Algoritmos de Regresión Lineal aplicados al mantenimiento de un Datacenter

Federico Gabriel D'Angiolo[1], Iván Federico Kwist[1] Matias Loiseau[1], David Exequiel Contreras[1], Fernando Asteasuain[1]

[1] Universidad Nacional de Avellaneda, Avellaneda, Argentina. Departamento de Tecnología y Administración. Ingeniería en Informática.

fdangiolo@undav.edu.ar, ivankwist@hotmail.com.ar, matiasloiseau@gmail.com, dcontreraspastorini@cev.undav.edu.ar, fasteasuain@undav.edu.ar

Resumen. En el presente Trabajo se describe la aplicación de Algoritmos de Regresión Lineal al estudio del comportamiento climático de un Datacenter. Este análisis permite comprender cómo varían la temperatura y la humedad, con el objetivo de predecir cuál es el grado de ventilación adecuada que debe haber dentro del recinto para mantener operativos a los servidores y sistemas de cómputo que se encuentran.

Palabras claves: Inteligencia Artificial, Regresión Lineal, temperatura, humedad.

1 Introducción

En la actualidad se tiene acceso a una gran cantidad de datos que facilitan la automatización de sistemas y agilizan las tareas que tal vez, hasta hace unos años, resultaban laboriosas. Con el caudal de datos que se puede conseguir, resulta necesario el procesamiento de los mismos con el objetivo de obtener conclusiones sobre un comportamiento en especial o determinar qué tarea debe ser ejecutada para lograr un fin determinado, por eso, es que hoy en día los algoritmos de Inteligencia Artificial (IA), cobran mayor peso. Ejemplo de esto se da en el Trabajo "A Review at Machine Learning Algorithms Targeting Big Data Challenges" [1], donde se describe un análisis comparativo de los algoritmos de Aprendizaje Automático basados en grandes cantidades de datos (Big Data). Dentro de la IA, uno de los campos importantes es el Aprendizaje Automático o Machine Learning el cual, a su vez, se encuentra dividido en tres ramas importantes: Aprendizaje Supervisado, Aprendizaje No Supervisado y Aprendizaje Reforzado. Un ejemplo de lo comentado, se da en el Trabajo "Machine Learning for Engineering" [2], en donde se estudian las técnicas de Apendizaje Automático para resolver problemas de producción dentro de la automatización. En particular, en el presente Trabajo, se estudia el comportamiento dinámico que tienen las variables como temperatura y humedad dentro de un Datacenter y para esto se recurre entonces a los algoritmos de Aprendizaje

Supervisado, en especial, a la Regresión Lineal Simple y a la Regresión Lineal Multivaribale ya que el objetivo es poder analizar la relación de la temperatura y la humedad en distintos puntos del Laboratorio. Este análisis permite estudiar cómo se propaga el flujo de calor dentro del Datacenter para mejorar o aumentar la ventilación dentro del mismo permitiendo una mayor vida útil de los servidores. Como base para el desarrollo de ese trabado, se toma el caso de "Multiple Linear Regression to Improve Prediction Accuracy in WSN Data Reduction" [3] en donde se comenta la utilización de las técnicas de Regresión Lineal Simple y Múltiple aplicadas a una red de sensores. Resulta interesante este artículo porque no se limita a describir cada una de estas técnicas sino que se detalla cómo la Regresión Múltiple mejora ciertos aspectos de precisión que la Regresión Lineal no llega a abarcar. Por otro lado, se estudió el trabajo "Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal" [4], en donde se comentan los modelos de Regresión Lineal y Regresión Múltiple para modelar relaciones entre variables como la temperatura, luz, ph y oxígeno disuelto, con el objetivo de analizar cómo influyen dichas variables en el crecimiento de las microalgas. Lo importante del trabajo mencionado, con el que se describe aquí, es cómo adaptar estos modelos matemáticos a las variables sensadas. Por último, para el análisis general de Regresión se estudia el Trabajo: "Caracterización térmica de edificios aplicando el modelo de regresión lineal múltiple" [5], en el cual se presenta una metodología para evaluar del comportamiento térmico de una construcción representativa del centro bonaerense.

Para aplicar estos algoritmos, previamente se realiza un Dataset con valores de temperatura y humedad los cuales fueron obtenidos mediante sensores ubicados dentro del Datacenter. Estos valores se muestrean en el intervalo de meses de septiembre 2018 – marzo 2019 (https://gitlab.com/datasets_gi/dataset) para lograr un conjunto de datos robusto. Con estos, el objetivo es realizar dos estudios: el primero consiste en obtener un modelo lineal que relacione las temperaturas entre dos puntos distintos de la habitación, y el segundo consiste en obtener un modelo lineal entre la temperatura de uno de los servidores con la temperatura que existe en distintos puntos del Datacenter.

Actualmente existen Trabajos relacionados a la toma de datos y análisis de los mismos para automatizar, por ejemplo, una residencia. El objetivo aquí es obtener modelos que permitan estudiar cómo se propagan las variables mencionadas para mantener el correcto funcionamiento de los servidores de forma que su vida útil no se vea disminuida. El Datacenter bajo estudio se encuentra dentro del Laboratorio de Redes de la Carrera de Ingeniería Informática de la Universidad de Avellaneda. Allí se encuentran dos servidores con dos sistemas de refrigeración. En la sección 3 de este Trabajo, se detalla cómo se distribuyen los servidores dentro del recinto.

La distribución de este Trabajo se puede describir de la siguiente forma: en la sección 2 se realiza una introducción al tema de Algoritmos de Regresión, luego en la sección 3 se describe la relación de temperatura entre dos puntos del ambiente con el objetivo de analizar cómo se distribuyen la temperatura y la humedad en el ambiente mediante la obtención de un modelo matemático aproximado y, por último, en la sección 4, se describe la relación de temperatura y de humedad, entre uno de los

servidores con el ambiente, con el objetivo de estudiar cómo se ve afectado dicho servidor por las variables mencionadas. Este último también se enfoca en la obtención de un modelo matemático.

2 Algoritmos de Regresión

Desde el punto de vista matemático, la Regresión tiene dos significados: uno surge de la distribución conjunta de probabilidades de dos variables aleatorias mientras que el otro significado es empírico y nace de la necesidad de ajustar alguna función a un conjunto de datos. [6]

En el caso del presente Trabajo, resulta interesante el segundo punto de vista de este concepto ya que el objetivo es encontrar un modelo lineal que permita predecir la relación entre los distintos valores de temperatura y humedad obtenidos. Previamente a estudiar la Regresión Lineal, resulta conveniente realizar un estudio de linealidad sobre el conjunto de datos para obtener un indicio de que las variables tengan un cierto grado de correspondencia. En base a este primer objetivo, se utiliza el Coeficiente de Correlación para el cual se toma como criterio de diseño que, si el mismo resulta ser un valor mayor o igual 0,7, se pueden ajustar los datos mediante Regresión Lineal, teniendo en cuenta que puede exixtir cierto error. En caso de que el porcentaje resulte menor a dicho valor, la linealidad no será evidente y se podrá proceder a buscar algún modelo no lineal que mejor ajuste [7].

Para darle un marco matemático al análisis, se comienza definiendo a un Dataset como un conjunto de valores, que se puede describir de la siguiente forma:

$$X=x_1,x_2,\ldots,x_n \quad \text{donde } x_n \in R^n \tag{1}$$

Cada vector de entrada está asociado a un valor real y_i :

$$Y=y_1,y_2,\ldots,y_n \quad \text{donde } y_n \in R \tag{2}$$

Un modelo lineal se basa en la suposición de que es posible aproximar los valores de salida a través de un proceso de regresión basado en la regla:

$$f(x_i)=\alpha_0+\alpha_1 . x_1+\alpha_2 . x_2+\ldots+\alpha . x_m \tag{3}$$

$$A=\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_n \quad \text{resultan constantes} \tag{4}$$

En la expresión anterior, las constantes se encuentran mediante el Algoritmo de Regresión Lineal. Si la ec (3) resulta tener las constantes α_2 , α_3 , ..., α_m nulas, el modelo resulta ser de Regresión Simple mientras que, si estas constantes no son nulas, el modelo se denomina Regresión Multivariable. Con esto queda definido el modelo de Regresión Lineal para explicar a continuación, cómo se relaciona con el objetivo de este Trabajo [8].

3 Relación de Temperatura entre dos puntos del Datacenter

El primer estudio a realizar consiste en comparar la medición de temperatura de dos puntos distintos del Datacenter. Para esto, se presenta el diagrama de la distribución de los sensores:

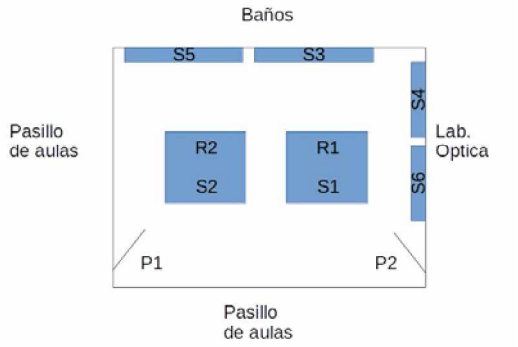

Figura 1. Layout del Datacenter y ubicación de los sensores

Siendo:

R1: Rack 1, donde se encuentra alojado el servidor N.º 1

R2: Rack 2., donde se encuentra alojado el servidor N.º 2

P1: Puerta que da al pasillo.

P2: Puerta que da al Laboratorio de Óptica.

S_1,S_2,S_3,S_4,S_5 y S_6: Sensores de temperatura y humedad, integrados.

Como se puede ver en la Fig.1, el sensor S_4, se encuentra ubicado de forma vertical con el S_6 con lo cual la finalidad de este estudio es encontrar, si existe, una relación lineal en el cambio de temperatura y de humedad de ambos sensores, es decir, si se observa una variación de temperatura (o humedad) en S_4 se debería ver la misma variación en S_6. Para lograr esto, se toma el dataset de los ocho meses comentados y se estudia un modelo de Regresión Lineal Simple.

Previo al estudio de Regresión Lineal, resulta conveniente analizar el coeficiente de correlación con el objetivo de verificar que existe una relación lineal entre los

datos. Al realizar este cómputo, se obtiene el siguiente valor para el coeficiente de correlación:

$$\rho = 0.696$$

El Coeficiente de Correlación toma un valor cercano a 0,7, para lo cual se puede admitir que el modelo mantiene cierta linealidad, es decir que mediante un análisis de Regresión Lineal Simple se obtendría un modelo como el siguiente:

$$T_6 = 10{,}78 + 0{,}64 \,.\, T_4 \quad \textbf{(5)}$$

El cual, gráficamente quedaría de la siguiente forma:

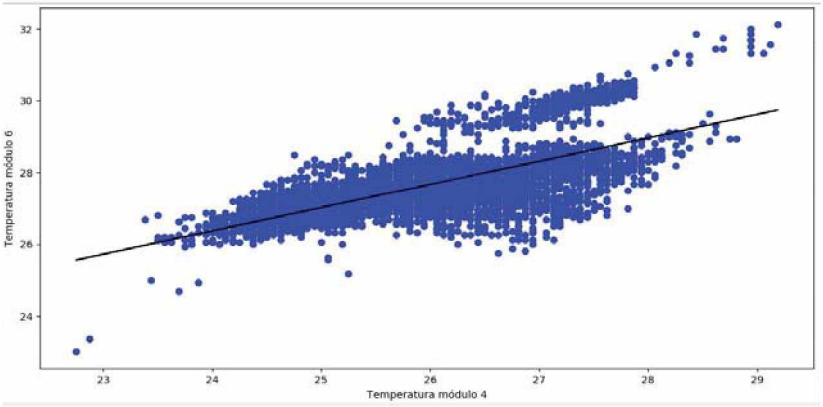

Figura 2. Regresión Lineal de temperatura entre el sensor S_6 y S_4

De la misma forma se realiza el estudio para el análisis de la humedad, obteniendo un coeficiente de correlación de:

$$\rho = 0.91$$

En este caso el Coeficiente de Correlación toma un valor de 0,91, para lo cual se puede admitir que el modelo mantiene una linealidad más estricta en comparación con el caso de la temperatura, es decir que mediante un análisis de Regresión Lineal Simple se obtendría un modelo como el siguiente:

$$H_6 = 7{,}37 + 0{,}79 \,.\, H_4 \quad \textbf{(6)}$$

El cual, gráficamente quedaría de la siguiente forma:

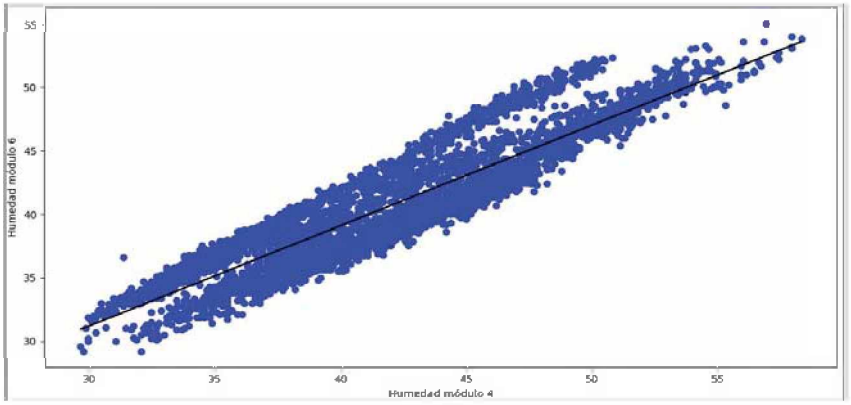

Figura 3. Regresión Lineal de humedad entre el sensor S_6 y S_4

Observando ambos análisis, se puede concluir que la relación de temperatura y humedad para ambos sensores mantiene cierta linealidad sobre todo en el caso de la humedad donde resulta ser más clara. Esto determina entonces que el flujo de calor se traslada linealmente desde S_4 a S_6. Es importante entonces tener en cuenta que una posible forma de ventilación resultaría estar cerca de S_6 (o en ese mismo lugar), para mantener refrigerado al servidor N.º 1.

4 Relación de temperatura entre el servidor y la del ambiente.

El objetivo de este segundo estudio es encontrar un modelo que permita predecir si la temperatura sobre uno de los servidores se ve afectada por la temperatura existente en los distintos puntos del Datacenter. Para esto, se toma como ejemplo el Servidor N°1, el cual tiene asociado el sensor S_1 y se propone el siguiente modelo de Regresión Multivariable:

$$T_1 = \alpha_0 + \alpha_2 . T_2 + \alpha_3 . T_3 + \alpha_4 . T_4 + \alpha_5 . T_5 + \alpha_6 . T_6 \quad (7)$$

El modelo propuesto por la ec. (7), supone que la temperatura del sensor S_1 depende linealmente de la temperatura que se encuentra alrededor del servidor la cual se encuentra cuantificada por los sensores S_2, S_3, S_4, S_5 y S_6 . Desde un punto de vista vectorial, se puede decir que T_1 es una combinación lineal de las otras temperaturas, tomadas a estas últimas como vectores. Si algunas de las constantes de la ec.(7),

resulta nula, evidencia que ese vector no tiene influencia sobre T_1, es decir, la temperatura en ese punto espacial no tiene ponderación suficiente.

Como se mencionó antes, previamente se observará el coeficiente de correlación entre las distintas variables para poder analizar si resulta eficiente aplicar Regresión lineal. Para esto, se tomando los coeficientes de correlación entre T_1 y cada una de las temperaturas mencionadas, se obtiene lo siguiente:

ρ 1,2 = 0 .77573
ρ 1,3 = 0 .45246
ρ 1,4 = 0 .71656
ρ 1,5 = 0 .81005
ρ 1,6 = 0 .78763

Para visualizar estas relaciones, se propone la siguiente imagen:

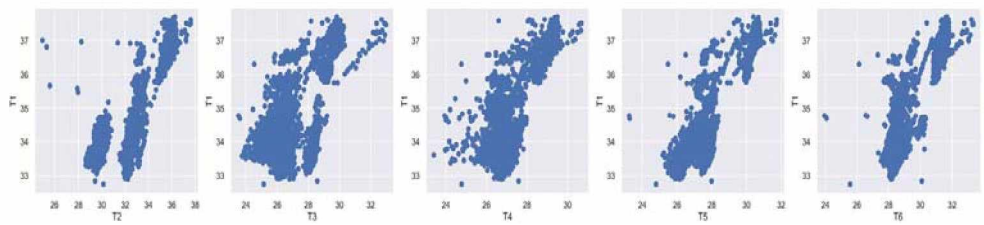

Figura 4. Gráficos de Correlación entre S_1 y los distintos sensores ubicados alrededor del servidor.

Como se puede observar, tanto por los coeficientes de correlación como por los gráficos, la relación no es exactamente lineal en todos los casos, lo cual se puede corroborar con los valores de los coeficientes obtenidos. Es decir, admitiendo que el modelo mantiene cierta linealidad para un coeficiente de ρ = 0,7, sería conveniente plantear un modelo de Regresión Lineal Múltiple.

Es importante contemplar el caso de la relación entre el sensor uno y el tres donde el factor de correlación no resulta tener un valor cercano al planteado como criterio.

En base a esta idea, la expresión matemática para el modelo, sería la siguiente:

$$T_1 = 9{,}49 + 0{,}18 \,.\, T_2 - 0{,}22 \,.\, T_3 + 0{,}07 \,.\, T_4 + 0{,}53 \,.\, T_5 + 0{,}28 \,.\, T_6 \qquad (8)$$

Luego para la humedad, se repite el mismo proceso obteniendo los siguientes valores del factor de correlación:

ρ 1,2 = 0 .73803
ρ 1,3 = 0 .91343
ρ 1,4 = 0 .58855
ρ 1,5 = 0 .90415

$$\rho\ 1{,}6 = 0\ .81064$$

Para visualizar estas relaciones, se propone la siguiente imagen:

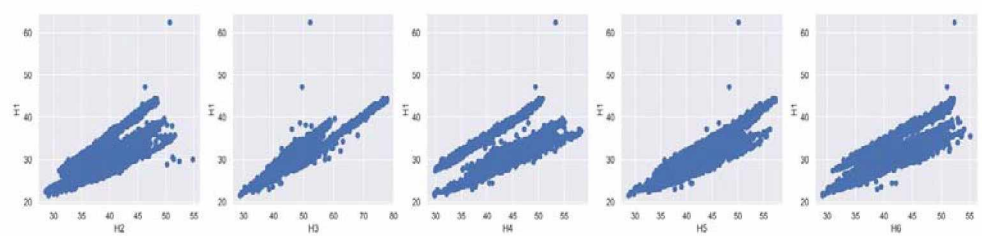

Figura 5. Gráficos de Correlación entre S_1 y los distintos sensores ubicados alrededor del servidor.

Como se puede observar, tanto por los coeficientes de correlación como por los gráficos, la relación de valores exhibe una linealidad más certera en comparación con el caso de la temperatura lo cual permite afirmar que sería coherente plantear un modelo de Regresión Lineal Múltiple.

En base a esta idea, la expresión matemática para el modelo, sería la siguiente:

$$H_1 = 3{,}16 - 0{,}01\,.\,H_2 + 0{,}43\,.\,H_3 + 0{,}28\,.\,H_4 - 0{,}26\,.\,H_5 + 0{,}16\,.\,H_6 \qquad (9)$$

Observando los modelos matemáticos de temperatura y humedad, se puede ver que el efecto del sensor números dos que pertenece al servidor dos, contiene constantes que no influyen sobre el servidor uno, es decir, que el servidor dos no impacta de manera importante en cuanto a temperatura y humedad se refiere, sobre el servidor uno.

Para corroborar el efecto de lo observado, se propone tomar el mismo Dataset y obtener un modelo de Regresión Lineal que no contemple el efecto del módulo dos. Con este objetivo, se repiten los mismos pasos y queda la siguiente expresión:

$$T_1 = 16{,}06 - 0{,}88\,.\,T_3 - 1{,}1\,.\,T_4 + 2{,}61\,.\,T_5 + 0{,}01\,.\,T_6 \qquad (10)$$

El cual comparado con el modelo original, el cual tiene en cuenta el efecto del sensor dos (sobre el servidor dos), se nota que el modelo modifica el peso de cada constante pero el efecto que tiene cada sensor sobre el del servidor uno, es similar al modelo obtenido anteriormente.

Este último análisis lleva a pensar que si se desea obtener un modelo más estricto, se puede tener en cuenta el efecto del módulo dos, sin embargo, se puede trabajar con un modelo menos robusto a los efectos de tener un menor costo computacional a la hora de obtener estas constantes aunque la aproximación no resulta tener la misma fidelidad que el modelo original.

Por último, es importante contemplar en el caso del modelo completo (donde interviene el módulo dos), se puede observar que tanto para la temperatura como para la humedad, los módulos tres y cinco son los que más influyen lo cual lleva a la conclusión de que si se deseara agregar ventilación al Laboratorio, se debería ubicar en lugares cercanos a donde se ubican estos módulos.

5 Conclusión

En base a estos estudios, se puede ver lo importante de analizar el Coeficiente de Correlación previamente al estudio de la Regresión Lineal. En el estudio de temperatura para el caso de Regresión Lineal Simple, el modelo obtuvo un Coeficiente de Correlación inferior comparado con el estudio para la humedad. En base a un criterio de diseño adoptado, se pudo concluir que las condiciones climáticas siguen un modelo lineal aproximado lo cual permite concluir que, para mejorar la ventilación del Datacenter, conviene agregar un sistema de refrigeración cerca del sensor $S_{6.}$

Continuando con el análisis de la Regresión Lineal Múltiple, también se hizo un análisis del Coeficiente de Correlación el cual fue analizado para cada una de las relaciones entre el servidor y los distintos puntos espaciales para observar dónde resulta haber un grado de linealidad mayor. En base a esto, se pudo ver que en el caso de la humedad el modelo resultó tener mayor linealidad en comparación con el caso de la temperatura, es decir, este estudio resultó ser similar al caso de Regresión Lineal Simple en el sentido de que la humedad proporciona un modelo totalmente lineal.

De este último estudio se desprende que, en caso de agregar una ventilación para mejorar el funcionamiento del servidor N.º 1, podría ser donde se sitúa el sensor S_6, para agregarle mayor énfasis a la climatización del Datacenter.

En el caso de ser más estrictos sobre el Coeficiente de Correlación en el caso del modelo de temperatura, resulta claro entonces que la Regresión Lineal no resulta adecuada para este análisis lo cual lleva a concluir que se puede realizar el estudio de otro tipo de Regresión que mejor se ajuste a estos datos.

Como trabajo a futuro, resulta interesante realizar una generalización de este tema aplicado a distintos escenarios, como por ejemplo, restaurantes, donde resulta importante una ventilación a los efectos de los clientes y el personal de trabajo, también en transporte donde resulta vital mantener determinada ventilación en las distintas épocas del año a los efectos de quienes adquieren estos servicios. Es decir, se podría sensar las variables estudiadas en este trabajo en cada uno de los escenarios mencionados para luego, mediante Regresión Lineal u otro tipo de Regresión, estudiar la generalización y obtener conclusiones. Esta comparación enriquecería la investigación dado que se extendería el campo a ambientes sociales donde la IA pueda tener mayor impacto.

Este Trabajo se realizó en base a los Proyectos de Investigación UNDAVCYT 2014 y PROAPI 2017, desde la Carrera de Ingeniería Informática de la Universidad

Nacional de Avellaneda, bajo el título de “Mantenimiento de parámetros del ambiente del Laboratorio de Redes y Sistemas de Computación mediante protocolos de IoT”.

Referencias

1. Abhinav Rathor., Manasi Gyanchandani.: A Review at Machine Learning Algorithms Targeting Big Data Challenges. ISBN: 978-1-5386-2361-9
2. Jeff Dyck.: Machine Learning for Engineering. Solido Design Automation. 978-1-5090-0602-1
3. Carlos Giovanni Nunes de Carvalho, Danielo Gonçalves Gomes, José Neuman de Souza and Nazim Agoulmine.: Multiple Linear Regression to Improve Prediction Accuracy in WSN Data Reduction. ISBN: 978-1-4577-1792-5
4. Arys Carrasquilla-Batista, Alfonso Chacón-Rodríguez, Kattia NúñezMontero, Olman Gómez-Espinoza, Johnny Valverde, Maritza Guerrero-Barrantes.: Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal . ISSN: 0379-3962
5. N. Muñoz., B.M. Marino., L.P.Thomas.: CARACTERIZACIÓN TÉRMICA DE EDIFICIOS APLICANDO EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE. ISBN: 978-987-29873-0-5
6. Canavos, G.: Probabilidad y Estadística. Aplicaciones y Métodos.. (1988). ISBN:968-451-856-0
7. Bonaccorso, G.: Machine Learning Algorithms. ISBN 978-1-78588-962-2.
8. Prateek, J.:Artificial Intelligence with Python.. BIRMINGHAM - MUMBAI. First published: January 2017. ISBN 978-1-78646-439-2