Revista Estudios en Educación.
Vol. 3, Nº 4, julio de 2020.
ISSN en línea 2452-4980 • pp. 97-124.

Modelos matemáticos de inteligencia artificial para analizar factores que afectan el desempeño escolar.

Mathematical models of artificial intelligence to analyze factors that affect school performance.

Ing. José Mauricio Baeza Díaz*

Resumen

El presente artículo analiza mediante modelos matemáticos de inteligencia artificial supervisada o Machine Learning distintos factores que afectan el desempeño escolar en un liceo técnico de la comuna de Talcahuano. El equipo de trabajo del establecimiento encargado de la calidad en la enseñanza estableció cinco preguntas claves para su interés, las cuales son respondidas mediante la analítica de algoritmos especializados que permiten la cuantificación de resultados, para esto se utilizó la información entregada de bases datos donde se incorporan encuestas de percepción y promedios de notas.

Palabras clave: inteligencia artificial, Machine Learning, modelos matemáticos, desempeño escolar.

Abstract

The present article analyzes, through mathematical models of supervised artificial intelligence or Machine Learning, different factors that affect school performance in a technical high school in the municipality of Talcahuano. The work team of the establishment in charge of teaching quality established five key questions for your interest, which are answered through the analysis of specialized algorithms that allow the quantification of results. For this, the information provided from databases where They incorporate perception surveys and grade point averages.

Keywords: artificial intelligence, Machine Learning, mathematical models, school performance.

* Chileno, Licenciado en Ciencias de la Ingeniería, Ingeniero Civil Industrial, Consultor de Inteligencia de Negocios (BI) y modelos de Inteligencia Artificial (AI), Concepción, Chile. Correo electrónico: josebaezad@gmail.com.

Introducción

Para generar una integración social adecuada, desde la etapa escolar a la etapa laboral o escolar a universitaria, se evalúa el nivel de aprendizaje, desarrollo personal y social de los alumnos en sus diferentes ciclos de escolaridad. Lo anterior lo evalúa la Agencia de Calidad de la Educación del Gobierno de Chile, pero las mejoras de calidad es tarea de cada establecimiento por esto el equipo de trabajo del Liceo Técnico de Talcahuano se embarcó en la tarea de encontrar las causas que afectar al rendimiento escolar, tomando los aspectos que ellos consideran importantes o relevantes según indicadores que se miden a nivel país y que forman parte de los instrumentos evaluativos.

Las evaluaciones estandarizadas son indicadores importantes, pero no son el único instrumento. Antes se contaba solo el Sistema de Medición de Calidad de la Educación, en adelante SIMCE, ahora existe también los Indicadores de Desarrollo Personal y Social conocidas por su sigla IDPS, estos miden áreas como autoestima académica, participación y formación ciudadana, entre otros, que reflejan una evaluación con una mirada más amplia de la calidad.

Estos indicadores tienen el propósito de ampliar la mirada de calidad y avanzar en el logro de una educación más integral para todos los niños, niñas y jóvenes del país. Por lo tanto, tienen un rol clave en la evaluación de calidad de la educación siendo un factor predominante para el desarrollo a nivel país (Henríquez, 2018: 9).

El equipo de psicólogos del liceo técnico en estudio, enfocan sus esfuerzos a la búsqueda de factores influyentes que guardan relación con los IDPS, por la experiencia profesional que poseen están convencidos de que un buen desarrollo personal y social dentro y fuera del establecimiento son aspectos relevantes a la hora de evaluar el desempeño escolar.

Se trabajará con la base de datos entregada por el establecimiento la cual está constituida por una encuesta de percepción realizada a los estudiantes y los promedios de notas Matemáticas, Lenguaje, Historia y promedios de notas general. Considerando todas las limitaciones que se puedan encontrar por falta de información o características especiales de los datos, como las limitaciones por la falta de variables que podrían ser significativas, con esto se intentará buscar la relación optima entre la muestra obtenida con el fin de dar respuestas a las interrogantes establecidas por el equipo de psicólogos, quienes para efectos de simplificar y optimizar la búsqueda de las causas en el bajo rendimiento escolar y relacionándolas con las temáticas de los IDPS, formularon las siguientes cinco preguntas que se responderán en el presente artículo; ¿Se puede dar uso a la

información extraída de la encuesta realizada a los estudiantes para analizar factores importantes que afectan la calidad de la educación en el establecimiento?, ¿Cuál es el promedio de notas que un alumno del liceo técnico debe obtener para ser considerado un buen rendimiento académico?, ¿Qué áreas debe mejorar el liceo técnico en estudio?, ¿El desempeño académico depende del sexo del alumno y el promedio de notas general?, ¿Cuál es la probabilidad de que la variable sexo influya en obtener un buen o mal rendimiento académico, y cuál es la probabilidad de que cada curso obtenga un buen rendimiento académico?.

En el desarrollo de esta investigación se usarán las técnicas de inteligencia artificial supervisada o Machine Learning mediante el software estadístico R.

Machine Learning (ML). Es decir, deseamos programar computadoras para que puedan "aprender" de la información disponible para ellos. En términos generales, el aprendizaje es el proceso de convertir la experiencia en experiencia o conocimiento. La entrada a un algoritmo de aprendizaje son datos de entrenamiento, que representan la experiencia, y el resultado es cierta experiencia, que generalmente toma la forma de otro programa de computadora que puede realizar alguna tarea (Shalev-Shwartz, 2014: 19).

Estableciéndose una relación directa entre los IDPS recopilados mediante la encuesta de percepción que se realizará a los alumnos, y promedios de notas por alumnos, como otra información que pueda ser relevante, para análisis en reuniones internas que apoyen el nuevo sistema de evaluación de aprendizaje interno, que se considerará métricas para toma de decisiones del establecimiento, de esta forma la ingeniería industrial será un apoyo relevante en soluciones de análisis de datos para el ámbito educativo. Por lo anterior los modelos matemáticos estarán en función de estadísticas aplicadas.

La estadística habrá de ser vista como un conjunto de métodos, técnicas y procedimientos para el manejo de datos, su ordenación, presentación, descripción, análisis e interpretación, que contribuyen al estudio científico de los problemas planteados en el ámbito de la educación y a la adquisición de conocimiento sobre las realidades educativas, a la toma de decisiones y a la mejora de la práctica desarrollada por los profesionales de la educación (Flores, 2003).

Definición de la problemática

El liceo técnico se encuentra en la categoría de insuficiente de acuerdo con la evaluación realizada por la Agencia de Calidad de la Educación en base a los IDPS y SIMCE. El director del establecimiento encomienda la tarea de buscar factores que influyan en el rendimiento escolar para establecer las bases de un diseño de trabajo que fortalezca el aprendizaje, por lo cual se establecerán bases para iniciar los trabajos de mejora en la calidad de la educación dentro del aula. Se busca responder analíticamente a cinco preguntas establecidas, las respuestas a estas servirán de bases para la toma de decisiones que utilizará el personal encargado con el fin de iniciar los cambios pertinentes y desarrollar su programa de trabajo interno. Considerando que el presente estudio se enfoca en responder y entregar las cinco preguntas establecidas, la toma de decisiones y medidas de cambios es exclusiva responsabilidad del equipo interno a cargo de área en el establecimiento.

Preguntas que se responderán en base a inteligencia artificial

Las siguientes preguntas son las que se estudiarán mediante modelos matemáticos y estadísticos avanzados de inteligencia artificial:

1. ¿Se puede dar uso a la información extraída de la encuesta realizada a los estudiantes para analizar factores importantes que afectan la calidad de la educación en el establecimiento?
2. ¿Cuál es el promedio de notas que un alumno del liceo técnico debe obtener para ser considerado un buen rendimiento académico?
3. ¿Qué áreas debe mejorar el liceo técnico en estudio?
4. ¿El desempeño académico depende del sexo del alumno y el promedio de notas general?
5. ¿Cuál es la probabilidad de que la variable sexo influya en obtener un buen o mal rendimiento académico, y cuál es la probabilidad de que cada curso obtenga un buen rendimiento académico?

Elementos de base de datos

Las bases de datos son el conjunto de datos que integran un mismo contexto dentro del estudio, pueden ser recolectados de una o varias fuentes, para el análisis en cuestión la base de datos posee 185 filas y se utilizarán 26 variables, extraídas desde una encuesta de percepción a los alumnos y los promedios de notas entregados. Estas se organizarán en grupos para una mayor comprensión, y facilitar el análisis.

Las siglas de denominación y el significado por cada variable se muestran a continuación:

CURSO: Curso.

SEXO: Sexo.

EDAD: Edad.

T.LICEO: Años en el Liceo Técnico de Talcahuano C-25.

X1: Siento que en mi liceo me tratan con respeto y se preocupan por mí.

X2: Siento que en mi liceo hacen lo posible por evitar o resolver los malos tratos, conflictos o peleas.

X3: En mi liceo, se destacan las cosas positivas tanto de mi curso como del resto de los cursos del establecimiento.

X4: En mi liceo, nos incentivan y nos apoyan en la realización de actividades constructivas, unidos como curso.

X5: En clases, profesores y profesoras nos hacen participar a todos y todas por igual.

X6: En mi liceo proponen metas exigentes, pero que yo soy capaz de lograr.

X7: Mis profesores y profesoras me guían para que yo pueda descubrir y proponer las soluciones a los problemas.

X8: Cuando me va mal en una prueba, el profesor o profesora me motiva para superarme.

X9: Cuando estoy teniendo dificultades en mis estudios, mis profesores y profesoras me ofrecen opciones o estrategias para poder superarlas.

X10: Cuando tengo algún problema o dificultad, se preocupan para ayudarme a encontrar la forma de solucionarlo.

X11: En mi liceo tengo la oportunidad de hacer actividades deportivas, artísticas o culturales que me gusten mucho.

X12: Pienso que en mi liceo podría realizar o desarrollar algo para lo que soy muy bueno o buena o en lo que me siento muy capaz.

X13: Las actividades que realizan los profesores y profesoras son atractivas para mí.

X14: Generalmente siento que tengo suficiente motivación y ganas de participar en clases.

X15: Creo que en mi liceo nos ofrecen apoyo de acuerdo a nuestras distintas necesidades.

X16: En mi liceo hay personas que me pueden apoyar si tengo algún problema personal.

suma.X: Sumatoria de todos los valores 1 de las variables Xi por alumno.

P.MAT: Promedio general de notas matemáticas.

P.LEN: Promedio general de notas lenguaje.

P.HIS: Promedio general de notas historia.

P.GEN: Promedio general de notas por alumno.

Distribución de variables por grupos

Al separar las variables por grupos estamos simplificando la comprensión de cada una de estas, según su importancia y características matemáticas. Los cuatro grupos definidos son:

Grupo 1: son las dieciséis variables extraídas de la encuesta de indicadores de autoestima académica y motivación escolar. Respuestas realizadas por los estudiantes donde seleccionan casillas según estén “Más en desacuerdo” o “Más de acuerdo”.

Son variables categóricas nominales, poseen distribución binomial.

Más en desacuerdo = 0

Más de acuerdo = 1

Grupo 2: Promedio de notas general, promedio de notas matemáticas, promedio de notas lenguaje, y promedio de notas historia.

Son variables del tipo cuantitativas continuas.

Promedio de notas general = P.GEN

Promedio de notas matemáticas = P.MAT

Promedio de notas lenguaje = P.LEN

Promedio de notas historia = P.HIS

Grupo 3: Variable dependiente, es la posibilidad de obtener buen rendimiento académico, y resultados óptimos en la prueba SIMCE.

Se conforma con la sumatoria del Grupo 1. El resultado de las respuestas por alumno es la suma de cada valor "Más de acuerdo" representada por el número 1, como máximo arrojará 16 puntos. Si las respuestas son mayores a 10 se considera como la probabilidad de obtener buenos resultados, en caso de que la sumatoria sea menor o igual a 10, existen posibilidades de fracaso a medida que YY se acerca a cero.

$\sum Xi > 10 \rightarrow YY=1$

$\sum Xi \leq 10 \rightarrow YY=0$

Siendo YY la variable dependiente que representa la posibilidad de obtener buen rendimiento académico, y Xi corresponde a cada una de las dieciséis variables binarias extraídas de la encuesta realizada a los alumnos.

Grupo 4: Es la información complementaria que los alumnos integran a la encuesta de indicadores de autoestima académica y motivación escolar.

Esta información corresponde a curso, sexo (Masculino, femenino), edad, años en el liceo.

Curso: Variable categórica nominal (compuesta de un valor entero y una letra).

SEXO: Variable binomial, donde Masculino = 1, y Femenino = 0.

EDAD: Variable cuantitativa discreta.

AÑOS EN EL LICEO: Variable cuantitativa discreta.

Modelos matemáticos seleccionados para análisis

Los modelos matemáticos son la base de la estructura estadística que se utilizará en los algoritmos de inteligencia artificial para dar respuestas de forma analítica a las interrogantes planteadas. Se usarán los siguientes modelos:

Regresión Lineal Simple

Considerando Y como variable dependiente y Xi como la variable independiente.

Una forma razonable de relación entre la respuesta Y, y el regresor Xi es la relación lineal.

$$Y = \beta 0 + \beta 1Xi+e$$

Donde, β0 es la intersección, β1 es la pendiente y e corresponde al error, distancia del punto a la recta de regresión.

Como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista, es decir, una Xi dada no siempre produce el mismo valor de Y. Como resultado, los problemas importantes en este caso son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de análisis de regresión se refiere a encontrar la mejor relación entre Y y Xi cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor Xi (Walpole, Myers, Myers & Ye, 2012: 389).

Regresión Lineal Múltiple

Considerando Y como variable dependiente y Xi como la variable independiente.

La complejidad de la mayoría de los mecanismos científicos es tal que, con el fin de predecir una respuesta importante, se requiere un modelo de regresión múltiple. Cuando un modelo es lineal en los coeficientes se denomina modelo de regresión lineal múltiple (Walpole, Myers, Myers & Ye, 2012: 443).

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 +\cdots+ \beta pXi + e$$

Donde Y es la variable dependiente, un conjunto de variables explicativas o regresoras X1, X2,...,Xi, y e corresponde a error, y β parámetros, miden la influencia que las variables explicativas tienen sobre el regrediendo.

Coeficiente de Correlación y Análisis P-Value

La correlación es en esencia una medida normalizada de asociación o covariación lineal entre dos variables. Esta medida o índice de correlación r puede variar entre -1 y +1, ambos extremos indicando correlaciones perfectas, negativa y positiva respectivamente. Un valor de r = 0 indica que no existe relación lineal entre las dos variables. Una correlación positiva indica que ambas variables varían en el mismo sentido. Una correlación negativa significa que ambas variables varían en sentidos opuestos (Vinuesa, 2016: 2).

El coeficiente de correlación con el que trabaja el software R Studio es el coeficiente de correlación de Pearson el cual calcula el valor de r.

$$\text{Coef. de correlación de Pearson}(r) = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(yi - \bar{y})}{(N-1)s_x s_y}$$

Donde (x, y) son variables cuantitativas continuas. “La correlación se define en términos de la varianza (s2) de las variables x e y, así como de la covarianza cov de x,y. Es por tanto una medida de la variación conjunta de ambas variables (cov(x,y))” (Vinuesa, 2016: 3).

Relación del valor de r:

Correlación despreciable: r < |0.1|

Correlación baja: |0.1| < r <= |0.3|

Correlación mediana: |0.3| < r <= |0.5|

Correlación fuerte o alta: r > |0.5|

Una vez que calculamos el coeficiente de correlación, es necesario identificar si es estadísticamente significativo, por lo que calculamos P-Value.

Si el P-Value es menor al nivel de significancia que nosotros escogemos, por ejemplo 5%, entonces el coeficiente es estadísticamente significativo.

Regresión Logística Simple

Considerando Y como variable dependiente y Xi como la variable independiente. “El valor predeterminado de la respuesta cae en una de dos categorías, Sí o No. En lugar de modelar esta respuesta Y directamente, la regresión logística modela la probabilidad de que Y pertenezca a una categoría particular” (Tibshirani, 2013: 130).

La regresión logística se utiliza para las tareas de clasificación: podemos interpretar h (x) como la probabilidad de que la etiqueta de x sea 1. La clase de hipótesis asociada con la regresión logística es la composición de una función sigmoide φsig: R → [0,1] sobre la clase de funciones lineales Ld. En particular, la función sigmoidea utilizada en la regresión logística es la función logística (Shalev-Shwartz, 2014: 126).

La fórmula de la función sigmoidal usada en regresión logística es:

$$y = \frac{1}{1 + e^{-x}}$$

Donde “y” representa la variable dependiente, “x” variable independiente, y “e” corresponde a la exponencial.

Naïve Bayes

El clasificador ingenuo de Bayes es un clasificador simple que se basa en el conocido teorema de Bayes. A pesar de su simplicidad, siguió siendo una opción popular para la clasificación de textos. Ahora en Naive Bayes, el algoritmo evalúa una probabilidad para cada clase, cuando se dan los valores predictores. E intuitivamente, podemos ir a la clase, que tiene la mayor probabilidad (Khan, 2017).

El clasificador Naive Bayes aplica el conocido teorema de Bayes para la probabilidad condicional, el cual es la base para la creación del clasificador, y requiere de variables categóricas.

$$P(A \cap B) = P(A, B) = P(A)P(B|A) = P(B)P(A|B)$$

$$\implies P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Donde: P(A) son las probabilidades a priori, P(B|A) es la probabilidad de B en la Hipótesis A, y P(A|B) son las probabilidades a posteriori.

Máquina de Soporte Vectorial

Las Máquinas de Vector Soporte se fundamentan en el Maximal Margin Classifier, que, a su vez, se basa en el concepto de hiperplano.

Clasificación binaria empleando un hiperplano: Cuando se dispone de n observaciones, cada una con p predictores y cuya variable respuesta tiene dos niveles (de aquí en adelante identificados como +1 y −1), se pueden emplear hiperplanos para construir un clasificador que permita predecir a que grupo pertenece una observación en función de sus predictores. Este mismo problema puede abordarse también con otros métodos (regresión logística, LDA, árboles de clasificación…) cada uno con ventajas y desventajas. La definición matemática de un hiperplano es bastante simple. En el caso de dos dimensiones, el hiperplano se describe acorde a la ecuación de una recta:

$$\beta 0+\beta 1x1+\beta 2x2=0$$

Dados los parámetros β0, β1y β2, todos los pares de valores x=(x1, x2) para los que se cumple la igualdad son puntos del hiperplano. Esta ecuación puede generalizarse para p-dimensiones:

$$\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp=0$$

y de igual manera, todos los puntos definidos por el vector (x= x1, x2, ... , xp) que cumplen la ecuación pertenecen al hiperplano.

Cuando x no satisface la ecuación:

$$\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp<0$$

o bien

$$\beta 0+\beta 1x1+\beta 2x2+...+\beta pxp>0$$

El punto x, cae a un lado o al otro del hiperplano. Así pues, se puede entender que un hiperplano divide un espacio p-dimensional en dos mitades. Para saber en qué lado del hiperplano se encuentra un determinado punto x, solo hay que calcular el signo de la ecuación (Amat, 2017).

Solución a preguntas realizadas por el equipo del liceo técnico

Pregunta 1.

¿Se puede dar uso a la información extraída de la encuesta realizada a los estudiantes para analizar factores importantes que afectan la calidad de la educación en el establecimiento?

Con esta primera pregunta, se formula la Hipótesis nula, la cual se debe cumplir para que sea factible la realización de todos los análisis posteriores, aquí se define la significancia de los datos recolectado. Si posee una alta significancia podemos decir que la base de datos entregada es útil para la realización del estudio.

Hipótesis nula

H0: Los indicadores de autoestima académica y motivación escolar están asociados con las notas de promedios entregados en la base de datos.

Sin la comprobación de H0 no será posible realizar análisis que puedan entregar alguna respuesta, ya que H0 determina la correlación entre los resultados de la encuesta aplicada a los alumnos y los promedios de notas.

La información que adquirimos de la encuesta Indicadores de autoestima académica y motivación escolar (Grupo 1 de variables) es netamente de apreciación, no representan datos duros, es la visión que tienen los alumnos sobre aspecto ambientales respecto a su educación. Para utilizar esta información es pertinente evaluar correlaciones con información concreta, para esto podemos utilizar los promedios de notas (Grupo 2 de variables), ya que este grupo de variables es el conjunto de valores que muestra de forma concreta y clara los bueno y malos resultados académicos.

Resumen de datos Grupo 2.

Figura 1. Gráfica box plot variables Grupo 2. Software estadístico R.

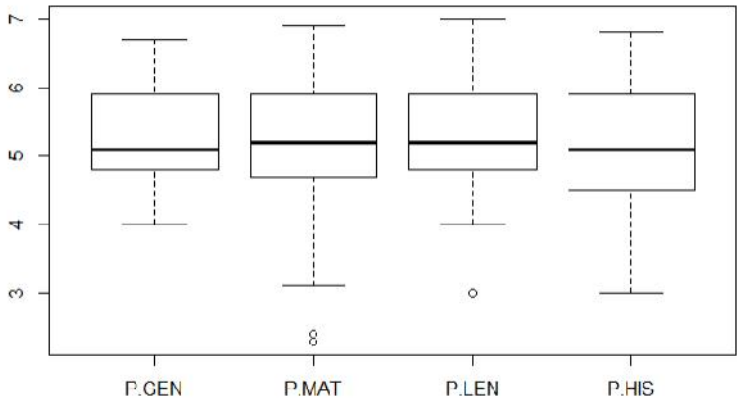


Figura 2. Resumen variables Grupo 2. Software estadístico R.

P.GEN	P.MAT	P.LEN	P.HIS
Min. :4.000	Min. :2.300	Min. :3.000	Min. :3.000
1st Qu.:4.800	1st Qu.:4.700	1st Qu.:4.800	1st Qu.:4.500
Median :5.100	Median :5.200	Median :5.200	Median :5.100
Mean :5.314	Mean :5.245	Mean :5.321	Mean :5.217
3rd Qu.:5.900	3rd Qu.:5.900	3rd Qu.:5.900	3rd Qu.:5.900
Max. :6.700	Max. :6.900	Max. :7.000	Max. :6.800

Si observamos el grafico Box-Plot las medias, son cercanas entre sí, desde 5.21 al 5.32, esta relación permite establecer otra posible conclusión para el Grupo 2 de variables.

Entonces si logramos determinar una fuerte correlación en este grupo, podríamos definir solo uno del promedio como el representante del conjunto completo, y así simplificar cálculos de análisis futuros. Se evalúa el promedio general (P.GEN) como postulante a este conjunto de datos.

Regresión Lineal Múltiple

La evaluación se desarrolla con el modelo de Regresión Lineal Múltiple, para este caso se considera P.GEN como variable dependiente, y P.MAT, P.LEN y P.HIS como variable independiente.

Figura 3. Resultados de algoritmo Regresión Lineal Múltiple. Software estadístico R.

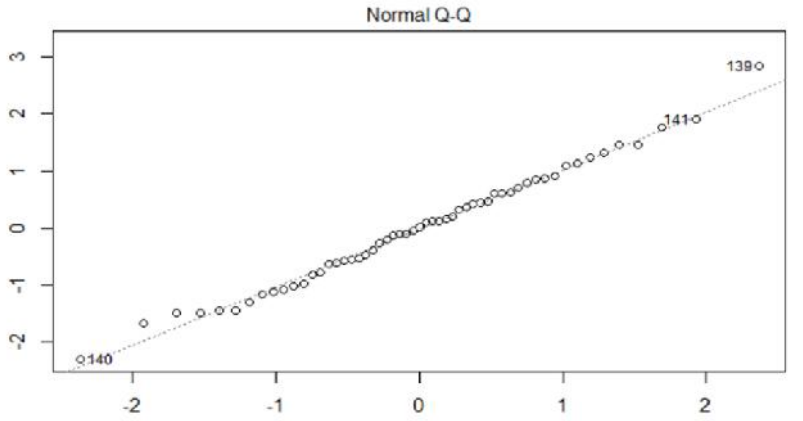
```
Call:
lm(formula = P.GEN ~ ., data = test_set)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48111 -0.17077  0.00626  0.16173  0.66097

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04603    0.23337   4.482 4.20e-05 ***
P.MAT        0.25361    0.05578   4.547 3.38e-05 ***
P.LEN        0.29664    0.06800   4.363 6.27e-05 ***
P.HIS        0.26861    0.06606   4.066 0.000166 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2496 on 51 degrees of freedom
Multiple R-squared:  0.8728,    Adjusted R-squared:  0.8653
F-statistic: 116.6 on 3 and 51 DF,  p-value: < 2.2e-16
```

Se observa que el P-Valor de todas las variables independientes es menos a 0.05, específicamente se encuentra entre 0 y 0.001 considerado con un alto nivel de significancia, esto se complementa con el valor de R^2=0.8728 lo que es muy cercano a 1, y el P-Valor general < 2.2e-16 respaldando la alta significancia anterior.

Gráfico Q-Q

Según el gráfico Normal Q-Q verifica que los residuos siguen una distribución normal, lo cual es una suposición de regresión lineal, muestra que los puntos están sobre la línea y=x esto representa una distribución normal de los residuos. En este caso muestra una relación aceptable.

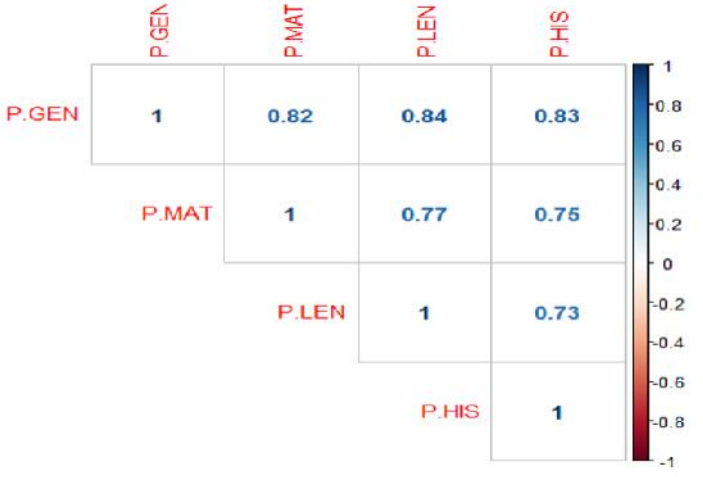
Figura 4. Gráfica de Normalidad Q-Q. Software estadístico R.



Correlación

Queda claro que las variables del Grupo 2 están altamente relacionadas, pero para asegurar que P.GEN es la variable más representativa se realiza un análisis de correlación, entre todas las variables.

Figura 5. Gráfica de correlación. Software estadístico R.



Se observa que P.GEN tiene la mayor correlación respecto a todas las otras variables (la que está más cercana al valor 1, relacionadas con las otras variables). Por lo que P.GEN puede representar claramente al Grupo 2, y de ahora en adelante se podrá tomar solo P.GEN para los distintos análisis.

Con los cálculos anteriores se establece, que las notas de promedio general (P.GEN), es la variable más significativa entre el grupo de variables de promedios (P.GEN, P.LEN, P.MAT, P.HIS), para utilizarla como representante entre estas variables, la cual servirá para analizar la variable dependiente de la hipótesis nula respecto al promedio general (P.GEN). Esto se realizará mediante la Regresión Logística Simple.

Regresión Logística Simple

Se puede determinar de forma fácil al aplicar Regresión Logística Simple, considerando que YY es la variable dependiente con distribución binomial (lo que hace a la regresión logística el modelo óptimo), y P.GEN es la independiente.

Figura 6. Resultados de algoritmo Regresión Lineal Simple. Software estadístico R.

```
Call:
glm(formula = vdependiente ~ P.GEN, family = "binomial", data = rlsimple)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1370  -0.8392   0.2633   0.7891   2.0758

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.0817     1.8936  -6.380 1.77e-10 ***
P.GEN         2.3373     0.3654   6.397 1.58e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 255.25  on 184  degrees of freedom
Residual deviance: 185.91  on 183  degrees of freedom
AIC: 189.91

Number of Fisher Scoring iterations: 5
```

Se observa que el P-Valor es menos a 0.05, específicamente se encuentra entre 0 y 0.001 considerado con un alto nivel de significancia entre ambas variables, de esta forma podemos concluir que YY y P.GEN guardan una alta relación, cumpliéndose la hipótesis nula.

Solución de hipótesis nula

H0: Los indicadores de autoestima académica y motivación escolar están asociados con las notas de promedios entregados en la base de datos.

Se concluye que están altamente asociadas, ya que el P-Valor de la Regresión Lineal se encuentra entre 0 y 0.001 lo que es una significancia importante entre las variables YY (posibilidad de obtener buen rendimiento académico) y promedio general de notas (P.GEN).

Pregunta 2.

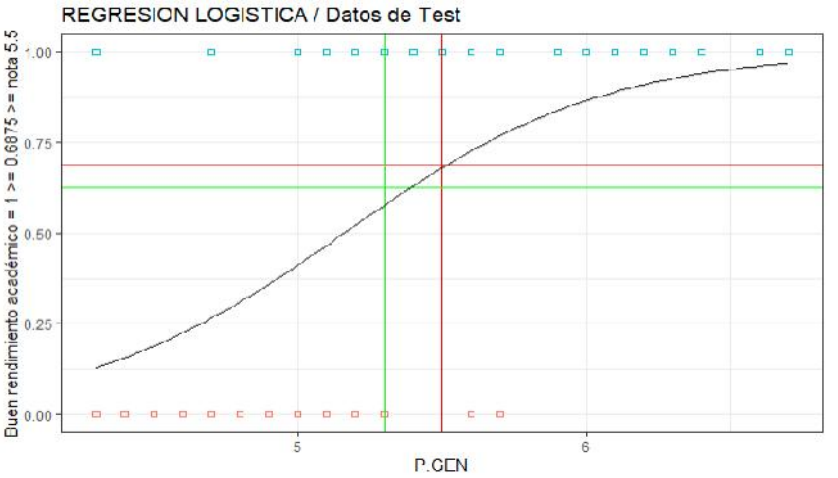
¿Cuál es el promedio de notas que un alumno del Liceo Técnico de Talcahuano debe obtener para ser considerado un buen rendimiento académico?

Existen varias formas para responder a esta pregunta, en este caso se puede aprovechar el uso de la regresión logística simple para delimitar ejes de intersección y encontrar el punto que determina el promedio notas.

Regresión Logística Simple con intersección de líneas a ejes paralelos

Si YY>10=1, las posibilidades de éxito se encuentran cuando se responde de forma positiva igual o sobre el 68.75%. YY≤10=0, las posibilidades de fracaso están desde el 62.5% hacia abajo. Donde: YY es la variable dependiente que corresponde a la posibilidad de obtener buen rendimiento académico, por los tanto 1= éxito y 0= fracaso.

Figura 7. Gráfica Regresión Logística Simple con intersección de líneas a ejes paralelos. Software estadístico R.



Las líneas horizontales verde y roja, representan YY= (0,1). Donde verde = 0 y rojo = 1, con 62.5% de límite máximo y 68.75% límite mínimo respectivamente.

Las líneas verticales verde y roja, se marcan según la intersección que generan las líneas horizontales respecto a la gráfica sigmoidal, el vertical rojo indica el promedio mínimo a obtener para "comenzar a tener buenos resultados o éxito en el aprendizaje", y la línea vertical de color verde indica desde "qué promedio hacia abajo, se considera un mal resultado". P.GEN de la base de datos tiene una media de 5.3, y P.GEN de la Regresión Logística muestra que se debe obtener como mínimo 5.5 para comenzar a obtener buenos resultados.

Desempeño de los resultados

Precisión: los datos clasificados correctamente son 73,9% lo que corresponde a un porcentaje aceptable para este tipo de análisis.

Sensibilidad: en esta métrica encontramos el 60% de los valores correctamente predichos positivos.

Especificidad: 90,4% corresponde a los correctamente predichos negativos.

Tasa de error: los datos clasificados de forma incorrecta son el 26%.

Pregunta 3.

¿Qué áreas debe mejorar el liceo técnico en estudio?

Considerando que H0 se ha comprobado, podemos establecer la significancia alta para obtener resultados desde la encuesta de indicadores de autoestima académica y motivación escolar. Mediante gráficos de barras, y ordenándolos por el método de Pareto, se establecen las áreas críticas.

En el siguiente gráfico de barras se observan las totalidades de respuestas 0 y 1 para todas las variables de la encuesta de percepción.

Figura 8. Gráfico de barras de variables críticas.

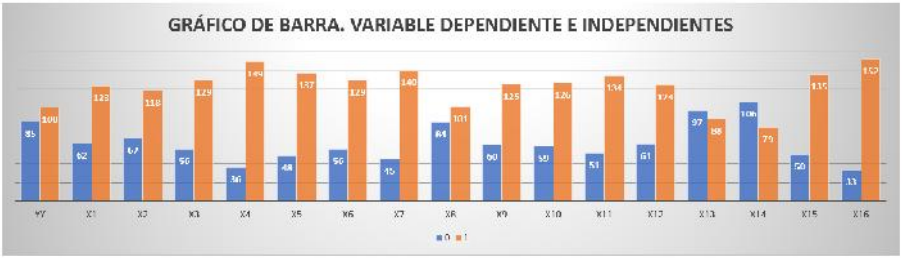


Figura 9. Diagrama de Pareto.

Según el orden realizado por el gráfico de Pareto, claramente tenemos tres tendencias marcadas de respuestas negativas o valor igual a cero.

Desde la más importante a menos importante, las variables a mejorar son:

X14: Generalmente siento que tengo suficiente motivación y ganas de participar en clases.

X13: Las actividades que realizan los profesores y profesoras son atractivas para mí.

X8: cuando me va mal en una prueba, el profesor o profesora me motiva para superarme.

Factores para mejorar

Los alumnos no encuentran atractiva la forma en que los profesores presentan las actividades, lo que generar desmotivación y malas notas, al mismo tiempo después de obtener malos resultados en las evaluaciones no sienten que los motiven para superarse. Factor de escases motivacional predominante.

Pregunta 4.

¿El desempeño académico depende del sexo del alumno y el promedio de notas general?

Se realiza la clasificación entre sexo y promedio de notas, respecto a la variable dependiente YY, esto se alcanza mediante la Regresión Logística Múltiple.

Figura 10. Regresión Logística Múltiple. Software estadístico R.

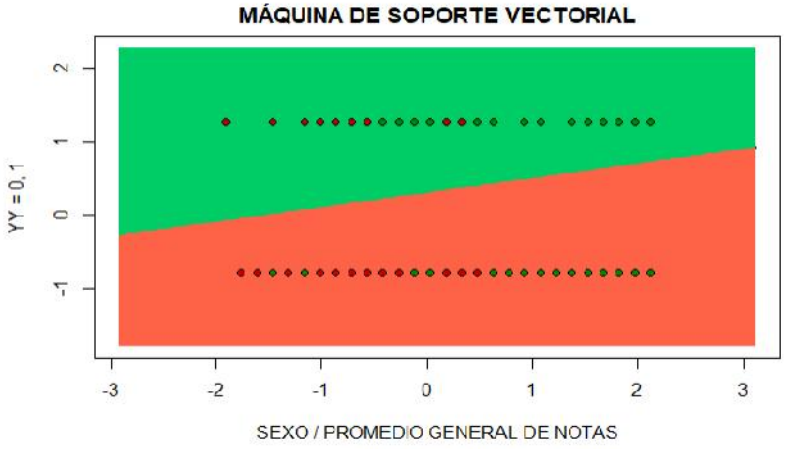
```
Call:  glm(formula = YY ~ ., family = binomial, data = training_set)

Coefficients:
(Intercept)        P.GEN         SEXO
     0.3272       1.5369      -0.3047

Degrees of Freedom: 138 Total (i.e. Null);  136 Residual
Null Deviance:      191.8
Residual Deviance: 143.3        AIC: 149.3

Call:
glm(formula = YY ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8591  -0.8987   0.2825   0.9061   1.9235

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3272     0.2141   1.528    0.127
P.GEN         1.5369     0.2898   5.303 1.14e-07 ***
SEXO         -0.3047     0.2065  -1.476    0.140
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 191.82  on 138  degrees of freedom
Residual deviance: 143.34  on 136  degrees of freedom
AIC: 149.34

Number of Fisher Scoring iterations: 5
```

Como anteriormente se explicó, la variable dependiente binomial que define el éxito o fracaso comparada con el promedio general de notas (YY y P.GEN) posee una alta significancia. Pero la significancia entre YY y sexo, es imperceptible, superior al 0.05 permitido, se encuentra entre 0.1 y 1. Por lo que no existe relación entre YY y sexo, consecuentemente tampoco se relaciona la variable sexo con el promedio general.

Gráfica Máquina de Soporte Vectorial

Resulta complejo poder mostrar gráficamente la regresión logística múltiple, por lo que se ha generado un gráfico de máquina de soporte vectorial (Support Vector Machines, SVM), que mediante la clasificación permite mostrar claramente la situación anteriormente expuesta.

Figura 11. Gráfica de Máquina de Soporte Vectorial. Software estadístico R.



El gráfico de SVM presenta dos grandes zonas, la superior de color verde y la inferior roja, estas zonas muestran la pertenencia de la variable dependiente YY respecto a las probabilidades de éxito o fracaso del desempeño escolar, si YY=1 la zona es de color verde, y si YY=0 la zona es de color rojo. Cada punto representa un alumno, dependiendo si es femenino o masculino, los colores de los puntos son rojo o verde respectivamente. Y el orden que se observa respecto al eje X es la posición de promedio general por alumno en orden ascendente.

En la gráfica se observan los puntos de colores (sexo femenino y masculino) mezclados en las dos áreas separadas, por lo que podemos concluir que el sexo no es un diferenciador para alcanzar un buen rendimiento académico o buenos promedio de notas, en caso de que los colores rojos y verdes se encontrasen ordenados y separados por color se definiría la existencia de significancia, lo que no ocurre en este caso.

Desempeño de los resultados

Precisión: los datos clasificados correctamente son el 76%, es un porcentaje aceptable, el modelo se puede utilizar.

Sensibilidad: los valores correctamente predichos positivos es el 64%.

Especificidad: los valores correctamente predichos negativos son el 90,4%, es un buen porcentaje.

Tasa de error: la tasa de datos clasificados incorrectamente es 23,9%, esta tasa es correcta ya que debe contener un porcentaje bajo.

Pregunta 5.

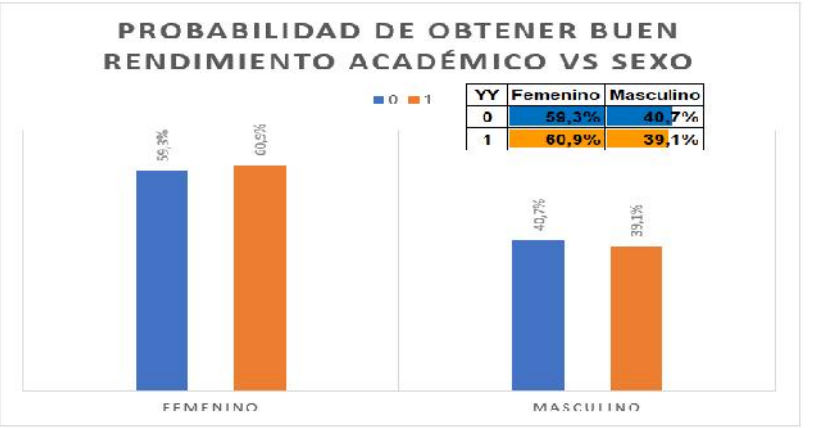
¿Cuál es la probabilidad de que la variable sexo influya en obtener un buen o mal rendimiento académico y cuál es la probabilidad de que cada curso obtenga buen rendimiento académico?

Tanto la variable dependiente, y las independientes son categóricas, una de las formas de poder dar respuesta a la pregunta formulada, es mediante el clasificador de Naïve Bayes, este es un modelo reconocido por la eficiencia y respuestas claras que entrega cuando debemos analizar solo variables discretas.

Análisis variable SEXO para su significancia en el rendimiento académico

El algoritmo de Naïve Bayes muestra en el Output la probabilidad de que el alumno sea de sexo femenino o masculino cuando la variable dependiente que representa la posibilidad de obtener buen rendimiento académico es de respuesta positiva o negativa, uno o cero respectivamente.

Figura 12. Gráfica Naïve Bayes para variable SEXO vs YY.



Lo observado radica netamente en el número de alumnos con sexo femenino (114) y sexo masculino (71), ya que estas diferencias son las que guardan relación con los resultados de Naïve Bayes, la variable SEXO no marca diferencial por el contexto de ser femenino o masculino, sino por la cantidad.

No existe significancia entre la variable sexo y la variable dependiente que representa la posibilidad de obtener buen rendimiento académico.

Como se observó en la gráfica, la fluctuación de la variable dependiente que representa la posibilidad de obtener buen rendimiento académico no genera un cambio relevante a lo que respecta el sexo, ya que esta variable en el caso del sexo femenino y sexo masculino no presenta cambios relevantes.

Análisis variable curso para su significancia en el rendimiento académico

Este análisis busca los cursos que entregan las probabilidades más altas en obtener buen rendimiento académico.

El equipo PIE consta de distintos tipos de profesionales relacionados al área de la educación, en su mayoría psicólogos, cada uno de estos se encuentran a cargo de dos o más cursos, quienes utilizarán esta información para realizar clusters internos con el fin de identificar aquellos cursos con buenos resultados, para analizar aquellos factores diferenciadores que se pueden replicar en aquellos cursos con probabilidades más bajas.

La siguiente ilustración muestra el output del algoritmo Naïve Bayes para los diferentes cursos involucrados en el estudio, para las probabilidades asociadas a respuestas positivas o negativas (uno o cero respectivamente) en relación con la variable dependiente que representa la posibilidad de obtener buen rendimiento académico.

Figura 13. Respuesta algoritmo Naive Bayes para variable CURSO. Software estadístico R.

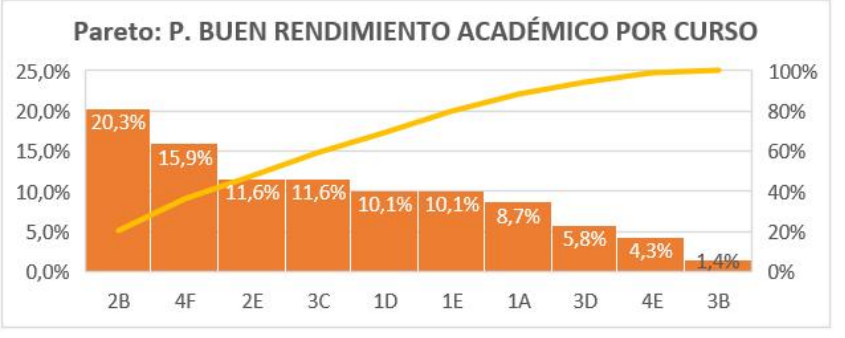
```
    CURSO
Y         1A         1D         1E         2B         2E
  0 0.08474576 0.11864407 0.10169492 0.05084746 0.15254237
  1 0.08695652 0.10144928 0.10144928 0.20289855 0.11594203

         3B         3C         3D         4E         4F
0.10169492 0.08474576 0.06779661 0.16949153 0.06779661
0.01449275 0.11594203 0.05797101 0.04347826 0.15942029
```

Las variaciones probabilísticas que entrega Naïve Bayes, identificando cursos con probabilidades cercanas en relación con la probabilidad de obtener buen o mal rendimiento académico, como en el caso de los cursos de primero medio (1A, 1D, 1E), como también aquellos cursos que presentan la probabilidad de obtener un mal rendimiento académico (3B, 4E), y en respuesta a la pregunta en cuestión se observan aquellos cursos que presentan probabilidad de éxito como es el caso del 2B, 3C y 4F.

Es importante, mediante otro análisis identificar cuáles serán aquellos cursos que se considerarán para observar los factores positivos que se pueden replicar.

Utilizando el gráfico de Pareto de forma inversa, para características con rasgos positivos, identificando las probabilidades que representa la posibilidad de obtener buen rendimiento académico, los cuales se muestran en el siguiente gráfico.

Figura 14. Gráfica de Pareto, Probabilidad de obtener un buen rendimiento académico por curso.



Para la variable CURSO, es importante considerar estos resultados en futuras mejoras, sobre todo en el caso de fracaso, en los cursos 4° Medio E, y 2° Medio E (ver figura 13) quienes, presentan el mayor aporte probabilístico para YY=0.

Desempeño del algoritmo de Neïve Bayes

Precisión: 98,9% es una alta precisión, casi la totalidad de los datos están clasificados correctamente.

Sensibilidad: los valores correctamente predichos positivos corresponden al 98%.

Especificidad: la totalidad de los valores predichos negativos está correctamente clasificado, esto es el 100%.

Tasa de error: el 1% de los datos fue clasificado incorrectamente.

Conclusiones

Tras el análisis de datos con los modelos matemáticos de inteligencia artificial supervisada buscando respuestas a preguntas que estableció el equipo de psicólogos del liceo técnico en estudio, se encontraron áreas para trabajar y mejorar, partiendo de temas como baja motivación, y baja percepción de esta por parte de los alumnos. Como parámetro cuantitativo, la necesidad de aumentar el promedio de notas general para todos los niveles, de 5,3 a un mínimo de 5,5, lo que podría establecerse como uno de los objetivos de mejora, en el cual deberán trabajar los profesionales del establecimiento, como también deberán aumentar, potenciar y expandir a todas las áreas la capacidad de desarrollar un ambiente motivador a los alumnos, buscar estrategias donde estos puedan percibir que están siendo apoyados en los distintos ámbitos por el Liceo y sus profesores. Esta tarea radica directamente en el aula, donde el profesor y alumno son los protagonistas.

Se debe prestar mayor énfasis a aquellos cursos de más alta tendencia a la desmotivación educacional, lo que desemboca en bajo rendimiento académico. Es importante crear cuadros comparativos entre cursos por nivel, y general. Los cursos con probabilidad de obtener mejor rendimiento analizados mediante Naïve Bayes son un claro ejemplo y de estos se puede extraer experiencias positivas a replicar.

En definitiva, el mayor valor que entrega el algoritmo desarrollado es obtener resultados de precisión al generar el entrenamiento, y asegurar estos resultados en base a la matriz de confusión. El desarrollo del algoritmo servirá como plantilla al momento del control (nuevos muestreos), que son de responsabilidad de los profesionales del liceo y quienes determinarán la necesidad de realizar este control de forma semestral o anual considerado como lo ideal. Las medidas posteriores en base a resultado, son de exclusiva responsabilidad de los psicólogos educacionales que están directamente relacionados con estas mejoras, así como lo fueron las preguntas realizadas para extraer su respuesta mediante el análisis de la base de datos estudiada.

Bibliografía

Acevedo, A. & Linares, M. (20/06/2012). El enfoque y rol del ingeniero industrial para la gestión y decisión en el mundo de las organizaciones. Revista de la Facultad de Ingeniería Industrial 15(1), 9 - 24.

Amat, J. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Recuperado de: https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines.

Gil Flores, J. G. (2003). La estadística en la investigación educativa. Revista de Investigación Educativa, 21(1), 231-248.

Henríquez, C. (2018). Los Indicadores de desarrollo personal y social en los establecimientos educacionales chilenos: una primera mirada. Recuperado de: http://archivos.agenciaeducacion.cl/estudios/Estudio_Indicadores_desarrollo_personal_social_en_establecimientos_chilenos.pdf.

Khan, R. (2017). Naïve Bayes Classifier: Theory and R example. South Dakota State University. Recuperado de: https://rpubs.com/riazakhan94/naive_bayes_classifier_e1071.

McAfee, A. & Brynjolfsson, E. (2017), Machine, Platform, Crowd: Harnessing Our Digital Future, Cambridge, Estados Unidos: Norton & Company.

Pande, P.; Neuman, R. & Cavanagh, R. (2004), Las claves prácticas de Seis Sigma, Madrid, España: McGraw-Hill.

Shalev-Shwartz, S. (2014), Understanding Machine Learning: from theory algorithms, Cambridge, Inglaterra: Cambridge University Press.

Statnikov, A.; Hardin, D. & Guyon, I. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Recuperado de: www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines.

Tibshirani, R. (2013), An Introduction to Statistical Learning with Applications in R, New York, Estados Unidos: Springer.

Vinuesa, P. (2016). Tema 8 - Correlación: teoría y práctica. Recuperado de https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.pdf.

Walpole, R. E.; Myers, R. H.; Myers, S. L. & Ye, K. (2012), Probabilidad y estadística para ingeniería y ciencias, Ciudad de México, México: Pearson.

Forma de citar este artículo

Baeza, J. (2020). Modelos matemáticos de inteligencia artificial para analizar factores que afectan el desempeño escolar. Revista Estudios en Educación, Vol. 3(4), 97–124, Santiago, Chile: Universidad Miguel de Cervantes. En: http://ojs.umc.cl/index.php/estudioseneducacion/index.

Fecha de recepción: 27/02/2020.

Fecha de aceptación: 08/05/2020.