Extracción de datos y estadísticas

Escuela: EETP N.612 “Eudocio de los Santos Giménez”. Coronda, Santa Fe.

Proyecto: ASISBIOM (Asistencia Biométrica)

Autor: Joaquín Gómez

Fecha de actualización 03 - 08 - 2024

Índice

Introducción

En el proyecto proponemos un sistema de asistencia electrónico a través de datos biométricos (implementación con huella digital), el cual presenta las siguientes ventajas:

- Conteo de asistencias automatizado.
- Exactitud en la extracción de horarios de entrada y salida.
- Automatización a la hora de la creación de planillas.
- Mejoras en el rendimiento de la escuela.
- Prevención de la falsificación o errores.
- Reducción de gastos relacionados a los medios tradicionales para contar asistencias, planillas, etc.

- Entre otras ventajas mencionadas en la documentación oficial.

Dentro de este documento se aprovechará una de las ventajas, que es el conocimiento de datos de entrada y salida de alumnos (horarios de llegada) para medir la puntualidad y rendimiento de la escuela en éste sentido.

Porcentaje de asistencias

El total de inasistencias de n alumnos es la suma de las inasistencias

$$I_n = \sum_{k=1}^{n} I_k$$

donde I_k es la inasistencia del alumno k. Un ciclo lectivo tiene (aproximadamente) 190 días de clase. Definimos C como la cantidad de días hábiles. La cantidad de asistencias es

$$K_a = n \cdot C - I_n$$

que sería la cantidad total de días hábiles por cada alumno, menos la suma de inasistencias. Entonces, el porcentaje de asistencias es

$$\%A = \left(\frac{K_a \cdot 100}{n \cdot C} \right)$$

Suponiendo que cada alumno tiene 10 inasistencias en la escuela, si un ciclo lectivo tiene 190 días hábiles y hay 500 alumnos, entonces:

$$K_a = n \cdot C - I_n = 500 \cdot 190 - 500 \cdot 10 = 90,000$$

$$\%A = \left(\frac{90,000 \cdot 100}{95,000} \right) \approx 94.736842105$$

Porcentaje de tardanzas

De la misma manera, podemos calcular el porcentaje de tardanzas. Deje que

$$T_n = \sum_{k=1}^{n} T_k$$

sea la suma de todas las tardanzas. Análogamente podemos definir

$$K_t = K_a - T_n$$

como la cantidad de asistencias puntuales sobre el total de asistencias. Entonces, el porcentaje de *puntualidad* es:

$$\%P = \frac{K_t \cdot 100\%}{K_a}$$

Para una escuela con 500 alumnos, suponiendo que cada alumno tiene 15 tardanzas, y la cantidad de asistencias total es la del ejemplo anterior (90,000) entonces:

$$K_t = 90,000 - 15 \cdot 500 = 82,500$$

$$\%P = \frac{82,500 \cdot 100\%}{90,000} \approx 91.666666667$$

Esto quiere decir que del total de asistencias, en el 91.7% los alumnos fueron puntuales.

Horario de llegada promedio

Para extraer el horario de llegada promedio, suponemos que tenemos un conjunto

$$H = \{h_{prom_1}, h_{prom_2}, h_{prom_3}, ..., h_{prom_n}\}$$

donde h_{prom_n} es el horario de llegada promedio del alumno n. Suponiendo que

$$h_{prom_n} = \frac{\sum_{k=1}^{K_{a_n}} h_k}{K_{a_n}}$$

donde h_k es el horario de llegada del alumno n y K_{a_n} es la cantidad de asistencias del alumno. El promedio de los horarios de llegada es

$$h_{prom} = \frac{\sum_{k=1}^{n} h_{prom_k}}{n}$$

Esto nos daría el horario de llegada promedio (hay que tener en cuenta que el horario debe estar en un formato de número entero, por ejemplo, minutos $[m]$ desde las 00:00$[hs]$).

Ordenamiento de alumnos por horario de llegada

Si definimos una función definida como la diferencia en el horario de llegada del alumno en el día respecto del horario de llegada esperado, llamémosla $h(t)$, donde t se mide en días. La función nos da como salida la diferencia entre el horario de llegada esperado, y el horario de llegada real del alumno

$$h(t) = H_t - H_{llegada}$$

donde H_t es el horario esperado de llegada en el día t. De esta manera, si el alumno llega tarde $h_n s(t) < 0$, si el alumno es puntual $h_n s(t) \approx 0$ y si llega temprano $h_n s(t) > 0$.

Si suponemos que $h_n s(t)$ es continua y es derivable en $0 \leq t \leq K_{a_n}$, donde K_{a_n} es el total de asistencias del alumno n. Entonces

$$I_h = \int_0^{K_{a_n}} h_n(t)dt$$

es el índice de puntualidad del alumno. Para visualizar esto, podemos graficar una función cualquiera, entonces el área bajo la curva será positiva si el alumno llega temprano; si el alumno llega temprano la mitad de los días y el resto de los días llega tarde, entonces $I_h \approx 0$; si el alumno llega tarde siempre, entones $I_h < 0$.

Como los datos que nosotros obtendremos no nos dan funciones continuas, y hacer una aproximación polinómica es incorrecto ya que la distribución de los horarios de llegada no es uniforme; optaremos por una suma que aproxime este valor:

$$I_n \approx \Delta x \cdot \sum_{k=1}^{K_{a_n}} h_n(k)$$

Donde $\Delta x = \frac{K_{a_n}}{C}$ donde C es la cantidad de días hábiles en el ciclo lectivo.

La definición de Δx es necesaria ya que, suponga que un alumno a_1 tiene 180 asistencias, entonces el puntaje de este alumno debe ser mayor al de un alumno a_2 con 170 asistencias. Entonces:

$$\Delta x_{a_1} > \Delta x_{a_2}$$

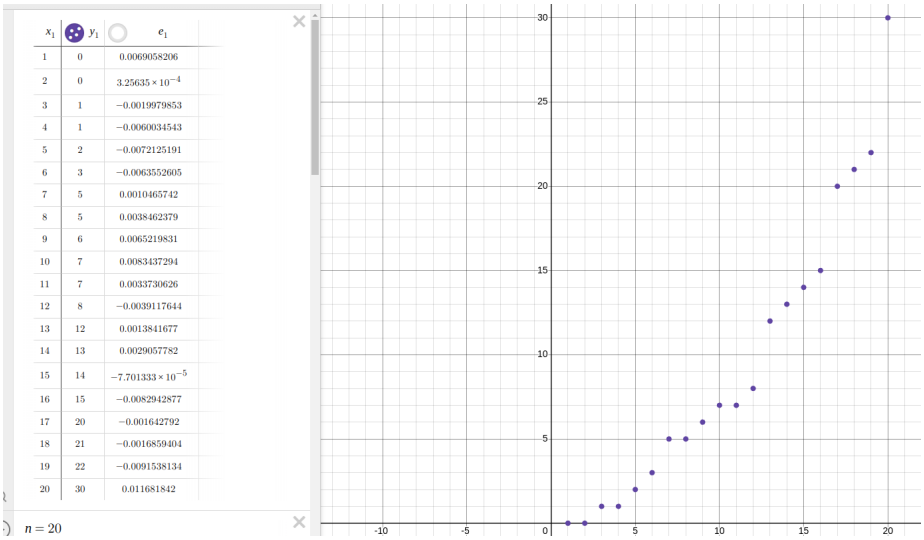
suponiendo que ámbos alumnos son igual de puntuales, el puntaje $I_1 > I_2$. En otras palabras, la calificación no solo se dará por puntualidad, sino que además por el número de asistencias totales.

Para ordenarlos bastaría con hacer una tabla donde utilizamos el valor de I_n como índice.

Distribución de puntualidad

En base a esto podemos graficar una curva de Lorenz.

Suponemos que tenemos una lista de valores, con alumnos de 1 al n, ordenados en base a I_n de menor a mayor, esto quiere decir que el alumno con el puntaje más bajo es el alumno 1, y así.



x_1	y_1	e_1
1	0	0.0069058206
2	0	3.25635×10^{-4}
3	1	−0.0019979853
4	1	−0.0060034543
5	2	−0.0072125191
6	3	−0.0063552605
7	5	0.0010465742
8	5	0.0038462379
9	6	0.0065219831
10	7	0.0083437294
11	7	0.0033730626
12	8	−0.0039117644
13	12	0.0013841677
14	13	0.0029057782
15	14	-7.701333×10^{-5}
16	15	−0.0082942877
17	20	−0.001642792
18	21	−0.0016859404
19	22	−0.0091538134
20	30	0.011681842

(los valores de y_1 corresponden I_n)

En este caso despreciamos valores negativos, suponiendo que todos los alumnos son puntuales o llegan temprano. Si tenemos valores negativos (alumnos que llegaron más tarde), transformaremos los valores de y_1 sumando $y_1 - S[1]$, siendo S la lista de valores, solo si el primer valor es negativo.

Realizamos la sumatoria acumulada de los valores de I_n

$$S_n = \sum_{k=1}^{n} I_k$$

si evaluamos la suma con $n = \{1, 2, 3, ..., n\}$ nos dará una lista S de números corespondiente a la suma acumulada hasta el alumno n de los valores de I_n

Alumno	Índice de puntualidad
1	I_1
2	I_2

Alumno	Índice de puntualidad
...	...
k	I_k

Donde I es el Índice de puntualidad y A es el alumno n, evaluando la suma:

$$S_1 = \sum_{k=1}^{1} I_k$$

$$S_2 = \sum_{k=1}^{2} I_k$$

$$S_3 = \sum_{k=1}^{3} I_k$$

...

$$S_n = \sum_{k=1}^{n} I_k$$

luego dividimos la lista obtenida por el valor más alto de I_n. Si graficamos los puntos $(\frac{k}{n}, \frac{S}{S_n})$ para cada alumno k formarán la curva de Lorenz.

La curva de Lorenz es la representación gráfica de la distribución de una variable de interés, como puede ser la renta o los ingresos, en una población, o en nuestro caso, la diferencia entre el horario de llegada de los alumnos y el horario de llegada estipulado.

en la imagen representamos la suma como una función de x, donde x es el número de iteraciones en la suma. Podemos graficar los puntos en $0 \leq x \leq 1$ dividiendo el índice del alumno x_1 por el número de elementos de la lista n.

Luego vamos a aproximar estos puntos, modelando una función cúbica.

Una curva perfectamente distribuida, es decir, cada alumno entra al mismo horario, se representaría con $f(x) = x$, o sea, una recta con pendiente 1. Si revisamos la definición anteriormente dada, esto es cierto, ya que si tenemos una lista $l = \{1, 1, 1, 1..., 1\}$, el resultado sería $r = \{1/n, 2/n, 3/n, ..., n/n\}$. Es decir, una recta a 45 grados. (Véase la figura 5 y 6).

La curva nos expresa el porcentaje de alumnos que llega a un porcentaje de esa diferencia de tiempo mencionada. Para explicarlo mejor, tenemos que el 50% de alumnos, representad0 en $g(0.5) = 0.147906270632$, esto quiere decir que el 50%

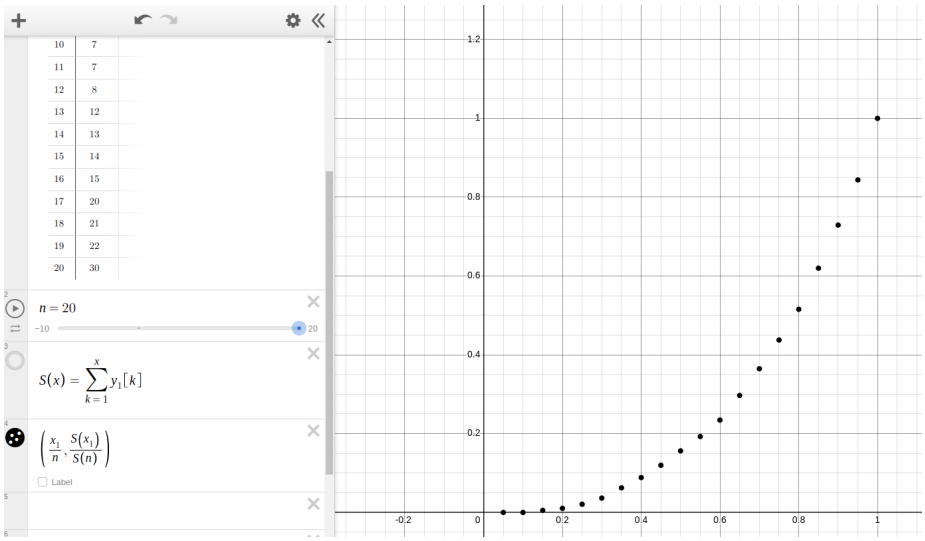


Figure 1: Conjunto de puntos normalizados

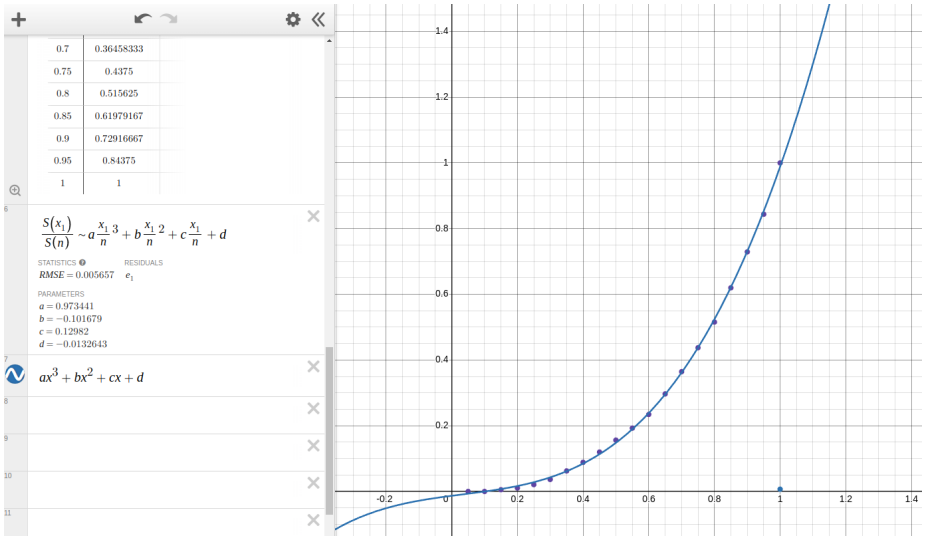


Figure 2: Modelado de una función cúbica para la curva de Lorenz

de los alumnos llega a un 14.7% de la diferencia de horario más alta de la tabla (30 min.). En otras palabras, el 50% de los alumnos llega con ~4.437 minutos de diferencia. En el caso analizado hay que tener en cuenta que ningún alumno llegaría tarde ya que los valores analizados de I_n son positivos. En este caso la parte más *baja* de la curva representaría el porcentaje de alumnos que llega más tarde, ya que son los que tienen menor puntaje, la parte más *alta* son aquellos que tienen una puntuación más alta (llegaron más temprano).

Podemos calcular la "desigualdad" existente entre los datos computados y una distribución perfecta, a continuación se explica cómo y qué significa.

Cómo medir la puntualidad en una escuela

Es importante dentro de una institución la puntualidad de sus alumnos. Para analizar estos datos podemos hacer uso de varias técnicas. Anteriormente se observó la distribución de la diferencia en los horarios de llegada de los alumnos. Por si solo esto nos dice el porcentaje de alumnos que llega a un porcentaje del horario de llegada, pero no nos vale para definir la puntualidad de la escuela.

Para realizar otras mediciones sobre la puntualidad, la distribución o desigualdad en la puntualidad, vamos a hacer uso del cálculo estadístico.

Distribución normal de puntualidad

La distribución normal de una variable aleatoria, como en este caso, la puntualidad o diferencia de horario de llegada de los alumnos, es muy común para este tipo de datos los cuales si bien presentan una aleatoriedad considerable aún conservan una cierta distribución uniforme.

La distribución normal se representa mediante una función de densidad de probabilidad que llamaremos f. Las propiedades de esta función de densidad de probabilidad son:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

La distribución normal de una variable aleatoria X es un miembro de la familia de funciones

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)š/2\sigma^2}$$

donde la letra griega σ (sigma) representa la desviación estandar y μ (mu) es el valor promedio de la variable. Lo que nos interesa analizar a nosotros es la distribución de la puntualidad de los alumnos, por lo que tomaremos $\mu = diferencia\ promedio\ en\ el\ horario\ de\ llegada$. La desviación estandar

σ nos indica qué tan dispersos están los datos con respecto al promedio, a continuación se realiza una explicación práctica.

Si tenemos un set de datos de la altura de personas del sexo masculino cuyo valor promedio es $\bar{a} = 175cm$ y queremos calcular cuánto "varían" las alturas, es normal tomar $\bar{a} - a_i$ donde a_i es la altura de la persona i. Hay dos posibilidades, o $\bar{a} - a_i < 0$ o $\bar{a} - a_i > 0$, si queremos calcular la media aritmética $\sigma = \sum_{i=1}^{N} (\bar{a} - a_i)$ el resultado sería incorrecto ya que algunos valores de a_i estarían por encima de la media, resultando en términos negativos, por lo que éste no nos reflejaría la "variación" que buscamos. Para ello tomamos la media cuadrática, lo que nos va a asegurar que los valores de $\bar{a} - a_i$ sean siempre positivos. Según la definición de media cuadrática, la desviación estándar estaría dada por:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\bar{a} - a_i)^2}$$

donde $\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a_i$ es la **media muestral**. Se introduce $N - 1$ en el denominador para evitar el sesgo en la estimación de la varianza. Si tenemos una función de probabilidad f entonces podemos hacer uso del cálculo diferencial.

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx}$$

donde $\mu = \int_{-\infty}^{\infty} x f(x) dx$ es la **media poblacional** y f es la función de densidad de probabilidad. Con la notación integral lo que hacemos es multiplicar el cuadrado de la variación de la variable respecto al promedio por la probabilidad de esa variable. En la notación sumatoria se realiza una sumatoria discreta, el término $1/N$ nos indica que cada punto tiene la misma probabilidad de ocurrir (una entre N), pero en la notación integral contamos con la función de densidad de probabilidad definida que agrega un "peso" a la variable. El valor promedio viene dado por la integral definida en el intervalo donde $f > 0$, del producto entre valor de la variable analizada y la probabilidad ligada a la misma.

La gráfica de una función de densidad de probabilidad tiene forma de "campana" donde el valor medio es $\max f(x) = f(\mu)$, es decir que la función está centrada en μ. El area bajo la curva de $f(x)$ indica la probabilidad de que X variable suceda, es decir, si tenemos una función de densidad de probabilidad $f(h)$ donde h es la diferencia en el horario de llegada de los alumnos, si tenemos dos valores h_1 y h_2, la probabilidad de que un alumno llegue con una diferencia de horario entre h_1 y h_2 se denota $P(h_1 < X < h_2)$:

$$P(h_1 < X < h_2) = \int_{h1}^{h2} f(h) dh$$

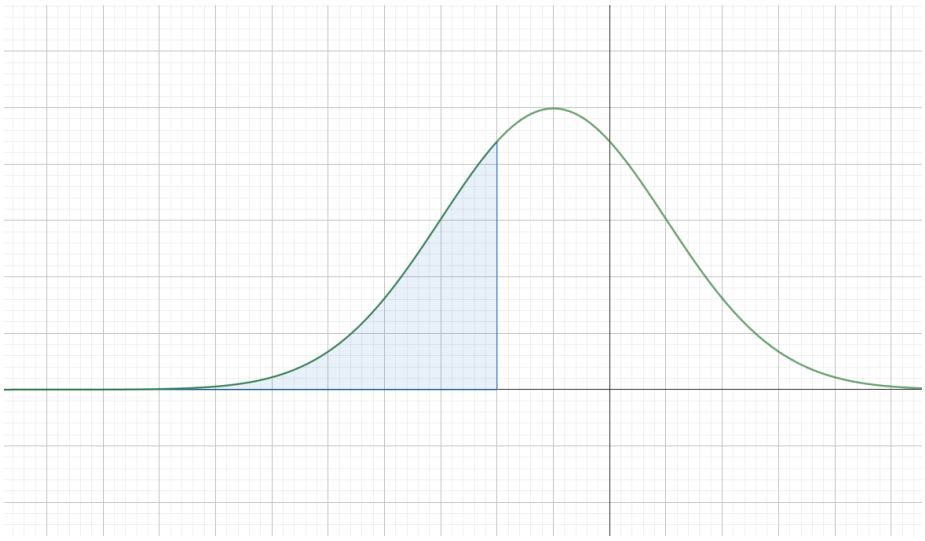


Figure 3: Representación gráfica de la integral dada

Podemos ver que la función está mayormente del lado negativo de la abcisa, esto debido a que la mayoría de los alumnos llega temprano. A su vez comprobamos que $\max f(x) = f(\mu)$ donde $\mu < 0$, es decir, en promedio los alumnos llegan temprano.

Índice de Gini

El índice de Gini compara la curva de Lorenz que generamos, con una distribución perfecta, lo que hacemos es hallar el área entre la curva $f(x)$ y $g(x)$:

En esta imagen comparan los ingresos de una población, sin embargo vamos a investigar la desigualdad en los horarios de llegada de los alumnos. Por ejemplo, en una escuela que está bien ordenada, el total de alumnos debería llegar un horario similar, en este caso la curva se va a asemejar a la "línea de igualdad", mientras que en el caso de una escuela que tiene un porcentaje de alumnos que llega muy temprano y otro que llega muy tarde, le corresponde una curva más empinada. Es necesario aclarar que esta curva no mide si una escuela es puntual o no, lo que nos indica es si los alumnos de dicha escuela llegan al mismo tiempo, es decir, si los alumnos llegan siempre tarde la curva será casi recta, y podemos pensar que la escuela tiene una buena puntualidad aunque esto no sea cierto. Por eso debemos comparar todos los datos, incluyendo sobre todo el promedio del horario de llegada, el cuál nos dará una condición para determinar la puntualidad de la escuela.

Continuando con lo anteriormente dicho, el índice de Gini nos sirve para medir la desigualdad, éste índice tiene un rango entre $0 \leq x \leq 1/2$ ya que puede ser a

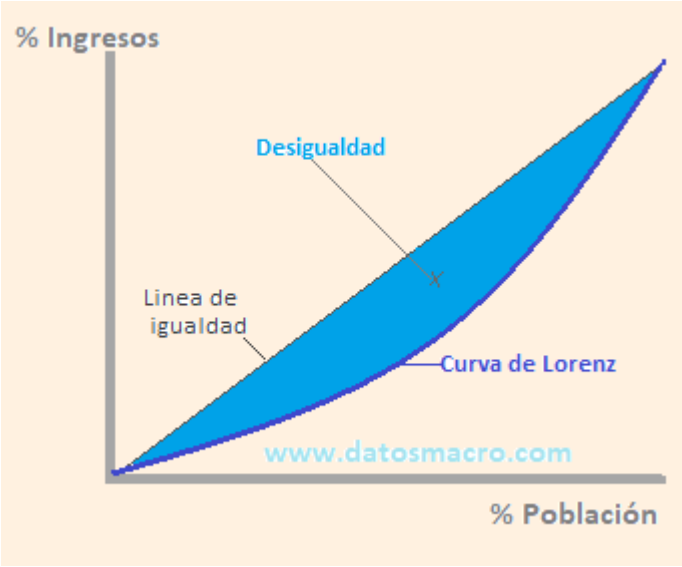


Figure 4: Representación gráfica de la desigualdad en la distribución

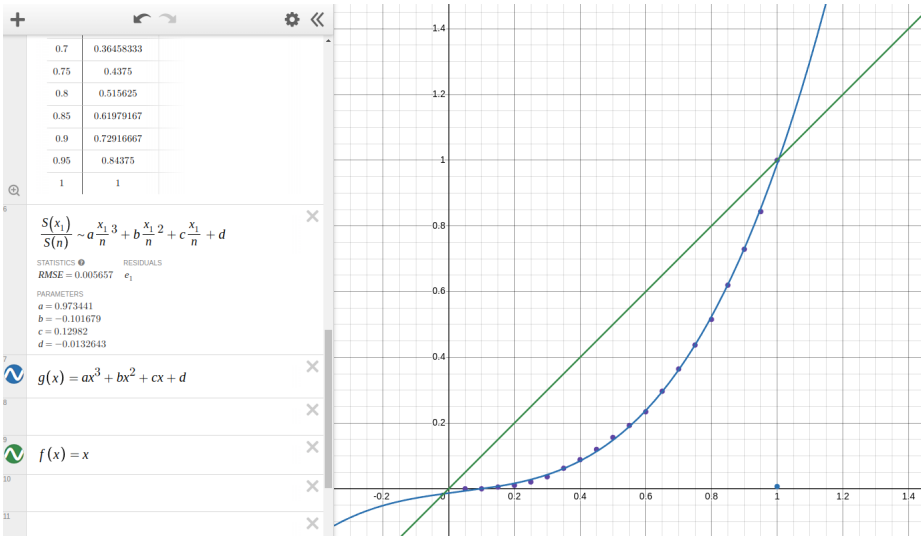


Figure 5: Curva de Lorenz en Desmos

lo más el área de un triángulo rectangulo cuyos catetos son 1 y su hipotenusa es $\sqrt{2}$, o cero.

Para calcular el índice de Gini hallamos el área entre $f(x) = x$ y $g(x)$ a través de una integral definida entre $0 < x < 1$.

$$G = 2 \cdot \int_0^1 [f(x) - g(x)]\, dx = 2 \cdot \int_0^1 [x - g(x)]\, dx$$

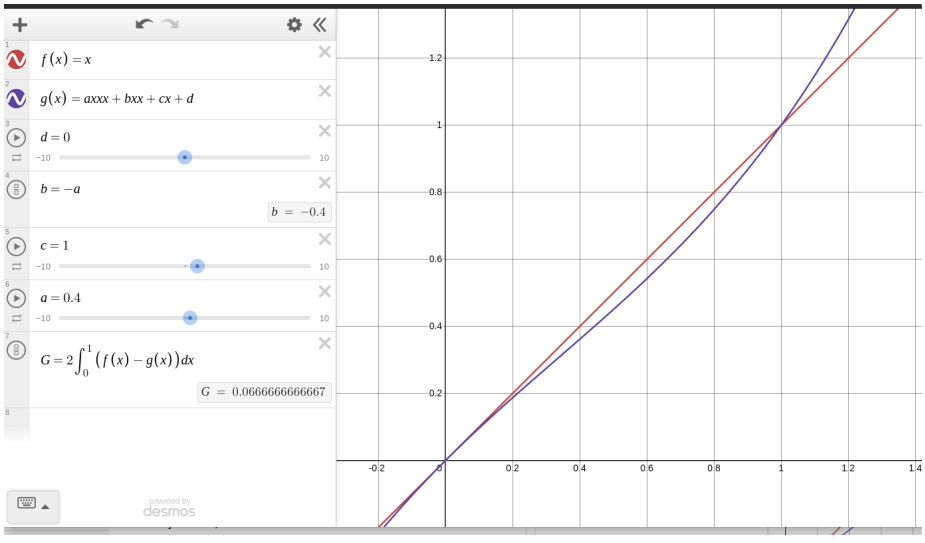
En base a nuestra aproximación:

$$g(x) = ax^3 + bx^2 + cx + d$$

$$G = 2 \cdot \int_0^1 [x - g(x)]\, dx \approx 0.477773693$$

Entonces, el índice de Gini es aproximadamente $G = 0.477773693$.

La interpretación de éste resultado nos puede dar una pista. Como se menciona, no nos detalla en absoluto si los alumnos son puntuales, pero vamos a suponer que una escuela tiene una diferencia en el horario de llegada, en promedio, de 5 minutos de anticipacion. Ahora, en base a este dato, podemos calcular el índice de Gini, por ejemplo, $G \approx 0.07$. En este caso se puede decir que la escuela es puntual ya que en promedio los alumnos llegan temprano, y además la distribución de sus horarios de llegada es bastante uniforme (una curva casi recta).



(Se aproximó el valor por debajo).

En resumen, si queremos saber cómo se distribuyen los horarios de llegada de una escuela, el método anterior puede ser útil.

Segmentación por grupos y probabilidades

Se supone que se desea estudiar el comportamiento de grupos de alumnos. Se comienza aprovechando una de las **ventajas principales de la aplicación**, que es el registro de los datos. Simplemente se busca un conjunto de alumnos que pertenezcan a una cierta característica, por ejemplo:

Se quiere explorar la probabilidad de que un alumno llegue tarde, dado que éste está cursando el segundo año de la escuela. La probabilidad condicional es el tema que se va a tratar, se ha de calcular con la siguiente regla

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

Donde $P(A \cap B) = P(B) \cdot P(A|B)$. Para comprobarlo se usa el mismo razonamiento, se tiene $P(A|B) = \frac{P(A \cup B)}{P(B)}$, de donde se puede despejar $P(A \cup B)$.

$P(B|A)$ se denota "Probabilidad de B dado A", es decir, qué probabilidad hay de que B suceda si A es cierto. El siguiente *diagrama de Venn* da una representación gráfica de la situación,

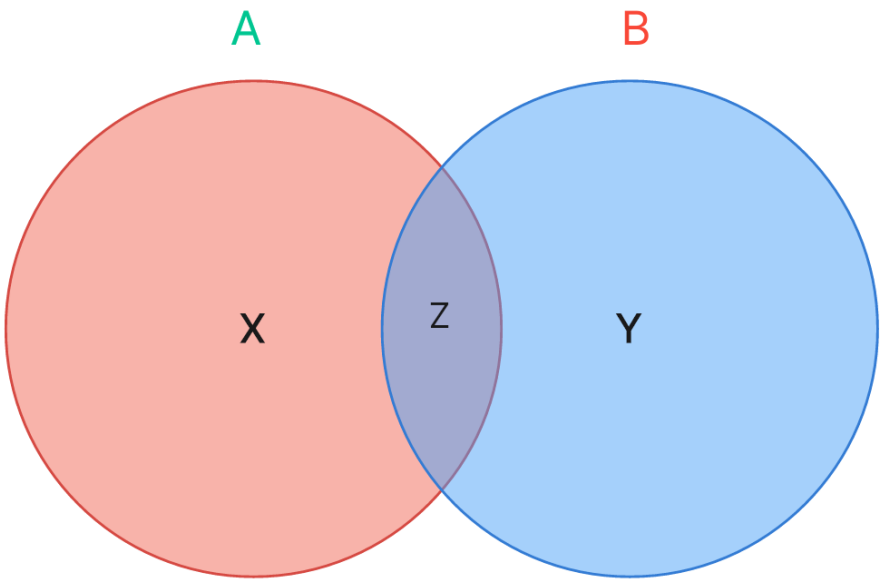


Figure 6: Diagrama de dos probabilidades A y B

El diagrama describe dos probabilidades A y B, la intersección ($Z = A \cap B$) indica la probabilidad de que ámbos eventos pasen (sean o no simultáneos, eso no se indica aquí). La formula anterior es el ratio entre Z e Y, es decir, informalmente: "si A está más «dentro» de B entonces la intersección es mayor, por lo tanto la probabilidad de ámbos eventos ocurriendo es mayor. Si se da que la intersección contiene la mayoría de los elementos de B, es decir $P(A \cap B) > P(B \setminus A \cap B)$ (o como en el diagrama, $Z > Y$), entonces sucede que B es un evento *más dependiente* de A, en el caso contrario es *menos dependiente*, y en el caso $A \cap B = \emptyset$ se dice que B *no depende de* A. Por lo que $P(B|A) = P(B)$".

Dada esa introducción, podemos estudiar dos eventos $T =$ "$Eventodetardanza$" y $E =$ "$Alumnocursasegundoa$ño". Si bien es discutible el hecho de que estos eventos sean independientes entre sí, se tiene que en la realidad puede pasar que un curso llegue tarde de manera habitual, por lo que se estudiarán como si fueran dependientes.

Se quiere hallar $P(T|E) = \frac{P(T) \cdot P(E|T)}{P(E)}$, es decir, la probabilidad de que un alumno llegue tarde *dado que* cursa segundo año. Si en una escuela se tiene una probabilidad total de tardanza (que se puede escribir $\frac{totaldetardanza}{totalasistencias}$) de $P(T) = 0.3$ y se tiene que $P(E) = \frac{Nřdealumnosdesegundoaño}{totalalumnos}$ y es $P(E) = 0.15$, se sabe que la probabilidad $P(E|T) = \frac{tardanzasdealumnosdesegundoaño}{tardanzastotales}$, y se tiene $P(E|T) = 0.18$.

Resumiendo:

$P(T) = 0.3$

$P(E) = 0.15$

$P(E|T) = 0.18$

Entonces, si calculamos la probabilidad final:

$P(T|E) = \frac{P(T) \cdot P(E|T)}{P(E)} = \frac{0.3 \cdot 0.18}{0.15} = 0.36$

Es decir, un 36% de probabilidad de que el alumno llegue tarde si cursa segundo año.

Más ejemplos así pueden darse para cualquier variable (tardanza, asistencia, inasistencia, retiros, etc.) y cualquier segmento o grupo (año, profesor, alumno, etc.).

Implementación técnica

La implementación de estos cálculos es bastante sencilla en el contexto de los lenguajes de programación actuales. El proyecto utiliza Java S.E junto con Spring Framework (como soporte para el backend), lo que nos va a permitir procesar los datos del lado del servidor.

Para la implementación de las funciones dentro de la aplicación habrá una clase llamada **StatsService.java** que se encargará de: 1. Recopilar la información relacionada 2. Ordenar los datos para su procesamiento 3. Procesar los datos y guardarlos en la base de datos

Para implementar funciones matemáticas haremos uso de la clase **Math.java**. Los algoritmos a implementar son bastante sencillos ya que constan de sumatorias. Se proporcionará una representación en formato JSON para que el frontend pueda representar los datos gráficamente.

Cuando el usuario desee consultar estos datos simplemente accederá a un método definido en el controlador **StatsController.java**.

Tiempo de ejecución

El tiempo de ejecución de los algoritmos se va a realentizar considerablemente según el número de alumnos. Para ejemplificar, si una escuela tiene $a = 500$ alumnos, y transcurrieron $d = 100$ días desde que se recopilaron datos de asistencia. Si cada alumno faltó $f = 10$ días en todo el ciclo lectivo, entonces se estima que en la base de datos deben haber $E_a = a \cdot (d - f) = 45000$ Entradas en la tabla `asistencias`. Por ende, la cantidad de datos a procesar crece linealmente, es decir, en el peor de los casos el tiempo de ejecución es *O(n)*. El sistema se ejecuta en *Java* y debe realizar las operaciones de busqueda en la base de datos, lo que ya crea un tiempo considerable, además de ello, si el servidor estuviera corriendo en una máquina virtual o en un *VPS*, hay que sumar tiempos de respuesta del servidor para las búsquedas. Para los algoritmos mencionados anteriormente lo mejor sería implementar una **memoria caché** (no incluido en el sistema actual).

Autorización y acceso a la información

Ningún cargo menor a **SECRETARIO** o **DIRECTIVO** tendrá acceso a las métricas mencionadas, pero esto no significa que la información no pueda ser compartida, sino que por motivos de cada escuela es posible que solo estos cargos quieran o puedan visualizarlos.

Aclaraciones

Es necesario aclarar que este documento es informal. Dentro del mismo se presentan distintas formas y conceptos que pueden utilizarse para conocer mejor a las instituciones de las que formamos parte. Además, las ideas comentadas caben dentro del contexto del proyecto como una forma de mejorar la institución educativa; ayudar a recopilar, analizar y comprender su comportamiento; facilitar el trabajo de aquellos individuos que trabajen con los datos proporcionados. En la documentación y presentación del proyecto damos a entender los objetivos y los medios para lograr los objetivos presentados, este documento es una rama

de la documentación principal dedicada especialmente al manejo de los datos que recibimos y que merece un documento aparte debido a que es una de las ventajas principales del proyecto.