

INFORME LABORATORIO 1 ANÁLISIS DE DATOS
TIC-TAC-TOE ENDGAME DATA SET

CARLOS CÁCERES
DANY RUBIANO

Profesor: Felipe Bello
Ayudante: Bryan Guzmán

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS.....	ii
ÍNDICE DE CUADROS	iii
CAPÍTULO 1. INTRODUCCIÓN.....	1
1.1 MOTIVACIÓN Y ANTECEDENTES	1
1.2 OBJETIVOS	1
1.3 ORGANIZACIÓN DEL DOCUMENTO	1
CAPÍTULO 2. DESCRIPCIÓN DEL PROBLEMA.....	3
2.1 DESCRIPCIÓN DE LA BASE DE DATOS	3
2.2 DESCRIPCIÓN DE CLASES Y VARIABLES	6
2.2.1 Clases	6
2.2.2 Variables	6
CAPÍTULO 3. ANÁLISIS ESTADÍSTICO E INFERENCIAL.....	9
CAPÍTULO 4. CONCLUSIONES.....	15
CAPÍTULO 5. REFERENCIAS	17

ÍNDICE DE FIGURAS

2.1	Posiciones representativas del tablero	5
3.1	Clases	9
3.2	Estado final del <i>tablero</i>	9
3.3	Esquina superior izquierda	10
3.4	Esquina superior central	10
3.5	Esquina superior derecha	10
3.6	Esquina central izquierda	11
3.7	Esquina central central	11
3.8	Esquina central derecha	11
3.9	Esquina inferior izquierda	12
3.10	Central inferior	12
3.11	Esquina inferior derecha	12
3.12	Resultados Totales por Jugada	14

ÍNDICE DE CUADROS

Tabla 2.1: Estructura del archivo tic-tac-toe.data	5
3.1 Frecuencia de las jugadas ganadoras para x y o	13

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN Y ANTECEDENTES

Tic-Tac-Toe, también conocido como *tres en línea* o *gato*, es un antiguo juego de lápiz y papel que millones de personas han jugado a lo largo de la historia. Algunos atribuyen el origen de este juego a los antiguos egipcios, otros piensan que evolucionó de un antiguo juego romano llamado Terni Lapilli, el cual es similar solo que tiene rejillas de 3x3. El Tic-Tac-Toe consiste en 9 casillas creadas a partir de 4 líneas que generan 3 columnas y 3 filas; el juego consiste en que cada jugador marca solo una casilla por turno con un símbolo que puede ser x o o ; x tiene el primer turno, luego se intercalan x y o tratando de alcanzar 3 marcas en una fila, sea horizontal, vertical o diagonal. *Tic-Tac-Toe* es un juego interesante matemáticamente pues se puede utilizar para enseñar principios matemáticos simples como probabilidad, por ejemplo existen 362.800 formas distintas de poner x y o en el *tablero* de las cuales 255.168 son posibles ganadoras. Sin embargo, si se eliminan todas las combinaciones simétricas, sólo hay 138 combinaciones ganadoras únicas. Con estas pocas combinaciones, *Tic-Tac-Toe* es un juego relativamente fácil de aprender para los niños por lo que es muy popular (Aha, 1991).

1.2 OBJETIVOS

El objetivo de este documento es estudiar e interpretar los datos presentes en un dataset provenientes del Tic-Tac-Toe Endgame. Para ello, se busca describir los atributos, clases y valores con el fin de comprender el dominio del problema y adquirir una idea clara y precisa del trabajo que se está desarrollando.

1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento distribuye su contenido de la siguiente forma, en primer lugar se encuentra una descripción del problema, donde se presenta la información considerada relevante para la comprensión de la problemática. En ese mismo capítulo, se realiza una descripción de la base de datos con su formato y composición, y la descripción de las clases y variables respectivas al dominio de esta.

Luego, se realiza un análisis estadístico e inferencial de las clases y variable más importantes, con el fin de esclarecer de una forma más profunda la base de datos abordada.

Por último, con lo desarrollado en los capítulos del documento, se realiza el análisis de interés el cual es presentado a en las conclusiones del presente informe.

CAPÍTULO 2. DESCRIPCIÓN DEL PROBLEMA

2.1 DESCRIPCIÓN DE LA BASE DE DATOS

La principal herramienta para obtener información del problema es la respectiva base de datos, en la cual sus características generales son:

Título: Tic-Tac-Toe Endgame database

Creador: David W. Aha

Fecha: 19 de Agosto de 1991

Usos conocidos en el pasado: Estos son:

- Matheus, C. J., & Rendell, L. A. (1989).
Construcción inductiva de árboles de decisión. En *procedimientos de la undécima Conferencia internacional conjunta sobre Inteligencia Artificial* (pp. 645-650). Detroit, MI: Morgan Kaufmann.
-CITRE fue aplicado a la formación de 100 instancias de entrenamiento y 200 instancias de conjuntos de prueba. En un estudio con varias cantidades de conocimiento específico del dominio, su exactitud promedio más alto fue 76.7 % (mediante el árbol de decisión final creado para pruebas).
- Matheus, C. J. (1990).
Agregando conocimiento del dominio SBL a través de la construcción de la característica. En *actas de la octava National Conference en Inteligencia Artificial* (pp. 803-808). Boston, MA: Prensa AAAI.
-Experimentos similares con CITRE, incluyeron las curvas de aprendizaje hasta la formación de 500 instancias de prueba de conjuntos, pero utilizaron todas las instancias en la base de datos para la prueba. Exactitudes alcanzaron por encima del 90 %, pero no se proporcionaron los valores específicos (véase la tesis doctoral de Chris para más detalles).
- Aha, D. W. (1991).
Inducción constructiva incremental: un enfoque basado en instancias. En *actas del octavo taller internacional sobre aprender de máquina* (pp. 117-121). Evanston, ILL: Morgan Kaufmann. -
Fue utilizado 70 % para la formación, y el 30 % de los casos para el análisis, se evaluaron los ensayos de más de 10. Los resultados reportados para seis algoritmos fueron:
 - NewID: 84,0 %
 - CN2: 98.1 %
 - MBRtalk: 88,4 %
 - IB1: 98.1 %

- IB3: 82,0 %

- IB3-CI: 99.1 %

Los resultados informaron también al agregar un atributo adicional 10 de valor ternario irrelevantes; se obtuvieron relativamente similares resultados, salvo que el rendimiento de IB1 degrada más rápidamente que los otros.

Información relevante: Esta base de datos codifica el juego completo de configuraciones posibles del juego tic-tac-toe, donde "x" es asumida como primera jugada. El concepto objetivo es "el triunfo para x" (p. ej., verdadero cuando "x" tiene uno de 8 modos posibles de crear "los tres en línea").

De modo interesante, esta base de datos cruda da un algoritmo de árbol de decisión denominado stripped-down. Sin embargo, el algoritmo a base de regla CN2, el algoritmo de estudio simple IB1 a base de instancias, y el algoritmo de árbol de decisión CITRE tienen un buen desempeño sobre esta.

El dataset completo contiene 2 archivos diferentes, tic-tac-toe.data y tic-tac-toe.names, los cuales se describen a continuación:

tic-tac-toe.names: Archivo que contiene información asociada a la base de datos, como el autor, los usos que se le ha dado anteriormente, información considerada relevante, significado de clases y atributos, entre otros.

tic-tac-toe.data: Archivo de principal predilección de la base de datos, que contiene los valores de cada juego registrado. Este distribuye su contenido tal como se puede observar en la siguiente tabla:

Tabla 2.1: Estructura del archivo tic-tac-toe.data

P1	P2	P3	P4	P5	P6	P7	P8	P9	Clase
x	x	x	x	o	o	x	o	o	positive
x	x	x	x	o	o	o	x	o	positive
x	x	x	x	o	o	o	o	x	positive
x	x	x	x	o	o	o	b	b	positive
.
.
.
o	x	o	x	x	o	x	o	x	negative
o	x	o	x	o	x	x	o	x	negative
o	x	o	o	x	x	x	o	x	negative
o	o	x	x	x	o	o	x	x	negative

Cada fila de la tabla representa una instancia de la base de datos, que a su vez interpreta cada juego registrado.

Cada columna representa un atributo, que también indica la jugada (x, y, b : posición vacía) en cierta posición del tablero. A modo de ejemplificación, en este documento se representa una posición específica con una P , seguida de un número, Véase Figura 2.1, y posteriormente se entrega el nombre de cada posición real descrita en el archivo *tic-tac-toe.names*. La ultima columna, es representativa a una situación final de cada juego o instancia, esto se describirá con más detalle en lo que sigue del documento.

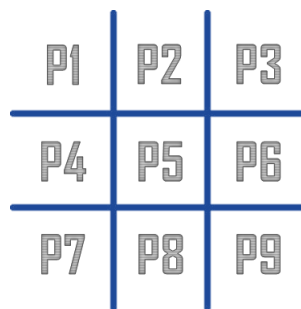


Figura 2.1: Posiciones representativas del tablero

- **P1:** top-left-square.
- **P2:** top-middle-square.
- **P3:** top-right-square.
- **P4:** middle-left-square.
- **P5:** middle-middle-square.
- **P6:** middle-right-square.
- **P7:** bottom-left-square.
- **P8:** bottom-middle-square.
- **P9:** bottom-right-square.

En total el numero de instancias presentes en este archivo corresponde a 958.

2.2 DESCRIPCIÓN DE CLASES Y VARIABLES

2.2.1 Clases

En cada instancia se indica un estado resultante del juego registrado. En sí, se presentan dos clases las cuales se definen a continuación:

- **Positive:** El jugador con el símbolo x ganó la partida.
- **Negative:** El jugador con el símbolo x no ganó la partida.

2.2.2 Variables

El número de variables o atributos total es de 9 , cada una correspondiente a una posición del tablero del tic - tac-toe. Cabe destacar que por su procedencia, todas las variables son del tipo cualitativa nominal, ya que presenta modalidades no numéricas que no admiten un criterio de orden.

- **top-left-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *superior-izquierda*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **top-middle-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *superior-media*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **top-right-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *superior-derecha*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **middle-left-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *media-izquierda*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **middle-middle-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *media-media (central)*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **middle-right-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *media-derecha*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.
- **bottom-left-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *inferior-izquierda*. Puede tomar el valor x si la casilla está tomada por el jugador x , el valor o si la casilla está tomada por el jugador o o b si la casilla no se ha utilizado.

- **bottom-middle-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *inferior-media*. Puede tomar el valor *x* si la casilla está tomada por el jugador *x*, el valor *o* si la casilla está tomada por el jugador *o* o *b* si la casilla no se ha utilizado.
- **bottom-right-square:** Variable de tipo *char* que indica el estado de la casilla en la posición *inferior-derecha*. Puede tomar el valor *x* si la casilla está tomada por el jugador *x*, el valor *o* si la casilla está tomada por el jugador *o* o *b* si la casilla no se ha utilizado.

CAPÍTULO 3. ANÁLISIS ESTADÍSTICO E INFERENCIAL

Según la información contenida en el data-set, existe una probabilidad de aproximadamente un 65,34 % de que el jugador que inicia el juego, en este caso el jugador que marca con el símbolo x , lo gane. O visto de otra forma, existe la misma probabilidad de que el jugador que marca con o pierda. El símbolo de indiferente para este estudio pues solo se necesita saber el resultado final, si gana o no, el jugador que inicia la partida.

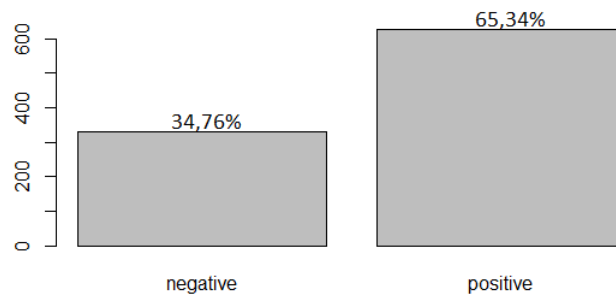


Figura 3.1: Clases

Así mismo, se muestra los porcentajes de partidas, donde x obtiene el triunfo, en las que se tiene el *tablero* completo o incompleto. Sin importar la secuencia de cada partida, se puede inferir con esto que en aproximadamente el 90 % de los casos, el jugador x en su primer turno posiblemente marcó una casilla que fue una de los tres símbolos que lo hizo ganar.

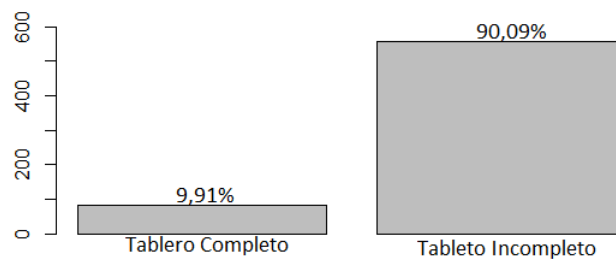


Figura 3.2: Estado final del *tablero*

A continuación se presentan 9 gráficos (uno por cada casilla del tic-tac-toe) en los cuales se indican los porcentajes de las frecuencias de cada símbolo en el estado final del juego (incluyen casillas que quedan sin utilizar). Estos gráficos solo contienen las partidas en que el jugador x gana, es decir, la clasificación es *positive*.

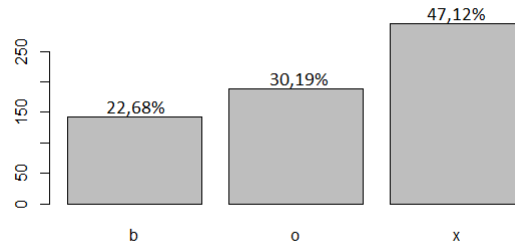


Figura 3.3: Esquina superior izquierda

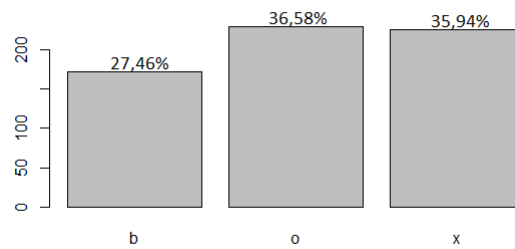


Figura 3.4: Esquina superior central

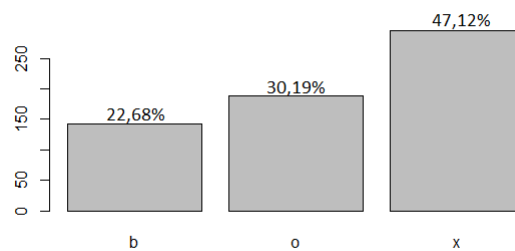


Figura 3.5: Esquina superior derecha

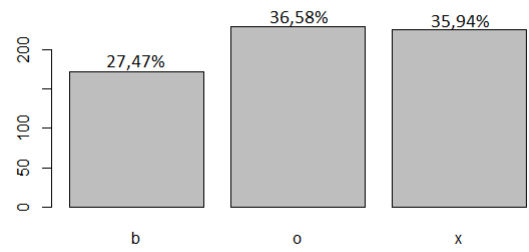


Figura 3.6: Esquina central izquierda

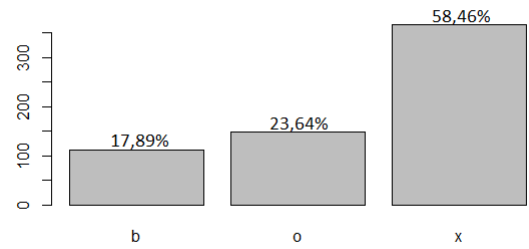


Figura 3.7: Esquina central central

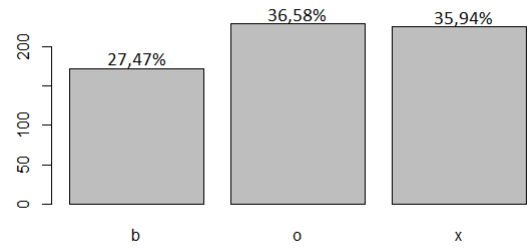


Figura 3.8: Esquina central derecha

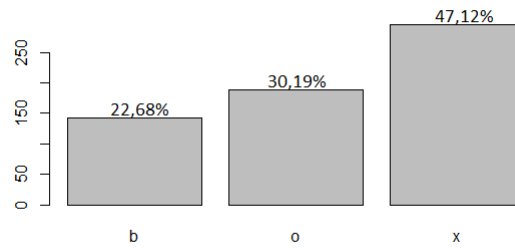


Figura 3.9: Esquina inferior izquierda

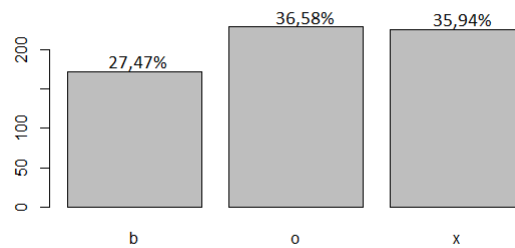


Figura 3.10: Central inferior

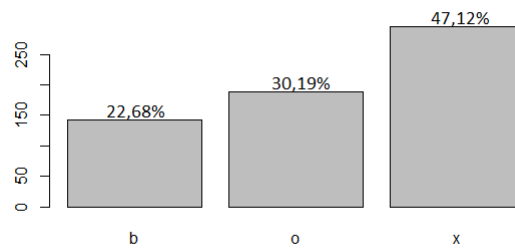


Figura 3.11: Esquina inferior derecha

Si enumeramos las casillas del 1 al 9, de izquierda a derecha y de arriba hacia abajo observamos ubicaciones 2, 4, 6 y 8 presentan los mismos porcentajes entre sí para cada símbolo posible. En estas casillas la frecuencia de ambos se aproxima bastante, con un diferencia de no más de 1 %.

El porcentaje de frecuencia del símbolo x aumenta en las esquinas (posiciones 1, 3, 7 y 9) disminuyendo a

su vez la cantidad de casillas marcadas con *o* y las en blanco.

Curiosamente la casilla central (5) se destaca por sobre las restantes ascendiendo a casi un 59 % su utilización en las partidas ganadoras. La casilla central, a diferencia de las otras, tiene 4 posibles *estados* ganadores; seguida por las esquinas que tienen 3 y por último las posiciones centrales de los costados que solo tienen 2. Es decir, se observa una tendencia a la utilización de las ubicaciones con más posibilidades de ganar y que efectivamente, según estas gráficas, en esta muestra se consigue la victoria.

Cuadro 3.1: Frecuencia de las jugadas ganadoras para *x* y *o*

	x ganador	o ganador
Diagonal (P1-P9)	83	50
Diagonal (P3-P7)	84	50
Fila Superior	78	36
Fila Centro	78	36
Fila Inferior	78	36
Columna Izquierda	75	36
Columna Centro	75	36
Columna Derecha	75	36

En el cuadro 3.1 se presentan las frecuencias de las posibles combinaciones ganadoras tanto para victorias del jugador *x* como el jugador *o*. Se observa que las columnas y filas (sea cual sea el símbolo) tienen probabilidades similares como combinaciones ganadoras; en contraste, las diagonales se llevan la mayoría de las victorias. Complementando a los gráficos, las diagonales contienen casillas de esquina y la central, que justamente son las con más posibles combinaciones ganadoras.

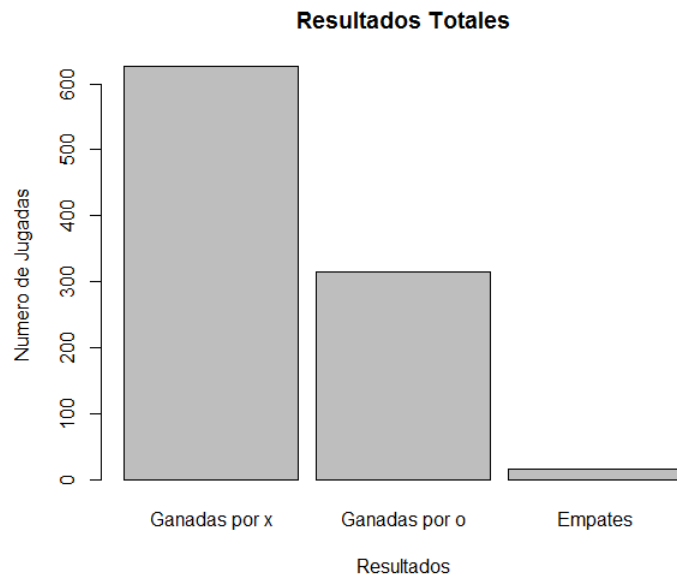


Figura 3.12: Resultados Totales por Jugada

Observando las jugadas posibles que concluyeron en una victoria para ambos jugadores, se percibe que hay un cierto numero de juegos o instancias pertenecientes a la clase "negative", que no se incluyen en aquellos en los que el jugador *o* no resulta ganador, por lo que se puede inferir que estas pertenecen a empates registrados. Por lo tanto, en la figura 3.12 se presenta un gráfico que muestra los resultados totales de los juegos, incluyendo lo analizado anteriormente.

CAPÍTULO 4. CONCLUSIONES

Como se ha mencionado antes (y como de ha aplicado en este presente informe), para obtener información adecuada del respectivo conjunto de datos es necesario conocer el origen de la información, es decir, el origen de las variables estudiadas. Específicamente en este caso las variables estudiadas son en su totalidad de tipo cualitativas, siendo una en particular de tipo nominal, la cual entregaba la clase de cada instancia, lo que permitía esclarecer el resultado obtenido en el tablero comprendido.

Una vez diferenciadas las clases, en particular se generaron pequeñas hipótesis respecto al *tic-tac-toe* que han sido aceptadas mediante un pequeño análisis estadístico-probabilístico. El análisis abarcó desde las combinaciones ganadoras más frecuentes, hasta las casillas que tienen más posibilidades de ganar. En esto se encontró que aquellas jugadas que tienen una mayor probabilidad de ganar son principalmente las que se generan en forma diagonal; lamentablemente, la base de datos no otorga información sobre el orden de las marcaciones que se ilustran en las diferentes posiciones del tablero en cada instancia, por lo que no se puede realizar un análisis más exhaustivo sobre la jugada ganadora en general.

Teniendo en cuenta las jugadas con más probabilidad de ganar y los datos obtenidos de la frecuencia de los jugadores en cada posición, se dilucidó que la posición central es la que más se ocupa. Generalmente según experiencias personales, esta es una casilla predilecta para el jugador que empieza, debido a su amplia gama de opciones de jugadas.

Si bien se pudieron establecer algunos análisis preliminares, la característica cualitativa de todas las variables no permitió realizar otro tipo de pruebas estadísticas que permitieran rescatar más hipótesis.

Finalmente con respecto a los objetivos inicialmente planteados, se ha logrado realizar un análisis preliminar acerca del tema de investigación, ya que a pesar de que el tic-tac-toe es un juego de conocimiento general, no se logra percibir las mismas inferencias que resultan a partir de la observación de un grupo general de juegos, por ende, para posteriores interpretaciones e inferencias acerca de los datos aquí estudiados, se tiene una base firme que permitirá desarrollar futuras investigaciones. En otras palabras, este documento contiene los cimientos que serán de gran importancia a lo largo del desarrollo del curso pues introduce a la base de datos asignada. Se espera que con esta base, en futuros trabajos, se pueda obtener conocimiento relevante que aporte estrategias en lo que es el juego en cuestión.

CAPÍTULO 5. 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/>.

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>