modelos asociados, para luego comparar los resultados obtenidos con el conocimiento alcanzado en las experiencias anteriores.

## 1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento distribuye su contenido de la siguiente forma, en primer lugar se encuentra un capitulo dedicado a un pequeño marco teórico en el cual se incluyen las definiciones de los conceptos y técnicas a utilizar en el desarrollo de la experiencia.
A continuación, se presenta todo lo referente a los resultados y su proceso de obtención. En lo que sigue, se exponen los análisis de los resultados obtenidos en el desarrollo de los capítulos anteriores, y por último, con lo desarrollado, se realiza una síntesis total, la cual es presentada en las conclusiones del presente documento.

# CAPÍTULO 2. MARCO TEÓRICO

## 2.1 REGRESIÓN LOGÍSTICA

La regresión es una técnica estadística utilizada para estudiar la relación entre las variables. Un caso particular es la regresión logística que es un tipo de análisis de regresión en el que la variable dependiente no es continua sino dicotómica (que puede tomar solo dos valores), mientras que las variables independientes pueden ser cualitativas o cuantitativas. Una de sus principales ventajas es que sus parámetros pueden interpretarse de forma sencilla en términos de *odds ratios*.

Se usa principalmente para medir la probabilidad de un suceso, como por ejemplo padecer o no una enfermedad (variable dependiente o resultado, codificada como 0 y 1) en función de una series de factores o variables independientes o explicativas.
*(de Madrid, s.f.).*

### 2.1.1 Odds y Odds ratio

- **Odds**: cociente entre la probabilidad de que un evento suceda frente a la probabilidad de que no ocurra (Guaicha, s.f.).
- **Odds ratios**: en estudios de casos y controles se refiere a la razón entre el odds de la exposición observada en casos y la odds de exposición en el grupo de control (Guaicha, s.f.).

## 2.2 P-VALOR

El p-valor corresponde al nivel de significación más pequeño posible que puede escogerse, para el cual todavía se aceptaría la hipótesis alternativa con las observaciones actuales. Cualquier nivel de significación escogido inferior al p-valor comporta aceptar $H_0$. Como es una probabilidad, el p-valor se encuentra en el intervalo $[0, 1]$.

El p-valor es una medida directa de lo verosímil que resulta obtener una muestra como la actual si es cierta H0. Los valores pequeños indican que es muy infrecuente obtener una muestra como la actual, en cambio, los valores altos que es frecuente. El p-valor se emplea para indicar cuánto (o cuán poco) contradice la muestra actual la hipótesis alternativa.
*(de Barcelona, s.f.).*

## 2.3 MÉTRICAS

### 2.3.1 Curva ROC

Las curvas ROC (Receiver Operating Characteristic) presentan la sensibilidad de una prueba que produce resultados contínuos, en función de los falsos positivos (complementario de la especificidad), o en otras palabras es una representación gráfica de sensibilidad frente a 1-especificidad, para distintos puntos de corte (de Málaga, s.f.).

Con estas variables se genera el gráfico ROC. El área bajo la curva (AUC) proporciona un parámetro para evaluar la bondad de la prueba puesto que esta área verifica desigualdades por lo que entre más área abarque, más confiable es la prueba. El área del ROC varía entre $0,5$ y $1$, siendo este último el óptimo.

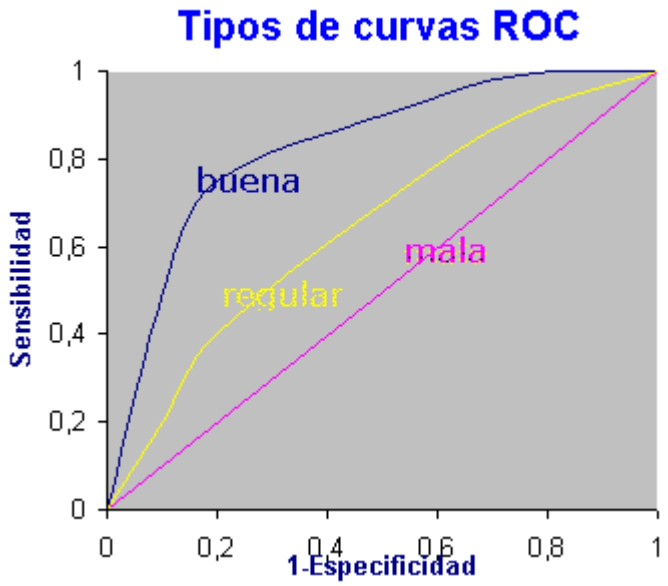



Figura 2.1: Ejemplo curva ROC

#### *2.3.1.1 Sensibilidad*

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como positivo, respecto a la condición de prueba (Deride Silva, 2010).
Matemáticamente se expresa de la siguiente forma:

$$sensibilidad = \frac{nro\ verdaderos\ positivos}{nro\ positivos\ reales} \tag{2.1}$$

*2.3.1.2 Especificidad*

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como negativo, respecto a la condición de prueba (Deride Silva, 2010).
Matemáticamente se expresa de la siguiente forma:

$$especificidad = \frac{nro\ verdaderos\ negativos}{nro\ negativos\ reales} \tag{2.2}$$

### 2.3.2 AIC

Una métrica que es utilizado comúnmente corresponde al Criterio de Información de Akaike (AIC del inglés Akaine Information Criterion), el cual corresponde a un índice que evalúa el ajuste del modelo a los datos así como la complejidad del modelo. Es útil para realizar comparaciones entre modelos similares, mientras menor sea el AIC, mejor es el ajuste del modelo (Seoane, 2013).

# CAPÍTULO 3. OBTENCIÓN DE RESULTADOS

Para cumplir con el objetivo propuesto para este laboratorio, en la aplicación de los modelos de regresión logística se usa la función *glm* en R, la cual determina una descripción simbólica del predictor lineal y una descripción de la distribución de error de las variables implicadas en el modelo. Así mismo se realiza el cálculo de los odss ratios, mediante la función *exp*, para cada modelo propuesto.

Antes de dicha aplicación, se establece un cierto preprocesamiento al dataset, en este caso distinto al de la experiencia anterior, ya que ahora se emplea una cierta codificación de los datos, de manera que estos se representen cuantitativamente asignándole un numero a los posibles valores de *x*, *o* ó *b*, así mismo para las clases *positive* y *negative*. Todo esto se hace necesario para poder hacer uso de las función *glm* en R.

Una vez listo el procedimiento, en primer lugar se realiza un modelo inicial aplicando la regresión logística con todas las variables, con el objetivo de obtener información que posibilite la determinación de cuáles son las que contribuyen en mayor medida la definición de la clase, y así poder realizar posteriormente otros modelos que refinen la información obtenida. El cálculo de los odds ratios correspondientes al modelo, ayudarán en la decisión anterior.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.45580    0.47408   0.961   0.3363
p1           0.11061    0.09460   1.169   0.2423
p2          -0.21897    0.09316  -2.350   0.0188 *
p3           0.11061    0.09460   1.169   0.2423
p4          -0.21897    0.09316  -2.350   0.0188 *
p5           0.51629    0.09773   5.283 1.27e-07 ***
p6          -0.21897    0.09316  -2.350   0.0188 *
p7           0.11061    0.09460   1.169   0.2423
p8          -0.21897    0.09316  -2.350   0.0188 *
p9           0.11061    0.09460   1.169   0.2423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.1: Modelo 1- Todas las variables.

Cuadro 3.1: Odds Ratios Modelo 1

| **(Intercept)** | **p1** | **p2** | **p3** | **p4** |
|---|---|---|---|---|
| 1.5774380 | 1.1169605 | 0.8033494 | 1.1169605 | 0.8033494 |
| **p5** | **p6** | **p7** | **p8** | **p9** |
| 1.6758005 | 0.8033494 | 1.1169605 | 0.8033494 | 1.1169605 |

A partir del modelo anterior, se propone la realización de otro modelo que incluya las variables p1, p3, p5, p7 y p9, ya que estas poseen un p-valor menor, sus valores estimados mayores a sus pares y además sus odss ratios son mayor que uno, indicando mayor probabilidad de incidencia en la clase.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.04453    0.30222  -3.456 0.000548 ***
p1           0.19406    0.09025   2.150 0.031535 *
p3           0.19406    0.09025   2.150 0.031535 *
p5           0.57973    0.09576   6.054 1.41e-09 ***
p7           0.19406    0.09025   2.150 0.031535 *
p9           0.19406    0.09025   2.150 0.031535 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.2: Modelo 2 - Casillas 1, 3, 5, 7 y 9.

Cuadro 3.2: Odds Ratios Modelo 2

| **(Intercept)** | **p1** | **p3** | **p5** | **p7** | **p9** |
|---|---|---|---|---|---|
| 0.3518562 | 1.2141689 | 1.2141689 | 1.7855588 | 1.2141689 | 1.2141689 |

Ya que la variable p5 es la que presenta un menor p-valor y un valor estimado mayor al de todos sus pares, se realiza un modelo con sólo esa variable.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.005297   0.133203   0.040    0.968
p5          0.493953   0.091948   5.372 7.78e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.3: Modelo 3 - casilla 5.

Cuadro 3.3: Odds Ratios Modelo 3

| **(Intercept)** | **p5** |
|---|---|
| 1.005311 | 1.638782 |

A modo de comparación se realiza un modelo que incluye todas las variables presentes en el segundo modelo, pero esta vez discriminando la variable p5.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03398    0.23793   0.143    0.886
p1           0.12370    0.08783   1.408    0.159
p3           0.12370    0.08783   1.408    0.159
p7           0.12370    0.08783   1.408    0.159
p9           0.12370    0.08783   1.408    0.159

(Dispersion parameter for binomial family taken to be 1)
```

Figura 3.4: Modelo 4 - casillas 1, 3, 7 y 9.

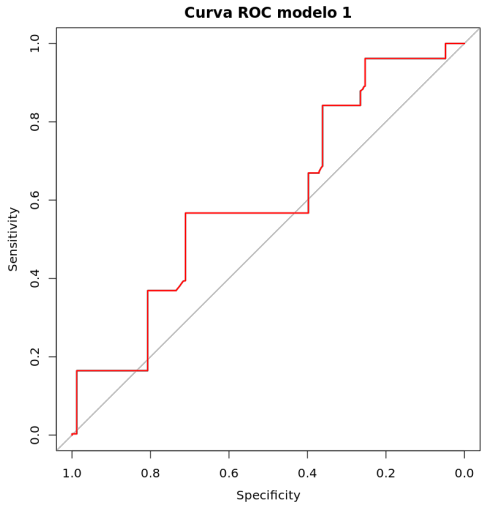
Cuadro 3.4: Odds Ratios Modelo 4

| **(Intercept)** | **p1** | **p3** | **p7** | **p9** |
|---|---|---|---|---|
| 1.034560 | 1.131678 | 1.131678 | 1.131678 | 1.131678 |

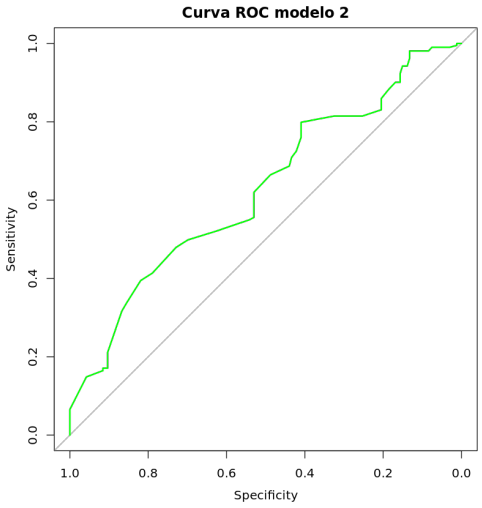
## 3.1 MÉTRICAS DE ANÁLISIS DE LA REGRESIÓN LOGÍSTICA

### 3.1.1 Curvas ROC y AIC

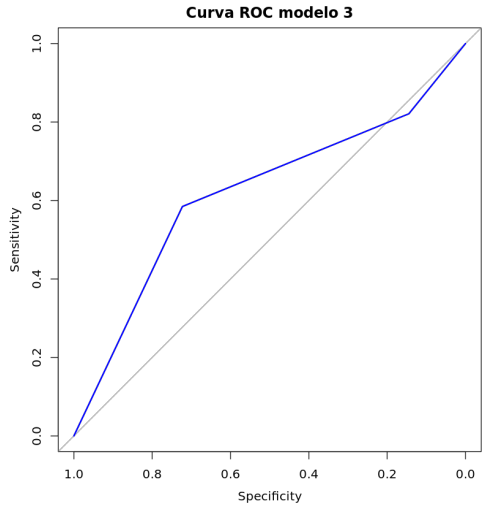
Para el análisis de los modelos de regresión logística se utiliza en primer lugar el cálculo de las curvas ROC y el análisis de el índice AIC, que relaciona la complejidad del modelo y la cantidad de variables presentes.



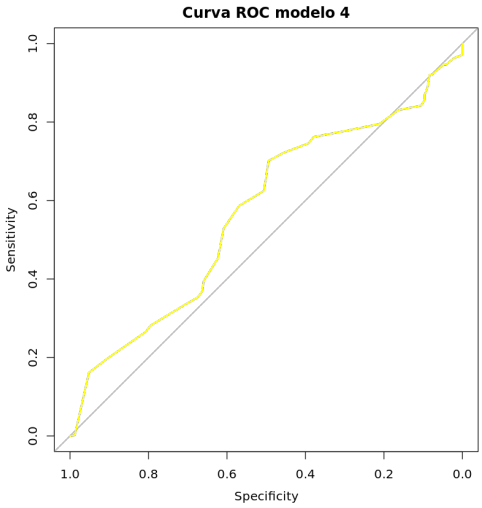

(a) Todas las casillas - AUC: 0.6046




(b) Casillas 1, 3, 5, 7 y 9 - AUC: 0.6266




(c) Casilla 5 - AUC: 0.6191




(d) Casillas 1, 3, 7 y 9 - AUC: 0.5717

Figura 3.5: Curvas ROC

AUC: Área Bajo la Curva.

*3.1.1.3 Visión General*

A continuación se presenta un gráfico de las curvas ROC superpuestas para poder ver las diferencias que presentan los modelos.

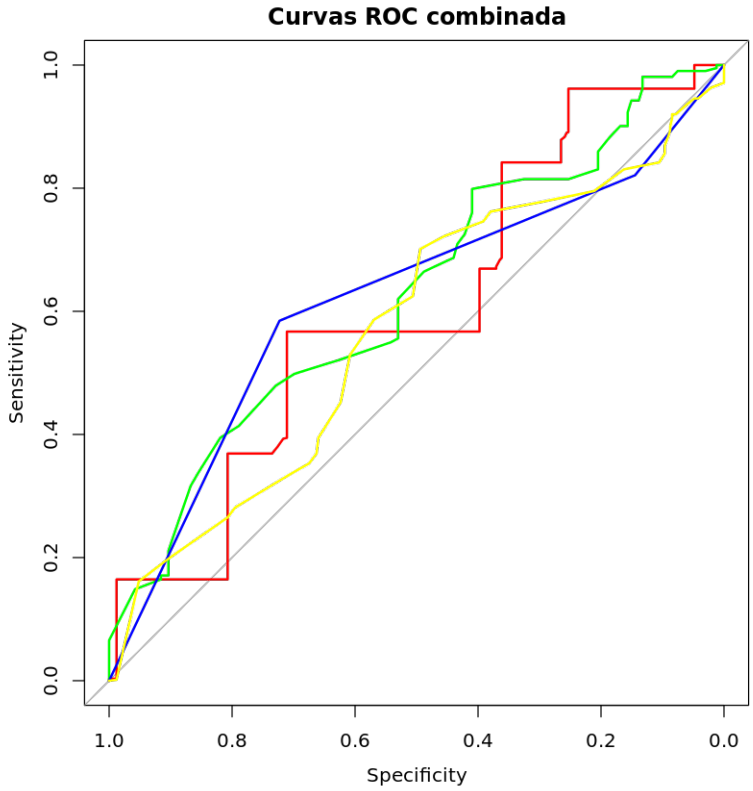



Figura 3.6: Cuvas ROC - Todos los modelos

### 3.1.2 AIC

Para una mayor ejemplaridad de la diferencia entre los modelos, se presenta una tabla resumen que muestra los índices AIC y el área bajo la curva (AUC) de las curvas ROC correspondientes a cada uno de los modelos.

Cuadro 3.5: AIC y AUC

| | **Modelo 1** | **Modelo 2** | **Modelo 3** | **Modelo 4** |
|---|---|---|---|---|
| **AIC** | 1194.8 | 1203.9 | 1211.1 | 1239.5 |
| **AUC** | 0.6046 | 0.6266 | 0.6191 | 0.5717 |

# CAPÍTULO 4. ANÁLISIS DE RESULTADOS

Recordando la representación que se establece en el dataset de la ubicación de las casillas, se presenta la figura 4.1 , de manera que el análisis sea mejor comprendido.

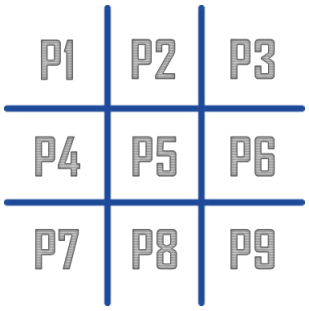



Figura 4.1: Posiciones representativas del tablero

Inicialmente se aplicó la regresión logística considerando la clase frente a todas las variables; la tabla de coeficientes obtenidos arrojó que la casilla 5 tiene un valor estimado de $0{,}5$ aproximadamente y se destaca por sobre todas las demás; le siguen las casillas 1, 3, 7 y 9 con un valor aproximado de $0{,}1$ cada una. Las variables anteriormente mencionadas presentan odds ratios todos mayores que 1, lo que indica que existe probabilidad de que estas incidan en la decisión de la clase, es decir, pueden incidir en que ocupar estas casillas en una jugada cualquiera determina la condición de victoria o pérdida.
Finalmente las casillas 2, 4, 6 y 8 muestran valores negativos en sus estimaciones, además los p-valores asociados a cada una de esas casillas son considerablemente superiores en comparación a los del resto y sus odds ratios representan lo contrario a lo que se indica para las otras variables mencionadas.

Con esto en cuenta, se generaron 4 modelos, el primero con todas las variables, el segundo con las variables correspondientes a las casillas 1, 3, 5, 7 y 9; el tercero solo con la casilla 5 y uno extra que consideró las casillas 1, 3, 7 y 9.

## 4.1 MODELOS

- **Modelo 1**: este modelo que considera todas las variables, al tener el menor valor en la métrica AIC podría considerarse a simple vista como el más exacto, sin embargo contiene variables en las cuales el p-valor llega a ser $0{,}2423$ que en comparación al

resto de las variables es demasiado grande y se puede pensar que existe un modelo que prediga mejor. El área bajo la curva ROC alcanzada por este modelo supera levemente el mínimo y tiene un valor de $0{,}6046$.

- **Modelo 2**: el modelo que considera las 5 variables que logran las mayores estimaciones y a la vez los p-valores que consideramos aceptables por ser inferiores a $0{,}05$, alcanza un valor de AIC de $1203{,}9$. Se esperaría que fuera menor al del modelo anterior al considerar solo variables *convenientes* (por lo dicho anteriormente); no fue el caso, pero logra superar el área bajo la curva alcanzando un valor de $0{,}6266$. No es un valor impresionante pero notamos que el modelo ha mejorado.

- **Modelo 3**: en este caso se consideró solo la variable 5 como la capaz determinar con más exactitud la clase a la que pertenecerá; escogida para esto tanto por su valor estimado que supera el $0{,}5$ como su p-valor considerablemente inferior a todas las restantes. En este modelo el valor de AIC vuelve a incrementarse levemente, pero esta vez el área bajo la curva ROC en vez de aumentar, disminuyó. Es decir que el modelo empeoró respecto a su antecesor.

- **Modelo 4**: este modelo, generado solo para comparar, consideró las variables que tenían los p-valores intermedios (los que no eran ni los mayores ni los menores). En este caso el valor de AIC vuelve a incrementarse y junto con ello el área bajo la curva ROC cae bruscamente 4 puntos alcanzando el valor de $0{,}5717$ que es el mínimo observado en lo que va de análisis.

## 4.2 COMPARACIÓN

De los 4 modelos generados se ha determinado que el segundo de ellos es el que entregó los mejores resultados. El valor de AIC para este modelo incrementó respecto al primer modelo generado pero si se considera que este incremento es de menos de 10 unidades y esto implica aumentar el área bajo la curva ROC de dicho modelo en poco más de $2$ unidades, se puede juzgar que es marginalmente mejor.

Le sigue el modelo 3 con el área bajo la curva ROC intermedia entre los 3 primeros modelos. En este caso el valor de AIC vuelve a aumentar y con ello se observa el decremento

del AUC por lo que se descarta inmediatamente como el mejor modelo y que además si consideramos que este modelo solo se encuentra la clase frente a la casilla 5 (la central), si lo llevamos a la realidad, esta tiene la mayor cantidad de combinaciones ganadoras posibles pero no determina el triunfo. En este mismo data-set se puede observar que solo 320 de las jugadas que tienen esta casilla marcada con *x* resultaron ganadoras, es decir, menos del 50 %.

Para finalizar se generó un último modelo que considera las casillas que tenían valores *intermedios* (ni buenos ni malos p-valores en general). En este caso se intentaba verificar si con estas 4 casillas con estas características, el modelo sería regular: ni el mejor, ni el peor. Se observó que si bien tiene valores de este tipo, el modelo obtenido es el peor de todos incrementando el valor de AIC considerablemente en comparación a los incrementos anteriores; y el área bajo la curva ROC menos valiosa de todas por lo que es un modelo desde ya descartado. Se debe considerar que este modelo contempla las casillas de las 4 esquinas, que sin una intermedia no tienen mayor valor en la realidad; necesitan de la central para ser ganadoras y se cree que por esto el mejor modelo es el que considera estas 4 esquinas más la central.

# CAPÍTULO 5. CONCLUSIONES

Finalmente se aplicó un análisis de regresión lineal al data-set objeto de estudio y se fueron afinando detalles a medida que avanzaba. Se generaron 4 modelos en total de los cuales el segundo que contemplaba las casillas esquina más la central fue declarado como el *mejor*. Llevado a la realidad y en comparación a lo visto en la primera experiencia, tiene mucho sentido esta determinación pues son las casillas que más combinaciones ganadoras tienen. No obstante, el resultado del área bajo la curva ROC no es motivo de alegría; si bien es el máximo que se pudo obtener, apenas llega al valor de $0{,}62$ que si consideramos que el óptimo es $1$ y $0{,}5$ es malo, su peso como un modelo confiable no es tan alto, pero como se dijo: es el mejor encontrado.

Respecto a los modelos restantes, el primero perfectamente pudo ser un candidato al mejor, salvo que el área bajo la curva ROC de este era levemente menor considerando $4$ variables más, es decir, todas las del tablero, aumentando su complejidad. El escogido logra mejores resultados (AUC - ROC), con menos variables.
Con el modelo 4 se pudo entender que índices intermedios no determinan una regresión logística con métricas promedio sino que se obtiene, el menos en este caso, peores resultados de los que se esperaban.

Respecto al objetivo principal de este informe no es posible hacer una comparación con la segunda experiencia puesto que la naturaleza del data-set impidió que aquella se realizara correctamente (hablando del algoritmo k-means). Aún así, en este informe se alude un par de veces al primer informe en donde el conocimiento del data-set era superficial lo que nos acerca más a lo que una persona común sin aplicar métodos puede observar. Esto nos sirvió como punto de comparación, en algunos casos, entre la regresión logística hecha en esta experiencia y lo observado en la realidad sin un análisis más profundo. Es así entonces, que se puede inferir con un fundamento más profundo que la casilla 5 (posición central), es determinante en el resultado final del juego, así mismo se da para las jugadas diagonales con sus respectivas casillas involucradas, pero con una menor incidencia.

# CAPÍTULO 6. REFERENCIAS


Aha, D. W. (1991, Agosto). Index of /ml/machine-learning-databases/tic-tac-toe. https://archive.ics.uci.edu/ml/machine-learning-databases/tic-tac-toe/.

Alonso Fernández, A. (s.f.). Introducción a la regresión logística. http://www.est.uc3m.es/amalonso/esp/bstat-tema9.pdf.

de Barcelona, U. (s.f.). El p-valor. http://www.ub.edu/stat/GrupsInnovacio/Statmedia/demo/Temas/Capitulo9/B0C9m1t18.htm.

de Madrid, U. C. (s.f.). Regresión logística. http://pendientedemigracion.ucm.es/info/dosis/Preventiva/doctorado/TEMA13.pdf.

de Málaga, U. (s.f.). Curvas roc: elección de puntos de corte y área bajo la curva (auc). https://www.bioestadistica.uma.es/analisis/roc1/.

Deride Silva, J. (2010). Curvas roc y regresión lineal. https://goo.gl/Ss55yE.

Guaicha, O. (s.f.). Odds ratio. http://es.slideshare.net/omargp100/odds-ratio-27849262.

Molinero, L. (2001). La regresion logistica (i). http://www.seh-lelha.org/rlogis1.htm.

Seoane, J. (2013). Análisis bioestadístico con modelos de regresión en r. http://www.uam.es/personal_pdi/ciencias/jspinill/CFCUAM2013/Multimodel_inference_CFCUAM2013.html.

UCI, M. L. R. (s.f.). Tic-tac-toe endgame data set. https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

wonderopolis.org. (Febrero 2016). How old is tic-tac-toe? Recuperado desde http://wonderopolis.org/wonder/how-old-is-tic-tac-toe

# CAPÍTULO 7. ANEXO: CÓDIGO FUENTE EN R

```
DB <- read.table("Lab3/tic-tac-toe.data", header=FALSE, sep=",",
col.names=c("p1","p2","p3","p4","p5","p6","p7","p8","p9","class"))

#Numerizacion

DB$p1<-as.numeric(DB$p1)
DB$p2<-as.numeric(DB$p2)
DB$p3<-as.numeric(DB$p3)
DB$p4<-as.numeric(DB$p4)
DB$p5<-as.numeric(DB$p5)
DB$p6<-as.numeric(DB$p6)
DB$p7<-as.numeric(DB$p7)
DB$p8<-as.numeric(DB$p8)
DB$p9<-as.numeric(DB$p9)
DB$class<-as.numeric(DB$class)

# Arreglo de la numerizacion
# b de 1 pasa a 0
# o de 2 pasa a 1
# x de 3 pasa a 2
# negative de 1 pasa a 0
# positive de 2 pasa a 1. Para poder aplicar glm sobre class.
DB[DB==1]<-0
DB[DB==2]<-1
DB[DB==3]<-2

#################################################################
#Modelos de Regresion Logística y odss ratios
```

```
reg1 <- glm(class ~., data = DB, family = binomial(link="logit"))
summary(reg1)
odsr1 <- exp(reg1$coefficients)
print(odsr1)

reg2 <- glm(class~p1+p3+p5+p7+p9, DB, family = binomial(link = "logit"))
summary(reg2)
odsr2 <- exp(reg2$coefficients)
print(odsr2)

reg3<- glm(class~p5, DB,family= binomial(link="logit"))
summary(reg3)
odsr3 <- exp(reg3$coefficients)
print(odsr3)

reg4 <- glm(class~p1+p3+p7+p9, DB, family = binomial(link = "logit"))
summary(reg4)
odsr4 <- exp(reg4$coefficients)
print(odsr4)



#Analisis de varianza, test de Chi-Cuadrado
anav <-anova(reg1,reg2,reg3,reg4,test = "Chisq")

#AIC
aic <-AIC(reg1,reg2,reg3,reg4)

#################################################################
```

```
#Curvas ROC

library(pROC)
prob <- predict(reg1,type=c("response"))
curva1 <- roc(class~prob, data = DB)
plot(curva1, col="red",  main="Curva ROC modelo 1")


prob <- predict(reg2,type=c("response"))
curva2 <- roc(class~prob, data = DB)
plot(curva2, col="green",  main="Curva ROC modelo 2")


prob <- predict(reg3,type=c("response"))
curva3 <- roc(class~prob, data = DB)
plot(curva3, col="blue",  main="Curva ROC modelo 3")

prob <- predict(reg4,type=c("response"))
curva4 <- roc(class~prob, data = DB)
plot(curva4, col="yellow",  main="Curva ROC modelo 4")

############################################################

prob <- predict(reg1,type=c("response"))
curva1 <- roc(class~prob, data = DB)
plot(curva1, col="red",  main="Curvas ROC combinadas")


prob <- predict(reg2,type=c("response"))
```

```
curva2 <- roc(class~prob, data = DB)
plot(curva2, col="green",  add=TRUE)


prob <- predict(reg3,type=c("response"))
curva3 <- roc(class~prob, data = DB)
plot(curva3, col="blue",  add=TRUE)

prob <- predict(reg4,type=c("response"))
curva4 <- roc(class~prob, data = DB)
plot(curva4, col="yellow",  add=TRUE)
```