PROYECTO FINAL GRUPO 2

Steven Escobar, Jacobo Gerrero, Yairis Córdova, Andrés Toala, Elkin Ramirez

2022-08-15

TÍTULO DE ESTUDIO

ESTUDIO Y ANÁLISIS DE VARIABLES QUE INFIEREN Y MODELEN EL RENDIMIENTO ACADÉMICO DE ESTUDIANTES

1. INTRODUCCIÓN

Mediante una base de datos proporcionados por la profesora Heidy Roa, fue factible el estudio y análisis de datos provenientes de estudiantes de 2 paralelos. De esta base de datos se pudo identificar tanto variables cuantitativas como cualitativas, como por ejemplo de ellas destacan para variables cuantitativas el promedio, las horas de estudio diarias, horas promedio diaria de sueño, entre otros.

Mientras que para las variables cualitativas fueron de interés si el estudiante trabajaba, el sexo, si despierta más de una vez en la noche y si el estudiante tiene computador de uso exclusivo. Todas estas variables a estudiar principalmente despertaron nuestra curiosidad ya que de cierta manera inciden en el desempeño del estudiante mostrado en el promedio.

De aquí en adelante es que se propuso encontrar la variable que incida más en el promedio mediante correlación de variables y pruebas de hipótesis para descartar o afirmar casos que vayamos encontrando en el transcurso del estudio.

2. DESCRIPCIÓN DEL PROBLEMA DE ESTUDIO

Una vez encontrado las medidas estadísticas con sus gráficos descriptivos para cada variable respectiva y en función de que variable corresponda (cualitativa o cuantitativa). Procedemos a calcular una matriz de correlación la cual arrojará información acerca de que variables se correlacionan brindando un nivel significativo haciéndolo posible de estudio para una regresión lineal.

Por otro lado, se tuvo como inferencia que la variable cualitativa de Trabajo puede incidir mucho en el rendimiento de un estudiante por lo cual se procede a realizar una prueba de hipótesis relacionando el promedio del rendimiento académico con los estudiantes que trabajan. Otra hipótesis a trabajar fue la diferencia de varianza entre mujeres y hombres para compararlos y conocer cual de los dos tienden más a promedios mayores a 8.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

- Identificar variables que expliquen el comportamiento del promedio académico de los estudiantes encuestados de dos paralelos de la Universidad Politécnica del Litoral.

3.2 OBJETIVOS ESPECÍFICOS

- Plantear hipótesis que nos ayuden a descartar o verificar la relación con respecto al comportamiento de una variable cuantitativa importante en nuestra investigación, con respecto a una variable cualitativa.
- Generar un modelo de regresión lineal que nos ayude a predecir el comportamiento del promedio académico de los estudiantes, mediante una ecuación que nos arrojará dicha gráfica de regresión.

4. MATERIALES

Como punto a seguir en nuestro proyecto, se debía escoger al menos 5 variables cuantitativas y 4 variables cualitativas. Como se mencionó anteriormente, nuestro enfoque es encontrar variables que modelen e infieran con el promedio académico de los estudiantes. Por consiguiente, se escogieron las siguientes variables:

Variables Cualitativas	Variables Cuantitativas
Sexo	Promedios
Trabaja	Horas promedio diaria de estudio
Computador de uso exclusivo	Materias promedio por término
Despierta más de 1 vez durante la noche	Horas promedio diarias en redes
	Horas promedio diarias de sueño
	Frecuencia semanal de actividad física

Posterior a definir las variables, se procedió con las medidas estadísticas para cada variable cuantitativa, pero la de esencial enfoque será la de la media del promedio académico cuyo valor es: 7.75.

Una vez definida la variables cuantitativas se podrá realizar una matriz de correlación la cuál mostrará que variables tienen una correlación significativa, las cuáles se podrán analizar posteriormente en un gráfico de regresión lineal. Y en especial donde 1 de las 2 variables sea el promedio académico.

Siguiendo uno de los objetivos específicos para descartar o verificar relaciones entre variables, de aquí se podrá plantear la hipótesis donde se enuncia que aquellos estudiantes que trabajan tienen un promedio menor a la media . Para esto se plantea la H0=7.75 y H1<7.75 donde

5. RESULTADOS

Al realizar una análisis y estudio de las variables que podrían inferir en el rendimiento académico de estudiantes, se pudo observar los siguientes resultados:

- **Encontramos que el promedio si sigue una distribución normal al categorizarlo y al no categorizarlo por sexo.**

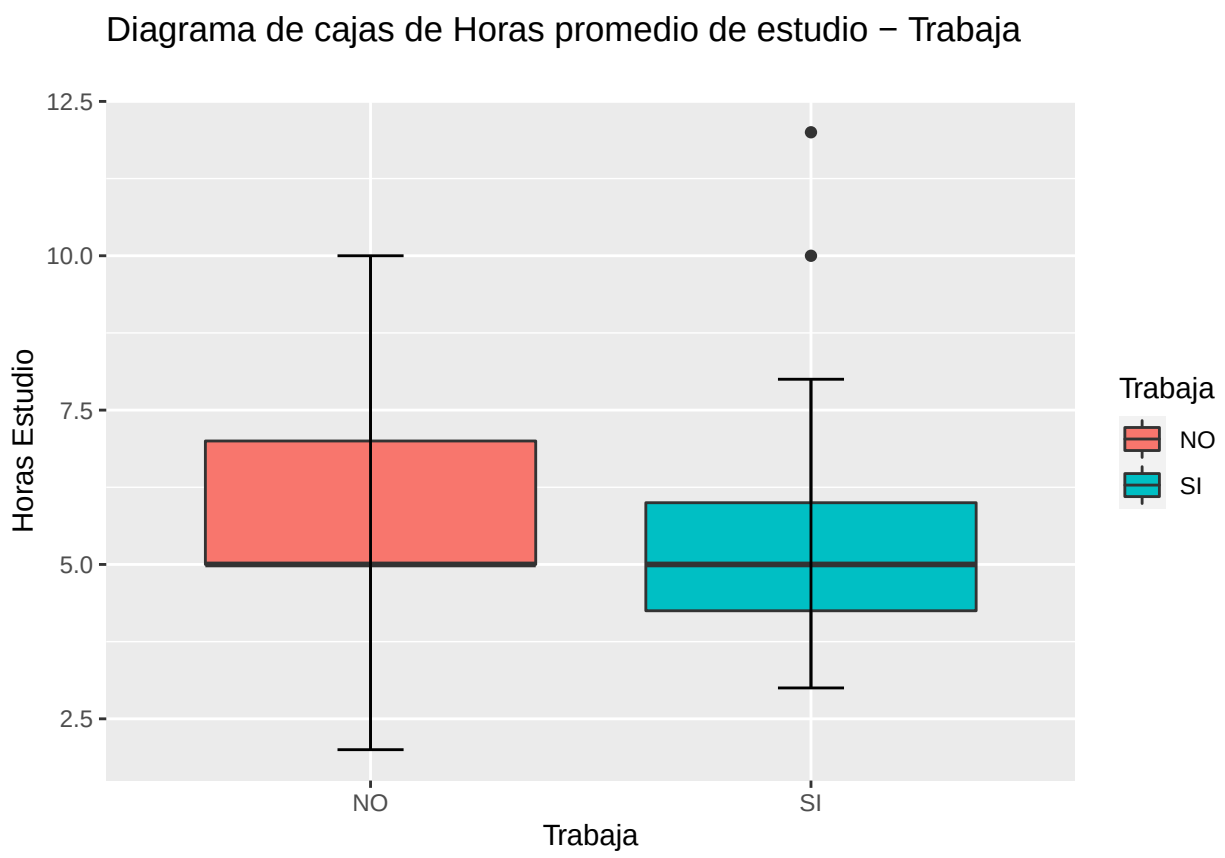
```
##
##  Shapiro-Wilk normality test
##
## data:  datos_est$Promedio
## W = 0.97742, p-value = 0.1868
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hombres$Promedio
## W = 0.9653, p-value = 0.1258
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mujeres$Promedio
## W = 0.96823, p-value = 0.6235
```

Ya que los valores p calculados en todas las pruebas anteriores son considerablemente altos (mayores que el nivel de significancia usual, $\alpha = 0.05$), se puede afirmar con evidencias suficientes que se sigue una distribución normal para todas las variables analizadas.

- **Realizamos una comparación en base a las Horas Promedio diarias de estudio y si el estudiante trabaja o no.**



Al realizar un diagrama de cajas comparativo entre si el estudiante se encuentra trabajando o no y sus horas promedio diarias de estudio empleadas, observamos que en sus representaciones en terminos de medias de cada sección son exactamente similares.

Debido a esto plantemos las correspondientes hipótesis para corroborar que esto sea cierto, $H_o : Var_1 = Var_2$ vs. $H_a : Var_1 \neq Var_2$.

```
##
##  F test to compare two variances
##
## data:  trabajan_Est$Horas_promedio_diarias_estudio and noTrabajan_Est$Horas_promedio_diarias_estudio
## F = 1.2542, num df = 21, denom df = 54, p-value = 0.4955
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.641847 2.754906
## sample estimates:
## ratio of variances
##           1.254196
```

Que mediante un F-Test podemos observar que $0.4955 > \alpha; \alpha = 0.05$ por lo cual hay evidencia suficiente para indicar que ambas observaciones poseen diferentes varianzas.

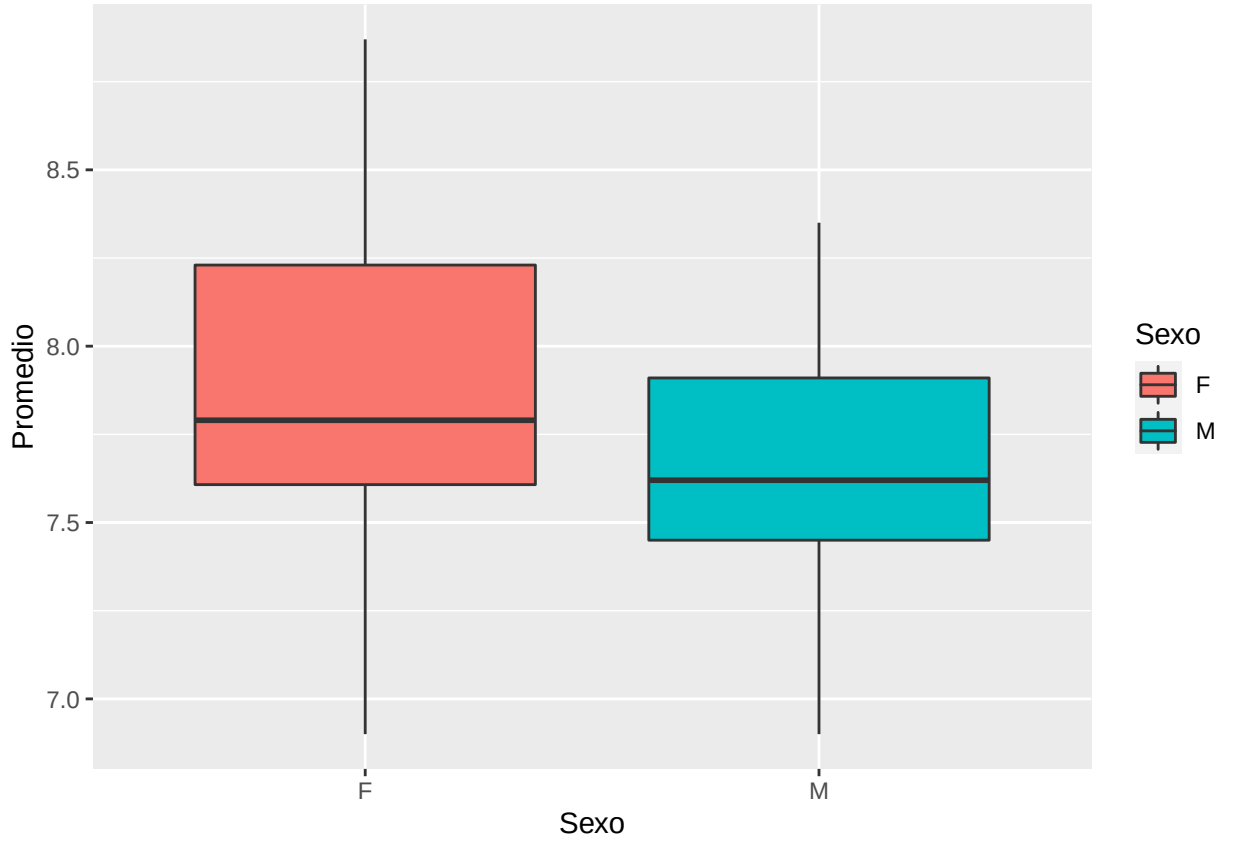
Ante lo cual planteamos las hipótesis para observar si los estudiantes que trabajan y los que no trabajan poseen iguales promedios de horas de estudio diarias, $H_o : \mu_0 = \mu_1$ vs. $H_a : \mu_0 \neq \mu_1$

```
##
##  Welch Two Sample t-test
##
## data:  trabajan_Est$Horas_promedio_diarias_estudio and noTrabajan_Est$Horas_promedio_diarias_estudio
## t = -0.26066, df = 35.141, p-value = 0.7959
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1982391  0.9255118
## sample estimates:
## mean of x mean of y
##  5.772727  5.909091
```

Y mediante un T-test para diferencia de medias con varianzas desiguales, obtenemos que $0.7959 > \alpha; \alpha = 0.05$.

Con esto indicamos que hay evidencia suficiente de que el promedio de horas de estudio de los estudiantes que trabajan y el de estudiantes que no trabajan son iguales.

- **Realizamos una comparación entre el rendimiento académico de hombres y mujeres. Mujeres > Hombres**



Al realizar un diagrama de cajas comparativo entre hombres y mujeres observamos una ligera diferencia en sus representaciones lo que indica que las mujeres están por encima de los hombres en términos de medias. Debido a esto planteamos las correspondientes hipótesis $H_o : Var_1 = Var_2$ vs. $H_a : Var_1 \neq Var_2$.

```
##
##  F test to compare two variances
##
## data:  hombres$Promedio and mujeres$Promedio
## F = 0.60913, num df = 52, denom df = 23, p-value = 0.1403
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2858068 1.1778309
## sample estimates:
## ratio of variances
##          0.6091323
```

Que mediante un F-Test podemos observar que $0.1403 > \alpha; \alpha = 0.05$ por lo cual no hay evidencia suficiente para indicar que ambas observaciones posean diferentes varianzas.

Ante lo cual planteamos las hipótesis para observar si las mujeres poseen mejor rendimiento académico que los hombres, $H_o : \mu_0 = \mu_1$ vs. $H_a : \mu_0 \geq \mu_1$

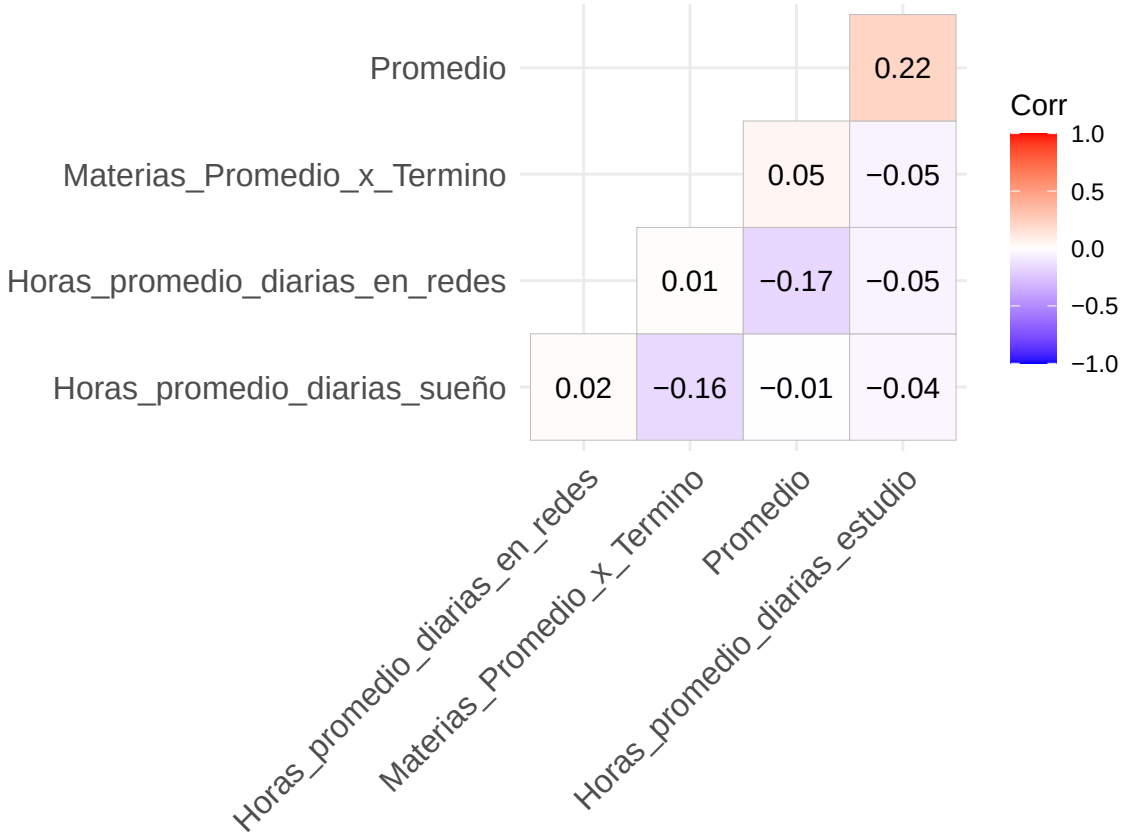
```
##
##  Two Sample t-test
##
```

```
## data:  mujeres$Promedio and hombres$Promedio
## t = 2.3715, df = 75, p-value = 0.01014
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0654962       Inf
## sample estimates:
## mean of x mean of y
##  7.897917  7.677925
```

Y mediante un T-test para diferencia de medias con varianzas iguales, obtenemos que $0.01014 < \alpha; \alpha = 0.05$.

Con esto indicamos que hay evidencia suficiente de que el promedio del rendimiento de las mujeres es mejor que las de los hombres.

- **Obtuvimos correlaciones muy bajas entre posibles variables predictoras para un modelo lineal de los datos.**



Al analizar la gráfica de correlaciones entre las variables predictoras podemos notar que la más alta de ellas es la variable **HORAS PROMEDIO DE ESTUDIO** con un valor de 0.22, debido a esto y a pesar de ser un valor pequeño realizamos un modelo de regresión lineal con esta variable.

- **Nos encontramos con un Modelo de regresión lineal (PROMEDIO ~ HORAS_ESTUDIO) no aceptable.**

A pesar de tener una correlación muy baja, realizamos un modelo de regresión lineal para verificar que estos valores sean o no linealmente dependientes y mediante la misma poder descartar todas las demás variables con menor correlación lineal que las seleccionadas.

```
##
## Call:
## lm(formula = Promedio ~ Horas_promedio_diarias_estudio, data = datos_est,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80867 -0.23604 -0.00214  0.17786  0.94396
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.49129    0.13788   54.33   <2e-16 ***
## Horas_promedio_diarias_estudio  0.04347    0.02229    1.95   0.0549 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3814 on 75 degrees of freedom
## Multiple R-squared:  0.04827,    Adjusted R-squared:  0.03558
## F-statistic: 3.804 on 1 and 75 DF,  p-value: 0.05488


## Analysis of Variance Table
##
## Response: Promedio
##                                Df  Sum Sq Mean Sq F value  Pr(>F)
## Horas_promedio_diarias_estudio  1  0.5532 0.55321  3.8035 0.05488 .
## Residuals                      75 10.9083 0.14544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenemos el modelo de regresión lineal denotado como $y = b_1(x) + b_0$ donde la variable dependiente es **"Promedio"** y la variable independiente es **"Horas_Promedio_Diarias_estudio"**. Por lo que el modelo de la ecuación quedaría como:

$$y = 0.04347x + 7.49129$$

Al realizar el análisis del modelo notamos varios factores que indican que el modelo no es aceptable.

1. El valor de R^2 es de 4.8% lo cual indica únicamente que solo el 4.8% de la variación observada de los datos de la variable predictora Horas_promedio_estudio puede ser explicada por la variable respuesta Promedio. (Baja fuerza de Asociación Lineal).

2. El valor p de la variable predictora (0.0549) es mayor que la significancia planteada ($\alpha = 0.05$) por lo cual no se puede afirmar que la variable β_1 contribuya al modelo planteado.

3. Al observar el análisis de las varianzas del modelo (ANOVA) podemos observar como la Suma Cuadrática de los valores residuales o Error es significativamente mayor a la Suma Cuadrática de la variable predictora lo cual denota una elevada cantidad de valores residuales que no son aceptables para un modelo de regresión lineal.

4. El valor p del F estadístico también nos indica que el modelo planteado no es apropiado para modelizar los datos.

6. CONCLUSIONES

- Dado el análisis realizado podemos concluir que la media de horas promedio de estudio que realizan estudiantes que trabajan es igual al de estudiantes que no trabajan, correspondiente a las observaciones analizadas dentro de la materia de Estadística.
- Logramos concluir de manera explicada que las mujeres dentro de las observaciones realizadas presentan (en promedio) mejor rendimiento académico que los hombres dentro de la materia de Estadística.
- Se pudo observar que la variable Promedio, principal variable de análisis presenta una distribución normal. Lo cual cumple con uno de los supuestos para que cualquier modelo de regresión lineal sea aceptada.
- Cabe denotar que de las variables seleccionadas como posibles variables predictoras en un modelo de regresión lineal mostraron valores de correlación lineal extremadamente bajos por lo cual no es posible obtener un modelo de regresión lineal aceptable que demuestre que estas variables individualmente sean factores que influyan en un mejor rendimiento académico.
- Al realizar un modelo de regresión lineal con la variable con mayor correlación lineal (Horas_promedio_diarias_estudio) se pudo corroborar que no es posible aceptar el modelo debido a diversos factores tales como que la modelización de los datos no es apropiada, no hay evidencias suficientes para determinar que la variable predictora contribuya al modelo planteado, elevados valores residuales que imposibilitan un modelo apropiado y además su baja fuerza de asociación lineal la cual esta por debajo del 5%. Debido a estos factores, se concluye que con un modelo de regresión lineal simple resulta difícil poder explicar o denotar factores que individualmente afecten al rendimiento académico.

Dado las conclusiones respectivas de esta investigación, podemos observar que un modelo de regresión lineal simple resulta poco efectivo para explicar los posibles factores del rendimiento académico de los estudiantes dentro de la materia de Estadística. Por lo cual es necesario que se lleve a cabo un análisis más profundo de los datos con más variables que únicamente no denoten las estrategias de aprendizaje o las horas de estudio u ocio, sino más bien con variables socioeconomicas, psicológicas, institucionales, etcétera, como lo denotan algunas investigaciones sobre el rendimiento académico tales como aquel que indica como la motivación del estudiante y otros factores como la estrategia de aprendizaje [1] pueden dar mejoras en el rendimiento académico de un estudiante. Tambien finalmente, podemos observar como en otros campos de investigación usan nubes de variables, minería de datos y formulaciones de modelos estadísticos para predecir si un estudiante tendrá o no un buen rendimiento durante un periodo de tiempo [2]–[4], demostrando así la complejidad de analizar y encontrar factores influyentes en el rendimiento académico de los estudiantes.

7. RECOMENDACIONES

- Procurar realizar encuestas con respuestas apegadas a la realidad y a gestionar las respuestas de los encuestados para que sean ingresadas correctamente, evitando así datos en blanco o datos inexplicables.
- Recomendamos realizar un correcto tratamiento de los datos dado que debido a fallas humanas o fallas computacionales es posible que ciertos factores sean ignorados debido a un mal ingreso de datos o a una mala interpretación.
- Al realizar las correspondientes pruebas de hipótesis tener en cuenta la hipótesis formulada y el correcto testeo a aplicar.

8. BIBLIOGRAFÍA

[1] M. Garrido Macías, N. Jiménez Luque, A. Landa Sánchez, E. Páez Espinar, and M. Ruiz Barranco, "Factores que influyen en el rendimiento académico: La motivación como papel mediador en las estrategias de aprendizaje y clima escolar." Proyecto de Innovación Docente "ReiDoCrea". Departamento de Psicología Social. Universidad de Granada. Universidad de Granada, 2013. doi: 10.30827/Digibug.27620.

[2] S. M. Merchan Rubiano and J. A. Duarte Garcia, "Formulation of a predictive model for academic performance based on students' academic and demographic data," in *2015 IEEE frontiers in education conference (FIE)*, 2015, pp. 1–7. doi: 10.1109/FIE.2015.7344047.

[3] F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," in *2018 third international conference on informatics and computing (ICIC)*, 2018, pp. 1–5. doi: 10.1109/IAC.2018.8780547.

[4] J. Jamesmanoharan, S. H. Ganesh, M. L. P. Felciah, and A. K. Shafreenbanu, "Discovering students' academic performance based on GPA using k-means clustering algorithm," in *2014 world congress on computing and communication technologies*, 2014, pp. 200–202. doi: 10.1109/WCCCT.2014.75.

9. ANEXOS