Título: Modelos predictivos para la estimación de tiempos de duración de los proyectos de automatización de la empresa Stanley Black & Decker

Autor
Luis Felipe Henao López

Para optar al título de Ingeniero Industrial título otorgado por UdeA

Asesoras

Olga Cecilia Usuga Manco
Ph.D en Estadística

Catalina Piedrahita Jiménez
Líder de equipo Análisis de Automatización

Universidad de Antioquia
Facultad de Ingeniería
Departamento de Ingeniería Industrial
Medellín
2022

Cita	(Henao, 2022)
Referencia	Henao López L.F (2022). *Modelos predictivos para la estimación de tiempos de duración de los proyectos de automatización de la empresa Stanley Black & Decker*, 2015 – 2022. Ingeniería Industrial. Universidad de Antioquia, Medellín.
Estilo APA 7 (2020)	

Agradecimientos

Agradecimientos a mi asesora interna Olga Úsuga por brindarme acompañamiento constante en la realización de este trabajo y a mi asesora externa Catalina Piedrahita, quien me brindo todo su apoyo al interior de la empresa y me dio la posibilidad de adquirir responsabilidades en mi cargo como practicante para encontrar el rumbo a mi carrera profesional. La Universidad de Antioquia me permitió llegar a un lugar que me enseñó mucho y Stanley Black & Decker con todo su talento humano me recibieron con los brazos abiertos y estuvieron dispuestos a enseñarme.

Tabla de contenido

Lista de tabla

Lista de figuras

Siglas, acrónimos y abreviaturas

UdeA	Universidad de Antioquia
PhD	Philosophiae Doctor
IAS	Intelligent Automation Solutions (Área de automatización de La Empresa)
SVM	Support Vector Machines (Máquinas de Soporte Vectorial)
SVR	Support Vector Regressor(Regresor de Soporte Vectorial)
DTR	Decision Tree Regressor (Regresor de Árboles de Decision)
GLM	General Lineal Models (Modelo Lineal Generalizado)
GLM-G	General Lineal Model-Gamma (GLM con distribucion Gamma)
GLM-IG	General Lineal Model-Inverse Gaussian (GLM-Gaussiana Inversa)
MSE	Mean Squared Error (Error Cuadrado Medio)
RMSE	Root Mean Squared Error (Raiz cuadrada del Error Cuadrado Medio)

Resumen

En el presente trabajo se analizan cuáles son las variables que intervienen en los tiempos de duración de los proyectos de automatizacion en la empresa Stanley Black & Decker para posteriormente probar 3 modelos predictivos. Para poder realizar el análisis se identificaron las etapas críticas del proceso en las que se dan mayores demoras, así como las causas potenciales para que esto ocurra. Identificado el problema de inexactitud en las estimaciones, se realizó la recolección de datos para las variables beneficio económico, número de sistemas usados, complejidad, Score o prioridad para ejecutar el proyecto y la variable tiempo de duración en semanas como variable explicativa. Recolectados los datos de las variables se construyeron los modelos computacionales Regresor de Sporte Vectorial y Regresor de Árbol de Decision, ambos modelos implementados en Python y un Modelo Lineal Generalizado con distribuciones Gamma y Gaussiana Inversa. Los modelos se compararon de acuerdo a los resultados de las medidas de desempeño Error Cuadrado Medio y Raíz Cuadrada del Error Cuadrado Medio, dando como resultados que el mejor modelo para predecir los tiempos de duración es el modelo computacional Regresores de Árboles de Decisión y en segundo lugar el Regresor de Maquina de Soporte, permitiendo llegar a la conclusión que los modelos predictivos computacionales usados en este trabajo son apropiados para la determinación de tiempos de duración de proyectos con varias actividades en el proceso y potencialmente aplicables a otros tipos de proyectos no necesariamente limitados a automatizaciones.

Palabras clave: Modelos predictivos, automatización de procesos, modelo lineal generalizado, modelo computacional.

Abstract

In this research the variables that have a direct impact in projects duration are analyzed within the automation projects in Stanley Black & Decker Inc, subsequently 3 predictive models are used. With the aim of performing the analysis, the critical stages of the projects are identified, which means when delays occur, as well as root causes of the problem. After the inaccurate time estimations are identified, the data extraction was carried for the variables monetary benefit, number of systems, complexity, score or execution priority, and duration in weeks as explanatory variable. After data gathering, the computational models were built: Support Vector Regressor and Decision Tree Regressor, both implemented in Python. The used statistical model was the Generalized Linear Model with the distributions Gamma and Inverse Gaussian, for this case implemented in R. The computational and statistical models were compared according to the performance measures Mean Squared Error and Root Mean Squared Error, as result it was obtained as the best predictive model for time duration the Decision Tree Regressor, in second place the Support Vector Regressor, reaching to the conclusion that the computational predictive models used in this research are appropriate for determining the project duration and they are potentially applicable to other kinds of projects, not necessarily limited to automations.

*Keywords***:** Predictive model, process automation, generalized linear model, computational model.

Introducción

Esta investigación se realizó en el área de Intelligent Automation Solutions (IAS), la cual consiste en que al automatizar un proceso se deben seguir una serie de etapas que van desde la recepción de la solicitud de automatización hasta su entrega. Estos pasos se deben seguir en su totalidad debido a que son la base para que un proyecto de automatización cumpla con todos los requerimientos necesarios que aseguren que se implemente la automatización exitosamente.

Se tiene como antecedente un proyecto que se inició desde Governance & Controls, que buscaba estimar mejor la exactitud de los tiempos de los proyectos, identificando las fases en las que se presentan los retrasos y las razones por las que se da, y que arrojó inicialmente que una de las principales razones de atraso es la falta de comunicación asertiva con la unidad de negocio. Lo anterior se da porque es muy importante que el área revise de forma oportuna la documentación de los procesos para dar retroalimentación sobre los resultados que arroja el Bot para indicar si están buenos o no. También está entre las motivaciones que llevaron a plantear esta propuesta, el hecho de dar el soporte metodológico sustentado en los datos recopilados de las automatizaciones para estimar con mayor exactitud las fechas de las etapas de los proyectos de automatización.

Inicialmente se cuenta con el registro de todas las automatizaciones realizadas, con la fecha de inicio y fin de los proyectos, además en la mayoría de las solicitudes se cuenta con el nivel de complejidad que tiene la automatización de acuerdo con la clasificación cualitativa que le dan los analistas y el beneficio monetario estimado que representa para la unidad de negocio como resultado de automatizar el proceso.

Planteamiento del problema

La gran dificultad que se presenta en el área IAS de la empresa Stanley Black & Decker es que desde el inicio de los proyectos se deben tener unas fechas estimadas para ejecutar cada una de las etapas, pero se da la limitante que a lo largo del proceso se deben correr las fechas debido a que ocurren retrasos con respecto a las fechas iniciales causados por diversas limitaciones, lo que causa que se forme una cadena de proyectos que no van a cumplir con las fechas iniciales estimadas, ya que el orden está definido de acuerdo a la prioridad asignada en la pipeline.

El hecho de tener que correr las fechas de las etapas del proyecto es algo común en el área, sumado a que existen solicitudes que no dependen netamente del equipo de IAS, sino también del equipo de Global Information Technology, quienes procesan solicitudes de características como accesos a los software a los cuales debe entrar el usuario del Bot para poder llevar a cabo las automatizaciones, es decir que si ocurren inconvenientes que no se identifican en la parte inicial de los proyectos y se evidencian justo antes de empezar la etapa de producción implica que se corran las fechas estimadas desde un principio debido a que el hecho de realizar una modificación llega a depender de la prontitud con la que el equipo de Global Information Technology realice las modificaciones necesarias. Es por ello por lo que la aparición de inconvenientes en cualquier etapa de la automatización, y más que todo justo antes de que el Bot pase a producción, deben ser considerados al momento de estimar las fechas. En la Figura 1 se presenta el gráfico de las etapas.

Figura 1 Etapas del proceso de automatización. Fuente. Intelligent Automation Solutions, Stanley Black & Decker Inc

Los analistas de automatización son los encargados de tener la comunicación entre el área del negocio que quiere automatizar un proceso y los desarrolladores que programan a los Bots que reemplazan la tarea repetitiva que realizan en la unidad del negocio. Esto implica que los analistas toman el rol de gestores de proyectos, por lo tanto, son los encargados de definir las fechas estimadas para cada etapa y de igual manera velar porque estas se cumplan de acuerdo con el cronograma estipulado. Basándose en un primer análisis exploratorio realizado, es poco común que las fechas planteadas al inicio de cada proyecto se cumplan con exactitud, es por ello que el hecho de poder definir las fechas de cada proyecto con mayor exactitud desde el inicio permitirá que dentro de IAS tengan una visión más realista sobre cuál será el tiempo de duración de un proyecto y de esta manera alinear los recursos con los que cuentan con el fin de mejorar el desempeño del área, esto es posible gracias a la identificación de las variables que impactan en mayor medida el tiempo de duración de los proyectos para llegar a estimaciones más realistas.

1 Objetivos

1.1 Objetivo general

Predecir los tiempos de duración de los proyectos de automatización.

1.2 Objetivos específicos

- Realizar un diagnóstico de los tiempos de duración de los proyectos de automatización
- Identificar las variables que pueden afectar los tiempos de duración de los proyectos
- Implementar los modelos computacionales y estadísticos apropiados para realizar el análisis de las variables.
- Hacer recomendaciones sobre la recolección de información, las variables que se deben tener en cuenta y la futura herramienta de estimación de tiempos de duración de proyectos.

2 Marco teórico

Para estimar la duración de las actividades de un proyecto se deben tener en cuenta los recursos que intervienen para que se lleve a cabo. Todo parte de identificar en el cronograma de las actividades de qué manera y en qué momento se van a utilizar cada uno de los recursos. A medida que se ejecuta el proyecto, se pueden presentar retrasos que no deben ser relevantes a la hora de cumplir con los tiempos de entrega si realmente se hizo una estimación inicial correcta que haya tenido en cuenta todas las implicaciones de poder desarrollar el proyecto exitosamente en los rangos estimados de tiempo.

En la Figura 2 se puede apreciar el Diagrama de Flujo de Datos del Proceso de estimar la duración de las actividades de un proyecto. En el proceso se dan como entradas la lista de actividades que se deben realizar para hacer satisfactoriamente el proyecto, los atributos de las actividades, los requerimientos de los recursos que vienen dados por la manera como se deben usar en cada una de las actividades, el calendario de los recursos que estimula el momento en que se usan cada uno de ellos, el enunciado del alcance del proyecto para tener clara cuál es la delimitación que tiene y no perder el enfoque hasta dónde va el proyecto, los factores del entorno empresarial y los activos del proceso organizacional que serán parte del proceso que junto a las herramientas y técnicas que se definen de acuerdo a las características del proyecto, a los datos que se tienen, al juicio de los expertos que están involucrados en cada una de las actividades y que luego de un análisis completo se tienen como salidas la estimación de la duración de las actividades que se deben realizar para obtener los resultados esperados del proyecto (PMI Project Management Institute, 2013).

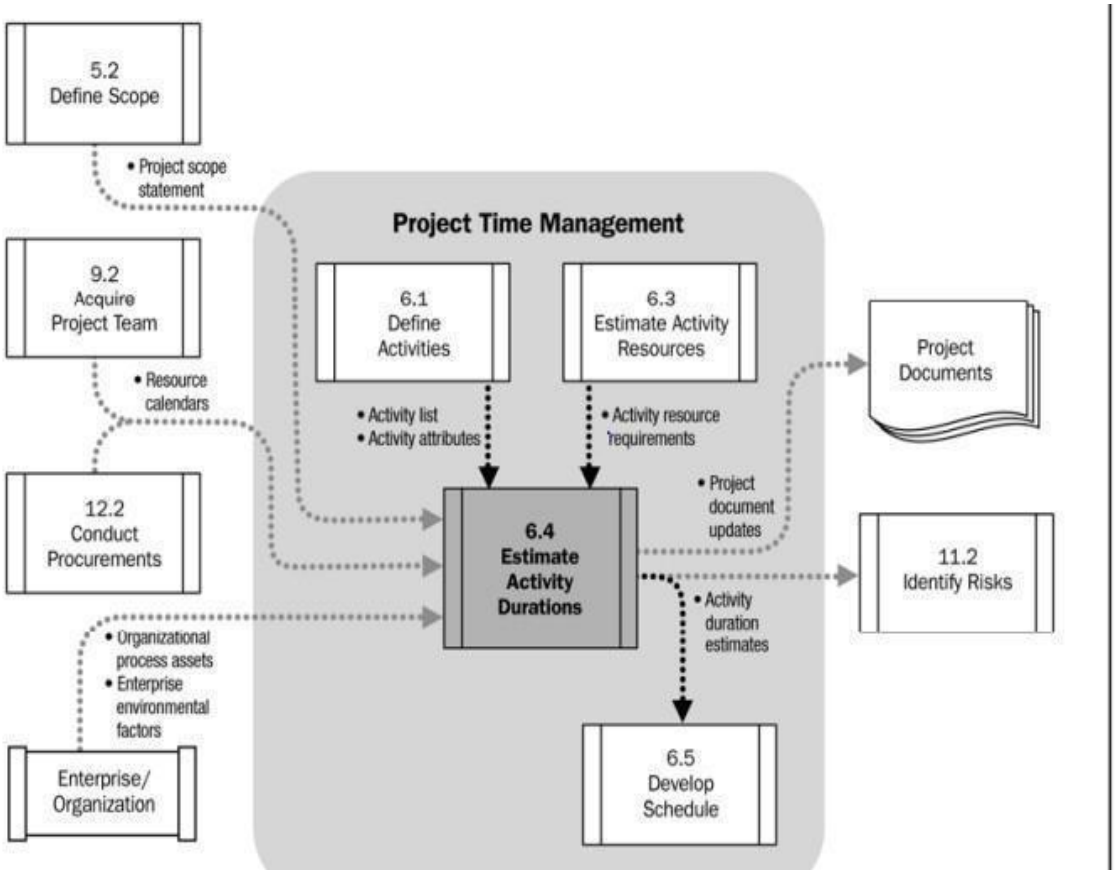

Figura 2 Diagrama de Flujo de Datos del Proceso de estimar la duración de las actividades de un proyecto. Fuente: PMI Project Management Institute

En la estimación de la duración de un proyecto se pueden dar calendarios inexactos y sobre optimistas como resultado del efecto de anclaje, que consiste en un sesgo cognitivo en el comportamiento humano causado por confiar demasiado en la información inicial a la que se tiene acceso y tomar decisiones que no se basan en su totalidad en información de fuentes confiables (Limón et al., 2014). Lo anterior se presenta comúnmente en la planeación de proyectos, y no desaparece incluso cuando hay antecedentes de estimaciones de tiempos de actividades que no se realizaron en los tiempos definidos y que tienen características similares a las que se deben considerar, y que se da en parte por parcialidades que se presentan por las estimaciones iniciales del encargado del proyecto (Lorko et al., 2019). Por lo anterior se identifica la necesidad de proveer información que sea útil para definir cuáles son las variables que generan un impacto mayor en los tiempos de duración de los proyectos de automatización con el fin de seguir basándose en

resultados verídicos que permitan llegar a estimaciones que se aproximen a la realidad de los proyectos.

Para identificar las variables más influyentes, se utilizarán los modelos predictivos que brindan los algoritmos de Machine Learning, los cuales mejoran el proceso de estimación a través de las distintas metodologías y técnicas de selección de variables para estimar modelos mejorando la eficiencia de los procesos (Management Solutions, 2018). Los algoritmos computacionales elegidos para aplicar en la metodología del presente proyecto son las Máquinas de Soporte Vectorial (SVM) y los Regresores de Arboles de Decisión (DTR). Las SVM "se caracteriza por el uso de núcleos, solución dispersa y control del número de vectores de soporte. Aunque menos popular que SVM, se ha demostrado que SVR es una herramienta eficaz en la estimación de funciones de valor real" (Awad & Khanna, 2015).

Otro de los algoritmos apropiados para aplicar en este análisis de los modelos computacionales de regresión son los DTR, "que son usados para variables dependientes que toman valores discretos ordenados o continuos, con el error de predicción típicamente medido por la diferencia al cuadrado entre los valores observados y predichos" (Loh, 2011). La forma como funciona este algoritmo es que, en la raíz, o sea la parte inicial del árbol, se toman los datos de las variables dependientes, y se van dando hacia la parte inferior las divisiones de las hojas de acuerdo con la condición de que se define desde el nodo raíz del árbol. En la Figura 3 se observa cómo se da la división binaria en cada nodo del árbol, de acuerdo con la condición prestablecida, y la representación de los valores bajos (en la escala de amarillos) y altos (en la escala de rojos), que toma la variable respuesta en los nodos que resultan luego de la división basada en la condición dada (Nerd, 2021).

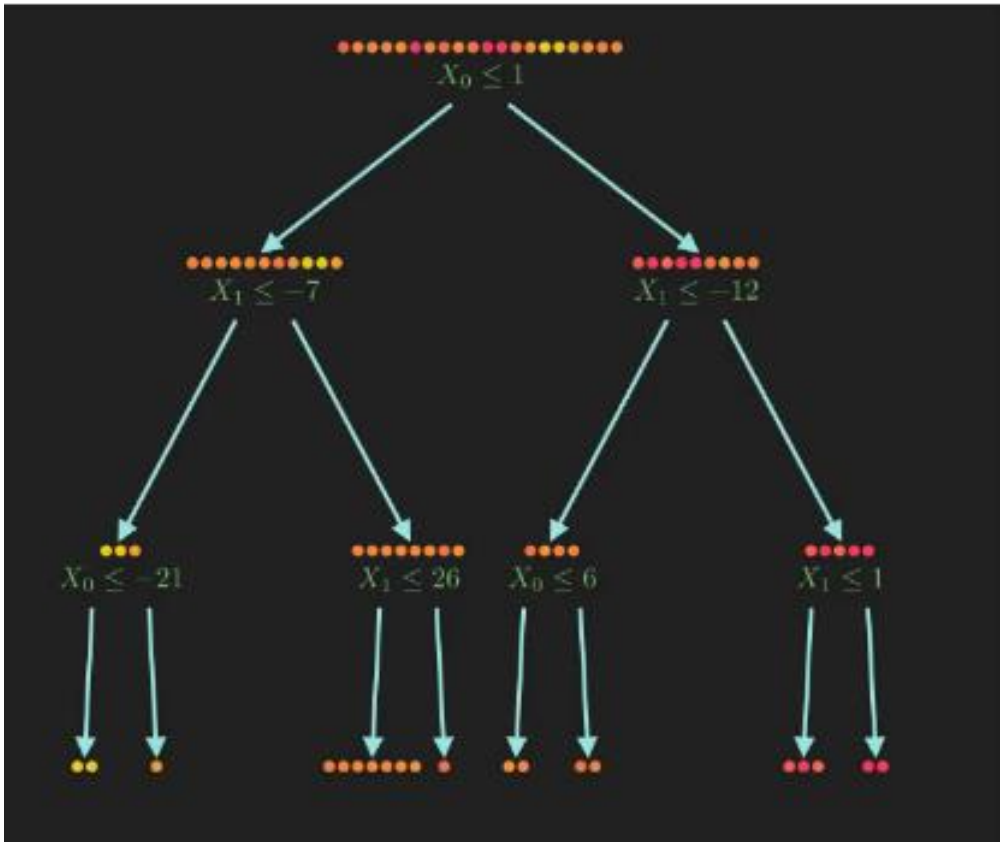

Figura 3 Representación de árbol de regresión Fuente: Canal de Youtube Normalizad Nerd

Con el fin de elegir la mejor división para realizar la predicción el algoritmo, se basa en la reducción de la varianza, en cada nodo se calcula la sumatoria de la varianza de todos los nodos de la ramificación elegida y el resultado se toma para restarlo a la varianza del nodo inicial , el menor resultado de la resta en cada una de las ramificaciones es el que indica cual es la ruta resultante que sirve como referencia para realizar la predicción (Nerd, 2021).

El modelo estadístico que se utiliza en el presente trabajo es el Modelo Lineal Generalizado (GLM), que consiste en una variable respuesta Y con observaciones independientes X1, X2,…, Xn. Este modelo se caracteriza porque los errores no deben ser obligatoriamente normales, o la varianza tampoco constante, permite realizar transformación a los errores y datos de la variable respuesta que no cumplen una distribución normal, además la variable respuesta tiene relación lineal con las otras variables explicativas (Cayuela, 2009). Entre las distribuciones que se encuentran para aplicar en el GLM hay dos que se pueden ajustar al problema que son la

distribución Gamma (GLM-G) debido a la variable aleatoria que representa el tiempo de duración de proyectos que se produce un determinado número de veces hasta que ocurre un suceso (Arroyo et al., 2014) y la distribución Gaussiana Inversa (GLM-IG) que se puede usar luego de la transformación de datos que no siguen una distribución normal. Si bien la distribución t-student brinda una forma similar a la de la Gaussiana Inversa, la t-student no es estable ante la operación de dos funciones que producen otra, lo cual computacionalmente no es apropiado debido a que el número de variables incrementa (Kalemanova et al., 2007), distinto a la Gaussiana Inversa que posee características más apropiadas de acuerdo con el tipo de datos que se tienen para la investigación.

Para medir el desempeño de los modelos estadísticos y computacionales planteados para realizar una posterior comparación de resultados, es apropiado usar la medida de desempeño del Error Cuadrático Medio (MSE) que mide la diferencia media entre los valores dados por y_i y sus respectivas estimaciones $\hat{y}_i$ como resultado del modelo usado, que en resumen sirve para medir la variabilidad de un estimador, o sea la precision, y el sesgo o exactitud de la estimacion (Zheng, 2019). La siguiente ecuación representa la medida de desempeño.

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

También se toma como medida de desempeño la Raíz Cuadrada del Error Cuadrático Medio (RMSE) para medir la precisión de la variable respuesta y así poder identificar si la finalidad del modelo es la predicción (Gonzalez, 2018). La siguiente ecuación representa la medida de desempeño de RMSE.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Para ambas medidas de desempeño se considera que el mejor modelo utilizado en la predicción es el que tenga los resultados más bajos de MSE y RMSE.

3 Metodología

El enfoque de la metodología llevada a cabo es cuantitativo, la problemática planteada al interior del área de la empresa se analizó minuciosamente con base a los datos recolectados sobre los proyectos de automatización llevados a cabo desde Enero del 2019 hasta Agosto de 2021. La construcción de la base de datos con las variables que intervienen en el proceso es parte fundamental de la realización de este proyecto. Como parte de la metodología se siguieron las siguientes Fases:

Fase 1:

Se hace un mapeo del proceso de automatización con los tiempos estimados de cada etapa de acuerdo con los proyectos llevados a cabo en el rango de tiempo establecido, el mapeo inicial permite caracterizar los proyectos de automatización y empezar a tener elementos que permitan identificar cuáles son las variables principales para tener en cuenta en la estimación de la duración de los proyectos. El mapeo se hace en el software Visio, en el que se hace un diagrama de flujo que contiene el proceso completo desde que se hace la solicitud de una automatización por las unidades de negocio hasta que se entrega dicha automatización.

Fase 2:

Se tiene de entrada una gran cantidad de información valiosa que es resultado de la documentación minuciosa de los procesos y la recopilación de los datos de cada una de las automatizaciones que se han realizado. Se lleva a cabo un análisis estadístico exploratorio que permita identificar las relaciones entre las variables cuantitativas iniciales que intervienen en la estimación de la duración de los proyectos. Para dicho análisis se sacan los tiempos promedio de duración de cada una de las 5 etapas del proceso de automatización, y se calcula el Pareto de donde se encuentra el 80% de los tiempos de duración de los proyectos y a partir de los resultados se realiza el filtro de las variables principales que intervienen en la estimación de los tiempos de duración. Posteriormente se eligen las variables cuantitativas con datos disponibles de los proyectos y se realiza la limpieza de la base de datos para poder ser leída en software. En este caso se utilizan los lenguajes de programación

Python y R, por ser apropiados para el análisis de datos gracias a las librerías que contienen como scikit-learn y gamlss respectivamente.

Fase 3:

Se implementan los modelos computacionales SVR y DTR en Python, las funciones utilizadas de la liberaría scikit-learn son SVR como alternativa de regresión para las SVM, una vez se plantea el modelo, se pasa a optimizar los hiperparametros debido a las múltiples posibilidades que hay para cada distribución, para ello se utiliza la función GridSearchCV que sirve para realizar la optimización automaticamente realizando todas las combinaciones posibles de los hiperparametros del modelo, en el caso de este algoritmo se usa un gamma fijo de 0.3, kernel rbf y poly, C de 20, 25 y 30, epsilon 0.1,0.01,0.001 y 0.0001. La función evalúa todas las posibles combinaciones comparadas en función de la medida de desempeño MSE para así definir el modelo que se ajuste a los datos de referencia y posteriormente realizar la predicción. El otro algoritmo computacional que se utiliza es el DTR, para este se utiliza la función DecisionTreeRegressor del mismo paquete scikit-learn, para este no se realiza optimización de hiperparametros, simplemente se toma la configuración que por defecto trae el paquete y similar al modelo SVR se realiza la predicción. En todos los casos se corrieron los modelos con todas las variables explicativas y la variable respuesta duración en semanas, luego de manera separada la variable explicativa Complejidad y por último la variable Score. Para los modelos de Python ambos inician de la misma forma para cargar los datos con la ayuda de la librería Pandas, de esa forma se crea un arreglo con la variable respuesta y otro con las variables explicativas, sean todas las variables iniciales o en los casos que se corre con una única variable. Los datos se dividen de manera aleatoria en un 80% datos de entrenamiento y 20% datos de prueba con la ayuda de la función train_test_split de la libreria sickit-learn, a partir de este paso es que cambia el algoritmo de SVR y DTR. Por último, se calcula el MSE y RMSE entre los datos de entrenamiento y de validación.

El modelo estadístico implementado en R es el GLM. Las distribuciones que se ajusten a los datos de tiempo de duración son usadas para correr el modelo. Posteriormente se comparan los resultados de las predicciones por medio de las medidas de desempeño MSE y RMSE.

Con base a los resultados obtenidos, se plantean pasos a seguir de acuerdo con variables que se deben tener en cuenta para recolectar información que puede ser importante para construir una herramienta de estimación de tiempos de duración de los proyectos, de esta manera será posible realizar estimaciones más exactas que cumplan con el nivel de servicio definido por los grupos de interés en la empresa.

4 Resultados

Resultados Fase 1:

Para la caracterización de los proyectos de automatización se realizó un diagrama de flujo en el software Visio (Anexo) el cual tiene cada una de las etapas en las cuales se pueden presentar más retrasos. Estas etapas fueron validadas con las personas del área encargadas de realizar el seguimiento de los proyectos. En mayor medida se dan los retrasos cuando hay aprobaciones que no dependen del personal del área IAS, al momento de presentar los resultados de cada etapa del proceso de automatización se debe recibir la aprobación de todas las partes involucradas, si una sola parte no está comprometida con hacer seguimiento al proyecto, los tiempos de entrega no van a ser cumplidos, también si el área de la empresa que solicitó la automatización no está al tanto de las necesidades del proyecto en cuanto a documentación, detalles que se deban tener en cuenta para la correcta ejecución y de las tareas de los analistas, se presentarán retrasos que dificultan el trabajo de los desarrolladores en las etapas de Build y UAT. La Figura 4 representa las etapas críticas en el proceso de automatización, es decir en las que es mas propenso experimentar retrasos con respecto a las fechas definidas inicialmente.

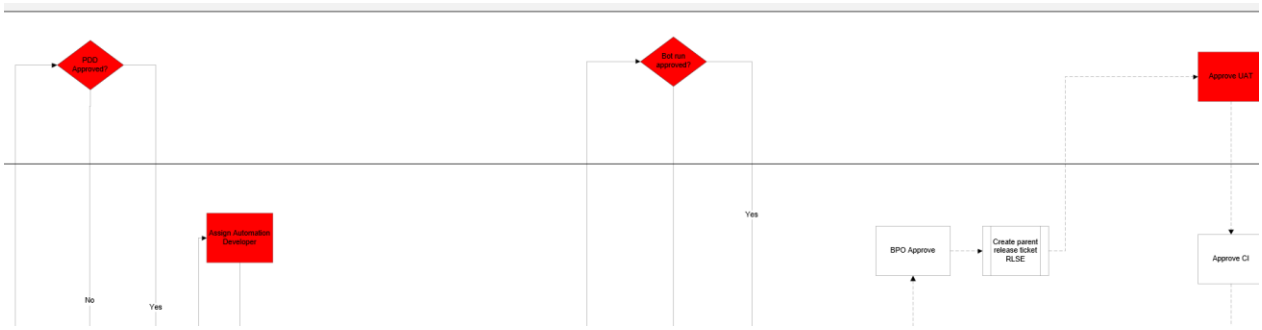

Figura 4 Etapas críticas para retrasos en Diagrama de Flujo de Automatizacion . Fuente: Elaboración propia

Desde los pasos que están a cargo del equipo IAS, en el que se generan retrasos es el de asignar desarrollador de automatización (*Assign Automation Developer*), paso que en algunas ocasiones puede tomar más tiempo del esperado debido a que ocurren situaciones inesperadas en otros

proyectos que van en etapas más avanzadas y que tienen prioridad para solucionar cualquier problema que pueda aparecer, causando así que se corran las fechas de entrega estimadas inicialmente, dando lugar a que en posteriores investigaciones la medición de los tiempos de espera en cola sea tenido en cuenta para la estimación de los tiempos totales de los proyectos.

Resultados Fase 2:

Para la identificación de variables que afectan la duración de los proyectos se llevó a cabo la recolección de datos de los proyectos de automatización que se encuentran en la etapa Live en el Bot Tracker de IAS, que es una hoja de cálculo en la que se lleva el registro de la información de todos los proyectos, desde enero de 2019 hasta agosto de 2021, con el filtro de los proyectos que ya fueron entregados satisfactoriamente a las unidades de negocio que los solicitaron en la empresa Stanley Black & Decker, es decir los que aparecen en etapa Live en el Bot Tracker. Se contó inicialmente con datos de 86 proyectos, no todos tenían datos de la fecha de inicio y fin del proyecto de automatización, pero finalmente fueron 75 los que tenían información completa de los tiempos de duración en semanas. Las variables que se consideraron para explicar el tiempo de duración son las variables cuantitativas: Número de sistemas utilizados para la automatización (sean softwares de organización de tareas, recepción y envió de información como correo electrónico o ERPs para la gestión de recursos del área) y beneficio en dólares para la empresa como estimación por llevar a cabo la automatización. También se identificaron las variables cualitativas ordinales de nivel de Complejidad siendo 1 el más bajo y 3 el más alto, asignado por el analista de manera subjetiva teniendo en cuenta el número de pasos que se deben llevar a cabo para realizar el proceso o algún otro factor que considere para definirlo de esa manera y por último la variable prioridad de entrega del proyecto o Score, que es calculada al interior del área teniendo en cuenta diversos factores y dependiendo de su valor siendo 1 la prioridad más alta y 5 la más baja, para ambos casos en las variables cualitativas cuando se dan valores de 0 significa que no se tienen datos y representan un porcentaje mínimo de los proyectos en la base de datos. En la Tabla 1 se muestran las primeras filas de la base de datos:

Tabla 1 Base de datos recolectada de los proyectos. Fuente: Elaboración propia

1	Bot ID	Complexity Level	Annual Benefit (K)	Final Score	Weeks	Number of Systems
25	2020.93	2	568	2	16	1
26	2020.94	2	43	3	38	3
27	2020.95	2	69	3	8	4
28	2020.96	1	5	3	16	3

Como se muestra en la Figura 5, los tiempos de duración de los proyectos se encuentran en su mayoría entre las semanas 1 y 30, los proyectos que estén por encima de este tiempo son casos a revisar minuciosamente debido a que la razón por la que están tomando más tiempo que el promedio de proyectos se puede dar por causas puntuales que varían de uno a otro, y que seguramente se encuentren en espera sin ser tenidos en cuenta por los analistas y desarrolladores.

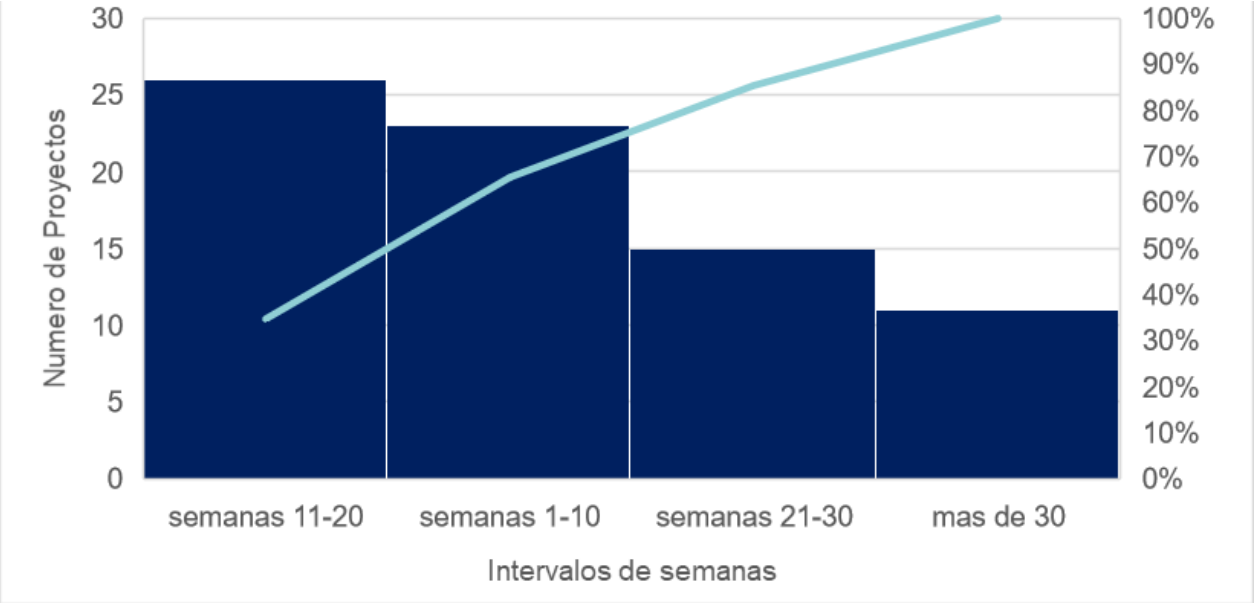

Figura 5 Diagrama de Pareto duración de proyectos en semanas Fuente: Elaboración propia

A continuación, en la Tabla 2, se muestran los resultados de estadística descriptiva con respecto a los tiempos de duración de los proyectos. Una desviación estándar de alrededor 13 semanas demuestra que la duración de los proyectos varía considerablemente entre los proyectos, dato al que no se le había prestado atención antes de la elaboración del presente trabajo, y es una

de las causas que impulsó a encontrar cuales son las variables que podrían influir en mayor medida en los tiempos de los proyectos.

Tabla 2 Estadística descriptiva variable Duración en semanas

Medida	**Resultado (Semanas)**
Media	18.73
Mediana	16
Moda	15
Desviación Estándar	12.48
Mínimo	2
Máximo	68
Total, de Datos	75

El histograma en la Figura 6 muestra una distribución en los datos para valores aleatorios no negativos y que se encuentran sesgados hacia la derecha . Con la función fitDist del paquete gamlss de R (Rigby & Stasinopoulos, 2005) se ajustan distribuciones paramétricas a los datos de interés, para este caso se encontró que las distribuciones que siguen los datos de tiempo de duración son Gamma y Gaussiana Inversa, y son las que se usaron a la hora de implementar el GLM

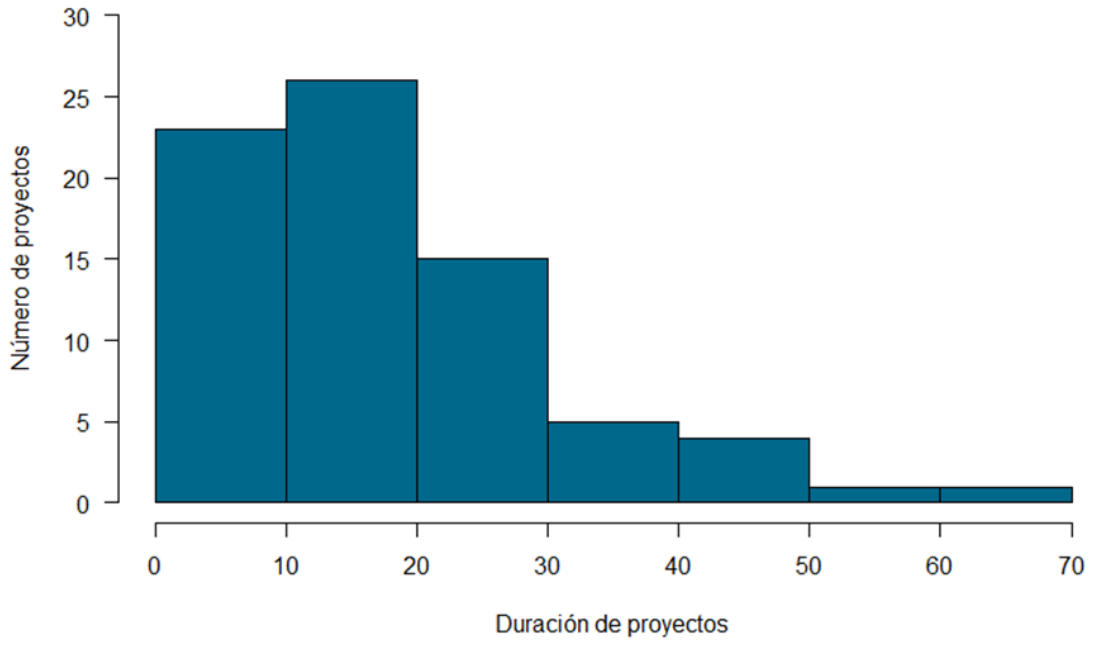

Figura 6 Histograma de Duración de proyectos. Fuente: Elaboración propia

También, como parte del análisis inicial a partir de la base de datos construida, se evaluaron las relaciones entre algunas variables que mostraron tener una incidencia directa en la variable duración en semanas de los proyectos. Las relaciones entre las variables Complejidad y Score fueron analizadas con el tiempo de duración de los proyectos, evidenciado en la Figura 7 y Figura 8respectivamente.

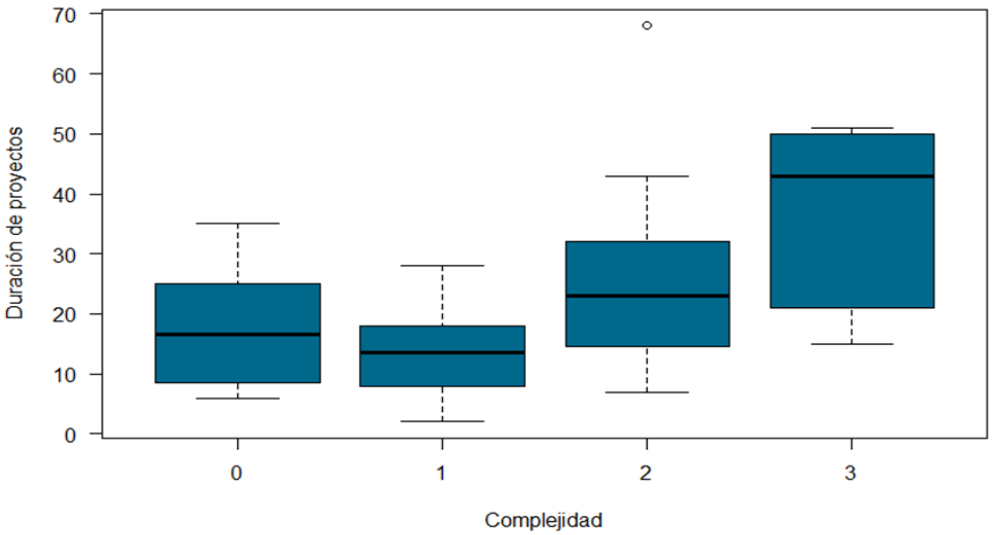

Figura 7 Boxplot Complejidad VS Duración

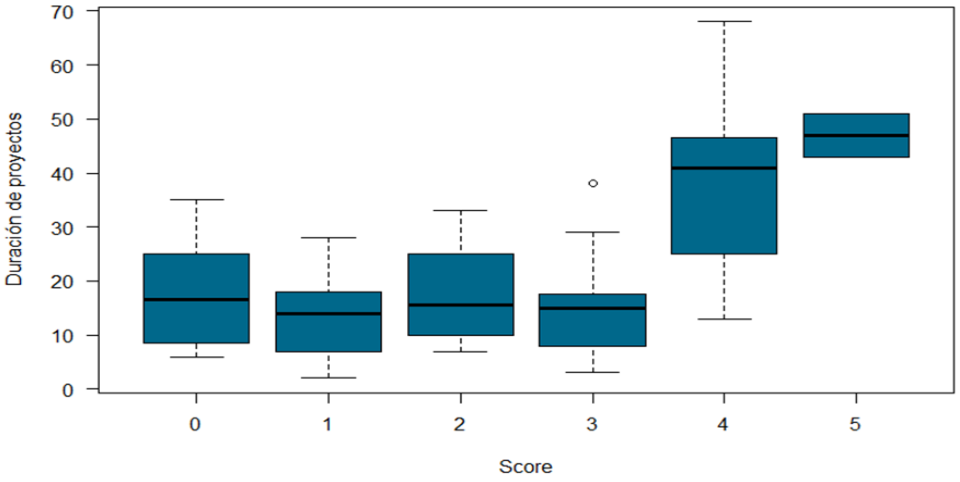

Figura 8 Boxplot Score VS Duración

Los resultados de las medidas descriptivas para ambas gráficas se presentan en la Tabla 3 y Tabla 4. Para la complejidad se observa que en todos los niveles de complejidad existe asimetría, así sean valores mínimos como en el caso de la complejidad 1 y 2. También es de resaltar que existe una dispersión considerable en los datos para el nivel de complejidad 3 que corresponde al más

alto, e hipotéticamente sería el nivel que más complejo seria poder buscar una forma estándar de explicar los tiempos de duración a partir de un nivel de complejidad alto. Un factor para considerar y que no se debe dejar pasar por alto es que el nivel de complejidad es asignado completamente de manera subjetiva, por lo tanto, los criterios de asignación de una escala varían entre los analistas que dan las valoraciones.

Tabla 3 Complejidad Vs Duración

Complejidad VS Duracion (semanas)						
Complejidad	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0	6	9.25	16.5	17.75	24	35
1	2	8.25	13.5	13.26	17.75	25
2	7	14.75	23	25.4	31.5	68
3	15	21	43	36	50	51

En el caso del Score que tienen los proyectos para los datos recopilados se encuentra que los datos están más dispersos para los Score de 4 (Tabla 4), los tiempos de duración de los proyectos no se encuentran en su mayoría alrededor de los mismos valores, su asimetría es negativa y un 50% de los proyectos con registros de Score 4 se encuentran entre 25 y 46 semanas según el primer y tercer cuartil, por tal motivo los proyectos que tienen esta duración y que según los resultados tienen tiempos de duración mayores a 30 semanas deben ser tomados en cuenta para la búsqueda de causas por las cuales está sucediendo dicha situación. Si bien son proyectos que se encuentran con las prioridades más bajas para trabajar en ellos junto a los que tienen Score de 5, es apropiado investigar qué es lo que está causando dicha duración en los proyectos para no generar acumulación de proyectos en espera.

Tabla 4 Score Vs Duración

Score VS Duración (semanas)						
Score	Mínimo	Cuartil 1	Mediana	Media	Cuartil 3	Máximo
0	6	9.25	16.5	17.75	24	35
1	2	7	14	13.43	18	28
2	7	10	15.5	17.95	24.5	33
3	3	8	15	14.07	17.5	38
4	13	25	41	37.86	46.5	68
5	43	45	47	47	49	51

En el número de sistemas no necesariamente los proyectos de automatización que tienen más sistemas son los que más se demoran. En los resultados no se encontró que el número de sistemas influye directamente en el tiempo de duración, lo que si se halló es que la mayoría de los proyectos necesitan de entre 1 y 3 sistemas para poder automatizar un proceso (Figura 9), y no necesariamente el aumento de ese número de sistemas haga que la fecha de entrega se extienda. Sería interesante para investigaciones posteriores analizar si alguna otra variable que permita medir la extensión del proceso en cuanto al número de pasos que un Bot debe ejecutar para un proceso pueda impactar la duración de los proyectos.

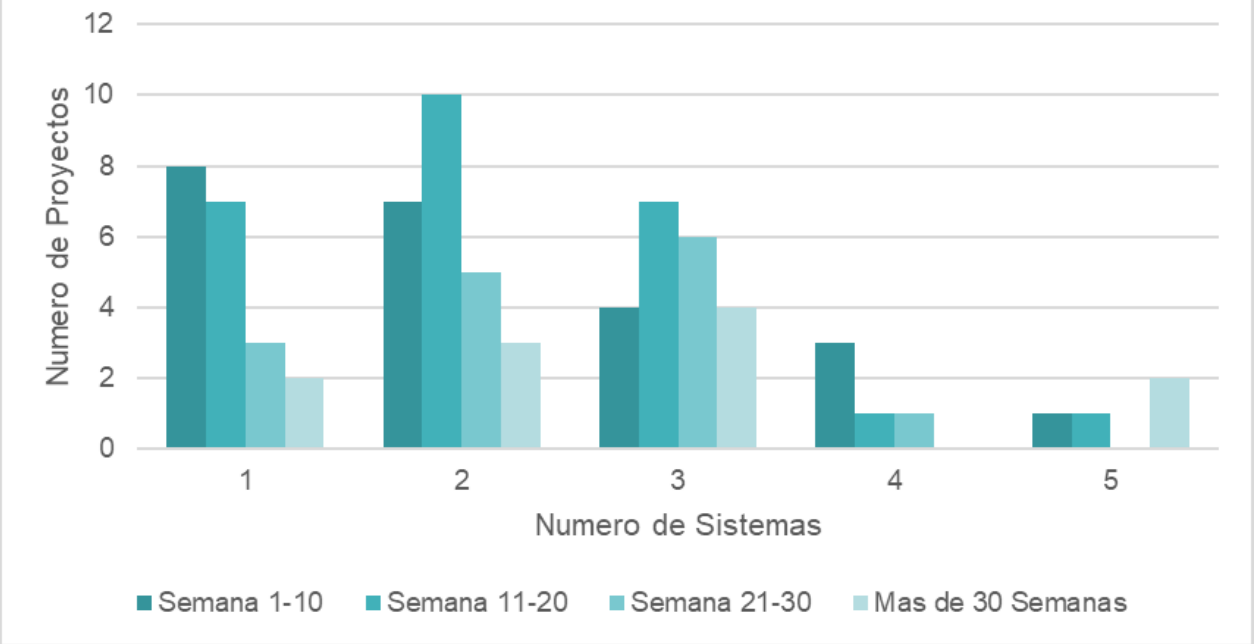

Figura 9 Número de Sistemas Vs Número de Proyectos en cada intervalo de semanas

Fuente: Elaboración propia

El beneficio monetario de los proyectos de automatización es calculado por el área de negocio que hace la solicitud, la forma de calcularlo varía y puede depender del ahorro en USD/hora-trabajador, el impacto que genera la mejora en el proceso gracias a la automatización o cualquier otro factor que se ajuste a las necesidades del negocio. Luego de que el negocio defina cuál es el beneficio, el analista de automatización encargado del proyecto debe sustentar la cifra ante el equipo de Governance & Controls encargado de realizar la auditoría de los proyectos. Una vez se aprueba el proyecto según los requisitos de beneficio, que sea un proceso estándar y automatizable con los sistemas disponibles, se pasa a la siguiente etapa de construcción de la documentación del proceso (PDD). En la Figura 10 se ve el beneficio promedio en dólares para cada intervalo de semanas.

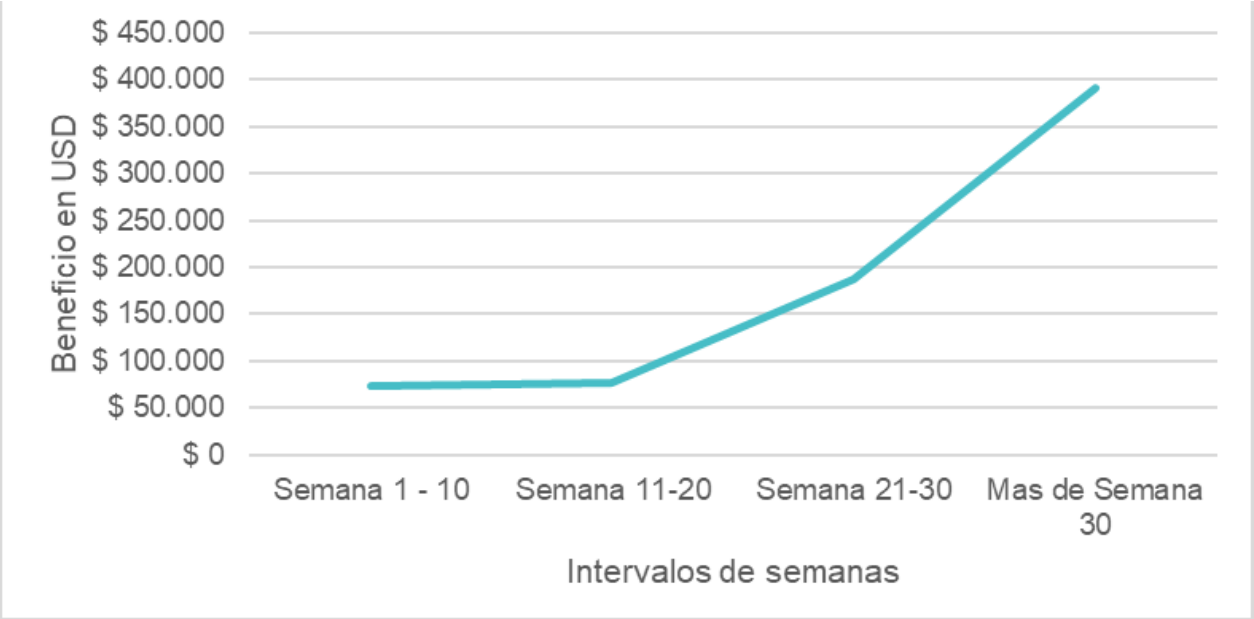

Figura 10 Beneficio promedio por intervalo de semanas Fuente: Elaboración Propia

A diferencia de lo esperado inicialmente, los proyectos que finalizan más pronto de acuerdo con la prioridad alta no son los que tienen un beneficio mayor, al contrario, se puede observar que el beneficio promedio aumenta a medida que el número de semanas necesarias para llevarlo a cabo lo hacen también.

Resultados Fase 3:

Lo que arrojó el proceso de optimización de hiperparametros de SVR con la ayuda de la función GridSearchCV fueron 11 combinaciones de hiperparametros, la mejor combinación está dada por:

gamma = 0.3

C = 30

epsilon = 0.0001

kernel = rbf

Para DTR se tomó desde la división de los datos de entrenamiento y prueba. Se tomaron los hiperparámetros que vienen por defecto en la función DecisionTreeRegressor que son (Scickit-learn, 2021):

criterion = squared_error

splitter = best

max_depth =None

min_samples_split=2

Para el GLM en R, el modelo se corrió con las distribuciones Gamma y Gaussiana Inversa con base a la siguiente información:

Tabla 5 . Distribuciones de probabilidad utilizadas en el Modelo Lineal Generalizado

Distribución	Función de densidad de probabilidad	Función de enlace	Media y varianza
GA	$f_Y(y\|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{\frac{1}{\sigma^2}-1} exp\{-y/(\sigma^2\mu)\}}{\Gamma(1/\sigma^2)}$ $y>0, \mu>0, \sigma>0$	*Log*	$E[Y] = \mu$ $Var[Y] = \sigma^2\mu^2$
IG	$f_y(y\|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}}\ exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$ $y>0, \mu>0, \lambda>0$	*1/y*	$E[Y] = \mu$ $Var[Y] = \mu^3/\lambda$

Los modelos se ajustaron y se realizó la predicción en los softwares respectivos, luego se compararon los resultados de la predicción con los datos iniciales del tiempo de duración de los proyectos a través de las medidas de desempeño MSE y RMSE. Los resultados arrojados se evidencian en la Tabla 6

Tabla 6 Resultados de MSE y RMSE para GLM-G, GLM-IG, SVR y DTR

Medida de Desempeño	**MSE**				**RMSE**			
Variables	**DTR**	**SVR**	**GLM-G**	**GLM-IG**	**DTR**	**SVR**	**GLM-G**	**GLM-IG**
Todas	0.14	0.67	10.43	9.89	0.38	0.82	108.75	97.82
Complejidad	114.30	71.33	11.08	11.08	10.69	8.45	122.84	122.84
Score	74.12	72.76	12.53	11.77	8.61	8.53	157.11	138.45

Se realizó la predicción con los modelos tanto con todas las variables explicativas como por aparte las variables Complejidad y Score. Todos los modelos se corrieron varias veces para cerciorarse de que se obtuvieron resultados confiables.

5 Análisis

Análisis Fase 1:

Entre las etapas críticas en las que son más propensos los retrasos en los tiempos estimados iniciales, se encuentra que la parte de recibir la aprobación por parte de las unidades de negocio, el envío de documentos necesarios para la ejecución de procesos y la realización de la retroalimentación pueden causar que los proyectos duren más de lo esperado inicialmente. Hay un factor que se debe considerar y es que en reiteradas ocasiones los compromisos adquiridos por el área de IAS no tienen algún tipo de presión por parte de las áreas que solicitan las automatizaciones debido a que en la mayoría de los casos se ve la automatizacion de los procesos como ayuda extra que reciben mas no vital para poder realizar los procesos, por lo tanto no existen niveles de servicio estándar previamente establecidos para que sean cumplidos por parte de las unidades de negocio y de IAS. Dicho panorama actualmente es así, pero teniendo en cuenta el crecimiento que ha tenido IAS al interior de Stanley Black & Decker y la necesidad emergente de automatizar más procesos, se espera que a futuro se establezcan niveles de servicio para que se realicen las etapas de los proyectos de manera más eficiente y con compromiso de los grupos de interés, de esta manera para cuando llegue ese momento las etapas críticas de los proyectos ya estarán identificadas y posiblemente se habrán realizado esfuerzos por prevenir retrasos en la operación.

Según testimonios de desarrolladores y analistas de automatizacion, las etapas críticas para considerar posibles retrasos son las de construcción del Bot (Build) y de pruebas (UAT). En dichas etapas es común que se den retrasos debido a complicaciones en el código desarrollado para la automatizacion, errores en los sistemas, comunicación interrumpida con las unidades de negocio o modificaciones de última hora. La alineación de las etapas con todos los responsables es un factor fundamental para la ejecución exitosa de los proyectos.

Análisis Fase 2:

Los tiempos de duración de los proyectos se ubican en un 80% entre 1 y 30 semanas, la desviación estándar se encuentra alrededor de las 13 semanas, llevado a la gestión de los proyectos es común encontrar que haya algunos que podrían ser terminados rápidamente pero que no tienen prioridad alta para ser ejecutados, por lo tanto es común ver proyectos que pasen algunas semanas en cola luego de terminar la documentación del proceso hasta que un desarrollador es asignado para construir el Bot, entonces no necesariamente el tiempo de duración del proyecto se refiera a semanas de trabajo continuo, en la cotidianidad es más probable que queden en espera algunas semanas si tienen prioridad baja de ejecución .

Entre las variables consideradas, la Complejidad y Score son las que tienen mayor impacto sobre la variable Tiempo de Duracion. Es de esperarse que los proyectos que tienen una complejidad 3 tomen más tiempo realizarlos, y para aquellos que se encuentran con complejidad 1 y tiempos de duración por encima del tercer cuartil de la Figura 7, su duración sea explicada por la relación con el Score que tiene el proyecto como criterio para darle prioridad. No deja de existir la incertidumbre de que la Complejidad sea una variable que toma valores subjetivos asignados por los analistas. Contrario a la variable Complejidad, en el caso de Score a medida que se cambia de nivel no ascienden o descienden de manera continua las estadísticas descriptivas que aparecen en la Tabla 4. Era de esperarse que los tiempos de duración aumentan a medida que disminuye el Score, pero en el caso del nivel 1 y 3 toma valores similares en los cuartiles, lo que lleva a la conclusión que no es una variable que genere un impacto considerable en los tiempos de duración.

Para el número de sistemas, en su mayoría se usan entre 1 y 3, no existe evidencia suficiente para afirmar que entre más sistemas mayor sea el tiempo de duración de los proyectos, pero como parte de las recomendaciones que se incluyen en este trabajo se encuentra que debería considerarse a futuro una variable que mida el volumen de pasos que deben ser ejecutados por un Bot para ver si influye de alguna manera en el trabajo llevado a cabo por los desarrolladores. Paralelamente, el beneficio económico es una variable que tiene una gran importancia al interior de área y que debería

ser siempre considerada para la priorización de los proyectos, pero no necesariamente debe ser una variable que por sí sola explica la duración de estos.

Análisis Fase 3:

Luego de correr los modelos, se encontró que para GLM-G y GLM-IG así como para los modelos computacionales SVR y DTR los resultados en las medidas de desempeño aumentan (ver Tabla 6), los valores más altos en MSE y RMSE son los que se obtuvieron para ambas distribuciones en el GLM llegando a estar hasta por encima de 100. No se obtuvieron resultados seleccionables en todos los modelos para los casos en que se corrieron solo con las variables Complejidad o Score, pero si se obtuvieron resultados a ser considerados en ambos modelos computacionales al ser corridos con todas las variables definidas desde el inicio. Se obtuvo una diferencia en los resultados de MSE entre SVR y DTR de solo 0.24 y en RMSE de 0.44, siendo DTR el mejor modelo basado en los resultados de ambos errores comparados con los otros. Los resultados permiten llegar a la conclusión que los modelos computacionales implementados en esta investigación son de gran importancia para ser parte del análisis que se lleva a cabo al interior de IAS para definir las variables que deben ser consideradas como el punto de partida en la mejora de la estimación de los tiempos de entrega de los proyectos. Los modelos tendrán mayor aplicación para predecir la duración en la medida que se recolecten y almacenen los datos de las variables en los servidores de SQL con los que cuenta IAS.

6 Conclusiones

1. Las etapas que más eventos inesperados presentan son las de Build y UAT, en dichas etapas las características de cada proyecto pasan a condicionar la duración, y de la misma forma condicionar la continuidad del resto de proyectos debido a que la fuerza de trabajo se enfoca en el proyecto que debe culminar, causando así colas de otros proyectos en espera de ser continuados. La comunicación en dichas etapas debe ser constante y directa, así como el compromiso de todas las partes involucradas por cumplir el nivel de servicio previamente acordado.

2. Las variables que mayor impacto generan son la Complejidad y Score, en la segunda toma gran relevancia el beneficio económico para su definición. El pilar fundamental del área IAS es mejorar los procesos de la empresa generando un beneficio económico, por lo tanto, es una variable que se debe tener en cuenta para investigaciones posteriores relacionadas con definir el tiempo que debe durar cada proyecto. También la Complejidad demostró tener un impacto considerable en el tiempo de duración de los proyectos, dicha variable cambia su valoración de un analista a otro, al ser una variable que está relacionada con la dificultad del proceso se debe hacer una redefinición por otra que tenga un método estándar para su cálculo y en la que se evidencian factores que podrían influir como el número de pasos que debe seguir el proceso para ser automatizado.

3. El uso de modelos computacionales para predecir el tiempo de duración de los proyectos de automatizacion brindó resultados satisfactorios según las medidas de desempeño MSE y RMSE principalmente en el modelo DTR. Dicho modelo debería ser tomado como base para los siguientes pasos en el proceso de mejora de las estimaciones de los tiempos de duración de proyectos de IAS. Los beneficios que llegue a tener el modelo dependerán de la consideración de otras variables que no tienen registro de datos actualmente, pero de las que se podría seguir recolectando datos de ahora en adelante. Este tipo de modelos predictivos computacionales podrían ser también aplicados en otras áreas de la empresa Stanley Black & Decker en mejoras de procesos relacionados con proyectos que tienen varias actividades con distintos grupos de interés y que deben cumplir determinados niveles de servicio.

7 Recomendaciones

Actualmente en el área no se encuentran datos de otras variables que pueden ser importantes para definir en qué se debe basar un analista de automatizacion para definir el tiempo de duración de un proyecto. Indudablemente el estado en el que se encuentre el pipeline con proyectos en cola debe ser considerado y además de eso el Score que es asignado. El área también se debería enfocar en la definición de una variable calculada con un método estándar para referirse a la complejidad de cada proyecto, la recolección de los datos de esta variable y de las otras se debería realizar de manera automática en los servidores de SQL del área, de esta manera se irá alimentando una base de datos que contenga las variables que se considere podrían interferir en el tiempo de duración de los proyectos. Una investigación más profunda sobre las causas potenciales de retrasos en los proyectos permitirá identificar nuevas variables nunca consideradas y de las cuales se puede empezar a recolectar datos para evaluar su comportamiento en el tiempo, de esta manera será factible construir una herramienta que permite estimar con exactitud los tiempos de duración de los proyectos con base a las variables pertinentes.

Referencias

Arroyo, I., Bravo, L. C., Llinás, H., & Muñoz, F. L. (2014). *Distribuciones Poisson y Gamma: Una discreta y continua relación. Prospectiva*, 12(1), 99–107.

Awad, M., & Khanna, R. (2015). *Support Vector Regression* (pp. 67–80). https://doi.org/10.1007/978-1-4302-5990-9_4

Cayuela, L. (2009). *Modelos lineales generalizados (GLM). Materiales de Un Curso Del R Del IREC.*

Gonzalez, L. (2018). *Evaluando el error en los modelos de regresión.* https://aprendeia.com/evaluando-el-error-en-los-modelos-de-regresion/

Kalemanova, A., Schmid, B., & Werner, R. (2007). *The normal inverse Gaussian distribution for synthetic CDO pricing.* The Journal of Derivatives, 14(3), 80–94.

Limón, A., Saavedra, B., & Gratta, F. (2014). *El Efecto Anclaje: Aplicación en negociación Brenda Gatica Saavedra. Forseti*, 2, 188. http://forseti.pe/media_forseti/revista-articulos/16_GATICA.pdf

Loh, W.-Y. (2011). *Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 14–23. https://doi.org/10.1002/widm.8

Lorko, M., Servátka, M., & Zhang, L. (2019). *Anchoring in project duration estimation. Journal of Economic Behavior & Organization*, 162, 49–65.

Management Solutions. (2018). *Machine Learning , una pieza clave en la transformación de los modelos de negocio.* https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Nerd, N. (2021). *Decision Tree Regression Clearly Explained!* https://www.youtube.com/watch?v=UhY5vPfQIrA

PMI Project Management Institute. (2013). *Guía de los fundamentos para la dirección de proyectos (Guía del PMBOK).* In Project Management Institute, Inc.

Rigby, R. A., & Stasinopoulos, D. M. (2005). *Generalized additive models for location, scale and shape.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 507–554.

Scickit-learn. (2021). *sklearn.tree.DecisionTreeRegressor.* https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor

Zheng, S. (2019). *Methods of Evaluating Estimators.* http://people.missouristate.edu/songfengzheng/teaching/mth541/lecture notes/evaluation.pdf

Anexos

Diagrama de Flujo del proceso de automatización