CASO DE ESTUDIO: ESTIMACIÓN DE HORAS PARA LA EJECUCIÓN DE PROYECTOS DE INGENIERÍA DE SOFTWARE A PARTIR DE MODELOS DE REGRESIÓN LINEAL

Case study: Estimation of hours for the execution of software engineering projects based on linear regression models

Luis Carlos Padilla Suárez
lcpadillas@libertadores.edu.co

Director:
José John Fredy González Veloza
jjgonzalezv02@libertadores.edu.co

Especialización en Estadística Aplicada, Facultad de Ingeniería y Ciencias Básicas, Fundación Universitaria Los Libertadores, Bogotá D.C, Colombia, 2021

RESUMEN

Para el desarrollo de este caso de estudio se exploró información histórica de proyectos de fábrica de software, fábrica de pruebas, servicios profesionales y consultoría de una compañía de ingeniería de software de Colombia con presencia en diferentes países de América. Se realizaron análisis descriptivos a la información disponible y se implementaron técnicas de transformación de datos para identificar y seleccionar variables predictoras para generar y comparar diferentes modelos de regresión lineal múltiple, a los cuales se les realizó análisis de supuestos de normalidad, homocedasticidad e independencia de los errores. Finalmente, se realiza la comparación de los diferentes valores R-cuadrado ajustado obtenidos de los modelos de regresión y se seleccionó el modelo que mejor representa la variabilidad observada en la variable objetivo. Como resultado de este caso de estudio, se obtiene un modelo de regresión lineal múltiple y un modelo de regresión robusto como las mejores opciones para estimar los proyectos de ingeniería de software de la compañía.

Palabras clave: Análisis descriptivo, Transformación de datos, Estimación de proyectos de ingeniería de software, Regresión lineal múltiple, Mínimos cuadrados lineales (OLS), Regresión lineal robusta (RLM).

ABSTRACT

For the development of this case study, historical information of software factory projects, test factory, professional services, and consulting of a software engineering company in Colombia with a presence in different countries of America was explored. Descriptive analyzes were carried out on the available information and data transformation techniques were implemented to identify and select predictor variables to generate and compare different multiple linear regression models, to which normality, homoscedasticity and independence assumptions were analyzed. mistakes. Finally, the different adjusted R-squared values obtained from the regression models are compared and the model that best represents the variability observed in the target variable was selected.

Keywords: Descriptive analysis, Data transformation, Estimation of software engineering projects, Multiple linear regression, Ordinary least squares (OLS), Robust linear regression (RLM).

INTRODUCCIÓN

De acuerdo con el portal oficial PROCOLOMBIA, en 2019, Colombia se posicionó como el cuarto mercado de TI más grande en Latinoamérica, después de Brasil, México y Chile, alcanzando los USD 8,2 mil millones en el mercado de software y servicios TI (Procolombia, 2020). Como consecuencia, a nivel mundial se ha venido identificado a Colombia como ubicación estratégica y altamente competitiva para las potencias en el desarrollo e ingeniería de software, control de calidad, pruebas de software y automatización, como lo es la compañía estadounidense Intertec International (Forbes-Colombia, forbes.co, 2021). A raíz de este constante y acelerado crecimiento en el mercado, las compañías colombianas han identificado la necesidad de fortalecer sus estructuras organizacionales

con el fin de garantizar productos y procesos de la más alta calidad, capaces de competir en el mercado global. Así mismo, incrementar inversión en factores relacionados con la retención, cultura empresarial, y motivación de los empleados (Forbes-Colombia, forbes.co, 2020), ya que estos son el factor clave de éxito de toda organización.

El CMMI Institute es un modelo organizacional de más de 25 años de evolución continua, que compila un conjunto de mejores prácticas orientadas a llevar la capacidad y desempeño de la compañía hacia un alto nivel de madurez (CMMI-Institute, 2018). Este modelo se presenta como una solución práctica para afrontar los nuevos retos originados por la globalización del mercado, ya que su objetivo es identificar y reducir los aspectos más importantes que generan incertidumbre en el cumplimiento de los objetivos estratégicos organizacionales.

Puntualmente para este caso de estudio, el modelo CMMI Development versión 2.0 ofrece una guía orientada a fortalecer los procesos de estimación de proyectos de desarrollo de soluciones técnicas, basada en el análisis estadístico de información histórica de la compañía. Para esto, se hace necesario establecer una línea base del proceso, la cual permite a la compañía identificar las métricas de software más relevantes para describir, comprender y predecir su comportamiento durante el desarrollo de sus diferentes proyectos, y de esta manera la productividad en sus entregables y la calidad en los procesos. La línea base del proceso se conforma con una muestra de proyectos ejecutados previamente, que contiene datos como línea de negocio, tamaño del equipo de trabajo, estimación inicial del proyecto, tecnología y lenguaje utilizados, entre otras variables descriptivas que puedan afectar el comportamiento de los proyectos analizados. Esta información debe ser precisa, abundante, consistente y similar en cuanto a la metodología de trabajo (Pressman, 2005), por ejemplo, no es posible generar línea base con una muestra de proyectos ejecutados bajo el marco de trabajo tipo Cascada y otros bajo el marco de trabajo tipo Scrum, ya que son metodologías totalmente diferentes que afectan

considerablemente el desarrollo de los proyectos y por lo tanto no son comparables en un análisis estadístico.

De acuerdo con Stephen H. Kan “Las métricas de software pueden ser clasificadas en tres categorías: métricas del producto, métricas del proceso y métricas del proyecto. Las métricas del producto describen características del producto tales como tamaño, complejidad, características de diseño, rendimiento y nivel de calidad. Las métricas del proceso pueden ser utilizadas para mejorar el proceso de desarrollo y mantenimiento del software. Algunos ejemplos son efectividad de la remoción de defectos durante el desarrollo y el tiempo de respuesta en el proceso de corrección de defectos. Las métricas del proyecto describen las características y ejecución del proyecto. Algunos ejemplos son: el número de desarrolladores de software, el comportamiento del personal durante el ciclo de vida de éste, el costo, el cronograma y la productividad. Algunas métricas pertenecen a múltiples categorías. Por ejemplo, las métricas de calidad del proceso de un proyecto son ambas métricas del proceso y métricas del proyecto” (Kan, 2002). Debido a que la compañía objeto de este caso de estudio no alcanza un nivel de madurez suficiente para generar las métricas anteriormente descritas, este caso de estudio se limita a las métricas del proyecto.

REFERENTES TEÓRICOS

Los proyectos de ingeniería de software ejecutados por la compañía se desarrollan bajo el marco de trabajo tradicional o también llamado método de cascada (Figura 1), en el que se ejecutan fases secuencialmente de manera ordenada iniciando con el levantamiento de información y finalizando con la fase de paso a producción. En la actualidad, se está implementando la metodología de trabajo ágil o Scrum, la cual integra las fases del marco tradicional y las distribuye en diferentes iteraciones de acuerdo con el número de funcionalidades planeadas para el proyecto. Para el presente caso de estudio se toman como referencia únicamente proyectos de marco tradicional.

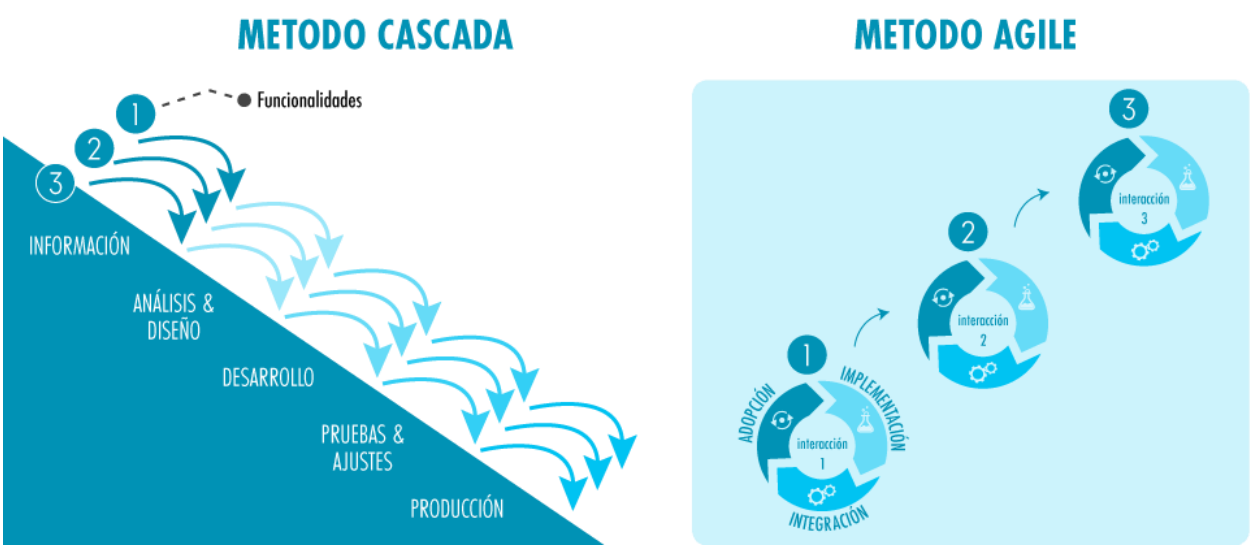

Figura 1. Metodologías de trabajo: Cascada vs. Scrum (Domé, 2018)

La compañía de estudio se encuentra valorada en CMMI DEV nivel 3, por lo tanto no tiene implementadas dentro de sus procesos técnicas de análisis cuantitativo para la estimación de proyectos que se describen en el nivel 4 y 5 de madurez (Figura 2.). Actualmente, la compañía determina la duración de sus proyectos haciendo uso de plantillas de estimación basadas en técnicas de análisis de juicio de expertos, en la que un grupo técnicamente especializado define las principales características del proyecto, y basados en su experiencia de proyectos pasados, establecen un total de horas estimadas para la ejecución de este, y de esta manera se genera la propuesta comercial al cliente.

Este método de estimación utilizado por la compañía genera riesgo en los proyectos en temas relacionados con el cumplimiento de los tiempos pactados con el cliente, la utilidad del negocio y la percepción del cliente en la relación calidad/precio (competitividad en el mercado). De acuerdo con análisis realizados por el área de calidad y procesos de la compañía se ha identificado que en una muestra de 174 proyectos ejecutados entre 2018 y 2020 (8% del total de proyectos de ese periodo de tiempo), la desviación en la estimación de los proyectos se encuentra en un rango aproximado de ± 440 horas, en la que una desviación positiva implica pérdidas económicas para la compañía y una

desviación negativa representa al cliente sobrecostos del proyecto. Para mayor entendimiento, la desviación en la estimación se define bajo la siguiente estructura.

$$Desviación\ en\ la\ estimación = \frac{Esfuerzo\ real\ ejecutado - Esfuerzo\ estimado}{Esfuerzo\ estimado}$$

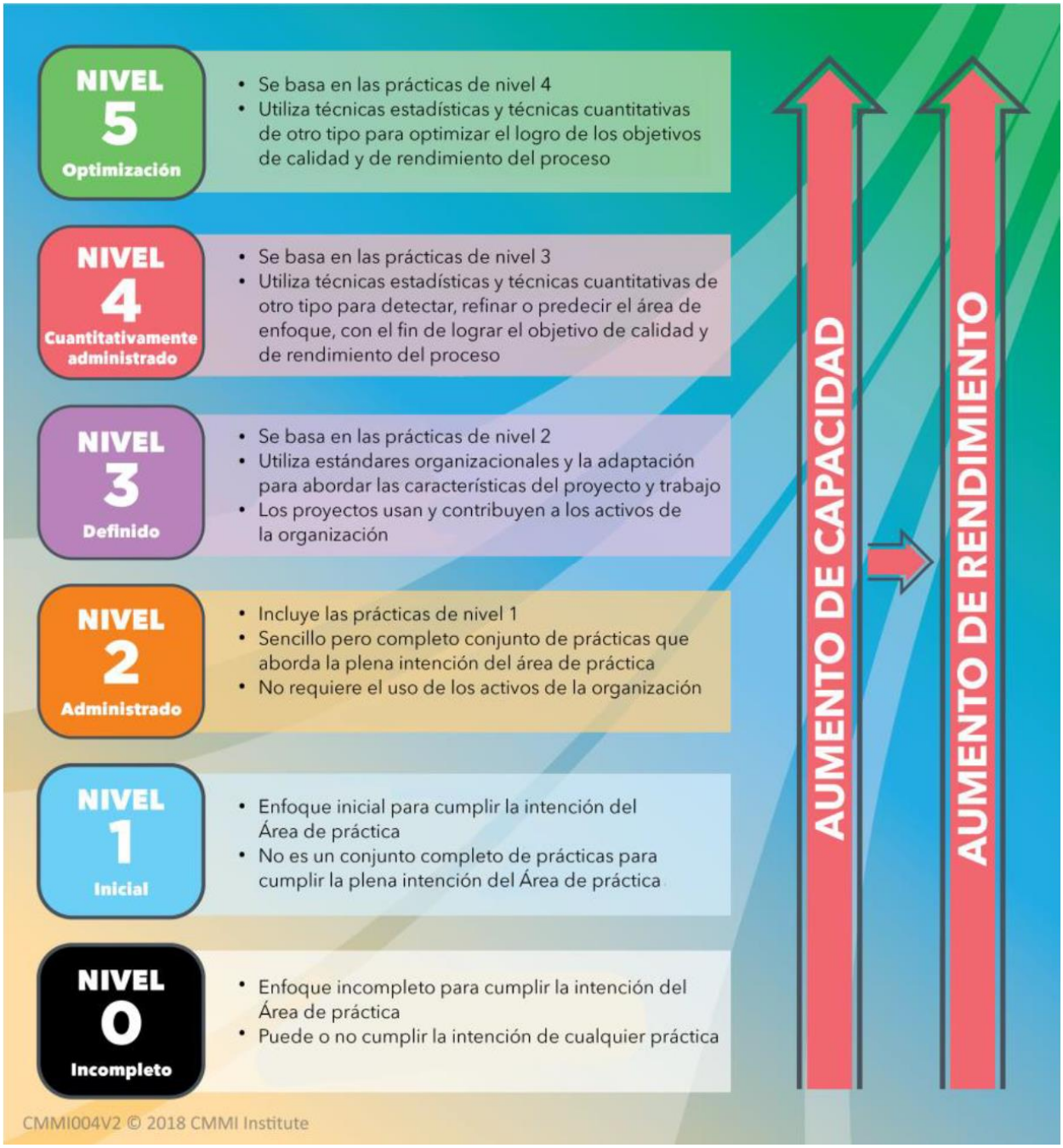

Figura 2. Definición de los niveles evolutivos del modelo CMMI (CMMI-Institute, 2018)

Adicionalmente, la organización cuenta con diferentes mecanismos para recolectar información de los proyectos que ejecuta, y esta información no está siendo aprovechada para fortalecer la planeación de proyectos a través de la implementación de técnicas de estimación y control basadas en la gestión cuantitativa.

De acuerdo con el texto Pronósticos en los negocios, todas las organizaciones, sin importar su tamaño o actividad económica, pueden reducir el grado de incertidumbre del entorno a través de la elaboración de pronósticos precisos y confiables que sean suficientes para satisfacer las necesidades en la planeación de la organización (Wichern, 2006).

El modelo de regresión lineal múltiple es aplicable en casos de estudio en los que existen dos o más variables predictoras para una única variable respuesta u objetivo. Este tipo de modelos se define bajo la siguiente estructura.

$$\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Donde γ hace referencia a la variable respuesta, X_n a las variables predictoras, ϵ al error aleatorio del modelo, β_n corresponden los coeficientes de regresión que transforman a las variables predictoras del modelo y β_0 es el parámetro que representa el intercepto del modelo, es decir, el valor esperado de la variable respuesta γ cuando todas las variables predictoras son iguales a cero (Montgomery, 2002).

Todo modelo de regresión lineal que se genere debe ser evaluado para determinar si los errores del modelo cumplen con los supuestos de no multicolinealidad, homocedasticidad e independencia (Montgomery, 2002).

En ocasiones las bases de datos utilizadas para generar los modelos de regresión lineal no presentan la distribución esperada. Por ejemplo, la variable respuesta debe tener una distribución normal, la cual se puede verificar a través de histogramas de frecuencia o mediante la prueba de hipótesis de normalidad.

Cuando la variable respuesta no tiene esta distribución, es posible realizar una transformación logarítmica a las variables predictoras y/o variable respuesta, con el fin de mejorar la distribución de estas y corregir algunos de los comportamientos que se espera se identifiquen con los supuestos descritos anteriormente (Tusell, 2011).

Finalmente, es importante evaluar el margen de error de los modelos generados, esto con el fin de identificar cual es la diferencia entre el valor pronosticado y el valor real en las muestras analizadas, el objetivo consiste en seleccionar aquel modelo que presente el menor valor en el error cuadrático medio, descrito bajo la siguiente estructura. Donde γ_i corresponde al valor real de la muestra N, mientras que $\hat{\gamma}_i$ corresponde al valor pronosticado por el modelo (Tusell, 2011).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\gamma_i - \hat{\gamma}_i)^2$$

METODOLOGÍA

La estrategia metodológica empleada para desarrollar el presente caso de estudio se base en cinco fases descritas a continuación, las cuales son factores clave para la generación adecuada de modelos de regresión lineal:

Fase 1. Caracterización del proceso y análisis de variables

Realizar el correspondiente análisis descriptivo de las potenciales variables cuantitativas y cualitativas predictoras y de la variable respuesta objetivo, con el fin comprender el comportamiento de las variables, su distribución e identificar posibles muestras atípicas.

Fase 2. Transformación de variables y detección de casos atípicos

Implementar técnicas para la transformación de datos, con el objetivo de moldear las variables hacia una distribución normal reduciendo posibles sesgos, y otras técnicas para la detección de muestras

atípicas. Esta modificación y limpieza de la base de datos es una etapa crucial para la generación de modelos de regresión de mayor precisión.

Fase 3. Selección de variables

Generar análisis de correlación de las variables que influyen en el pronóstico del esfuerzo real requerido para la ejecución de proyectos de ingeniería de software. Este análisis provee la información necesaria para identificar las variables más representativas que impactarán en el modelo de regresión.

Fase 4. Creación del modelo de regresión lineal múltiple

Implementar métodos estadísticos para generar el mejor modelo de regresión lineal múltiple que estime el número de horas requeridas para la ejecución de un proyecto, tomando como referencia las variables predictoras analizadas en fases previas.

Fase 5. Validación del modelo

Realizar las validaciones pertinentes al modelo de regresión generado, correspondientes a los supuestos de normalidad, homocedasticidad e independencia de los errores.

RESULTADOS

Análisis descriptivo de variables

La compañía objeto de este caso de estudio tiene la necesidad de generar un modelo de regresión que estime el número de horas totales requeridas para la ejecución de proyectos de ingeniería de software, usando como variables predictoras la información disponible en sus repositorios documentales, entre las cuales se destacan:

Variables categóricas

- **Cliente:** Define el nombre del cliente con el cual se ejecutó el proyecto de ingeniería de software, la base de datos contiene 82 diferentes clientes situados en 5 diferentes ciudades/locaciones.

- **Línea de negocio:** Determina el tipo de proyecto que se ejecutó, actualmente la organización cuenta con 4 líneas de negocio: Fábrica de software, Fábrica de pruebas, Servicios profesionales y Consultoría, las dos primeras bajo la metodología de trabajo tradicional (cascada) y las dos últimas bajo metodología del servicio definida por la organización.

Variables numéricas enteras

- Colaboradores: Representa la cantidad de personas que hicieron parte de equipo de trabajo del proyecto.
- Cargos: Los equipos de trabajo están conformados por cargos de operación, entre los cuales se encuentra Consultor, Arquitecto, Desarrollador y Analista, y cargos de apoyo como Gerente y Otros, estas variables de cargo representan la cantidad de personas que hicieron parte del equipo de trabajo especificados cada cargo.
- Días de ejecución: Hace referencia al número de días estimados para la ejecución total del proyecto.
- Año: Representa el año en el que se dio inicio al proyecto, la base de datos de este caso de estudio comprende proyectos iniciados y ejecutados entre 2012 y 2020.

Variable objetivo

- Total de horas consolidadas: Variable que representa el número total de horas reportadas por el equipo de trabajo para la ejecución total de un proyecto de ingeniería de software.

La base de datos está conformada por 2.243 muestras de proyectos de ingeniería de software ejecutados entre 2012 y 2020, distribuidos entre las cuatro diferentes líneas de negocio como se observa en la Figura 3.

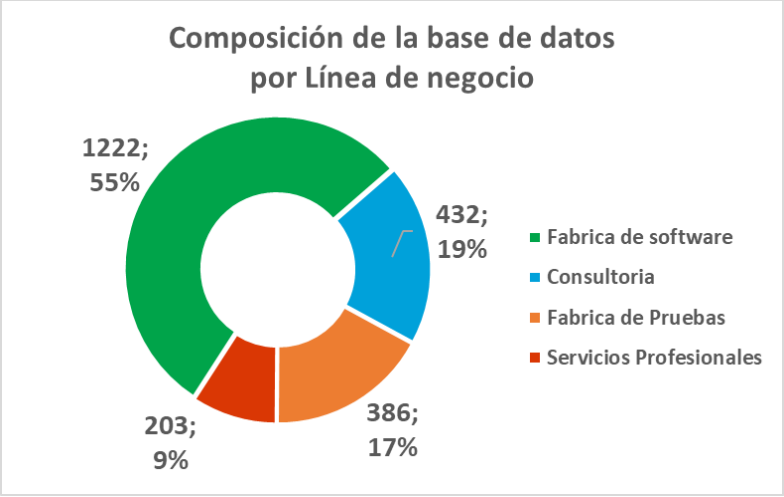

Figura 3. Distribución de proyectos por línea de negocio

En la Tabla 1 se detallan los valores resultados del análisis descriptivo de las variables predictoras cuantitativas.

Análisis descriptivo variables cuantitativas	Días de ejecución	Colaboradores	Gerente	Otros
Muestra	2243	2243	2243	2243
Media	369,5	5,1	0,3	0,1
Desviación Estándar	425,9	6,7	0,6	0,4
Valor mínimo	0,0	1,0	0,0	0,0
Cuartil 1 (25%)	122,0	1,0	0,0	0,0
Cuartil 2 (50%)	232,0	3,0	0,0	0,0
Cuartil 3 (75%)	424,0	6,0	0,0	0,0
Valor máximo	4720,0	75,0	4,0	3,0
Análisis descriptivo variables cuantitativas	**Consultor**	**Arquitecto**	**Desarrollador**	**Analista**
Muestra	2243	2243	2243	2243
Media	3,2	0,1	0,2	1,2
Desviación Estándar	4,3	0,3	0,8	3,4
Valor mínimo	0,0	0,0	0,0	0,0
Cuartil 1 (25%)	1,0	0,0	0,0	0,0
Cuartil 2 (50%)	2,0	0,0	0,0	0,0
Cuartil 3 (75%)	4,0	0,0	0,0	1,0
Valor máximo	36,0	9,0	13,0	56,0

Tabla 1. Análisis descrptivo variables predictoras cuantitativas

La Tabla 2 muestra el análisis descriptivo realizado a las variables predictoras cualitativas, y la Tabla 3 especifica el análisis descriptivo de la variable objetivo del modelo de regresión lineal. Se puede observar que las variables predictoras cuantitativas presentan un sesgo hacia la derecha, ya que las muestras ubicadas en el último cuartil se encuentran significativamente más alejadas de la mediana que las muestras ubicadas en el primer cuartil.

Análisis descriptivo variables cualitativas	Cliente	Línea de negocio
Muestra	2243	2243
Categorías Ciudad	5	4
Moda	Bogotá	Fábrica de software
Frecuencia moda	2096	1222

Tabla 2. Análisis descrptivo variables predictoras cualitativas

Análisis descriptivo variable objetivo	Total de horas consolidadas
Muestra	2243
Media	2399,5
Desviación Estándar	6298,1
Valor mínimo	10
Cuartil 1 (25%)	149,3
Cuartil 2 (50%)	545
Cuartil 3 (75%)	2003,7
Valor máximo	107020,5

Tabla 3. Análisis descrptivo variable objetivo

En la Figura 4 se puede evidenciar que la variable objetivo no sigue una distribución normal, lo mismo ocurre con las demás variables cuantitativas de la base de datos, por lo tanto se hace necesario realizar una transformación logarítmica para modificar estos valores hacia una distribución normal.

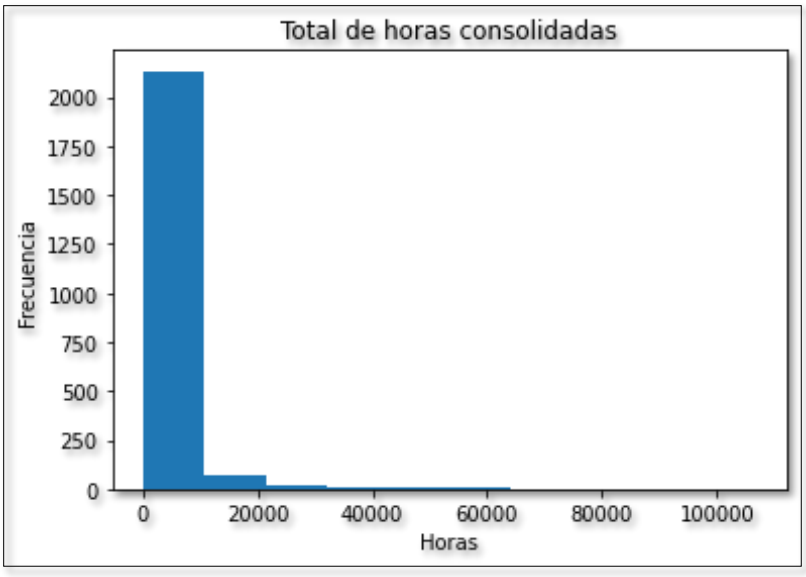

Figura 4. Distribución de la variable objetivo Total de horas consolidadas

A través de la Figura 5, se puede observar que la variable objetivo aumenta a medida que aumenta el número de colaboradores que conforman el equipo de trabajo, al realizar el gráfico de dispersión diferenciado por línea de negocio se puede identificar visualmente una pequeña variación en la pendiente de las muestras, por ejemplo, la pendiente de la línea de negocio relacionada con servicios profesionales resulta ser menor a la pendiente de los proyectos tipo consultoría, al igual que los proyectos de fábrica de software tienden a tener una pendiente menor a los proyectos de fábrica de pruebas.

Por otro lado, mediante la ejecución de la librería Sweetviz de Python, su pudo observar que las variables Colaboradores, Analista, Consultor y Cliente presentan una correlación superior a 0.55 con respecto a la variable objetivo. Pero al realizar un análisis más detallado sobre las variables relacionadas con el número de colaboradores (Colaboradores vs. Cargos) se pudo identificar que se presenta una alta correlación entre estas, por lo tanto se identifica un incumplimiento al supuesto de no multicolinealidad. Este incumplimiento requiere de la eliminación de alguna de las variables predictoras que presentan alta correlación entre sí.

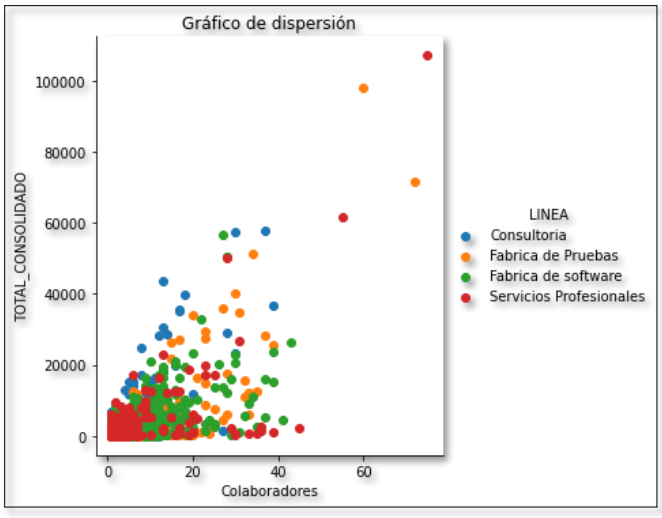

Figura 5. Distribución de la variable objetivo Total de horas consolidadas

Transformación logarítmica de las variables predictoras

Se realiza la transformación logarítmica a la variable objetivo para normalizar su distribución. Para las variables predictoras de días de ejecución, colaboradores y cargos, de igual manera se realiza la transformación logarítmica, con la diferencia que para las últimas dos variables se realiza una previa adición de una unidad a todos sus datos, con el fin de eliminar de la base aquellas muestras iguales a cero.

La Figura 6 muestra que al transformar la variable objetivo, ahora presenta una distribución normal. Por otro lado, en la Figura 7 se presenta la distribución de las variables cuantitativas con la transformación logarítmica descrita.

Por otro lado, de manera paralela a la transformación logarítmica, se realiza una transformación a las variables relacionadas con los cargos: Desarrollador, Analista, Consultor y Arquitecto, se agrupan estos dos primeros para formar una única variable y se realiza el mismo procedimiento con los dos últimos, ya que se tiene conocimiento que esta agrupación puede relacionarse a las funcionalidades que ofrecen estos dos grupos, los primeros siendo operativos y los segundo estratégicos.

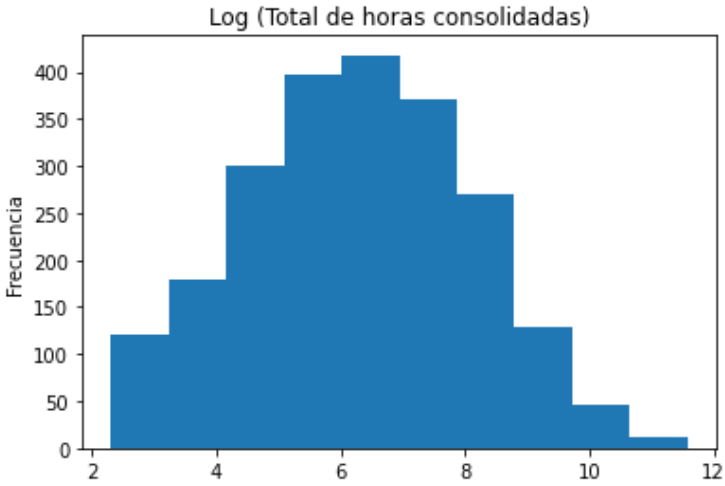

Figura 6. Distribución Total de horas consolidadas con transformación logarítmica

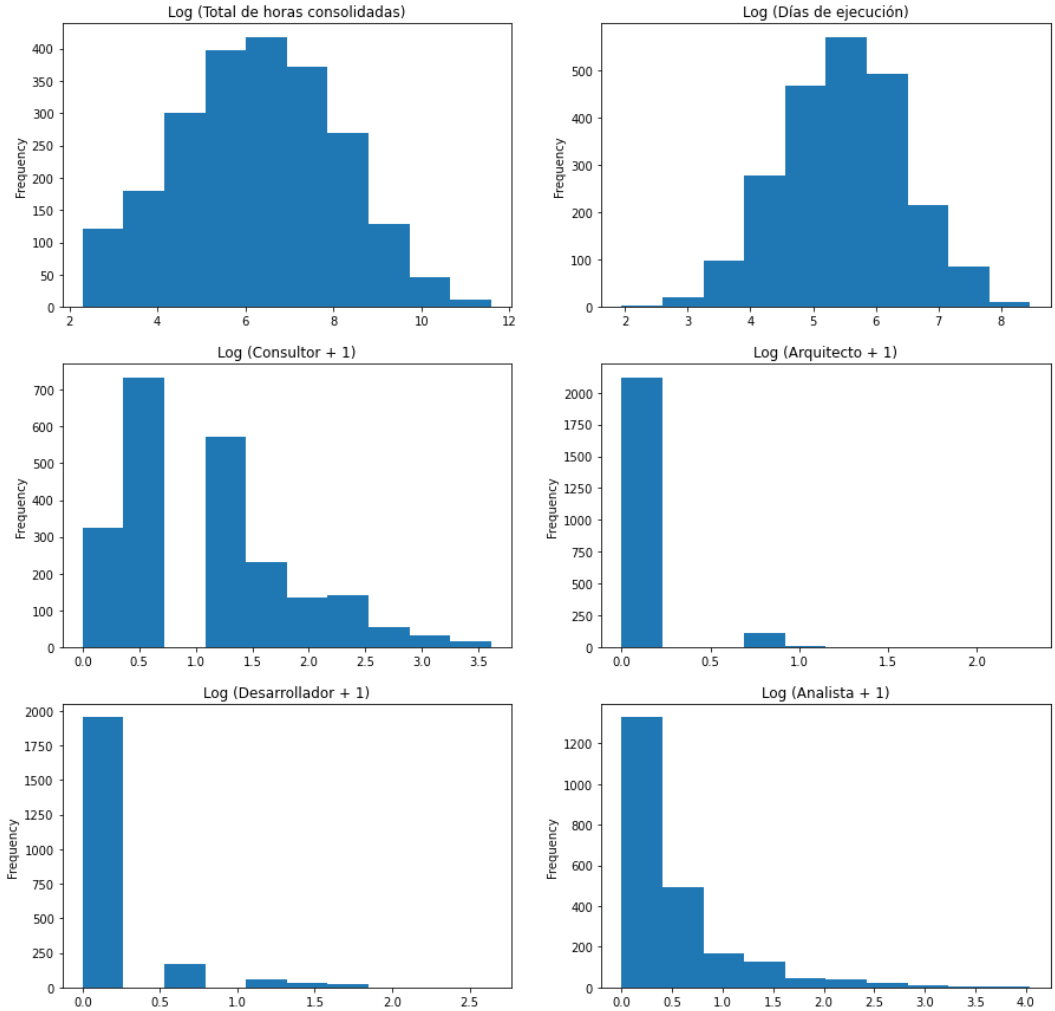

Figura 7. Distribución variables cuantitativas con transformación logarítmica

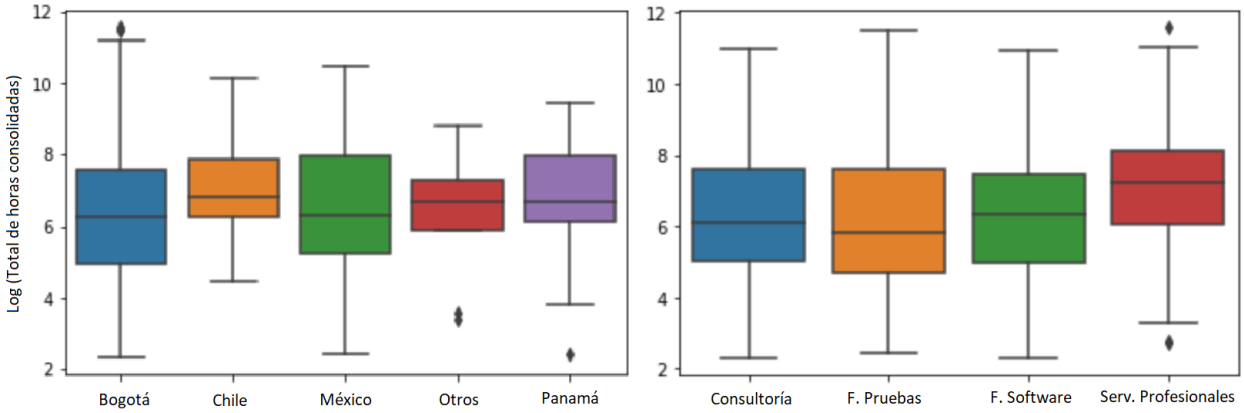

Figura 8. Boxplot Variables cualitativas vs. variable objetivo

La Figura 8 muestra el comportamiento de las variables cualitativas con respecto a la transformación de la variable objetivo, se puede evidenciar gráficamente que las variables Ciudad y Línea de negocio presentan comportamientos diferentes en cada una de sus categorías.

Selección de variables predictoras para el modelo de regresión

Para determinar la correlación de las variables predictoras con la variable objetivo, se hace uso de la matriz de correlación que se describe en la Tabla 4. Consultor, Días de ejecución, Analista y Desarrollador son las variables más representativas para predecir el comportamiento de la variable objetivo.

Variable 1	Variable 2	r
Consultor	**Total horas consolidadas**	0,4941
Total horas consolidadas	**Consultor**	0,4941
Total horas consolidadas	**Días de ejecución**	0,4879
Días de ejecución	**Total horas consolidadas**	0,4879
Días de ejecución	**Consultor**	0,3813
Consultor	**Días de ejecución**	0,3813
Analista	**Total horas consolidadas**	0,3635
Total horas consolidadas	**Analista**	0,3635
Desarrollador	**Total horas consolidadas**	0,2638
Total horas consolidadas	**Desarrollador**	0,2638

Tabla 4. Matriz de correlación entre variables cuantitativas

Al realizar la prueba ANOVA entre la variable predictora Línea de negocio y la variable objetivo Total de horas consolidadas (Ver Tabla 5), se obtiene que las medias de las cuatro categorías de la variable predictora son diferentes entre sí, por lo tanto es posible afirmar la variable Línea de negocio debe ser incluida en el modelo de regresión lineal múltiple, por lo que existe evidencia estadística significativa que demuestra que las diferentes líneas de negocio que maneja la compañía se comportan de manera diferente.

Variables analizadas	Estadístico F	P Valor
Categorías de la variable Línea de Negocio	16.28	$1.8e^{-10}$
Categorías de la variable Cliente	3.62	0.006

Tabla 5. Resultados prueba ANOVA Variables cualitativas vs. Total consolidado de horas

Como se observa, el P Valor en ambos casos es significativamente inferior a 0.05, por lo tanto es posible afirmar que las medias del total consolidado de horas por cada línea de negocio y por cada cliente son diferentes entre sí, lo que hace a estas variables cualitativas se consideren importantes para la generación de un modelo de regresión.

Modelos de regresión lineal múltiple

Teniendo en cuenta los resultados obtenidos en la sección anterior, se generan modelos de regresión lineal múltiple usando como variables predictoras: Línea de negocio, Cliente, Días de ejecución, Consultor, Desarrollador y Analista. Inicialmente se genera un modelo de regresión tomando la base de datos original, posteriormente se generan modelos secundarios a partir de las transformaciones descritas anteriormente, transformaciones logarítmicas y agrupación de variables. Finalmente, de genera un modelo de regresión lineal robusto. Para esta etapa, se utilizaron las librerías sklearn.linear_model y statsmodels.api con los modelos Ordinary Least Squares OLS y Robust Linear Model Regression RML.

Para el primer modelo de regresión, generado a partir de la base de datos original y con la librería sklearn.linear_model, se obtiene un error medio cuadrático de 4045.8 horas, lo que en la práctica puede percibirse como un error significativo ya que puede representar una fracción considerable del total de horas estimadas para la ejecución de un proyecto.

Al generar un segundo modelo de regresión usando la librería statsmodels.api y tomando con base la misma base de datos original, es decir, generar el mismo modelo de regresión pero con librerías diferentes, se obtiene un R-cuadrado ajustado de 0.397, lo que implica que el modelo es capaz de explicar solamente 39.7% de la variabilidad observada en el total de horas necesarias para la ejecución de un proyecto, pero se alcanza un P Valor de 3.30e-187, lo que representa que el modelo puede presentar un efecto significativo en la transformación de la variable objetivo en relación a los coeficientes calculados en las variables predictoras.

Como se observa en la Figura 9, las variables predictoras utilizadas arrojan un P Valor cercano a cero, inferior a 0.05 exceptuando las variables dummy relacionadas a proyectos ejecutados en México y Otras locaciones y la variable relacionada a proyectos de servicios profesionales. Esto implica que los coeficientes de estas variables no son significativos para la predicción de la variable objetivo, esto puede deberse a que la base de datos de la compañía no cuenta aún con información suficiente para incluir este tipo de proyectos en un análisis estadístico.

OLS Regression Results

Dep. Variable:	TOTAL_CONSOLIDADO	R-squared:	0.401
Model:	OLS	Adj. R-squared:	0.397
Method:	Least Squares	F-statistic:	107.3
Date:	Sun, 04 Apr 2021	Prob (F-statistic):	3.30e-187
Time:	21:55:49	Log-Likelihood:	-17226.
No. Observations:	1778	AIC:	3.448e+04
Df Residuals:	1766	BIC:	3.454e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-79.7500	247.533	-0.322	0.747	-565.238	405.738
DIAS_EJEC	1.6734	0.261	6.409	0.000	1.161	2.185
Consultor	491.3789	25.353	19.381	0.000	441.653	541.105
Desarrollador	824.7268	138.595	5.951	0.000	552.899	1096.555
Analista	460.0043	40.602	11.330	0.000	380.371	539.638
LINEA_Fabrica de Pruebas	-1256.6773	336.327	-3.736	0.000	-1916.319	-597.036
LINEA_Fabrica de software	-680.5469	257.974	-2.638	0.008	-1186.514	-174.580
LINEA_Servicios Profesionales	-438.6814	396.506	-1.106	0.269	-1216.353	338.990
Cliente_Chile	1608.5414	628.820	2.558	0.011	375.231	2841.852
Cliente_Mexico	567.6936	656.541	0.865	0.387	-719.985	1855.372
Cliente_Otros	54.1312	1390.043	0.039	0.969	-2672.171	2780.433
Cliente_Panama	962.5826	816.063	1.180	0.238	-637.969	2563.135

Omnibus:	1454.612	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	77603.669
Skew:	3.434	Prob(JB):	0.00
Kurtosis:	34.628	Cond. No.	8.39e+03

Figura 9. Resultados del primer modelo generado con statsmodels.api

Para este modelo se obtiene la siguiente ecuación:

$$Total\ horas\ consolidasdas = -79.75 + 1.67(días) + 491.4(Consultores) + 824.7(Desarrolladores) + 460.0(Analistas) - 1256.7(F.Pruebas) - 680.5(F.Software) + 1608.5(Proyecto\ Chile)$$

Se generan otros cuatro modelos de regresión lineal a partir de las transformaciones realizadas anteriormente. En la Tabla 6 se encuentran definidos los modelos y los resultados correspondientes, los cuales permiten identificar el modelo con mayor aproximación a la realidad de la compañía. El valor R-cuadrado ajustado la capacidad del modelo de explicar la variabilidad observada en la variable objetivo y el P Valor indica si al menos uno de los coeficientes de regresión del modelo es significativo, valores inferiores a 0.05 indican la presencia de variables significativas para el modelo de regresión.

Modelo	Descripción	R-cuadrado ajustado	P Valor
Modelo 1	Modelo OLS: Base de datos original.	0.397	3.30e-187
Modelo 2	Modelo OLS: Base de datos realizando agrupación por cargos operativos y estratégicos.	0.427	6.02e-208
Modelo 3	**Modelo OLS: Base de datos sin realizar agrupación pero con transformación logarítmica.**	**0.459**	**5.48e-229**
Modelo 4	Modelo OLS: Base de datos realizando agrupación por cargos y con transformación logarítmica.	0.441	5.99e-217
Modelo 5	**Modelo RLM: Base de datos sin realizar agrupación pero con transformación logarítmica.**	**NA**	**NA**

Tabla 6. Resultados modelos de regresión lineal

Como se puede evidenciar en los resultados anteriores, el modelo 3 generado a partir de la base de datos con transformación logarítmica (marcado en verde), obtuvo el mayor R-cuadrado ajustado y por lo tanto es el que representa de mejor manera la variabilidad observada en el total de horas necesarias para la ejecución de un proyecto (Ver Figura 10).

Adicionalmente, es importante recalcar que el modelo de regresión lineal robusto (Ver Figura 11) no puede ser comparado con los demás modelos, ya que fue generado a partir de una técnica diferente a los demás modelos. Esta técnica RLM se caracteriza por generar modelos de regresión utilizando bases de datos que contienen alta presencia de valores atípicos, por lo tanto no es adecuado evaluar el ajuste a partir del R-cuadrado ajustado, para este tipo de modelos la mejor opción es más adecuado medir la diferencia entre los resultados predichos y los resultados reales. Para este modelo específico, al realizar la sumatoria de los valores residuales del modelo se obtuvo un total de -102.25 horas de diferencia entre las predicciones y los resultados reales, lo que puede representar un resultado prometedor teniendo en cuenta que actualmente, mediante la estimación de juicio de expertos, se tiene una desviación en la estimación de aproximadamente ± 440 horas, esta última obtenida de la diferencia de la estimación por juicio de expertos y el número de horas reales de una muestra de 174 proyectos.

Resultados Modelo de regresión #3. Modelo OLS - Base de datos con transformación logarítmica

OLS Regression Results

Dep. Variable:	TOTAL_CONSOLIDADO	R-squared:	0.463
Model:	OLS	Adj. R-squared:	0.459
Method:	Least Squares	F-statistic:	138.3
Date:	Sun, 04 Apr 2021	Prob (F-statistic):	5.48e-229
Time:	22:12:36	Log-Likelihood:	-3008.6
No. Observations:	1778	AIC:	6041.
Df Residuals:	1766	BIC:	6107.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.5237	0.215	7.079	0.000	1.102	1.946
DIAS_EJEC	0.6430	0.038	17.012	0.000	0.569	0.717
Consultor	0.6454	0.047	13.592	0.000	0.552	0.739
Desarrollador	0.8633	0.096	9.008	0.000	0.675	1.051
Analista	0.7192	0.055	13.160	0.000	0.612	0.826
LINEA_Fabrica de Pruebas	-0.4821	0.116	-4.159	0.000	-0.709	-0.255
LINEA_Fabrica de software	0.2319	0.089	2.619	0.009	0.058	0.406
LINEA_Servicios Profesionales	0.3865	0.133	2.914	0.004	0.126	0.647
Cliente_Chile	1.4332	0.215	6.661	0.000	1.011	1.855
Cliente_Mexico	0.4075	0.226	1.800	0.072	-0.037	0.852
Cliente_Otros	0.5580	0.501	1.114	0.265	-0.424	1.540
Cliente_Panama	1.1079	0.276	4.011	0.000	0.566	1.650

Omnibus:	65.574	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	69.517
Skew:	-0.465	Prob(JB):	8.03e-16
Kurtosis:	2.731	Cond. No.	92.6

Figura 10. Resultados modelo de regresión #3

Resultados Modelo de regresión #5. Modelo RLM - Base de datos con transformación logarítmica

Robust linear Model Regression Results

Dep. Variable:	TOTAL_CONSOLIDADO	No. Observations:	1778
Model:	RLM	Df Residuals:	1766
Method:	IRLS	Df Model:	11
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Mon, 05 Apr 2021		
Time:	03:09:09		
No. Iterations:	21		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7142	0.223	7.680	0.000	1.277	2.152
DIAS_EJEC	0.6053	0.039	15.401	0.000	0.528	0.682
Consultor	0.6737	0.050	13.590	0.000	0.577	0.771
Desarrollador	0.7885	0.098	8.088	0.000	0.597	0.980
Analista	0.7993	0.057	14.104	0.000	0.688	0.910
LINEA_Fabrica de Pruebas	-0.6528	0.119	-5.497	0.000	-0.886	-0.420
LINEA_Fabrica de software	0.1887	0.091	2.082	0.037	0.011	0.366
LINEA_Servicios Profesionales	0.4023	0.138	2.913	0.004	0.132	0.673
Cliente_Chile	1.4098	0.224	6.282	0.000	0.970	1.850
Cliente_Mexico	0.1978	0.234	0.844	0.399	-0.262	0.657
Cliente_Otros	0.6670	0.563	1.186	0.236	-0.435	1.770
Cliente_Panama	0.8504	0.286	2.974	0.003	0.290	1.411

Figura 11. Resultados modelo de regresión #5

Validación del modelo de regresión

Se realiza el correspondiente análisis de supuestos de normalidad, homocedasticidad e independencia a los residuales obtenidos del modelo #3 generado a partir de la transformación logarítmica de la base de datos, adicionalmente se realizan pruebas estadísticas para sustentar de manera más clara estos resultados.

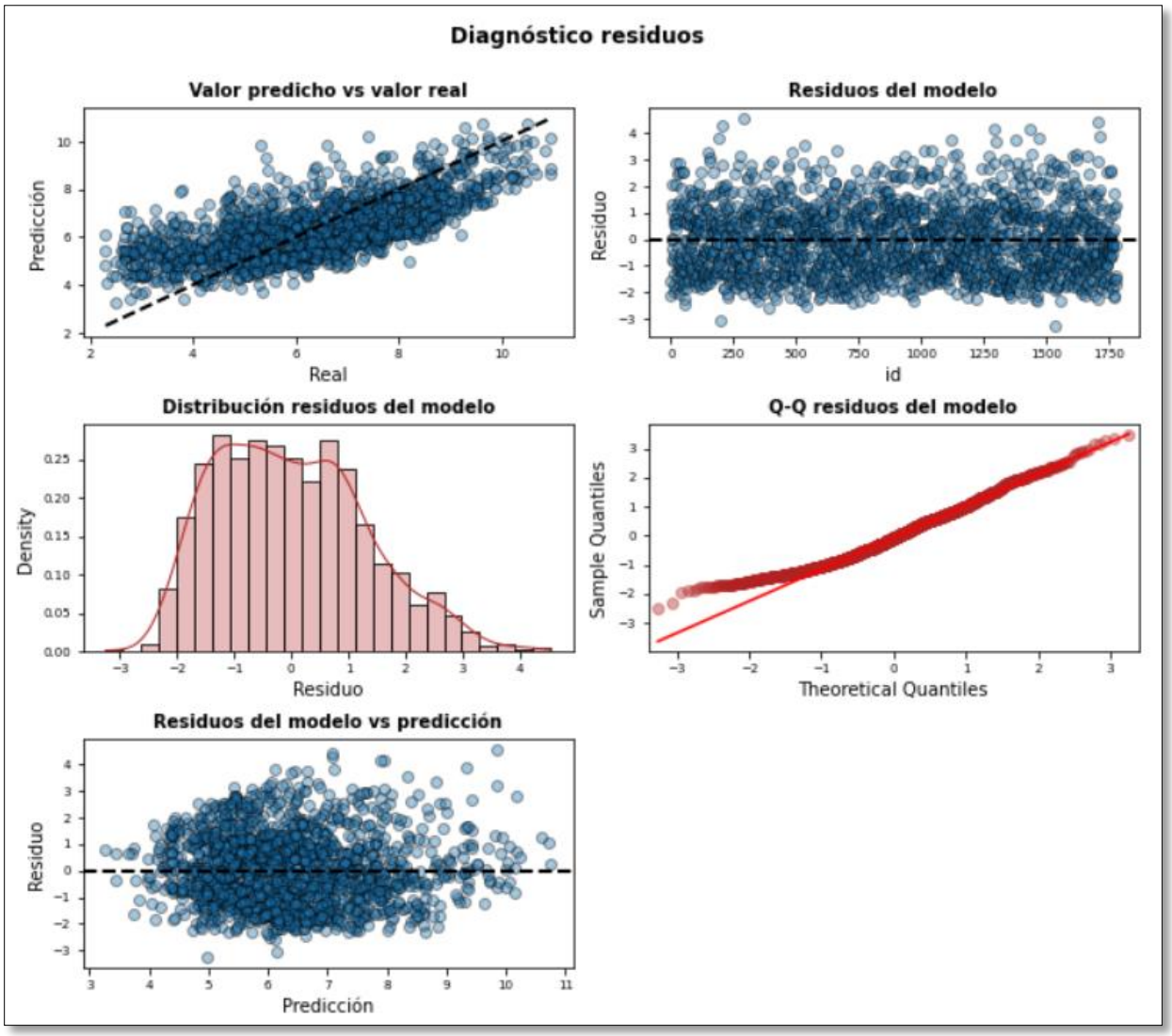

Figura 12. Diagnóstico de los residuos del modelo de regresión lineal #3

El ejecutar la prueba Shapiro-Wilk a los residuos del modelo de regresión, se obtiene un P Valor de $3.61e^{-16}$, lo que concluye dichos residuos no se distribuyen de forma normal. Al realizar inspección visual al comportamiento de los residuos del modelo, es posible afirmar que los residuos presentan un sesgo hacia la izquierda, como se puede observar en el histograma de frecuencia y en la gráfica Q-Q plot de la Figura 12, lo que implicaría la imposibilidad de obtener una distribución normal. Al incumplir con el supuesto de normalidad, es importante realizar un análisis más profundo a los datos, para tratar de identificar la presencia de más muestras atípicas, o identificar transformaciones de variables más adecuadas, o considerar la implementación de modelos de regresión que contemplen distribuciones no normales.

Para validar la independencia de los residuos del modelo, se utiliza la prueba de Durbin-Watson, la cual arroja un valor de 1.97, ya que este valor se encuentra dentro del rango de 1.5 y 2.5, se puede afirmar la existencia de independencia entre los residuos.

Por último se realiza el análisis de igualdad de varianza, relacionando los residuos del modelo frente a los valores ajustados de predicción, a nivel gráfico no se encuentra evidencia que se presente una tendencia en particular, puesto que los residuos parecen estar distribuidos de forma aleatoria en torno al valor cero.

CONCLUSIONES

La implementación de técnicas de análisis estadístico para analizar el desempeño de los proyectos ejecutados por la compañía de ingeniería de software es de vital importancia para fortalecerse en el mercado globalizado, y a su vez representa un gran reto en cuanto al fortalecimiento para la definición, recolección y análisis de métricas de calidad, de proceso y de desempeño de los proyectos. El análisis realizado para determinar un modelo de regresión lineal múltiple evidenció que es posible explicar aproximadamente el 46% de la variabilidad observada en el total de horas necesarias para la ejecución

de un proyecto, a través del uso de variables predictoras como el número de días estimados para el proyecto, el número de colaboradores necesarios para conformar el equipo de trabajo, la línea de negocio y la ciudad/país de ejecución del proyecto, toda estas, exceptuando las últimas dos variables, con transformación logarítmica.

De igual manera, es posible concluir que al definir nuevas variables más enfocadas a factores específicos de los proyectos, podría mejorarse significativamente la bondad de ajuste del modelo de regresión. A manera de ejemplo, podrían implementarse las variables relacionadas a la tecnología usada en el proyecto, fortalecer la estimación a través de plantillas fraccionadas por etapas del ciclo de vida de los proyectos, definir criterios para la caracterización de los clientes, incluir métricas relacionadas con el análisis de riesgos potenciales del proyecto, y de esta manera fortalecer progresivamente el desempeño de la compañía al igual que la precisión en futuros modelos estadísticos de estimación.

El modelo de regresión lineal múltiple logrado en este caso de estudio mediante el método OLS, funciona como punto de partida inicial para determinar el plan de mejora de la compañía en relación con obtener datos de mayor precisión y especificación. Adicionalmente, es posible implementar procesos de análisis por componentes principales en las variables de colaboradores por cargo, ya que se logró identificar que los cargos técnicamente especializados como Consultor, Desarrollador y Analista fueron los que finalmente se incorporaron como variables predictoras, mientras que los cargos estratégicos como Gerente, Arquitecto y Otros cargos administrativos no resultaron significativos para predecir el comportamiento de la variable respuesta.

Ecuación del modelo de regresión lineal #3 usando el método OLS:

$$Total\ horas\ consolidasdas = 1.52 + 0.64(días) + 0.65(Consultores) + 0.86(Desarrolladores) + 0.72(Analistas) - 0.48(F.Pruebas) + 0.23(F.Software) + 0.39(Servicios\ Profesionales) + 1.43(Proyecto\ Chile) + 1.11(Proyecto\ Panamá)$$

El modelo de regresión robusto puede resultar bastante prometedor para la compañía, puesto que en la prueba realizada se logró evidenciar que reduce la deviación en la estimación de 440 horas (valor actual mediante la estimación por juicio de expertos) a 102.25 horas aproximadamente.

Por último, también se logró identificar que los proyectos ejecutados en países como México, Panamá y demás nuevos mercados extranjeros no cuentan por ahora con suficientes muestras para generar modelos de predicción estables, por el contrario, si se cuenta con una gran variedad de proyectos ejecutados en casa matriz de Bogotá/Medellín.

RECOMENDACIONES

1. Implementar Análisis de Componentes Principales (PCA) a las variables relacionadas con el numero de colaboradores por cargo, es posible que se puedan identificar nuevas variables que agrupen los cargos por tipo de responsabilidades, por ejemplo separando los cargos estratégicos de los cargos operativos.
2. Enfocar el análisis estadístico a proyectos desarrollados en Colombia, ya que representan el 93.5% de os proyectos capturados en la muestra de 2012 a 2020. Posteriormente incluir los demás países cuando se tenga una mayor muestra.
3. Definir métricas que aporten más información acerca de los proyectos, relacionadas con aspectos técnicos, riesgos del proyecto, riesgos del cliente, entre otras. De igual manera, estudiar la posibilidad de implementar métodos de estimación fraccionados por etapas del ciclo

de vida de los proyectos, por ejemplo, estimación para las fases de diseño, desarrollo y pruebas, esto con el fin de tener la posibilidad de realizar un análisis estadístico más robusto y tal vez más preciso que el análisis desarrollado en este caso de estudio.

4. Definir métricas de desempeño de los proyectos, tales como calidad de los entregables, capacidad de identificación y remoción de defectos, con el objetivo de dar mayor variedad en las estimaciones de estos, no solamente abarcando el enfoque de estimación de horas de ejecución.

REFERENCIAS BIBLIOGRÁFICAS

CMMI-Institute. (2018). *CMMI Institute - Modelo CMMI versión 2.0.* Illinois: ISACA.

Domé, T. (2018). Metodologías ágiles para emprendedores. *Latinpreneur*.

Forbes-Colombia. (7 de Julio de 2020). *forbes.co*. Obtenido de La globalización de las oportunidades en tecnología: https://forbes.co/2020/07/02/red-forbes/la-globalizacion-de-las-oportunidades-en-tecnologia/

Forbes-Colombia. (5 de Marzo de 2021). *forbes.co*. Obtenido de Empresa de software llega a Colombia bajo modelo de relocalización: https://forbes.co/2021/03/05/negocios/empresa-de-software-llega-a-colombia-bajo-modelo-de-nearshoring/

Kan, S. H. (2002). Metrics and models in software quality engineering. Addison-Wesley.

Montgomery, D. (2002). *Applied Statistics and Probability for Engineers 3rd ed.* Nueva York: John Wiley & Sons, Inc.

Pressman, R. S. (2005). Ingeniería de software: un enfoque práctico. McGraw–Hill.

Procolombia. (Diciembre de 2020). *investincolombia.com.co*. Obtenido de Tecnología de la información e industrias creativas: https://investincolombia.com.co/es/sectores/tecnologia-de-la-informacion-e-industrias-creativas/software-y-servicios-de-ti

Tusell, F. (2011). *Análisis de regresión. Introducción teórica y práctica basada en R.* Bilbao: Universidad del País Vasco.

Wichern, J. E. (2006). *Pronósticos en los negocios.* México D.F.: Pearson Educación.