UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PREDICCIÓN DE RENUNCIA VOLUNTARIA DE COLABORADORES CON PERFIL TECNOLÓGICO DE UNA ENTIDAD FINANCIERA UTILIZANDO *REGRESIÓN LOGÍSTICA BINARIA*"

TRABAJO DE SUFICIENCIA PROFESIONAL
PARA OPTAR TÍTULO DE
INGENIERO ESTADISTÍCO INFORMÁTICO

RENZO RUBÉN ROMERO MONTOYA

LIMA – PERÚ

2023

TSP RENZO

Dr. Rino Sotomayor
Asesor

INFORME DE ORIGINALIDAD

16%	15%	4%	9%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.lamolina.edu.pe Fuente de Internet	2%
2	Submitted to Universidad Nacional Agraria La Molina Trabajo del estudiante	1%
3	hdl.handle.net Fuente de Internet	1%
4	www.armadilloamarillo.com Fuente de Internet	1%
5	www.coursehero.com Fuente de Internet	1%
6	Submitted to Pontificia Universidad Catolica del Peru Trabajo del estudiante	1%
7	peru21.pe Fuente de Internet	<1%
8	Submitted to University College London Trabajo del estudiante	<1%

UNIVERSIDAD NACIONAL AGRARIALA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

“PREDICCIÓN DE RENUNCIA VOLUNTARIA DE COLABORADORES CON PERFIL TECNOLÓGICO DE UNA ENTIDAD FINANCIERA UTILIZANDO *REGRESIÓN LOGÍSTICA BINARIA*”

PRESENTADO POR

RENZO RUBÉN ROMERO MONTOYA

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERO ESTADISTÍCO INFORMÁTICO

SUSTENTADO Y APROBADO ANTE EL SIGUIENTE JURADO

……..….	….……
Dr. Fernando René Rosas Villena	Dr. Rino Nicanor Sotomayor Ruiz
PRESIDENTE	ASESOR
.....	.....
Dr. Iván Dennys Soto Rodríguez	Dr. Raphael Félix Valencia Chacón
MIEMBRO	MIEMBRO

Lima – Perú

2023

DEDICATORIA

Para mis padres, hermanas y a mis dos sobrinos gigantes J y M.

AGRADECIMIENTO

Agradezco a mis padres y hermanas que siempre estuvieron apoyándome e impulsándome en todo momento.
Agradezco a los distintos compañeros que conozco y conocí durante lo que llevo en mi vida laboral ya que de ellos aprendí y sigo aprendiendo muchas cosas.
Agradezco al Profesor Rino Sotomayor, por su apoyo en la realización de este trabajo.

INDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

Las entidades financieras necesitan estar a la vanguardia de la tecnología para poder competir contra otras organizaciones del mismo rubro, por lo cual es fundamental no perder y retener a los colaboradores con perfiles tecnológicos, ya que ello significaría una gran pérdida de conocimientos y dinero, lo que pondría en riesgo la continuidad del negocio. Una forma de prever estas posibles pérdidas es mediante técnicas estadísticas como la regresión logística. En el presente trabajo de suficiencia profesional se describe la manera en la que se aplicó la regresión logística apoyándose en la metodología CRISP, para así obtener una clasificación correcta del 60% en la ocurrencia trimestral de la decisión de renunciar del colaborador con perfil tecnológico y reducir su indicador de rotación voluntaria trimestral hasta casi un 1.5%.

Palabras clave: Renuncia, Rotación, Perfil tecnológico, Regresión logística, Probabilidad de renuncia

ABSTRACT

Financial entities need to be at the forefront of technology to be able to compete against other organizations in the same field, which is why it is essential not to lose and retain collaborators with technological profiles, since this would mean a great loss of knowledge and money, which that would put business continuity at risk. One way to predict these possible losses is through statistical techniques such as logistic regression. This professional sufficiency work describes the way in which logistic regression was applied based on the CRISP methodology, to obtain a correct classification of 60% in the quarterly occurrence of the decision to resign of the collaborator with a technological profile and reduce its quarterly voluntary turnover indicator up to almost 1.5%.

Keywords: Resignation, Turnover, Technological profile, Logistic regression, Probability of resign.

I. INTRODUCCIÓN

1.1. Problemática

Uno de los problemas que afecta a toda empresa es la renuncia o fuga del personal, este hecho puede causarle pérdida de conocimientos y dinero a la organización lo cual desembocaría en que esta pierda terreno frente a la competencia y finalmente desaparezca, es por ello que cuando existe una rotación o tasa de abandono alta es de gran importancia para los departamentos de recursos humanos el poder enfrentar dicho problema, descubrir las causas de que un colaborador renuncie y también sus características, de tal manera que incluso se puedan adelantar a la decisión de la renuncia.

Uno de los puestos más importantes en varios rubros empresariales de acuerdo con Sánchez (2021) es el del empleado con perfil tecnológico, ya que ayuda a abordar la adopción y transformación tecnológica en las empresas con sus diversos conocimientos especializados, por ejemplo, en ingeniería de software, automatización de procesos, auditoria de ciberseguridad, aprendizaje automático, gestión de datos, ciencia de datos y desarrolladores de código abierto. En el Perú, según PageGroup Perú (2021, como se citó Chávez, 2020) la demanda de estos perfiles aumentó a inicios de la pandemia en un 60%, siendo la oferta bastante reducida. Debido a ello la entidad financiera en la que se aplicó lo expuesto en este trabajo optó por priorizar el hecho de mantener a los empleados que ya tenía con este perfil, siendo su principal problema el de reducir su rotación, específicamente lograr que se redujera el indicador de renuncia o rotación voluntaria trimestral, que si bien no contaba con un valor muy grande (apenas un 3% aproximadamente) se traducía en un costo monetario considerable, ya que los procesos de selección para poder conseguir un solo empleado de este perfil son altos por la baja oferta.

Por el problema mencionado, es que se decidió analizar la información histórica relacionada con la demografía y trabajo de los empleados con este perfil, para finalmente obtener probabilidades de renuncia mediante una regresión logística binaria que ayude a los líderes de equipo y al área de recursos humanos de la entidad a generar iniciativas de retención para aquellos colaboradores con mayor probabilidad de renuncia.

1.2. Objetivos

1.2.1. Objetivo general

Proveer de una herramienta que permita al área de recursos humanos y a los líderes aplicar estrategias que puedan retener a aquellos colaboradores con perfil tecnológico más propensos a renunciar y así reducir el indicador de rotación voluntaria trimestral.

1.2.2. Objetivos específicos

- Desarrollar un modelo predictivo capaz de poder predecir la propensión a la renuncia voluntaria en una ventana de tiempo de tres meses mediante el uso de regresión logística binaria.
- Identificar aquellas variables que influyen en la decisión de renuncia de un colaborador.

II. REVISIÓN DE LITERATURA

2.1. Definiciones de Recursos Humanos

2.1.1. La renuncia

Según Cortés (s.f.), no existe una única forma de denominar al término del contrato por voluntad propia del trabajador o la disposición de no continuar en el empleo dependiendo del país, por ejemplo, dimisión, retiro y abandono, sin embargo, en el Perú se le denomina "renuncia".

En el caso de la entidad financiera de nuestro interés, se entiende como renuncia al rompimiento de la relación laboral entre el colaborador y la organización por decisión propia del empleado, de forma similar a lo que menciona Pacori (2021).

2.1.1.1. Indicador de rotación trimestral

El indicador de rotación o tasa de rotación es una métrica de recursos humanos que ayuda a medir la eficacia de las organizaciones al momento de fidelizar a sus empleados INTERIM GROUP (2022). Es el porcentaje de colaboradores que dejaron la organización en un periodo de tiempo determinado. Para la entidad bancaria en estudio se maneja el indicador de rotación voluntaria trimestral, el cual se calcula como el número de colaboradores (en nuestro caso con perfiles tecnológicos) que decidieron terminar su relación laboral con la empresa en un periodo de tres meses, entre el promedio de los colaboradores que se mantienen activos durante esos mismos tres meses.

$$\textit{Indicador de rotación voluntaria trimestral} = \frac{\text{Número de empleados que renuncian en de tres meses}}{\text{Promedio de empleados activos en tres meses}}$$

2.1.2. Los perfiles tecnológicos

Los perfiles tecnológicos son aquellos cuyas habilidades están relacionadas con el desarrollo y transformación tecnológica en las organizaciones. Según Ostrowski (2022), los conocimientos de estos profesionales que abarcan la ciencia, ingeniería, tecnología y matemáticas ayudan a las empresas a seguir innovando, creando y mejorando herramientas para que puedan continuar en el mercado.

De acuerdo con Cuadrado (2021), algunas de las especialidades de perfiles tecnológicos pueden ser, por ejemplo, Ingenieros de software, que son los responsables de la codificación, administración ágil de sistemas y colaboración con otros equipos de tecnología; Técnicos de automatización de procesos, que aceleran el trabajo y reducen los errores en las implementaciones tecnológicas; Diseñador UX, quienes se encargan de garantizar una buena navegación y coherencia las interfaces de cualquier sistema; Auditores de seguridad, encargados de proteger la información de la organización mediante procedimientos, hacking técnico y análisis de vulnerabilidades; Expertos en la nube, que proporciona capacidades de almacenamiento, aceleración de transformación y soluciones escalables; Arquitectos de Big Data, que se especializan en la gestión de datos, como la ciencia de datos, entre muchos otros.

2.1.2.1. Demanda y Fuga

La búsqueda de los perfiles tecnológicos continúa en aumento al igual que el avance de la tecnología, sin embargo, su oferta es aún escasa a nivel mundial. Sánchez (2021) incluso menciona que la unión europea tiene escasez de especialistas en áreas como la Inteligencia Artificial y la Ciberseguridad. Hay que resaltar también los datos que proporciona el Barcelona Digital Talent Overview (2022) sobre la situación en España, que nos comenta que la demanda por este tipo de profesionales había crecido en un 43% en el 2021, pero la oferta era solo del 11% y para el Perú según el estudio Prospección del mercado de TI (Apuy, 2020) ya éramos el quinto país de América Latina con mayor demanda de estos perfiles y contábamos con un déficit de 17 mil especialistas.

Toda esta alta demanda convierte a las empresas no solo en competidoras por el negocio, sino también por los talentos en perfiles tecnológicos, Por ejemplo, Vargas (como se citó en Perú 21, 2023) comenta que tanto en Perú como Latinoamérica, al ser especialidades que

pueden trabajar de forma remota, son solicitadas también por organizaciones extranjeras capaces de otorgarles mayores beneficios salariales, por lo cual se está optando por motivar el desempeño, crear programas de reconocimiento, bonificaciones y oportunidades de desarrollo profesional, construyendo así una cultura laboral que hagan que el empleado se sienta a gusto trabajando con su líder y equipo, de manera que se frene la fuga de estos profesionales.

2.2. Conceptos estadísticos

2.2.1. La regresión logística binaria

La regresión logística es una técnica estadística que se utiliza para estimar la probabilidad de ocurrencia de un evento con varios posibles resultados o categorías en función de otras variables o predictoras numéricas o también categóricas. La regresión logística binaria hace alusión a que el suceso de interés tiene solo dos posibles resultados.

Según comenta González (2019), la regresión logística discrepa de su similar la regresión lineal en que se usa cuando la variable dependiente es una categoría y no una cantidad. Buitrago (2020) nos menciona, por ejemplo, que la regresión lineal no funciona en problemas donde se requiera predecir clases binarias, a diferencia de la regresión logística binaria que se apoya en la función logit o sigmoide.

Ahora, como nos muestra Saini (2021) lo que deseamos es estimar la probabilidad *P* de la ocurrencia de un evento *y* en función de variables independientes *x,* y como sabemos las probabilidades van desde o hasta 1, por lo que utilizaríamos la siguiente fórmula:

$$\text{Log}\left(\frac{\text{P (Y=1)}}{\text{1-P(Y=1)}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Donde *Y=1* indicaría la ocurrencia del evento y $\beta_1, \beta_2, ..., \beta_k$ son los parámetros desconocidos asignado a cada variable x_i (β_0 está asignada a la constante).

Entonces como deseamos la ecuación en función de *P (Y=1)* transformarnos y nos resulta

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}}$$

Cabe comentar que esta técnica de clasificación es una de las más famosas y sencillas de aplicar de acuerdo con Rodríguez (2018) y González (2019), ésta funciona mejor si es que se consideran como variables predictoras a aquellas que no estén relacionadas, es decir que no presenten multicolinealidad (Chique, 2020) porque puede afectar a los coeficientes resultantes.

2.2.1.1. Odds, odds ratio y coeficientes

En la regresión logística los odds o razón de probabilidad se definen como el cociente entre la probabilidad de que un evento ocurra o sea verdadero *P(Y=1)* y que no ocurra *P(Y=0)* o sea falso, supongamos entonces que *P(Y=1) / P(Y=0) = a*, entonces podemos decir que el *odds=a*, es decir se esperan *"a"* eventos verdaderos por cada evento falso (Amat, 2020). De acuerdo con López y Fachelli, (2015), el odds ratio o razón de ventajas es el cociente entre dos odds y también se representa como la exponencial de β_i *(exp* (β_i*))*, entonces cuando *exp* (β_i*)* es mayor que 1, esto señala que un aumento de la variable independiente i, aumenta los odds que ocurra el evento (variable dependiente) y cuando el *exp* (β_i*)* es menor que 1, un aumento de la variable independiente i, reduce los odds que ocurra el suceso (variable dependiente).

2.2.1.2. Multicolinealidad

Como mencionamos antes, la multicolinealidad es uno de los aspectos a revisar al momento de realizar el modelado de la regresión logística, su presencia afecta a los coeficientes y por ende a la interpretación de los resultados. La existencia de ella en los modelos se puede evaluar de las siguientes formas:

a. **Coeficiente de correlación:** Mide la fuerza de asociación entre dos variables numéricas y su dirección, sus valores fluctúan entre -1 y 1, tomando el signo negativo cuando la relación es inversa, positivo cuando es directa y 0 cuando no hay relación. Dos de los más utilizados son el coeficiente de correlación de Pearson y Spearman (Universidad de Guanajuato, 2022).

    - **Coeficiente de correlación de Pearson:** Mide la relación lineal entre dos variables que poseen una distribución normal y se calcula de la siguiente manera:

$$\rho = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Donde $\bar{x}$ y $\bar{y}$ son las medias de las variables analizadas y $-1 \leq \rho \leq 1$

- **Coeficiente de correlación de Spearman:** Es unan medida de asociación utilizada cuando las variables numéricas no se distribuyen normalmente

$$\boldsymbol{\gamma = 1 - 6\left(\sum_{i=1}^{n} d_i^2\right) / n(n^2-1)}$$

Donde d es la diferencia de rangos del elemento n y $-1 \leq \boldsymbol{\gamma} \leq 1$

b. VIF generalizado: Conocido también como *GVIF* mide el aumento en la varianza de la estimación de los coeficientes de un grupo de regresores en relación con la varianza de la estimación de un coeficiente en un modelo que contiene solo un regresor. Para los predictores numéricos continuos, el *GVIF* es el mismo que el *VIF*, pero para los predictores categóricos, el *GVIF* proporciona un solo número para todo el grupo de coeficientes codificados por contraste asociados con un predictor categórico (Fox y Monette, 2012). Su valor tiene el mismo significado que el *VIF* utilizado normalmente en regresiones lineales, es decir que cuando las variables tengan valores iguales a 1 no hay efecto de multicolinealidad, entre 1 y 5 que hay un efecto moderado y mayor a 5 que el efecto de la relación entre variables afecta al modelo.

2.2.1.3. Bondad de ajuste de modelo y variables

Las pruebas utilizadas para evaluar la bondad este algoritmo, así como la importancia de las variables serán las siguientes:

a. Prueba de razón de verosimilitud: Es una prueba que compara un modelo reducido, usualmente sin ningún regresor contra otro más complejo, como nos comenta Hernández (2020) en general nos dice sí el modelo más pequeño proporciona un mejor ajuste que el más complejo o con más variables predictoras, entonces la hipótesis nula nos indica que los coeficientes del modelo son cero y la alterna que al menos uno no lo es, teniendo así:

$$H_0: \beta_i = \beta_{i+1} = \ldots \beta_k = 0$$

H_1: Al menos uno de los $\beta_j \neq 0$, donde $j = i+1, i+2 \ldots\ldots k$

$$LR = -2ln(L1/L2)$$

Siendo *L1* y *L2* los valores del Log-likelihood para los modelos reducido y complejo respectivamente y donde *LR* se distribuye como un chi cuadrado con *k-1* grados de libertad.

b. Prueba de wald: En esta prueba se evalúa si los coeficientes β_i del modelo son distintos de 0, donde $i \neq 0$, es decir si la variable independiente en cuestión aporta de una forma estadísticamente significativa a la explicación de la variable dependiente mediante el estadístico de Wald que sigue una distribución normal estándar (López y Fachelli, 2015).

H_0: $\beta_{i....n} = 0$

H_1: $\beta_j \neq 0$ donde $j = i+1, i+2 \ldots\ldots n$

$$Wald = \frac{\hat{\beta}_i}{desviación\ estándar\ (\hat{\beta}_i)}$$

Donde $\hat{\beta}_i$ es el estimador de máxima verosimilitud para β_i

2.2.1.4. Evaluación del poder clasificatorio del modelo

Las herramientas que generalmente se usan para contrastar los resultados de clasificación del modelo de regresión logística binaria son:

a. Matriz de confusión: Es una tabla que nos permite visualizar los resultados de la aplicación de un algoritmo. Sirve para mostrar de forma explícita cuándo una clase es confundida con otra (Recuero, 2018).

Figura 1

Matriz de confusión

Matriz de confusión		Estimado por el modelo	
		Negativo (N)	*Positivo (P)*
Real	*Negativo*	a: (TN)	b: (FP)
	Positivo	c: (FN)	d: (TP)

Fuente: Recuero (2018)

a: es el número de predicciones clasificadas correctamente como negativas.

b: es el número de predicciones clasificadas incorrectamente como positivas y que en realidad son negativas.

c: es el número de predicciones clasificadas incorrectamente como negativas y que en realidad son positivas.

d: es el número de predicciones clasificadas correctamente como positivas.

- **Exactitud:** Es la proporción del número total de predicciones clasificadas correctamente es llamada exactitud y se determina de la siguiente forma:

$$Exactitud = \frac{a+d}{a+b+c+d}$$

- **Precisión:** Es la proporción entre los positivos reales predichos correctamente por el algoritmo y todos los casos positivos.

$$Precisión = \frac{d}{b+d}$$

- **Sensibilidad:** Conocido como tasa de verdaderos positivos, es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$Sensibilidad = \frac{d}{c+d}$$

- **Especificidad:** Son los casos negativos que el algoritmo ha clasificado correctamente.

$$Especifidad = \frac{a}{a+b}$$

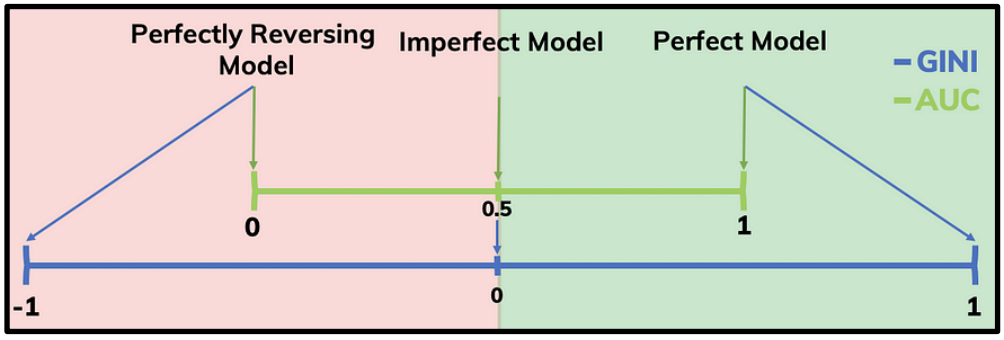
b. Curva ROC: La Característica Operativa del Receptor (Receiver Operating Characteristic o *ROC* en inglés), es usada para examinar la razón o ratio de verdaderos positivos frente a la razón o ratio de falsos positivos (Lantz, 2015)

mediante el cálculo del área de su curva o *AUC* que va del 0 a 1 y mientras más grande mejor, describiendo el grado de discriminación de nuestro modelo.

c. **Coeficiente GINI:** Cuantifica la desigualdad entre los valores de una distribución de frecuencia y se suele utilizar para medir la calidad de un clasificador binario. Un índice de *Gini* de 0 nos indica una clasificación totalmente mala, mientras que un índice de *Gini* de 1 expresa una perfecta (h2o.ai 2018). Se puede derivar directamente de la curva *AUC*, como su valor multiplicado por 2 y finalmente restado por 1 variando esta vez entre -1 y 1. De acuerdo con El Khal (2021), el índice de Gini es una forma de ajustar el valor de *AUC* y nos ejemplifica el poder clasificatorio del modelo mediante la relación entre el valor de *AUC* y el *GINI*, como se muestra a continuación:

Figura 2

Calidad de clasificación del modelo



FUENTE: El Khal (2021)

De acuerdo con lo ejemplificado en el gráfico, el modelo perfectamente inverso (Perfectly reversing model) nos indica que estamos clasificando todas las observaciones positivas como negativas, por lo cual este resultado nos diría que hemos tenido algún posible error. El modelo imperfecto nos señala que nuestro modelo no tiene ninguna capacidad clasificatoria, ya que no es capaza de diferenciar una categoría de la otra, y el modelo perfecto que si cumple con la tarea de correcta discriminación.

2.2.2. Balanceo de datos

En muchas ocasiones al tratar de realizar modelos de clasificación en data real puede ser posible observar que nuestra variable respuesta tiene un bajo porcentaje de ocurrencia para la clase que queremos clasificar, por ejemplo, del 3% y menores, lo que finalmente conlleva a que nuestro modelo no funcione correctamente, ya que no posee suficientes categorías de éxitos para poder entrenarse (Maggi, 2022). Para enfrentar estos problemas podemos usar varias alternativas brindadas por softwares estadísticos, entre los más sencillos se encuentran el submuestreo y el sobremuestreo. El submuestreo elimina los casos de la categoría mayoritaria hasta igualar o acercarse al porcentaje de ocurrencia de la clase menos frecuente, pero la desventaja de hacer esto es probablemente que se elimine información relevante, aún peor cuando se trata de muestras grandes, por otro lado, el sobremuestreo aumenta aleatoriamente el contenido de la categoría minoritaria repitiendo aleatoriamente lo registros existentes, sin embargo, podríamos estar agregando información irrelevante.

2.3. Proceso Estándar de Industria Cruzada de Minería de Datos (CRISP-DM)

La metodología utilizada en el presente trabajo se basará en el CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es uno de los modelos de proceso para proyectos de minería de datos más populares entre personas especializadas en analítica, ciencia de datos y aprendizaje automático. Lo anterior queda demostrado por la encuesta realizada por el sitio KDnuggets, del 2014 realizado a expertos en proyectos de minería de datos superando por mucho a otras como SEMMA o KDD. CRISP-DM fue propuesta a mediados del año 1999 por un consorcio de empresas compuesto por NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata y Daimer-Chrysler, organizando el procedimiento en seis etapas cuyo orden es flexible, dichas fases según Rodríguez *et al*. (2003), se describirán a continuación:

1. **Fase de comprensión del negocio o problema. -** Es la fase probablemente más importante, en ella es necesario entender de la mejor manera posible el problema que se desea resolver para aprovechar la minería de datos.
2. **Fase de comprensión de los datos. -** En esta fase se entenderán los datos, su significado y sus relaciones más evidentes. Esta etapa junto a las siguientes dos, son las que necesitan de mayor tiempo y esfuerzo.

3. **Fase de preparación de los datos. -** En esta fase de preparan a los datos para adaptarlos a las técnicas de minería de datos, esta etapa va de la mano con la de modelado, ya que según ella los datos requieren ser procesados de distintas formas.
4. **Fase de modelado. -** Se usan las técnicas de modelado de acuerdo con los objetivos.
5. **Fase de evaluación. -** Se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema.
6. **Fase de Implementación. -** En esta fase se documentan y difunden los resultados obtenidos, así como también se generan recomendaciones en relación con el problema.

2.4. Antecedentes

En los últimos años una de las bases de datos más estudiadas en cuanto a la predicción de la renuncia de empleados es la brindada por el data set de IBM HR Analytics, que contiene 1470 registros únicos con treinta cinco variables como por ejemplo, la edad, sexo, nivel educativo, años desde la última promoción, años de permanencia en la empresa, número de empresas donde el colaborador trabajó y sus tiempos de capacitación en el último año, cabe mencionar también que no hay mayor detalle en la temporalidad de la información extraída o si hubo un muestreo. Uno de los trabajos que analiza este conjunto de datos mencionado es el de Qutub *et al*. (2021), en el cual compara distintos algoritmos como el adabost, random forest, regresión logística, gradient boosting y modelos de ensamble para predecir si un empleado se queda o se va, mostrando que la regresión logística pudo superar a todas las demás técnicas por poca diferencia en los indicadores de clasificación del modelo.

Henao (2021) aplica modelos predictivos para determinar a aquellos empleados de una caja de compensaciones que cesarán voluntariamente, al igual que el caso anterior presenta registros únicos (3414) con información que abarca desde el 2010 hasta el 2021 y detallando si renunciaron o no, sin considerar el momento del evento. Recogió alrededor de ochenta y tres variables entre demográficas como la edad y el género y referentes a su puesto, por ejemplo, el grado salarial y el tiempo en la organización. En contraste con el estudio anteriormente mencionado, los resultados para la regresión logística fueron superados por la

técnica xgboost, sin embargo, la diferencia en los indicadores de clasificación es muy pequeña.

Adicionalmente en cuanto a la muestra de entrenamiento cabe destacar el trabajo de Abdel-Rahmen *et al*. (2021), que, aunque no use una regresión logística aplica una metodología propia de muestreo, de tal manera que no haya un solo registro por colaborador en los datos de entrenamiento, porque considera que la información de un empleado cambia a lo largo del tiempo y no debemos perderla. Para ello repite hasta cuatro veces registros de colaboradores que no renunciarán y tres veces para aquellos que sí lo harán en una ventana de 6 meses. Los resultados que obtienen aplicando esta forma singular de entrenamiento resultan tener indicadores de precisión y clasificación mucho más altos que si el entrenamiento hubiese tenido una observación por individuo.

Todo lo anterior mencionado ayudó a poder determinar que la regresión logística es una técnica útil para este caso de problemática, además de probar formas de muestreo en la cual repitamos los registros de un colaborador para observar la variación de sus características en el tiempo.

III. DESARROLLO DEL TRABAJO

3.1. Aplicación de la metodología

Como se comentó anteriormente este trabajo siguió la metodología CRISP-DM, y el desarrollo de este se describe de acuerdo con los pasos mencionados.

3.1.1. El problema de la renuncia y el aumento de la rotación trimestral

Como se comentó previamente la entidad bancaria, cuenta con los denominados perfiles tecnológicos en su planilla, específicamente comenzó a nombrarlos así a partir de enero del 2021, es decir la organización a la vez que iba contratando nuevos empleados comenzaba a llamar a algunos pocos colaboradores ya existentes y especializados con esta denominación, alcanzando para el mes de junio del 2022 la cantidad de 2133 profesionales con este perfil. A continuación, en la siguiente tabla se mostrará el detalle mensual de la composición de empleados con perfiles tecnológicos en el banco:

Tabla 1

Composición mensual de los perfiles tecnológicos en la entidad e indicador de rotación trimestral

Periodo	Total	Activos	Renunciantes	Rotación trimestral voluntaria
ENE-21	422	413	9	
FEB-21	440	436	4	
MAR-21	453	451	2	3.5%
ABR-21	477	468	9	3.3%
MAY-21	481	478	3	3.0%
JUN-21	491	489	2	2.9%
JUL-21	578	573	5	1.9%
AGO-21	1621	1618	3	1.1%
SET-21	1688	1675	13	1.6%
OCT-21	1804	1787	17	1.9%
NOV-21	1843	1824	19	2.8%
DIC-21	1884	1870	14	2.7%
ENE-22	1929	1908	21	2.9%
FEB-22	2013	2000	13	2.5%
MAR-22	2059	2040	19	2.7%

Periodo	Total	Activos	Renunciantes	Rotación trimestral voluntaria
ABR-22	2103	2062	41	3.6%
MAY-22	2130	2115	15	3.6%
JUN-22	2133	2114	19	3.6%
JUL-22	2167	2150	17	2.4%
AGO-22	2147	2124	23	2.8%
SEP-22	2184	2169	15	2.6%

Fuente: Elaboración propia

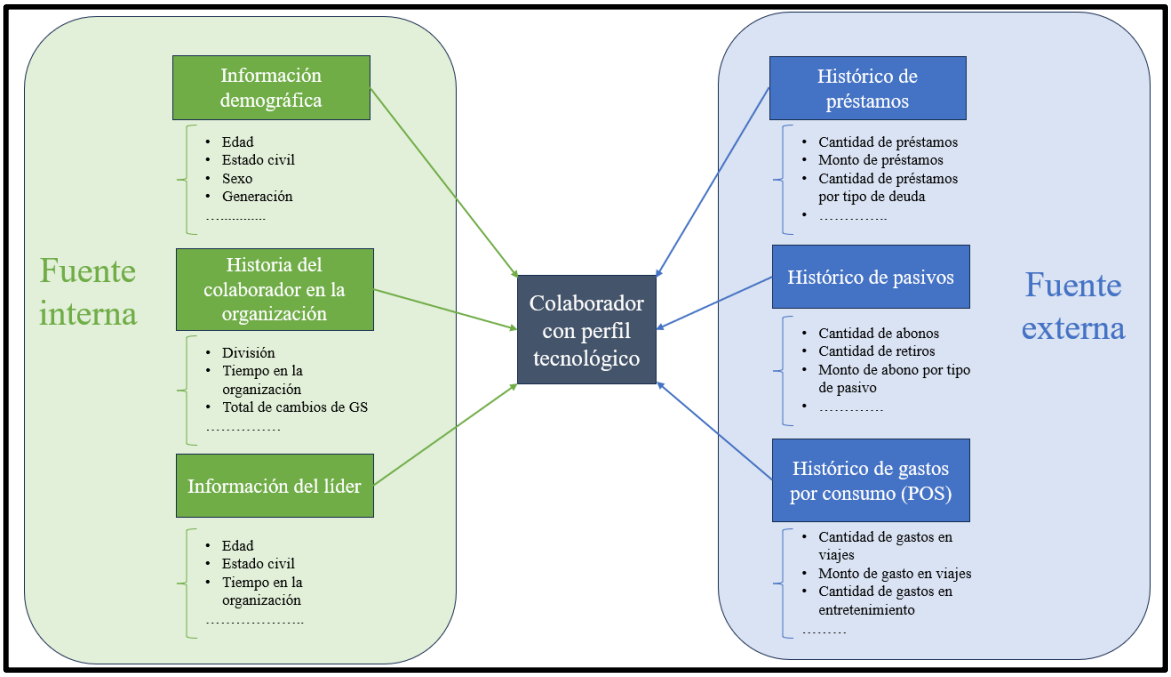
En la Tabla 1 se puede observar que el indicador de rotación trimestral voluntario tuvo un valor promedio de 2.8% en el último año e incluso llegó a un máximo de 3.6%, por lo que se generó una alarma y se solicitó por parte de la jefatura del área de recursos humanos buscar la forma de reducir este indicador mediante la generación, de un modelo predictivo sencillo que ayude a los líderes de equipo a poner especial a atención a aquellos colaboradores con una probabilidad considerable de renunciar y disminuir hasta un 1.5% el indicador en los próximos 6 meses (a partir de octubre 2022), fecha en la cual todos los equipos ya deberían de disponer de los resultados del modelo.

3.1.2. Obtención y comprensión de la información disponible

El área de recursos humanos ya contaba con los datos históricos de los movimientos en la entidad para todos los colabores con perfil tecnológico, como por ejemplo tiempo en la entidad, área o división de pertenencia, cambios en el área o en la división, grado salarial y cambios en el grado salarial, así como información demográfica básica, como la edad, el sexo y el estado civil del empleado. Toda esta información se encuentra recolectada en una base de datos en SQL server, que se alimenta de forma mensual, aunque el acceso a ella está restringido solo a algunos colaboradores de recursos humanos por contener datos sensibles del empleado. Otra fuente de información valiosa fue la del área de analítica y ciencia de datos de la entidad, ya que posee datos financieros del colaborador, por ejemplo, la cantidad de deudas, el tipo de deudas, el monto, los movimientos en los pasivos e información de consumo con POS en establecimientos diversos. Para este último caso se tuvieron que pedir permisos a dichos equipos, ya que también es información sensible y había que agregarlos a la base ya construida en recursos humanos.

Figura 3

Estructura de información



Fuente: Elaboración propia

Se creyó en la importancia de estas variables ya que se sostenían hipótesis como por ejemplo, que el colaborador es menos propenso a renunciar si él posee un grado salarial alto, o si tiene mucho tiempo en la entidad bancaria, o tuvo muchos cambios en grados salariales, o cambios en los grupos de trabajo, lo cual significaría una mejora en el desarrollo del colaborador y por ende mayor satisfacción; A su vez, la edad del jefe o su tiempo en dicha función indicaría que la experiencia influiría en el manejo y satisfacción del líder para con el empleado. En cuanto a los datos financieros algunas premisas eran que el empleado era más conservador al tomar la decisión de renunciar, por lo que la cantidad de sus préstamos disminuirían, al igual que gastos de consumo (en entretenimiento y viajes) y los movimientos en los pasivos. Hay que mencionar que también se pensó en incluir información de encuestas de experiencia, pero la baja participación de los perfiles tecnológicos y la distancia temporal entre los despliegues (cuatro meses) hizo que se descarte esta alternativa.

La ventana de tiempo a analizar partirá desde enero del 2022 hasta junio del 2022, ya que como se ve en el total de colaboradores representa un 88% del total de junio y además en enero ingresó en el estudio una de las divisiones más importantes de la entidad.

3.1.3. Preparación de datos y modificación de variables

En nuestros datos el evento objetivo queda definido como si la renuncia de un colaborador con perfil tecnológico sucederá en los siguientes 3 meses, por ello se colocó una marca de "si", por ejemplo, si el empleado se encuentra activo en enero del 2022, pero sabemos que renunciará en abril del mismo año. Es decir, ya en enero tendrá la etiqueta de que renunciará dentro de los próximos 3 meses y de forma sucesiva en febrero y marzo, sin embargo, este colaborador ya no aparecerá en abril dado que fue el mes en que renunció. Este mismo colaborador tendrá la marca de "no" para diciembre del 2021 y periodos anteriores. La variable objetivo será llamada "Ren3mes", de esta forma la distribución de nuestra variable dependiente en los periodos analizados quedó de esta forma:

Tabla 2

Composición mensual de la variable objetivo

Periodo	Ren3mes (NO)	Ren3mes (SI)
ENE-22	1779	69
FEB-22	1864	72
MAR-22	1907	72
ABR-22	1956	49
MAY-22	1998	59
JUN-22	2010	55

Fuente: Elaboración propia

Nótese que la distribución y la cantidad de empleados según la variable objetivo cambió un poco en la Tabla 2 con respecto a la Tabla 1, debido a que en la Tabla 2 se marca como se mencionó a aquellos colaboradores que renunciarán o no en los próximos 3 meses. Hay que recalcar que no hay mucho detalle sobre la parte de construcción de los datos para el modelo en esta clase de problemática, una crítica que resalta Qutub (2021) y en cuyo trabajo se repiten registros, pero solo en ciertos periodos ya que se quiso recoger la variación de las independientes en el tiempo. Para nuestro caso se hizo algo similar, sin embargo, se consideró la aparición de cada colaborador en cada periodo hasta antes de renunciar.

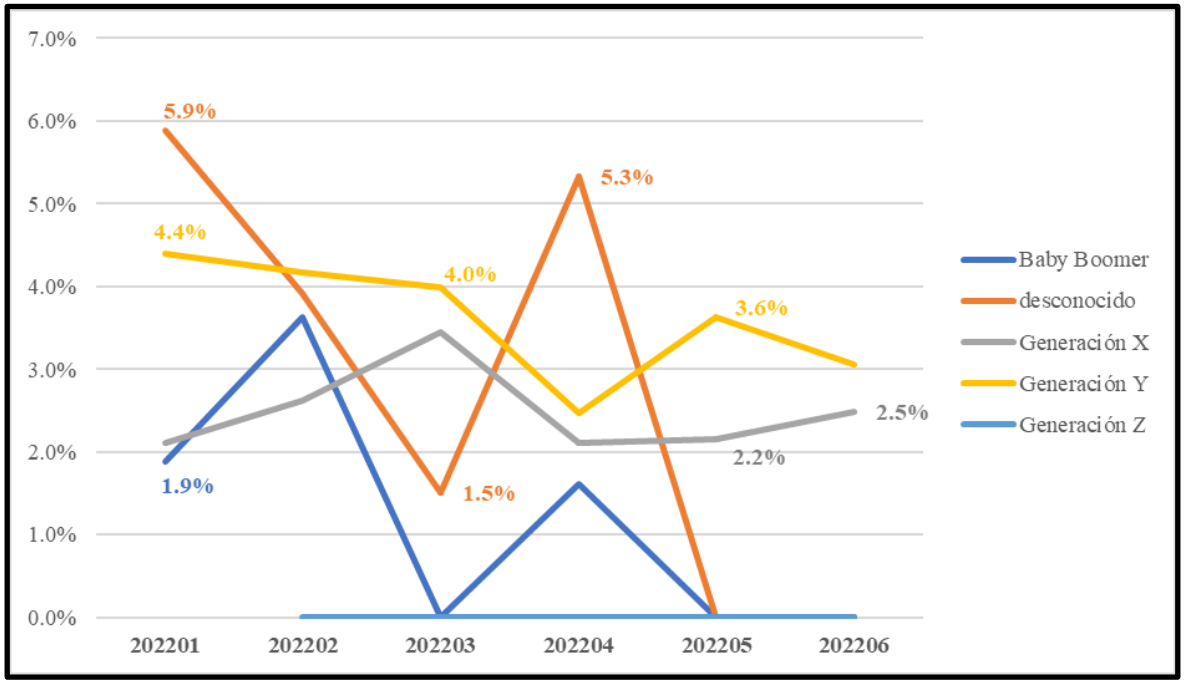
Una vez recolectada la información demográfica de historia en la entidad financiera correspondiente al mes de actividad del colaborador (antes que renuncie), se procedió a crear más variables basadas en las originales, por ejemplo, a partir del número total de ascensos o cambios en el grado salarial se crearon las variables ascensos en los últimos 6 meses y

ascensos en los últimos 12 meses, de la misma forma se construyeron más variables a partir de las financieras, como máximo monto de deuda de préstamo de consumo en los últimos 6 meses y así sucesivamente. Con todo esto se llegaron a construir más de 350 variables.

Se modificaron las variables cualitativas de tal forma que pudieran diferenciarse por el porcentaje de rotación obtenido, como la variable generación del líder que se mostrará a continuación:

Figura 4

Porcentaje de variable objetivo presente categoría de generación de líder por periodo

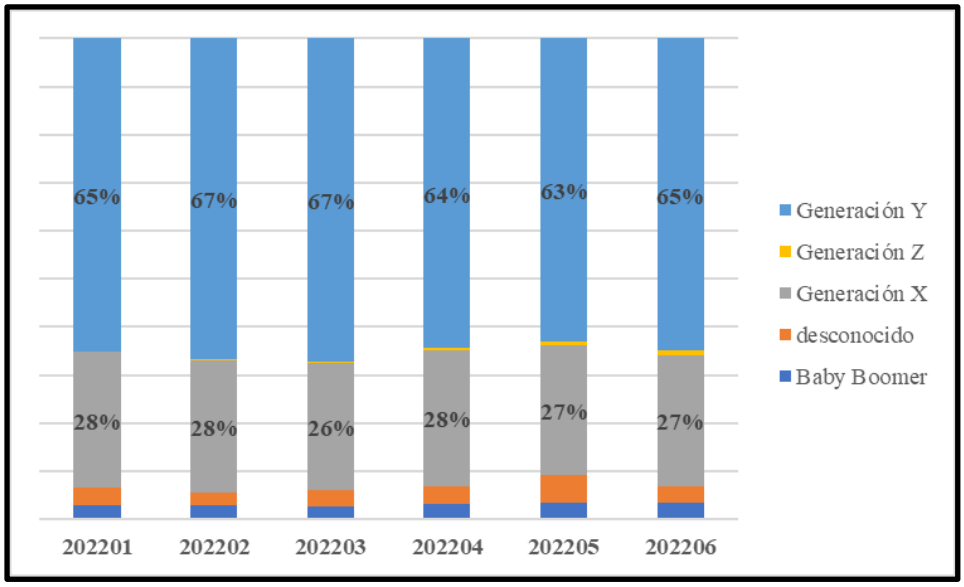


Fuente: Elaboración propia

Como se observa al principio, las categorías de esta variable no muestran un patrón de comportamiento en cuanto al indicador de rotación trimestral, pero también debemos de considerar la proporción por categoría:

Figura 5

Proporción de categoría de generación de líder por periodo

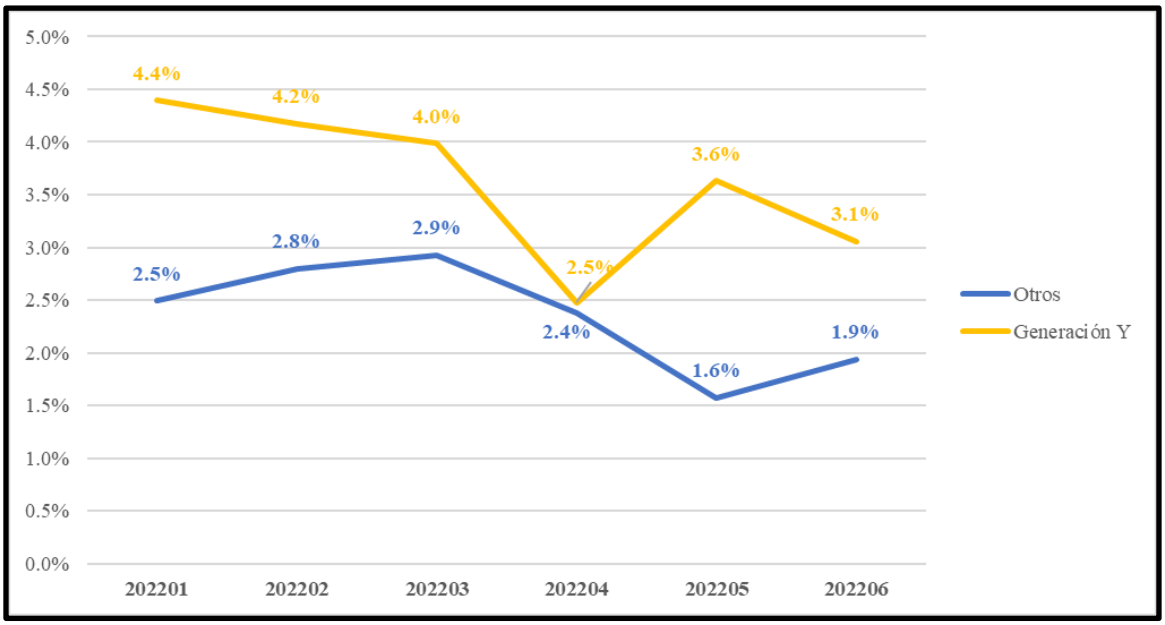


Fuente: Elaboración propia

De acuerdo con la imagen lo que nos convendría seria agrupar la Generación X con el resto de las categorías minoritarias y dejar la Generación Y por separado, al hacer esto revisamos nuevamente el indicador de rotación trimestral y tendremos finalmente:

Figura 6

Porcentaje de variable objetivo presente por categoría de generación de líder agrupado por periodo



Fuente: Elaboración propia

Como se ve ahora ya se observa una diferenciación más clara entre categorías de generación del líder, de esta forma se trabajó con el resto de las variables cualitativas y con algunas cuantitativas que se consideró necesario categorizar. Estas transformaciones se hicieron en el software R Project en su versión 3.6.2 y fue utilizado también en las siguientes dos fases.

3.1.4. Modelamiento con regresión logística

Luego de haber creado las variables y haber transformado algunas de ellas, se procedió a modelar la propensión a la renuncia en los próximos 3 meses mediante la técnica de la regresión logística binaria. Se tomó como base para entrenamiento y prueba los registros pertenecientes a los colaboradores desde enero hasta mayo 2022 y se dividió aleatoriamente en un 70% y 30% respectivamente. Junio finalmente se utilizó como validación.

Una de las dificultades principales al momento de modelar estos casos, es el desbalance en la muestra, por lo cual se aplicó el sobre muestreo en la base de entrenamiento hasta que ésta quedara en porcentajes iguales para las frecuencias del éxito (que renuncie en los próximos 3 meses) y fracaso (que no renuncie en los próximos 3 meses) en nuestra variable objetivo

y así mejorar los resultados. Para la selección de variables se consideró utilizar indicadores como la d de cohen que ayuden a identificar que variables numéricas estaban más relacionadas con la ocurrencia de nuestro evento de interés y con el mismo objetivo, la prueba chi cuadrado para el caso de las variables cualitativas. Cabe mencionar que también se revisó el impacto del a multicolinealidad con el uso de la correlación y el VIF generalizado.

3.1.5. Evaluación del modelo

En esta fase se aplicaron matrices de confusión para medir el poder clasificatorio del modelo, se obtuvieron indicadores como el AUC y el GINI, además de realizarse las pruebas de bondad de ajuste y revisar el impacto de la multicolinealidad. El código en R utilizado en esta etapa y las dos anteriores se podrá observar en el anexo.

3.1.6. Presentación del modelo y distribución de la información

En primer lugar, se presentaron los resultados del modelo a los líderes del área de recursos humanos que solicitaron el pedido, luego al equipo de data analytics de la entidad financiera, quienes aconsejaron y dieron sus sugerencias para afinar la metodología, finalmente se hizo la presentación a algunos de los líderes más importantes de los equipos con perfiles tecnológicos. Con estos últimos se acordó compartir un archivo Excel de forma mensual a partir de noviembre del 2022 con las probabilidades de renuncia de los colaboradores para que comiencen a aplicar medidas de retención a aquellos con probabilidades más altas, de tal manera que no se hicieran sobregastos. Algunas de las iniciativas que se aplicaron fueron mejorar la confianza del líder con el colaborador, ser flexibles con los horarios, dar capacitaciones, mejorar el reconocimiento mediante ascensos y bonos especiales por desempeño, potenciar la cultura y acuerdos del equipo y desplegar una serie de beneficios como acceso a plataformas de desarrollo personal y descuentos en gimnasios.

3.2. Contribución en la solución de situaciones problemáticas

Se contribuyó en reducir el indicador de rotación trimestral a partir de noviembre del 2022 y esto dio visibilidad frente a otras áreas e incluso demostró que un equipo del área de recursos humanos era capaz de realizar esta clase de proyectos utilizando solo técnicas sencillas y con las limitaciones que tiene el equipo de poder acceder a distintas fuentes de información y herramientas computacionales más poderosas. Este trabajo y otros realizados

contribuyeron a que actualmente el área de recursos humanos empiece a cambiar de mentalidad y que los equipos confíen en el análisis de datos.

3.3. Análisis de la contribución en términos de competencias y habilidades

El análisis y resultados de las diversas problemáticas relacionadas con los recursos humanos en la banca y en otros rubros son poco difundidos por contener datos confidenciales que podrían comprometer a la organización, e incluso hay muy poca referencia bibliográfica que tenga enfoque estadístico, lo que genera muchos retos y discusiones aún sin resolver que se pueden encontrar en internet.

El análisis del comportamiento de los colaboradores en una organización hace uso de distintas habilidades analíticas y también de conocimientos informáticos para la construcción y entendimiento de la información. Es a su vez muy variable por los distintos eventos que pueden afectar a una empresa y no solo eso, sino por el hecho de conocer y entender los motivos personales de cada empleado, por lo que es necesario que los investigadores cuenten con capacidades blandas.

3.4. Beneficio obtenido por el centro laboral

Como se comentó al reducir la rotación se redujeron también los costos en convocatorias para los reemplazos de los renunciantes y por ende reducción de tiempos de aprendizaje y dinero por capacitación. Se acertó por parte de las áreas de recursos humanos encargadas, en brindar iniciativas de salario emocional que tampoco generen gastos excesivos de acuerdo con lo mencionado anteriormente.

IV. RESULTADOS Y DISCUSIÓN

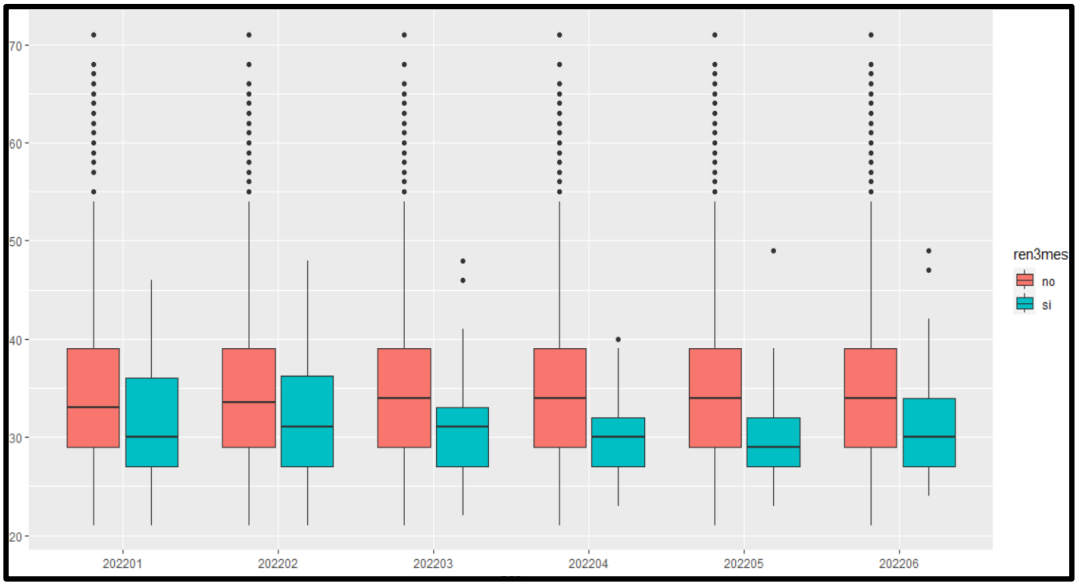
4.1. Variables en el modelo y significancia

Las variables finales incluidas en el modelo se describirán a continuación:

- **Edad:** La distribución de esta variable nos indica que los colaboradores más jóvenes son los que tienen mayor propensión a la renuncia en los siguientes tres meses. La edad no se categorizó quedando como una cuantitativa.

Figura 7

Distribución de la edad por variable objetivo por y por periodo

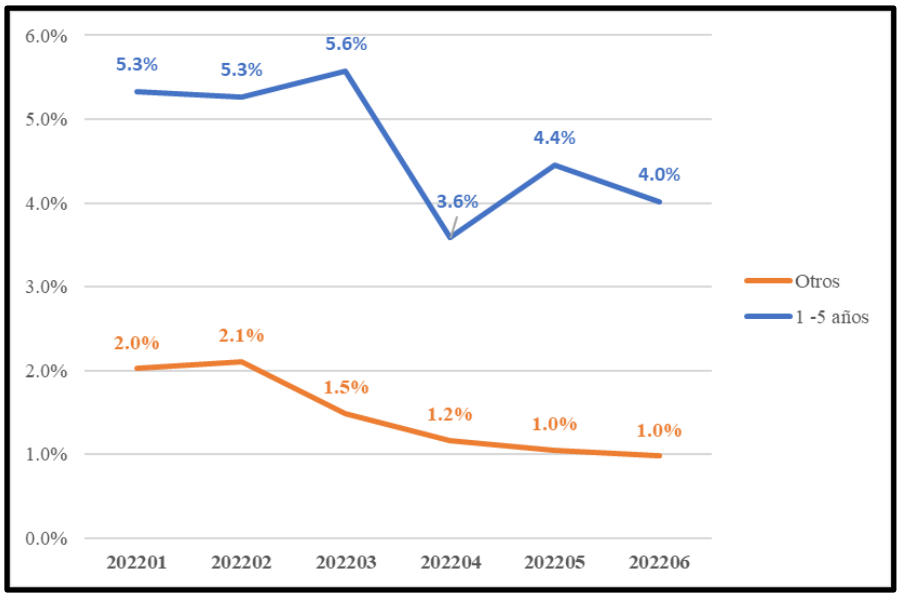


Fuente: Elaboración propia

- **Tiempo trabajando en la entidad:** Esta variable originalmente cuantitativa fue categorizada, en ella se muestra que el porcentaje más alto de la variable objetivo se encuentra a partir del año 1 hasta el 5, quedando el resto como otra agrupación.

Figura 8

Porcentaje de variable objetivo presente categoría de tiempo en la entidad y por periodo

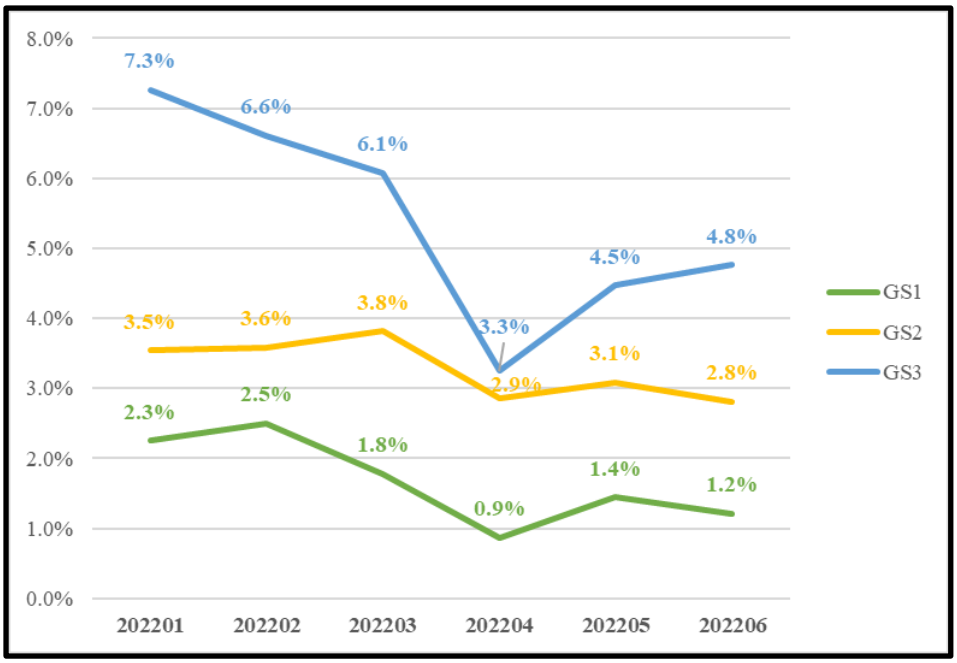


Fuente: Elaboración propia

- **Grado salarial:** Originalmente teniendo once categorías, las cuales fueron agrupadas según la proporción contenida de colaboradores que renunciaron en los siguientes tres meses. El grupo con más proporción (“GS3”) está conformado en su mayoría por los grados salariales más bajos, seguido de los grados medios (“GS2”) y finalmente los grupos más altos (“GS1”), con menor porcentaje de éxito en la variable objetivo.

Figura 9

Porcentaje de variable objetivo presente categoría de grado salarial y por periodo

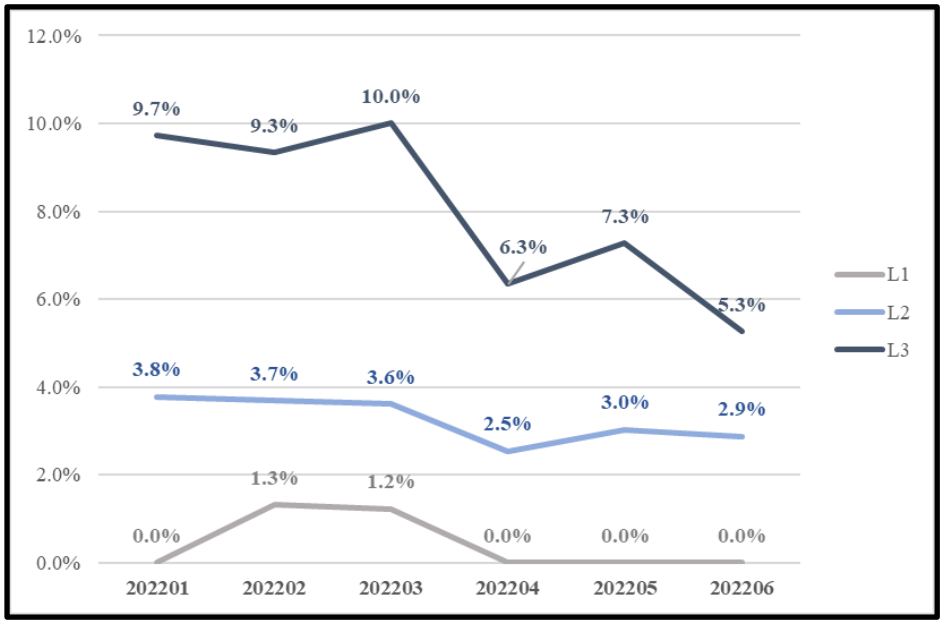


Fuente: Elaboración propia

- **Tipo de líder:** Muestra si el colaborador analizado es un líder o no y su tipo, en principio teniendo ocho categorías se agrupó en 3, nombrándolas según el aumento de la proporción contenida de la variable respuesta como “L1”, “L2” y “L3”.

Figura 10

Porcentaje de variable objetivo presente categoría de líder y por periodo

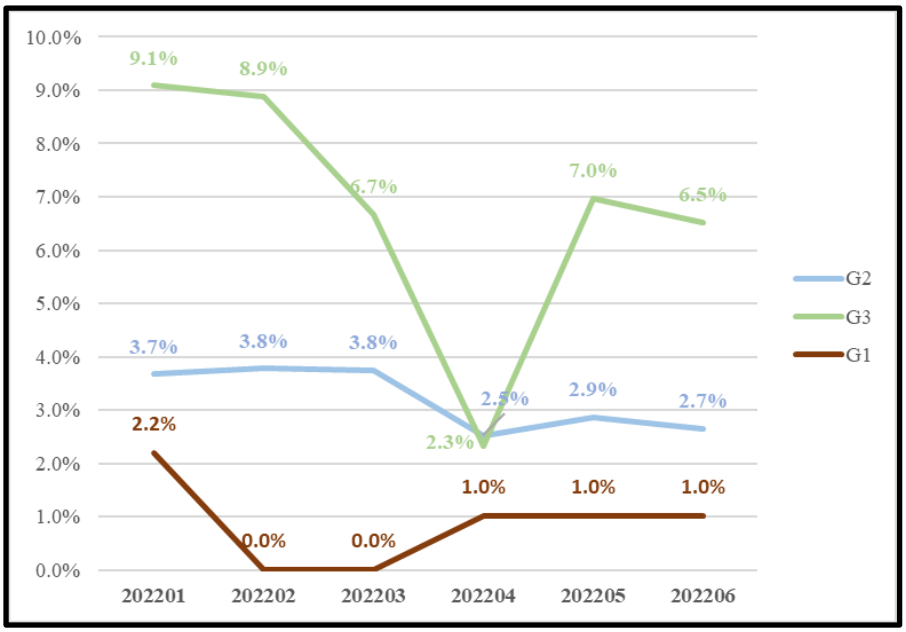


Fuente: Elaboración propia

- **Gerencia:** Nos da la información sobre la gerencia a la que pertenece cada colaborador. Para este caso existen diez gerencias las cuales también fueron divididas en tres grupos, el primero de ellos con menos proporción de éxito en la variable respuesta contiene gerencias como las de recursos humanos (“G1”), la siguiente área de informática y analítica (“G2”) y por último de áreas relacionadas con los negocios (“G3”) con más porcentaje de colaboradores que renunciaron en los próximos tres meses.

Figura 11

Porcentaje de variable objetivo presente categoría de gerencia y por periodo



Fuente: Elaboración propia

El detalle de la significancia de las variables mencionadas se presenta a continuación:

Tabla 3

Coeficientes estimados del modelo y significancia

VARIABLE	COEFICIENTE	Z VALUE	P VALOR
EDAD	-0.070	-19.660	0.000
TIEMPO EN LA ENTIDAD (1-5 AÑOS)	0.743	17.561	0.000
TIPO DE LÍDER (L3)	2.377	15.741	0.000
TIPO DE LÍDER (L2)	1.291	10.033	0.000
GERENCIA (G3)	1.478	8.694	0.000
GERENCIA (G2)	1.058	8.021	0.000
TIPO GS (GS3)	0.587	7.723	0.000
TIPO GS (GS2)	0.455	7.677	0.000

Fuente: Elaboración propia

Se visualiza en la Tabla 3 que todas las variables resultaron significativas, ahora de acuerdo con los gráficos y resultados mostrados se observa una tendencia a que los colaboradores más jóvenes renuncien, posiblemente porque aún no buscan establecerse en una sola organización y quieren continuar explorando otras ramas y especialidades o tal vez no poseen mayores responsabilidades a diferencia de los más maduros, esto también se puede observar si transformamos la edad en generación, en la cual se verá que la generación Z es

la que tiene más porcentaje de empleados que renunciarán en los próximos tres meses, seguidos por los Y, los X y Baby Boomers con una proporción bastante baja. En el caso del tiempo en la entidad al parecer el colaborador con perfil tecnológico se siente optimista de pertenecer a la organización, pero al cumplir un año hasta los cinco esto cambia, luego a partir de los 6 ya comienza a acostumbrarse, o a resignarse al ritmo de sus labores. Para el tipo de líder curiosamente, los que contienen mayor proporción de renunciantes son aquellos que si son considerados como líderes de equipo seguidos por los que no ostentan ningún puesto de liderazgo y el menos propenso contiene en su mayoría a Gerentes. Con respecto de la gerencia, como se comentó los menos propensos están relacionados con gerencias de recursos humanos, cuyos equipos posiblemente manejen mejor la cultura por estar en dicho grupo, seguidos por los de analítica o tecnologías y finalmente aquellos con trabajos que se enfocan en el negocio de la banca y que poseen más rotación, pudiendo ser el efecto de la presión para llegar a las metas. Por último, en cuanto a los grados salariales, los más altos son los que menos proporción de renuncias tienen, tal vez por la estabilidad económica que les brinda el sueldo.

4.2. Comprobación de la bondad de ajuste del modelo y efecto de la multicolinealidad

- Para probar la bondad de ajuste se construyó un modelo sin variables dependientes y se comparó con el modelo final.

  H_0: $\beta_i = \beta_{i+1} = \ldots \beta_k = 0$

  H_1: Al menos uno de los $\beta_i \neq 0$, donde $i=1$ y $k=8$

Tabla 4

Significancia global de los coeficientes

PRUEBA	VALOR	P VALOR
RAZÓN DE VEROSIMILITUD	2205.3	0.000

Fuente: Elaboración propia

El resultado fue significativo por lo que podemos decir al menos uno de los coeficientes es distinto de 0.

- Para e caso del efecto de la multicolinealidad usamos calculamos el vif generalizado para cada variable obteniendo:

Tabla 5

Valores de GVIF para las variables en el modelo final

VARIABLE	GVIF
Edad	1.123
Tiempo en la entidad	1.048
Tipo de líder	1.019
Gerencia	1.005
Tipo gs	1.056

Fuente: Elaboración propia

El resultado nos muestra valores muy cercanos a uno por lo cual podemos decir que no hay un impacto considerable de la multicolinealidad y además los signos de los coeficientes estimados coinciden con los gráficos que muestran las relaciones entre la variable dependiente y sus regresores.

4.3. Eficacia predictiva del modelo

- Para medir la eficacia del modelo mostramos los indicadores de clasificación global, sensibilidad, especificidad, AUC y GINI.

Tabla 6

Indicadores de clasificación y discriminación para el modelo final

Base	Clasificación correcta	Sensibilidad	Especificidad	AUC	GINI
Entrenamiento	69%	78%	60%	0.69	0.39
Prueba	63%	75%	62%	0.69	0.37
Validación	60%	75%	60%	0.67	0.35

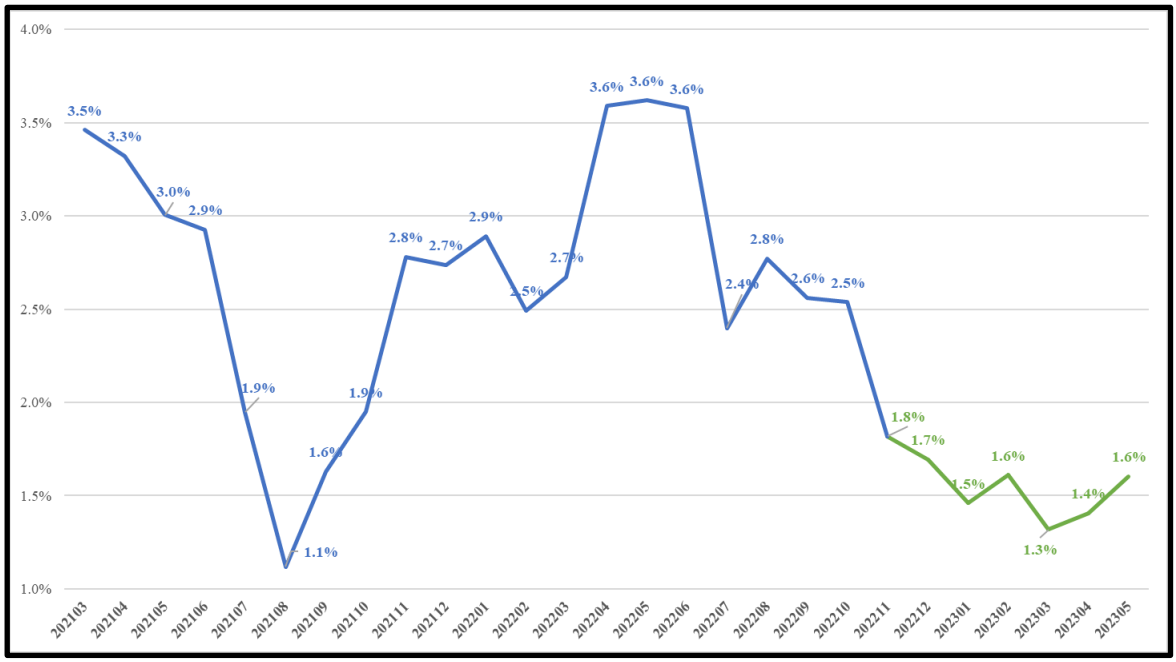
Fuente: Elaboración propia

De acuerdo con el criterio mostrado por El Khal (2021) sobre la relación entre los valores del AUC y GINI, se podría considerar que el modelo cuenta con un poder predictivo aceptable. Hay que comentar que un después del lanzamiento y despliegue se volvió a comprobar el mismo modelo con otro periodo fuera de tiempo (octubre 2022), obteniendo una clasificación correcta o exactitud del 60%, sensibilidad del 72%, especificidad de 60%, AUC de 0.66 y GINI de 0.32.

A continuación, se mostrarán los resultados del indicador de rotación trimestral voluntaria que fueron producto de las iniciativas de los líderes de los perfiles tecnológicos apoyándose en los resultados de este modelo.

Figura 12

Indicador de rotación trimestral de perfiles tecnológicos en la entidad bancaria por periodo



Fuente: Elaboración propia

Como se observa el indicador de rotación trimestral voluntaria disminuye a partir de la fecha de implementación (línea verde) hasta llegar a un 1.6% de 2684 colaboradores para el último mes de mayo del 2023.

V. CONCLUSIONES

1. El evento de la renuncia de un perfil tecnológico en los próximos 3 meses (mes +1, +2, +3) desde su observación (Mes 0) queda explicado por el modelo de regresión logística binaria y este cumple con tener mejor ajuste que un modelo reducido de acuerdo con la prueba de razón de verosimilitudes y a su vez los predictores en el modelo final son significativos.
2. El modelo presentó un valor de AUC de 0.67 y un GINI de 0.35 lo que nos dice que el modelo es aceptable y es capaz de proporcionar una buena discriminación al momento de clasificar a los colaboradores con perfil tecnológico que renunciarán o no en los siguientes 3 meses.
3. De acuerdo con las variables finales en el modelo: A menor edad mayor probabilidad de renuncia, si el tiempo en la organización se encuentra entre 1 y 5 años se es más propenso a renunciar, cuando se es líder de equipos muy pequeños es más probable la renuncia, los colaboradores más propensos a la renuncia pertenecen en su mayoría a la gerencia de negocios y finalmente se cumple que a menor grado salarial mayor propensión a la renuncia.
4. Los resultados de este modelo sirvieron como herramienta para los líderes de las áreas de perfil tecnológico, para que puedan aplicar iniciativas de retención que redujeron el indicador de rotación trimestral voluntaria y cuyo valor llega a acercarse a la meta de 1.5%.

VI. RECOMENDACIONES

1. Para poder observar si se pueden mejorar la predicción se recomienda utilizar otras técnicas de modelamiento, por ejemplo, random forest o xgboost.
2. Analizar si se puede mejorar la predicción ampliando la ventana de tiempo a 6 meses o más.
3. Probar otras formas de diseño de muestra para ver si es que también esto puede provocar una mejora en las predicciones.
4. Agregar más variables como, por ejemplo, aquellas referidas a encuestas de experiencia, clima y liderazgo, que podrían ser muy útiles y precisas en cuanto a las razones de la renuncia del colaborador.

VII. REFERENCIAS BIBLIOGRÁFICAS

Abdel-Rahmen, K., Kheddouci, H., J West, D. (2021). *Predicting employee attrition with a more effective use of historical events*. https://www.esann.org/sites/default/files/proceedings/2021/ES2021-110.pdf

Amat, J. (noviembre de 2020). *Regresión logística con Python*. https://cienciadedatos.net/documentos/py17-regresion-logistica-python

Apuy, E. (setiembre de 2020). *Prospección del mercado de TI en Perú*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://sistemas.procomer.go.cr/DocsSEM/B882B8FA-3A4E-4BB2-BAE8-285FFFDFD807.pdf

Buitrago, B. (12 de agosto de 2020). *Modelo de Regresión Logística*. https://medium.com/@csarchiquerodriguez/modelo-de-regresi%C3%B3n-log%C3%ADstica-a17ffe204ade

Barcelona Digital Talent. (2022). *Digital Talent Overview 2022*. https://barcelonadigitaltalent.com/en/report/digital-talent-overview-2022/

Chávez, C. (12 de octubre de 2021). *¿Qué perfiles tecnológicos escasean en Perú y cuánto pueden llegar a ganar*?. Forbes Perú. https://forbes.pe/capital-humano/2021-10-12/que-perfiles-tecnologicos-escasean-en-peru-y-cuanto-pueden-llegar-a-ganar

Chique, C. (17 de setiembre de 2020). *Regresión Logística I — Machine Learning*. https://medium.com/iwannabedatadriven/regresi%C3%B3n-log%C3%ADstica-i-machine-learning-84ffe9d6be15

Cortés, J. (s.f.). *Temas sobre la regulación de la renuncia en el Perú*. https://www.spdtss.org.pe/wp-content/uploads/2021/11/Luis-Aparicio-Homenaje-full-001-424-331-359.pdf

Cuadrado, C. (3 de marzo de 2021). *Los 8 perfiles tecnológicos más demandados en las empresas en 2021*. Armadillo Amarillo. https://www.armadilloamarillo.com/blog/los-8-perfiles-tecnologicos-mas-demandados-en-las-empresas-en-2021/#perfilestecnologicosdemandados

El Khal, Y. (18 de marzo de 2021). *Confusion matrix, AUC and ROC curve and Gini clearly explained.* https://yassineelkhal.medium.com/confusion-matrix-auc-and-roc-curve-and-gini-clearly-explained-221788618eb2

Fox, J., Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183

https://doi.org/10.2307/2290467

González, L. (12 de agosto de 2023). *Regresión Logística – Teoría*. https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/

H2o.ai. (28 de noviembre de 2018). *Performance and prediction*. https://h2o-release.s3.amazonaws.com/h2o/rel-xia/2/docs-website/h2o-docs/performance-and-prediction.html

Henao Ríos, C. (2021). *Modelo de Medición de la Rotación de Personal como Variable de Decisión Estratégica*. [Trabajo final presentado como requisito parcial para optar al título de: Magister en Ingeniería-Analítica, Universidad De Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/81925/1033646947.2021.pdf?sequence=3&isAllowed=y

Hernández, F. (13 de julio de 2020). *Prueba razón de verosimilitudes en GLM's.* https://rpubs.com/fhernanb/lrt_glm

Interim Group. (18 de noviembre de 2022). *Índice de rotación de personal: qué es y cómo calcularlo.* https://interimgrouphr.com/blog/indice-rotacion-personal/

KDnuggets. (2018). What main methodology are you using for your analytics, data mining, or data science projects? Poll https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Lantz, B. (2015). Evaluating model performance. En Packt Publishing (Ed.), Birmingham, Reino Unido. Machine Learning with R

López-Roldán, P y Fachelli, S. (2015). Análisis de Regresión Logística. En Creative Commons (Ed.), España. Metodología de la investigación social cuantitativa. https://ddd.uab.cat/pub/caplli/2016/163564/metinvsoccua_a2016_cap1-2.pdf

Maggi, F. (10 de julio de 2022). *Clases desbalanceadas y su tratamiento en R.* https://www.linkedin.com/pulse/clases-desbalanceadas-y-su-tratamiento-en-r-felipe-maggi/?originalSubdomain=es

Ostrowski, V. (23 de mayo 2022). *Cuáles son los perfiles tech que te ayudarán a incrementar tu negocio*. https://www.multiplicatalent.com/blog/transformacion-digital/perfiles-tech-para-potenciar-negocio/

Pacori, J. (28 de febrero de 2021). *La renuncia como acto jurídico laboral [modelos de escritos]*. Pasión por el derecho. https://medium.com/@csarchiquerodriguez/modelo-de-regresi%C3%B3n-log%C3%ADstica-a17ffe204ade

Pavansubhash. (2017). *IBM HR Analytics Employee Attrition & Performance. Kaggle.* https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Perú21. (26 de julio de 2023). *Desafíos en el sector TIC: Rotación laboral aumenta ante la creciente demanda de perfiles especializados.* https://peru21.pe/peru/desafios-en-el-sector-tic-rotacion-laboral-aumenta-ante-la-creciente-demanda-de-perfiles-especializados-noticia/

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R y Alghamdi, H. (2021). Prediction of Employee Attrition Using Machine Learning. International Journal of Machine Learning and Computing. 11(2). https://www.researchgate.net/publication/351911311_Prediction_of_Employee_Attrition_Using_Machine_Learning_and_Ensemble_Methods

Recuero, P. (23 de enero de 2018). *Machine Learning a tu alcance: La matriz de confusión.* Telefónica. https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

Rodríguez, D. (23 de julio de 2018). *La regresión logística.* Analytics Lne. https://www.analyticslane.com/2018/07/23/la-regresion-logistica/

Rodríguez, M., Álvarez, J., Mesa, J. y González, A. (8 de octubre de 2003). *Metodologías para el Desarrollo de Proyectos en Minería de Datos* https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Sánchez, P. (2 de diciembre de 2021). *Por qué los perfiles tecnológicos especializados son clave en la digitalización*. Visionarios. https://blogvisionarios.com/impulsa-tu-negocio/perfiles-tecnologicos-especializados-son-clave-digitalizacion/

Saini, A. (3 de agosto de 2021). *Conceptual Understanding of Logistic Regression for Data Science Beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

Universidad de Guanajuato. (29 de marzo de 2022). *Clase digital 10. Correlación de Pearson y de Spearman*. Recursos educativos abiertos. Universidad de Guanajuato. https://blogs.ugto.mx/rea/clase-digital-10-correlacion-de-pearson-y-de-spearman/

VIII. ANEXOS

Anexo 1 Código R utilizado

```
library(RODBC)
library(nnet)
library(psych)
library(caret)
library(ROSE)
library(pROC)
library(plyr)
library(dplyr)
library(ggplot2)
library(corrplot)
library(ggpubr)
library(readxl)
library(scales)
library(MASS)
library(DescTools)
library(rstatix)
library(ltm)
library(rpart)
library(rpart.plot)
library(effsize)
library(glmtoolbox)
##conexión R con sql
cn <- odbcDriverConnect(connection="Driver={SQL Server Native Client 11.0}
;server=DESKTOP-DKN5U2T\\SQLEXPRESS;database=rota")
#carga de bases para entrenamiento, prueba y validación
data<-sqlQuery(cn, "select * from dato where FECHA>202112 and
FECHA<202206",stringsAsFactors=TRUE)
val1<-sqlQuery(cn, "select * from dato where FECHA=202206",stringsAsFactors=TRUE)
#agrupaciónde categorias
levels(data$GS)[levels(data$GS)=="GRADS0"]<-" GS1"
```

```
levels(data$GS)[levels(data$GS)=="GRADS4"]<-" GS1"
levels(data$GS)[levels(data$GS)=="GRADS9"]<-" GS1"
levels(data$GS)[levels(data$GS)=="GRADS10"]<-" GS1"
levels(data$GS)[levels(data$GS)=="GRADS1"]<-" GS1"
levels(data$GS)[levels(data$GS)=="GRADS3"]<-"GS1"
levels(data$GS)[levels(data$GS)=="GRADS5"]<-"GS2"
levels(data$GS)[levels(data$GS)=="GRADS6"]<-"GS2"
levels(data$GS)[levels(data$GS)=="GRADS2"]<-"GS2"
levels(data$GS)[levels(data$GS)=="GRADS7"]<-"GS3"
levels(data$GS)[levels(data$GS)=="GRADS8"]<-"GS3"
levels(data$Gerencia)[levels(data$Gerencia)=="GR1"]<-"otros"
levels(data$Gerencia)[levels(data$Gerencia)=="GR2"]<-"otros"
levels(data$Gerencia)[levels(data$Gerencia)=="GR3"]<-"otros"
levels(data$Gerencia)[levels(data$Gerencia)=="GR4"]<-"G1"
levels(data$Gerencia)[levels(data$Gerencia)=="GR5"]<-"G1"
levels(data$Gerencia)[levels(data$Gerencia)=="GR6"]<-"G1"
levels(data$Gerencia)[levels(data$Gerencia)=="GR7"]<-"G1"
levels(data$Gerencia)[levels(data$Gerencia)=="GR7"]<-"G1"
levels(data$Gerencia)[levels(data$Gerencia)=="GR8"]<-"G2"
levels(data$Gerencia)[levels(data$Gerencia)=="GR9"]<-"G2"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="TIPL1"]<-" L1"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="TIPL2"]<-" L1"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="TIPL3"]<-" L1"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="TIPL4"]<-"L1"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="TIPL5"]<-"L1"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=” TIPL6"]<-"L2"
levels(data$Tipo_lid)[levels(data$Tipo_lid)=="No lider"]<-"L2"
levels(data$Tipo_lid)[levels(data$Tipo_lid)==" TIPL0"]<-"L3"
levels(val1$GS)[levels(val1$GS)=="GRADS0"]<-" GS1"
levels(val1$GS)[levels(val1$GS)=="GRADS4"]<-" GS1"
levels(val1$GS)[levels(val1$GS)=="GRADS9"]<-" GS1"
levels(val1$GS)[levels(val1$GS)=="GRADS10"]<-" GS1"
levels(val1$GS)[levels(val1$GS)=="GRADS1"]<-" GS1"
levels(val1$GS)[levels(val1$GS)=="GRADS3"]<-"GS1"
```

```
levels(val1$GS)[levels(val1$GS)=="GRADS5"]<-"GS2"
levels(val1$GS)[levels(val1$GS)=="GRADS6"]<-"GS2"
levels(val1$GS)[levels(val1$GS)=="GRADS2"]<-"GS2"
levels(val1$GS)[levels(val1$GS)=="GRADS7"]<-"GS3"
levels(val1$GS)[levels(val1$GS)=="GRADS8"]<-"GS3"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR1"]<-"otros"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR2"]<-"otros"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR3"]<-"otros"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR4"]<-"G1"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR5"]<-"G1"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR6"]<-"G1"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR7"]<-"G1"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR7"]<-"G1"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR8"]<-"G2"
levels(val1$Gerencia)[levels(val1$Gerencia)=="GR9"]<-"G2"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="TIPL1"]<-" L1"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="TIPL2"]<-" L1"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="TIPL3"]<-" L1"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="TIPL4"]<-"L1"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="TIPL5"]<-"L1"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)==" TIPL6"]<-"L2"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)=="No lider"]<-"L2"
levels(val1$Tipo_lid)[levels(val1$Tipo_lid)==" TIPL0"]<-"L3"
##la variable de tiempo en la entidad ya estaba categorizada de acuerdo a lo indicado
##partición de muestras de entrenamiento y prueba
tammuestra <- floor(0.70 * nrow(data))
set.seed(38)
entrind<- sample(seq_len(nrow(data)), size = tammuestra)
train <- data[entrind, ]
test <- data[-entrind, ]
table(train$ren3mes)
tablatrain<-table(train$ren3mes)
nmuestra<-tablatrain[1]*2
##balanceo de datos
```

```
set.seed(45)
dataover <- ovun.sample(ren3mes~.,data =train,method = "over",nmuestra)$data
##establecimiento de categoría de referencia
dataover$Gerencia<-relevel(dataover$Gerencia,ref="otros")
data$ren3mes<-relevel(data$ren3mes,ref="no")
train$ren3mes<-relevel(train$ren3mes,ref="no")
test$ren3mes<-relevel(test$ren3mes,ref="no")
dataover$ren3mes<-relevel(dataover$ren3mes,ref="no")
val1$ren3mes<-relevel(val1$ren3mes,ref="no")
#construcción de modelo
modelo<- glm(ren3mes~tiempo_banco_anios_cat+
GS+Gerencia+Tipo_lid+Edad,data=dataover,family = "binomial")
##significancia de variables
summary(modelo)
##construcción table cruzada e indicadores AUC y GINI
a<-predict(modelo,data=dataover,type="response")
pred <- ifelse(a>0.5,"si","no")
MC<-table(pred,dataover$ren3mes)
clasificadoscorrec<-(MC[1]+MC[4])/(MC[1]+MC[2]+MC[3]+MC[4])*100
especificidad<-(MC[1]/(MC[1]+MC[2]))*100
sensibilidad<-(MC[4]/(MC[3]+MC[4]))*100
preda<-as.numeric(revalue(pred, c("no"= 1,"si"= 2 )))
obsea<-as.numeric(revalue(dataover$ren3mes, c("no"= 1,"si"= 2)))
auca<-auc(obsea,preda)
ginia<-(2*auc(obsea,preda))-1
cbind(clasificadoscorrec,especificidad,sensibilidad,auca,ginia)
b<-predict(modelo,newdata=test,type="response")
predb <- ifelse(b>0.5,"si","no")
pru<-table(predb,test$ren3mes)
clasificadospru<-(pru[1]+pru[4])/(pru[1]+pru[2]+pru[3]+pru[4])*100
especificidadpru<-(pru[1]/(pru[1]+pru[2]))*100
sensibilidadpru<-(pru[4]/(pru[3]+pru[4]))*100
predb<-as.numeric(revalue(predb, c("no"= 1,"si"= 2)))
obseb<-as.numeric(revalue(test$ren3mes, c("no"= 1,"si"= 2)))
```

```
aucb<-auc(obseb,predb)
ginib<-(2*auc(obseb,predb))-1
cbind(clasificadospru,especificidadpru,sensibilidadpru,aucb,ginib)
c<-predict(modelo,newdata=val1,type="response")
predc <- ifelse(c>0.5,"si","no")
va<-table(predc,val1$ren3mes)
clasificadosva<-(va[1]+va[4])/(va[1]+va[2]+va[3]+va[4])*100
especificidadva<-(va[1]/(va[1]+va[2]))*100
sensibilidadva<-(va[4]/(va[3]+va[4]))*100
predc<-as.numeric(revalue(predc, c("no"= 1,"si"= 2)))
obsec<-as.numeric(revalue(val1$ren3mes, c("no"= 1,"si"= 2)))
aucc<-auc(obsec,predc)
ginic<-(2*auc(obsec,predc))-1
cbind(clasificadosva,especificidadva,sensibilidadva,aucc,ginic)
##Bondad de ajuste
modelo0<- glm(ren3mes~ 1 ,data=dataover,family = "binomial")
anova(modelo0, modelo, test="Chisq", dispersion=1)
##GVIF
gvif(modelo)
```