ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

Desarrollo e implementación de un modelo estadístico para pronosticar el overhead en una compañía de desarrollo de software.

JUAN DAVID LUNA QUINTERO

Universidad® Católica de Manizales
VIGILADA MINEDUCACIÓN
Obra de Iglesia de la Congregación
Hermanas de la Caridad Dominicas de La Presentación de la Santísima Virgen

DESARROLLO E IMPLEMENTACIÓN DE UN MODELO ESTADÍSTICO PARA PRONOSTICAR EL OVERHEAD EN UNA COMPAÑÍA DE DESARROLLO DE SOFTWARE

Trabajo de grado presentado como requisito para optar al título de Especialista en Estadística Aplicada.

Asesor

Mg. Luis Hernando Carmona

Autor:

Juan David Luna Quintero

UNIVERSIDAD CATÓLICA DE MANIZALES

FACULTAD DE EDUCACION DISTANCIA

ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

MANIZALES

2022

Tabla de contenido

Resumen

El objetivo de esta investigación fue desarrollar e implementar un modelo estadístico predictivo bajo la técnica de regresión lineal, donde se buscó determinar qué variable o variables pueden predecir el comportamiento del gasto u "overhead" al interior de una empresa de desarrollo de software. Inicialmente, se estudiaron diferentes técnicas de predicción estadística bajo regresión lineal: regresión lineal simple y regresión lineal múltiple. Con base en estas técnicas, se determinó la forma para obtener el grado de ajuste en términos de correlación y causalidad, a través de los diferentes tests estadísticos para el tipo de distribución que presentaron los datos estudiados. Entre los resultados, se encontró que la regresión lineal múltiple no es óptima para la predicción de la variable dependiente Overhead, y que al ajustar el modelo a una sola variable explicativa es posible crear un modelo de regresión lineal simple que explique la varianza de la variable dependiente. Se concluye que, al crear un modelo de regresión lineal múltiple, se debe evaluar la multicolinealidad, heterocedasticidad y autocorrelación, para determinar el grado de bondad y ajuste.

Palabras clave: overhead, gastos, regresión lineal, pronóstico.

Abstract

The objective of this study was to develop and implement a statistical predictive model under the technique of lineal regression, searching for evidence of one or multiple variables that can predict the behavior of the overhead cost inside a company dedicated to software development. Initially there was a research of different linear regression predictive techniques: simple linear regression, multiple linear regression. The findings stablished, a way to obtain an accurate adjustment degree in terms of correlation and causation, through different statistical tests for the specific type of distribution found in the dataset. Among the results, it was found that multiple linear regression model was not suitable for explaining the variance of the dependent variable Overhead, and that adjusting the model to just one predictive variable to perform a simple linear regression model, was optimal to predict the dependent variable. In conclusion, to apply a multiple regression model, is mandatory to test aspects such as multicollinearity, heteroscedasticity, and autocorrelation, to determine the adjustment degree.

Key words: overhead, expenses, lineal regression, forecast.

Introducción

La administración financiera de la compañía es una disciplina necesaria y pertinente sin importar el tamaño del negocio. De la correcta toma de decisiones financieras depende el éxito o fracaso de cualquier empresa. Es por esto por lo que los gerentes financieros de la compañía deben esmerarse en construir un sólido conocimiento en torno a la utilización de métodos y herramientas que optimizan y aceleran la toma de decisiones oportunas, que a su vez se apoyan en el estudio de los datos financieros.

Uno de los temas más importantes del manejo financiero es sin lugar a duda la administración de costos y gastos, siendo este último uno de los menos tenidos en cuenta a la hora de fijar una política de precios eficiente. En la práctica se brinda una mayor atención al control de costos dada su importancia dentro del giro ordinario de negocios y de la gran dependencia que tienen las ventas de estos, sin embargo, el control de los gastos u "overhead" (en inglés) juega un papel fundamental para la rentabilidad de una compañía. El desconocimiento de esta realidad ha llevado a muchas empresas a la quiebra debido a la desestimación del impacto del gasto y cómo este se imputa en el precio final de venta.

El overhead no es entonces otra cosa que aquellos gastos requeridos para mantener operando un negocio. A diferencia de los costos de producción o costos de venta que se encuentran directamente relacionados a la operación del negocio, el overhead no se encuentra ligado de manera directa a la producción debido a que, no obstante, se produzca o no, son gastos que se deben cubrir para seguir operando como compañía y su valor no es directamente proporcional al nivel de ventas como si lo son los costos directos de producción.

A manera de ejemplo si se tiene una compañía dedicada a la fabricación de instrumentos musicales, se incurrirá mensualmente en una serie de costos tales como: la compra de materias primas para realizar la fabricación, el pago de los trabajadores que realizan la fabricación, el mantenimiento de las máquinas empleadas en dicha fabricación, entre otros. Estos costos irán directamente a la sección del costo del estado de pérdidas y ganancias mensual, mientras que

erogaciones que corresponden al pago de servicios públicos de la sede administrativa, la nómina administrativa, la suscripción al software contable, intereses, impuestos, entre otros, son gastos que irán a la sección de gastos del estado de pérdidas y ganancias, y que deberán ser cubiertos aunque aumente o disminuya la producción e incluso en los meses en que no se produjo absolutamente nada.

Pregunta de investigación

¿Es posible encontrar una relación estadística lineal entre la variable overhead y otras variables financieras dentro del negocio, tales como, ventas, costos, gastos, número de desarrolladores, entre otras, para posterior a esto predecir el comportamiento de esta?

Objetivo general

Estudiar el grado de correlación que existe entre la variable overhead y las demás variables que se encuentran en la base de datos del negocio, con el fin de establecer un modelo lineal que pronostique dicha variable y disminuya la laboriosidad actual para fijar la política de precios dentro de la compañía objeto de estudio.

Referente teórico

Cuando un gerente financiero toma una decisión con respecto a un proyecto cercano en su compañía, bien sea una compra de activos, disminución de gastos, o la definición del método de depreciación de los activos con los que cuenta la compañía, entre otros, está incurriendo en un riesgo inminente. Este riesgo puede ser alto, medio o bajo, de acuerdo con el impacto que generaría en la rentabilidad y estabilidad de la operación del negocio, y en ocasiones producir una situación adversa por causa de la decisión que se tomó. Para hacer frente a los diferentes riesgos corporativos, es necesario que el gerente adquiera las destrezas que puede brindar la gestión del riesgo financiero, lo cual no es otra cosa que el grado óptimo de proyección y anticipación a eventos probables futuros, y que de su correcto manejo dependerá el éxito del negocio.

La modelación financiera es una de las herramientas de mayor uso en el área de planeación financiera, que permite gestionar el riesgo. Tal vez es un término que aversión y desconfianza entre algunos gerentes, lo cual ha provocado que la construcción de modelos sea encargada a los expertos en analítica y matemática, sacando del escenario a los mismos gerentes financieros y cegando de alguna manera, los procesos numéricos que se desarrollan al emplear dichos modelos (Pérez, 2019).

Cuando se habla de un modelo financiero, se hace referencia a un sistema numérico que representa la situación económica de una compañía, y es utilizado como una herramienta que

facilita la visualización del estado actual financiero, como medida de control y posibles futuros resultados que impactarán positiva o negativamente el negocio. (Pérez, 2019)

En esta medida, la predicción de estos sucesos en el tiempo se convierte en una necesidad que debe ser atendida por la empresa, como también la constante capacitación en el manejo de herramientas que faciliten la implementación y desarrollo de modelos financieros. Es ahí donde la estadística como ciencia que se dedica a la organización de los datos y a la presentación de los resultados obtenidos a partir de estos, (Salazar y Del Castillo, 2018) adquiere un papel categórico como herramienta que brinda a la administración financiera, soluciones óptimas y apropiadas.

De esta manera, el análisis predictivo de datos ofrece una variedad de métodos empleables al interior de los departamentos financieros toda vez que al hacer uso de la información que brindan los datos y el debido procesamiento de estos, se puede llegar a elaborar modelos apropiados para la constante supervisión y toma de decisiones. Una de estas técnicas estadísticas conocida como el análisis de regresión, estudia la relación que puede existir entre las variables observadas, este término fue acuñado por el estadístico inglés Francis Galton y fue producto de un estudio llevado a cabo para descubrir las correlaciones que tenían las alturas de los padres con los hijos de mas de mil grupos familiares. Es así como el análisis de regresión es útil para estudiar y explicar cuantitativamente, cuál es el grado de relación entre una variable dependiente, o variable cuya variación y comportamiento es causado por otra, y una variable predictora o variable que genera alteraciones en la distribución numérica de otra variable. La explicación final da como resultado un modelo estadístico llamado ecuación lineal (Pereira, 2009).

En caso de que los datos presenten una relación lineal, es decir, cuando al graficar los datos de las variables en un plano cartesiano, se observa que los puntos se concentran formado una especie de óvalo apuntalado, se puede decir que existe algún tipo de correlación entre las variables graficadas, para plasmar esa relación entre las variables existe la línea de mínimos cuadrados, la cual se traza ajustando de la mejor manera las distancias de los puntos hacía la misma recta. El objetivo de la línea de mínimos cuadrados es fijar una distancia mínima de cada

punto hacia la recta, minimizando la varianza de los datos, por tal motivo se llama mínimos cuadrados, ya que la varianza es la suma de los errores al cuadrado.

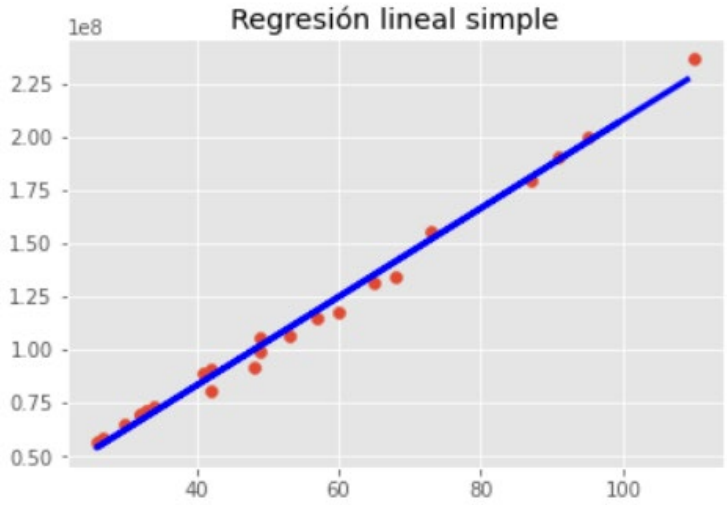

Figura 1: Regresión lineal simple

Fuente: Elaboración propia

Existen diferentes tipos de técnicas de regresión estadística útiles para realizar predicciones, sin embargo, cada técnica es utilizada de acuerdo con determinadas características especiales que poseen las distribuciones de datos.

Regresión Lineal Simple

La regresión lineal simple es ciertamente la técnica predictiva más cómoda de usar, ya que sólo estudia la correlación de dos variables, una independiente y otra dependiente. Como se mencionó anteriormente, la variable dependiente, basa su comportamiento en la fluctuación de la variable independiente, por lo que al variar esta última, ocasiona que la variable dependiente también varíe. El modelo estadístico de regresión lineal simple a menudo es expresado de la siguiente forma:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

Donde y es la variable dependiente, β0 es el intercepto al eje de las y, β1 es la pendiente, x es la variable independiente y ε es el error estándar. La variable dependiente muchas veces también es conocida como la variable de respuesta y la variable independiente también es

llamada comúnmente como variable explicativa, además, es normalmente asumido que el error estándar siempre es una constante igual a cero.

Uno de los métodos utilizados para explorar la posibilidad de encontrar una relación lineal simple entre dos variables, es la gráfica de dispersión. Esta ayuda a determinar de manera visual, en qué forma se distribuyen los puntos de cada variable a lo largo del plano, teniendo en cuenta que cada punto es una representación de los valores observados de (x,y), donde la variable independiente se grafica a lo largo del eje x, y la variable dependiente se grafica en el eje y. (Acosta et al, 2013)

Dependiendo de la formación de los puntos a lo largo del plano, se puede decir que es una relación lineal, curvilínea o que los datos representados no poseen ninguna relación tal como se presenta en la siguiente imagen.

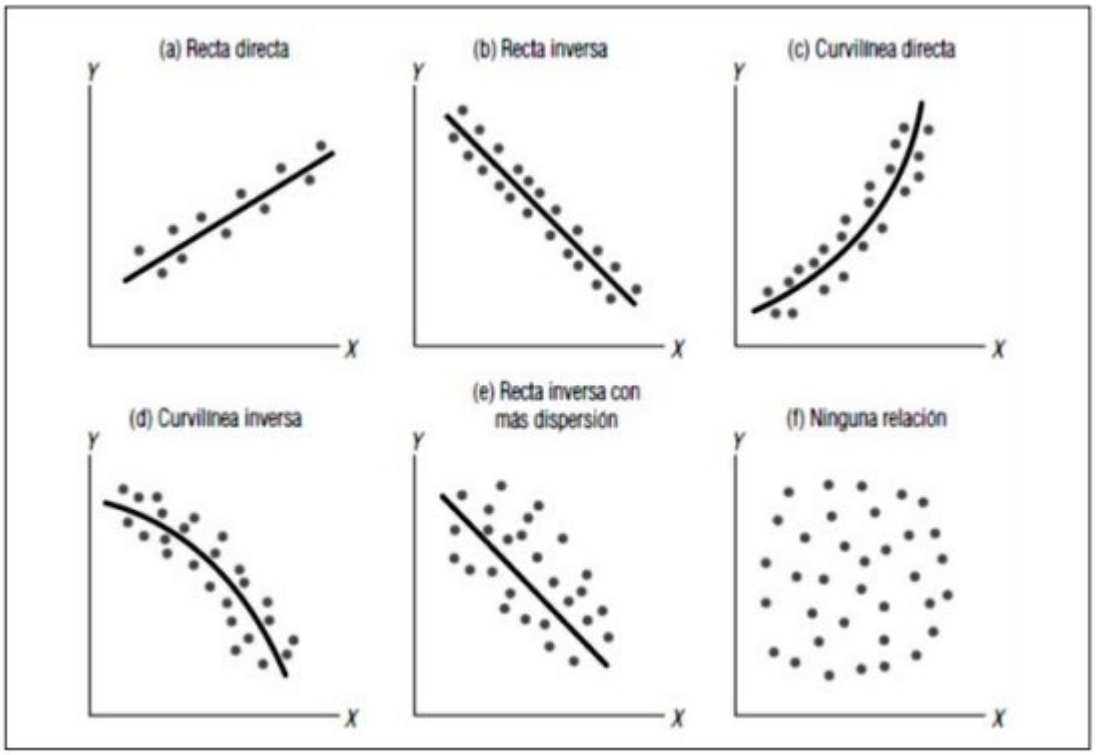

Figura 2: Tipos de relación lineal

Fuente: Levin, R. I. y Rubin, D. S. (2004). Estadística para administración y economía. Pearson Educación.

Ahora bien, una vez analizados los datos de forma visual, se pueden utilizar técnicas estadísticas que determinan de forma matemática el grado de correlación entre ambas variables, una de ellas es el coeficiente de correlación. El cual es un indicador que mide la asociación entre las dos variables, dependiente e independiente, calificando esta relación entre un intervalo [-1,1].

Si al medir la correlación, el coeficiente se encuentra cercano a cero o es igual a cero, quiere decir que no existe relación alguna entre las variables, mientras que si el coeficiente es cercano a -1 o a 1 se está indicando un grado de relación bien sea inversamente en caso de acercarse a -1 o directamente en acercarse al 1.

La fórmula para encontrar el coeficiente de correlación es la que se muestra a continuación:

r = (signo de b) √(r2) (17) (Cardona et al, 2013)

Otra de las medidas importantes que se estiman para encontrar el grado de ajuste de la ecuación de regresión lineal es el coeficiente de determinación o más conocido como R^2, el cual evalúa el porcentaje de variación total que explica la ecuación de regresión lineal y su resultado se encuentra entre los porcentajes 0% y 100%.

$$R^2 = 1 - \frac{SCR}{SCT}$$

(Montero, 2016, p. 39)

Donde, SCReg es la suma de los cuadrados de la regresión y SCTotal es la sumatoria de los cuadrados del total.

Regresión Lineal Múltiple

Ahora bien, cuando se trata de varias variables que predicen una variable, se esta hablando de una relación lineal múltiple. La Regresión lineal múltiple viene dada por la siguiente fórmula:

$$y = \beta_o + \beta_1 x1 + \beta_2 x2 + u$$

(Montero, 2016, p. 54)

Donde y es la variable dependiente que se desea pronosticar, β0, β1,. ., βp son los parámetros del modelo, x1, x2,... , xp, son las variables independientes y ε es el error aleatorio. (Acosta, et al. 2013)

> En regresión lineal múltiple sólo suele haber una variable endógena y puede haber varias variables exógenas. Es decir, se individualiza el fenómeno observado. También puede darse el caso de la existencia de varias variables endógenas, pero su solución es difícil por lo que no es el caso general. (Montero, 2016)

Las variables que se estudian en estos tipos de modelos pueden distribuirse en dos grupos, variables continuas y variables discretas. Las variables continuas son aquellas que contienen datos representados por números reales, y que pueden tener o no decimales. Algunos ejemplos de estas variables son el peso y la edad, como se puede entender una persona puede pesar 70 kg mientras que otra persona puede pesar 70,5 kg. Dentro de las variables continuas se pueden encontrar frecuentemente los porcentajes (Montero, 2016).

Por otro lado, las variables discretas son aquellas que por lo general contienen datos expresados por números enteros, por ejemplo, el número de hijos, habitantes de una ciudad, número de aviones producidos mensualmente, etc. Un elemento importante en la ecuación de regresión es el coeficiente de regresión, este busca explicar el cambio promedio que tiene una variable dependiente en la medida que varía la variable independiente, teniendo como regla la constancia en el resto de las variables (Acosta et al, 2013).

Para determinar si un modelo de regresión posee un ajuste óptimo, es necesario realizar pruebas como la del coeficiente de determinación múltiple (R^2) el cual tiene la bondad de medir qué porcentaje de variación puede explicar una variable independiente sobre una variable dependiente. Si ese valor se acerca a uno, quiere decir que el modelo hallado, explica en gran medida la variabilidad de la variable dependiente, mientras que, si se acerca a cero, esto indica que el modelo no explica o explica muy poco de la variabilidad de la variable dependiente. No obstante, cuando se ingresan demasiadas variables al modelo, el valor de R^2 tiende a aumentar, por lo cual se prefiere estimar el R^2 ajustado, el cual está expresado en la siguiente ecuación:

$$\bar{R}^2 = 1 - \frac{SCR\ /(n-k)}{SCT/(n-1)}$$

(Montero, 2016, p. 39)

Pruebas de hipótesis.

Cuando se ha terminado la labor de seleccionar la muestra de forma aleatoria, se han medido las diferentes variables que hacen parte de la base de datos objeto de análisis y se han revisado los resultados de la matriz de correlación, seleccionando así las variables que realmente tienen una incidencia positiva o negativa en la varianza de la variable en estudio. Se procede a determinar cuál es la ecuación que tiene la capacidad de predecir con mayor exactitud a la variable dependiente (Acosta et al, 2013).

Multicolinealidad.

Los valores t son objeto de un interés particular por parte de los investigadores, toda vez que estos coadyuvan a detectar la multicolinealidad entre las variables, en este orden de ideas, en caso de ser estos suficientemente grandes, se dice entonces que la correlación entre las variables independientes no presenta inconveniente y pueden ingresar juntas al modelo. Sin embargo, si al calcular dichos valores, son menores a los valores t de las tablas, se estaría presenciando un caso de multicolinealidad (Acosta, et al. 2013).

En el caso de que exista la multicolinealidad entre las variables predictoras, se hace complejo determinar que porción de la variación de la variable dependiente es generado por una variable predictora en particular. Si esto es así, las variables predictoras con multicolinealidad estarían proporcionando prácticamente la misma información de predicción (Acosta, et al. 2013).

En resumen, para seleccionar las variables que harán parte del modelo de regresión múltiple se deben en primer lugar, hallar una variable dependiente que tenga una fuerte correlación con una variable independiente, en segundo lugar las variables independientes no

deben alta correlación entre sí, es decir, que la correlación hallada entre estas debe ser muy por debajo que la correlación hallada entre la variable dependiente e independiente, por último, si existe multicolinealidad, se debe realizar el estudio previo para determinar qué variables ingresan al modelo o por lo menos cuáles variables pueden ingresar juntas y cuales no. A continuación, se ilustra cómo se puede determinar la multicolinealidad entre variables.

Matriz de correlación

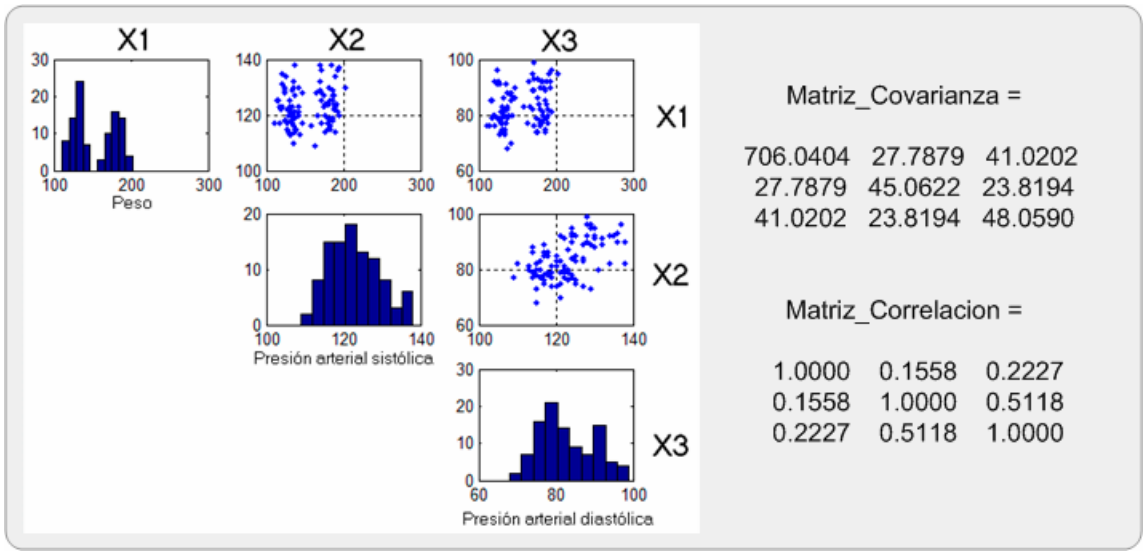

Figura 3: Matriz de correlación

Fuente: Pereira (2009-2010). Análisis predictivo de datos mediante técnicas de regresión estadística.

Metodología

El presente estudio se basa en una investigación de índole cuantitativa donde se observará el comportamiento de la base completa y no de una muestra, de los datos de la venta, costos, gastos, número de desarrolladores y porción del overhead de una empresa dedicada al desarrollo de software.

Por consiguiente la presente investigación desarrolla las etapas propias del enfoque cuantitativo, empezando por la fase conceptual, donde se determinó y se formuló cuál es el problema de investigación, seguido de una revisión de referentes teóricos y conceptuales que abarquen propiamente el tema de la regresión lineal y del manejo del overhead, con lo cual se procede a construir un marco teórico que reúna los conceptos principales concerniente al tema en asunto y por último para esta fase se procede a la presentación de una hipótesis, la cual se resume en si el comportamiento de las variables exógenas, ventas, costos, gastos y número de desarrolladores puede explicar la variación de la variable dependiente overhead a través de un modelo de regresión lineal simple o un modelo de regresión múltiple.

Para la fase de planeación y diseño del proyecto de investigación, se identificaron los datos originales de la compañía desarrolladora de software y se realizaron algunas pruebas de estadística descriptiva sobre estos, obteniendo un análisis preliminar que se convierte en insumo para empezar a esbozar en el imaginario un tipo de modelo que posiblemente sea congruente con la distribución y tipo de datos, en este caso financieros.

Resultados

Para el análisis de los resultados se procede a importar el conjunto de datos al aplicativo Python a través de la herramienta Google Colab, no sin antes haber importado las librerías necesarias para el correspondiente manejo del dataset y para su modelación estadística y visualización.

```
# Tratamiento de datos
# ==============================================================================
import pandas as pd
import numpy as np

# Gráficos
# ==============================================================================
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

# Preprocesado y modelado
# ==============================================================================
from scipy.stats import pearsonr
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.stats.anova import anova_lm
from scipy import stats
```

```
df=pd.read_excel("/content/drive/MyDrive/Dataset Dev Juan Luna.xlsx")
pd.set_option('display.float_format', lambda x: '%.f' % x)
df
```

	Year	Month	Sales	Cogs	Expenses	TRM	Developers	Overhead_portion
0	2020	January	149032333	79403765	116016612	3411	27	58401000
1	2020	February	162557267	83555978	52249458	3540	32	69216000
2	2020	March	172357500	84229658	55668445	4065	32	69216000
3	2020	April	152822180	81043865	48409538	3983	26	56238000
4	2020	May	150192817	72968187	48285942	3719	26	56238000
5	2020	June	175768833	93391527	55111805	3759	33	71379000

Figura 4. Importación de librerías y carga de dataset

Fuente: Elaboración propia

Una vez importados los datos, se inicia lo que se conoce como EDA o "Análisis explotario de datos" por sus iniciales en inglés. Se busca entonces, obtener una primera impresión de la forma y distribución de las variables que forman parte del dataset en estudio.

```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Year              29 non-null     int64
 1   Month             29 non-null     object
 2   Sales             29 non-null     float64
 3   Cogs              29 non-null     float64
 4   Expenses          29 non-null     float64
 5   TRM               29 non-null     float64
 6   Developers        29 non-null     int64
 7   Overhead_portion  29 non-null     float64
dtypes: float64(5), int64(2), object(1)
memory usage: 1.9+ KB
```

```
df.describe()
```

	Year	Sales	Cogs	Expenses	TRM	Developers	Overhead_portion
count	29	29	29	29	29	29	29
mean	2021	344464967	194742465	115667299	3782	57	118310856
std	1	179156451	115491970	91675905	168	26	53676972
min	2020	149032333	72968187	48285942	3411	26	56238000
25%	2020	188277833	93391527	68358943	3715	34	73542000
50%	2021	267498000	149350045	88820310	3760	49	99498000
75%	2021	454374667	273752978	132069281	3911	73	155482215
max	2022	734018250	454191010	535375482	4065	110	236313723

Figura 5. Tipos de variables y Estadística descriptiva

Fuente: Elaboración propia

Con la función info(), es posible conocer el tipo de dato que se tiene en cada variable, en el caso del presente estudio, las variables disponibles en el dataset son de tipo entero, objeto y flotante o contínua.

Seguidamente, con la función describe(), se obtiene la estadística básica descriptiva de cada variable, teniendo en cuenta que el año no es una variable que aporte para el análisis se omite y se revisa cuál es la distribución del resto de variables. Lo primero que salta a la vista es

que no se tienen datos nulos, algo que es importante al analizar un dataset, debido a que la incompletitud de los datos en una variable puede generar imprecisiones en la modelación.

Modelación.

```
[ ]  model = smf.ols('Overhead_portion ~ Sales + Cogs + Expenses + Developers', df)
     result = model.fit()
     result.summary()
```

OLS Regression Results

Dep. Variable:	Overhead_portion	**R-squared:**	0.994
Model:	OLS	**Adj. R-squared:**	0.993
Method:	Least Squares	**F-statistic:**	929.8
Date:	Sat, 23 Jul 2022	**Prob (F-statistic):**	6.24e-26
Time:	10:26:01	**Log-Likelihood:**	-483.58
No. Observations:	29	**AIC:**	977.2
Df Residuals:	24	**BIC:**	984.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.714e+06	4.49e+06	0.828	0.416	-5.55e+06	1.3e+07
Sales	0.0489	0.055	0.887	0.384	-0.065	0.163
Cogs	0.0139	0.066	0.211	0.835	-0.122	0.150
Expenses	-8.272e-05	0.012	-0.007	0.995	-0.026	0.026
Developers	1.668e+06	4.09e+05	4.077	0.000	8.23e+05	2.51e+06

Omnibus:	3.038	**Durbin-Watson:**	0.967
Prob(Omnibus):	0.219	**Jarque-Bera (JB):**	1.962
Skew:	-0.423	**Prob(JB):**	0.375

Figura 6. Modelo de regresión múltiple.

Fuente: Elaboración propia

Luego de visualizar la base de datos en un primer acercamiento descriptivo, se procede a indagar si el modelo de regresión múltiple es óptimo para explicar el comportamiento de la variable Overhead_portion, teniendo en cuenta la cantidad de variables que se considera que aportarían a la explicación de la variable dependiente.

Se crea el modelo múltiple con las variables Overhead_portion como dependiente y Sales, Cogs, Expenses y Developers como explicativas, para revisar el summary del modelo y visualizar cuál es la calidad de los estadísticos que arroja.

Al revisar detalladamente, se encuentra que el R cuadrado arroja un valor cercano a 1, lo mismo que el R ajustado. En primera instancia eso sería un buen indicio, sin embargo, al revisar los p-valor de cada variable se encuentra que la única variable que posee un valor menor a 0.05

es la variable Developers. Esto de por si genera algún tipo de incertidumbre ya que no es congruente con el valor del R cuadrado y ajustado.

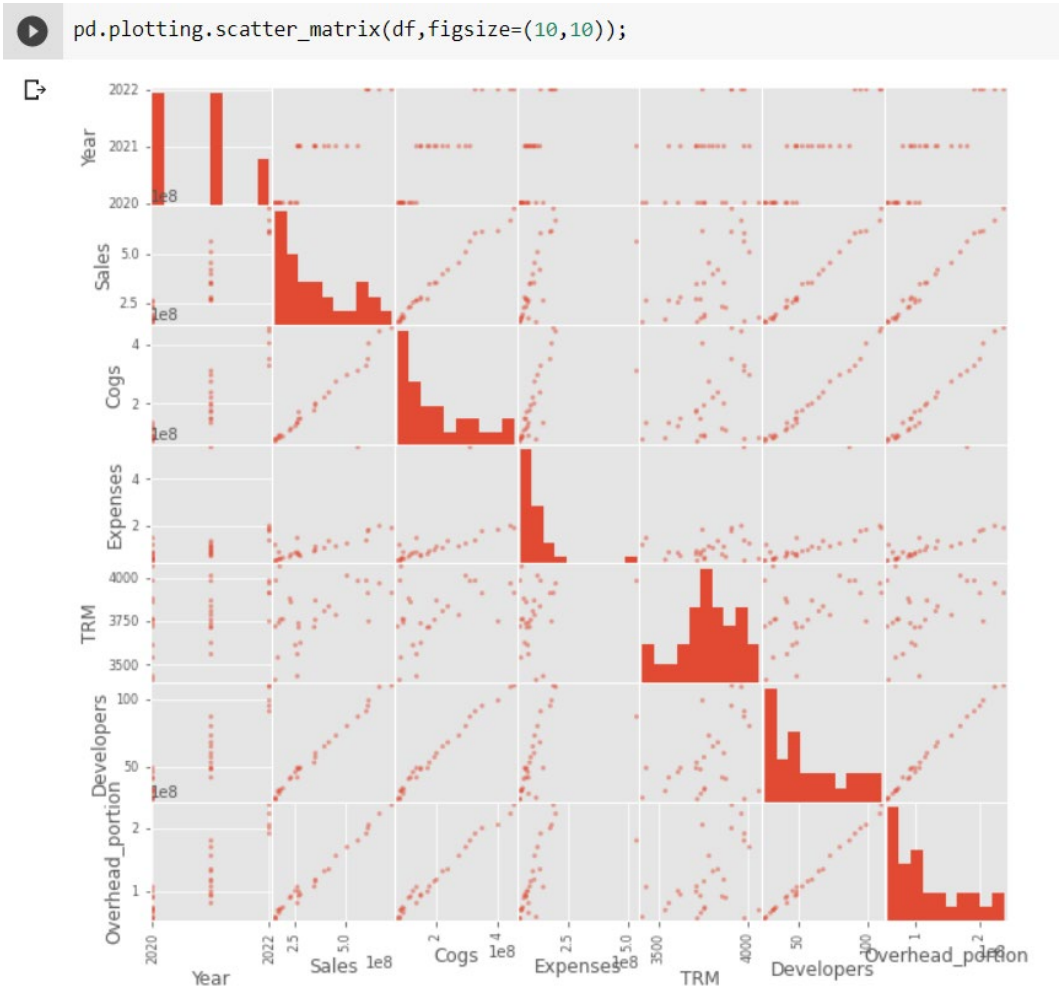

Figura 7. Matriz de correlación e histogramas.

Fuente: Elaboración propia

Posterior a esto, se grafica una matriz de correlación entre las variables con el fin de conocer su distribución a través de un histograma y un diagrama de dispersión para cada variable. Se encuentra que la mayoría de las variables con excepción de la variable TRM (la cual no hace parte del modelo) tienen asimetría positiva con un alargamiento de cola a la derecha, este fenómeno es común en los datos que tienen un orden cronológico y con tendencia creciente como las bases de datos financieros.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
from patsy import dmatrices
y, X=dmatrices('Overhead_portion ~ Sales + Cogs + Expenses + Developers', df,return_type = "dataframe")
vif = pd.DataFrame()
vif["VIF"]= [variance_inflation_factor(X.values,i)for i in range (X.shape[1])]
vif["Variables"]= X.columns
vif.round(3)
```

	VIF	Variables
0	27	Intercept
1	127	Sales
2	75	Cogs
3	2	Expenses
4	146	Developers

Figura 8. Test de multicolinealidad.

Fuente: Elaboración propia

Para el test de multicolinealidad se importa de la librería statsmodels el factor inflacionario de la varianza o "VIF" por sus siglas en inglés. El cual arroja un puntaje que, si llega ser mayor a 10, se entiende que esa variable explicativa tiene problemas de multicolinealidad con otras variables también explicativas.

En este caso todas las variables del modelo presentan multicolinealidad con excepción de la variable Expenses. Al analizar la teoría, cuando se estudian variables que posiblemente tengan datos que dependen unos de otros, en este caso como datos de ventas, costos y gastos, es muy posible que surja el problema de multicolinealidad. Hasta cierto punto la colinealidad es esperada entre las variables predictoras, sin embargo, cuando existe una relación lineal entre variables explicativas, se presenta un problema de distorsión en los resultados esperados. (Montero, 2016)

▾ Test de White (Prueba de Heterocedasticidad)

Ho : Existe Homocedasticidad en los residuos de las observaciones

H1 :Existe heterocedasticidad en los residuos

```
[ ] from statsmodels.stats.diagnostic import het_white
```

```
[ ] white_test = het_white(result.resid,result.model.exog)
    labels = ['LM Statistic','LM-Test p-value','F-Statistic','F-Test p-value']
    print(dict(zip(labels,white_test)))
```

```
{'LM Statistic': 10.776702804848858, 'LM-Test p-value': 0.46215572371944436, 'F-Statistic': 0.91393473729869, 'F-Test p-value': 0.5483420294360417}
```

Figura 9. Test de heteroscedasticidad.

Fuente: Elaboración propia

Para toda base de datos se tiene el supuesto (Ho) de que la varianza o dispersión de los residuos del modelo son constantes, sin embargo, cuando este supuesto no es real, existe heterocedasticidad, es decir, que la varianza de los residuos es diferente para los valores de x.

En este caso a través del test de heterocedasticidad de White, se puede comprobar si existe o no este fenómeno a través del LM-test p-value.

Para este caso el valor arrojado es se aleja de cero marcando un 0.5 en el LM-test p-value, por lo cual se entiende que no hay suficiente evidente para determinar que existe heterocedasticidad, dado que el valor no es menor a 0.05.

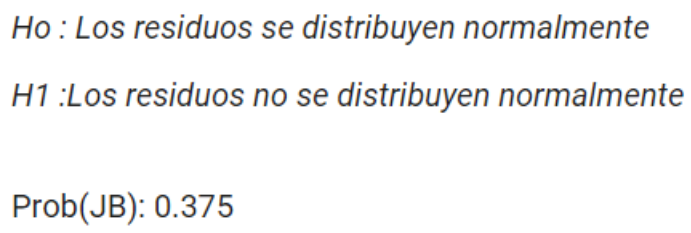

Figura 10. Test de normalidad de los residuos.
Fuente: Elaboración propia

El estadístico de probabilidad de Jarque-Bera arroja un valor de 0.375, por lo cual al ser mayor a 0.05 se acepta la Ho que dice que los residuos se distribuyen normalmente o que tienen a la normalidad. Como se puede apreciar también a continuación en el gráfico de residuos.

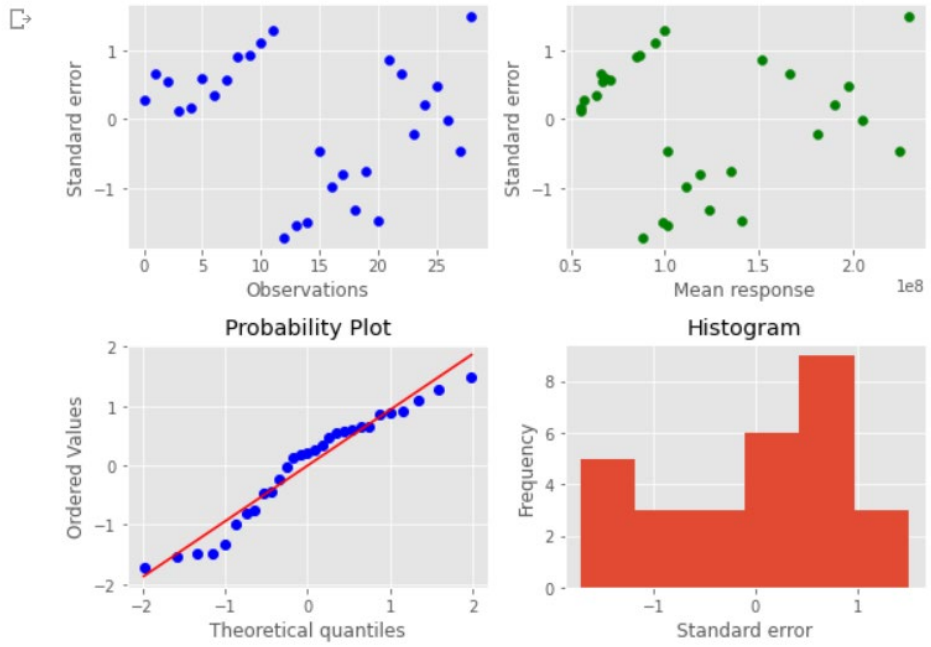

Figura 11. Gráfico de residuos.

Fuente: Elaboración propia

Ho: p=0

Ha: p>0

Si d < dL Rechazar Ho

Si d > dU No Rechazar Ho

Si dL > d > dU La prueba no es concluyente

Durbin-Watson: 0.967

dL = 1,124

dU = 1,743

Figura 12. Test de autocorrelación

Fuente: Elaboración propia

Para la prueba de autocorrelación se utiliza el test de Durbin-Watson, que establece un supuesto de que el valor p es igual a cero en la distribución de las variables, y que, si este supuesto no es real, entonces el valor p será mayor a cero.

Para lo cual se calcula el estadístico d, el cual es calculado en la función summary() del modelo. Por lo cual, teniendo en cuenta que el valor arrojado es 0.967 y teniendo en cuenta que

es menor al límite inferior para la tabla de valores Durbin-Watson, no se acepta la hipótesis nula y se acepta la hipótesis alternativa que dice que la correlación o valor p es mayor a cero entre las variables, por lo cual se determina que existe autocorrelación.

Teniendo en cuenta los resultados de las pruebas de bondad y ajuste, se establece que existe multicolinealidad, no hay suficiente evidencia para determinar heterocedasticidad y existe autocorrelación en el modelo de regresión múltiple, para lo cual se toma la decisión de depurar el modelo y explorar la idea de que un modelo de regresión lineal simple sería óptimo para explicar la variable dependiente.

Se realiza entonces la depuración de variables teniendo en cuenta el p-valor que arroja la función summary() y que en este orden de ideas la variable apropiada sería Developers. Por lo cual se procede a establecer un nuevo modelo predictivo en esta ocasión de naturaleza lineal simple entre la variable dependiente Overhead_portion y la variable predictora, Developers.

```
[ ]  model = smf.ols('Overhead_portion ~ Developers',df)
     result= model.fit()
     result.summary()
```

OLS Regression Results

Dep. Variable:	Overhead_portion	**R-squared:**	0.993
Model:	OLS	**Adj. R-squared:**	0.993
Method:	Least Squares	**F-statistic:**	4011.
Date:	Sat, 23 Jul 2022	**Prob (F-statistic):**	6.66e-31
Time:	10:27:53	**Log-Likelihood:**	-484.18
No. Observations:	29	**AIC:**	972.4
Df Residuals:	27	**BIC:**	975.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.61e+05	2.04e+06	0.276	0.785	-3.62e+06	4.74e+06
Developers	2.066e+06	3.26e+04	63.335	0.000	2e+06	2.13e+06

Omnibus:	2.287	**Durbin-Watson:**	0.909
Prob(Omnibus):	0.319	**Jarque-Bera (JB):**	1.825
Skew:	-0.466	**Prob(JB):**	0.401
Kurtosis:	2.199	**Cond. No.**	153.

Figura 13. Modelo de regresión lineal simple

Fuente: Elaboración propia

El modelo arroja indicadores R cuadrado y R ajustado tendientes a 1, lo cual es buen indicio y un p-valor menor a 0.05, por consiguiente se procede a particionar los datos para graficarlos y empezar a aplicar el algoritmo de machine learning X_train,X_test,y_train,y_test con el fin de entrenar el 80% de los datos y luego probarlos en el 20 % restante para ver si la predicción es consistente a través del score del modelo.

```
[ ]  X = df[['Developers']]
     y = df['Overhead_portion']
```

```
[ ]  plt.scatter(X,y)
     plt.xlabel('Developers')
     plt.ylabel('Overhead_portion')
     plt.show()
```

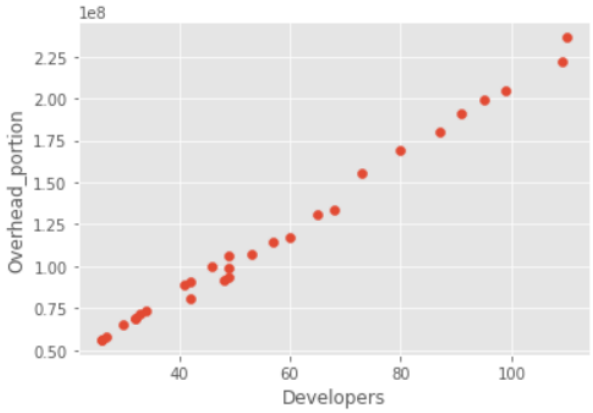

Figura 14. Partición de datos y gráfico de dispersión.

Fuente: Elaboración propia

Se observa que al graficar ambas variables, Overhead_portion y Developers, su distribución intuye una relación lineal ajustada y positiva.

```
[ ]  X_train,X_test,y_train,y_test= train_test_split(X,y,test_size=0.20)
```

```
[ ]  regresion = linear_model.LinearRegression()
     regresion.fit(X_train,y_train)

     LinearRegression()
```

Figura 15. Entrenamiento del algoritmo X_train, X_test, y_train, y_test.

Fuente: Elaboración propia

Se ajusta el algoritmo con los datos ya particionados para empezar el entrenamiento y prueba de los datos estudiados.

```
[ ]  print(regresion.score(X_train,y_train))

     0.9920560313957136
```

Figura 16. Precisión del modelo de regresión lineal simple.

Fuente: Elaboración propia

El modelo tiene una precisión excelente de 0.99 lo cual indica que ambas variables tienen una alta correlación.

```
plt.scatter(X_train,y_train)
plt.plot(X_test,y_pred,color= "blue",linewidth =3)
plt.title("Regresión lineal simple")
plt.show

<function matplotlib.pyplot.show>
```

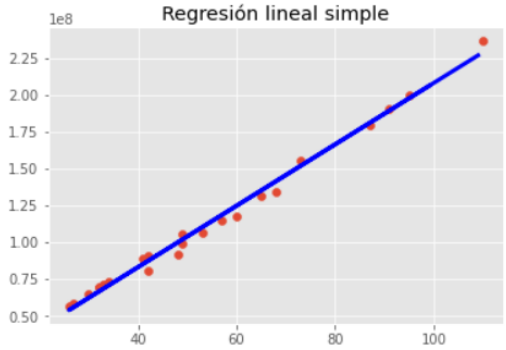

Figura 17. Gráfico de dispersión y recta de ecuación lineal.

Fuente: Elaboración propia

```
[ ]  sns.residplot(x='Developers', y='Overhead_portion', data=df)

     plt.show()
```

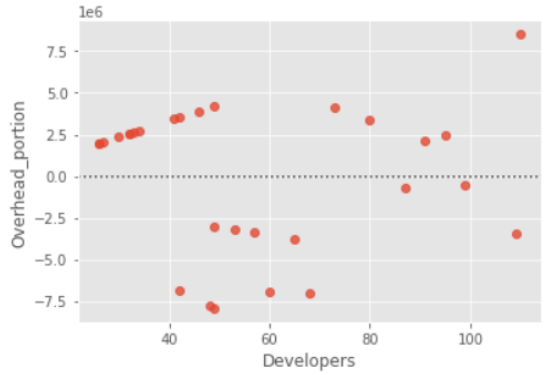

Figura 18. Gráfico de dispersión de residuos.

Fuente: Elaboración propia

Se grafica la línea de regresión y el diagrama de dispersión de los residuos del modelo, lo cual muestra que existe homocedasticidad dado que se encuentran distribuidos a lo largo del eje de las X.

```
[ ]  print('Y = ',regresion.coef_,'X + ',regresion.intercept_)

     Y =  [2079619.99371779] X +  -71938.56284120679
```

Figura 19. Ecuación del modelo de regresión lineal simple.

Fuente: Elaboración propia

Comparación de los resultados del entrenamiento de los datos con los datos reales del dataset.

```
[ ] data_tested = pd.DataFrame(y_test)
    data_tested
```

	Overhead_portion
26	204508247
22	169232214
4	56238000
10	99498000
13	93846667
27	222299021

```
[ ] data_predicted = pd.DataFrame(y_pred,columns=['Overhead_portion'])
    data_predicted
```

	Overhead_portion
0	205810441
1	166297661
2	53998181
3	95590581
4	101829441
5	226606641

Figura 20. Comparación de datos reales con datos entrenados.

Fuente: Elaboración propia

En vista de que se evidencia similitud y un score cercano a uno en el algoritmo de machine learning, se toma como buena evidencia, que el modelo es óptimo.

```
[ ] print(regresion.score(X_test,y_test))

    0.9949990810860785
```

Figura 21. Precisión de la predicción.

Fuente: Elaboración propia

Conclusiones

Al hacer un análisis de regresión múltiple, se debe tener en cuenta que el R cuadrado y el R ajustado no son determinantes a la hora de establecer si el modelo es adecuado para predecir la variable dependiente.

Las pruebas de multicolinealidad, heterocedasticidad y autocorrelación en conjunto, conforman un paquete de herramientas esenciales a la hora de probar la bondad de ajuste de un modelo de regresión múltiple y pueden dar indicios de un muestro pobre o de que existe la necesidad de tener estimadores robustos.

Eliminar las variables con altos índices de multicolinealidad no siempre es adecuada, sin embargo, dadas las características de los valores P para las variables Sales, Cogs y Expenses, luego de eliminarse de la ecuación de regresión arrojaron buen resultado.

Python es un excelente lenguaje de programación con amplia variedad de librerías disponibles para aplicar algoritmos de machine learning y aprendizaje estadístico.

Referencias

Acosta, S., Laines, B. y Piña, G. (2013) *Estadística inferencial*. Universidad Peruana de Ciencias Aplicadas (UPC). http://hdl.handle.net/10757/292942

Cardona, D., González, J., Rivera, M., Cárdenas, E. (2013) *Inferencia estadística Módulo de regresión lineal simple. Documento de investigación* No. 147. Editorial Universidad del Rosario.

Levin, R. I. y Rubin, D. S. (2004). *Estadística para administración y economía.* Pearson Educación.

Montero, R. (2016): *Modelos de regresión lineal múltiple.*
Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España.

Pereira, A. (2009-2010). *Análisis predictivo de datos mediante técnicas de regresión estadística* [Tesis de Maestría, Máster en Investigación en Informática, Universidad Complutense de Madrid].

Pérez, R. (2019) *Modelación Financiera. Conceptos y aplicaciones*. Universidad Piloto de Colombia.

Salazar, C. y Del Castillo, S. (2018). *Fundamentos Básicos de Estadística.* Primera Edición.