UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

PREDICCIÓN DE LA ROTACIÓN DE PERSONAL EN LA EMPRESA SITEL COSTA RICA

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Matemática para optar al grado y título de Maestría Profesional en Métodos Matemáticos y Aplicaciones

SERGIO RAMÍREZ RODRÍGUEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica

2025

Dedicatoria

A Dios, por brindarme salud y muchas bendiciones a lo largo de toda mi vida, permitiéndome llegar hasta este día. A Tati, mi mayor soporte y fuente de inspiración para superar las distintas pruebas que he enfrentado a nivel académico, profesional y personal. A mis padres, por siempre brindarme su apoyo incondicional en cada meta que me he propuesto. A Mía y Romina, por acompañarme durante el proceso de esta investigación y la maestría. Finalmente, a mí mismo, por nunca rendirme en la búsqueda de mis sueños hasta cumplirlos.

Agradecimientos

A todo el equipo asesor: los profesores Álvaro Guevara, Maikol Solís y el director de Recursos Humanos, Jonathan Benavides, por sus valiosas recomendaciones que contribuyeron a que este proyecto obtuviera mención honorífica. Al profesor Pedro Méndez, por motivarme a ingresar a la maestría y guiarme durante mi paso por Ciencias Actuariales.

Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Matemática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Métodos Matemáticos y Aplicaciones.

JOSE ALEXANDER RAMIREZ GONZALEZ (FIRMA)
Digitally signed by JOSE ALEXANDER RAMIREZ GONZALEZ (FIRMA)
Date: 2025.03.17 15:50:06 -06'00'

Dr. Alexander Ramírez González

Representante de la Decanatura

Sistema de Estudios de Posgrado

ALVARO JOSE GUEVARA VILLALOBOS (FIRMA)
Firmado digitalmente por ALVARO JOSE GUEVARA VILLALOBOS (FIRMA)
Fecha: 2025.03.18 10:39:32 -06'00'

Dr. Álvaro Guevara Villalobos

Profesor Guía

MAIKOL SOLIS CHACON (FIRMA)
PERSONA FISICA, CPF-01-1206-0753.
Fecha declarada: 20/03/2025 03:28:22 PM
Esta es una representación gráfica únicamente, verifique la validez de la firma.

Dr. Maikol Solís Chacón

Lector

JONATHAN BENAVIDES JIMENEZ (FIRMA)
Firmado digitalmente por JONATHAN BENAVIDES JIMENEZ (FIRMA)
Fecha: 2025.03.25 11:41:25 -06'00'

M.B.A. Jonathan Benavides Jiménez

Lector

DARIO ALBERTO MENA ARIAS (FIRMA)
Firmado digitalmente por DARIO ALBERTO MENA ARIAS (FIRMA)
Fecha: 2025.03.17 10:34:05 -06'00'

Dr. Darío Alberto Mena Arias

Director del Programa de Posgrado en Matemática

SERGIO VINICIO RAMIREZ RODRIGUEZ (FIRMA)
Firmado digitalmente por SERGIO VINICIO RAMIREZ RODRIGUEZ (FIRMA)
Fecha: 2025.03.26 12:00:31 -06'00'

Sergio Ramírez Rodríguez

Sustentante

Índice

Resumen

El entorno empresarial actual, caracterizado por una competencia intensa en mercados globalizados y la necesidad imperativa de innovación para garantizar una ventaja competitiva, ha resaltado la importancia del análisis de datos como una herramienta fundamental para optimizar recursos, aumentar las ganancias y mejorar los indicadores de desempeño en las empresas. Un indicador crucial para cualquier empresa es la minimización de la rotación de empleados, especialmente en sectores como el de Business Process Outsourcing (BPO).

Una alta rotación conlleva un aumento significativo en los costos de reclutamiento, el tiempo requerido para encontrar reemplazos, así como los gastos asociados con la capacitación de nuevos empleados y la pérdida del conocimiento acumulado por el personal que se marcha, entre otros efectos negativos. Por lo tanto, es esencial que las empresas mitiguen este riesgo mediante el uso de análisis de datos y metodologías de *machine learning* para desarrollar modelos predictivos que identifiquen la probabilidad de rotación de empleados dentro de un período específico.

Se evaluaron diversos modelos, incluidos *XGBoost*, Bosques Aleatorios, Regresión Logística y Consensus, comparando métricas como el área bajo la curva ROC, la precisión global, la sensibilidad y la especificidad. *XGBoost* demostró ser superior debido a su alta capacidad predictiva, aprovechando un enfoque de ensamblado con árboles de decisión, técnicas de regularización para prevenir el sobreajuste, optimización de hiperparámetros para una configuración óptima y escalabilidad para manejar conjuntos de datos grandes y complejos. Además, se evaluó la sensibilidad del modelo mediante pruebas de estrés tanto en las observaciones como en las variables predictoras. Desde su implementación, el modelo ha generado ahorros de millones de dólares en los aspectos mencionados.

Palabras Clave: Rotación de personal, *Machine Learning*, *XGBoost*, Bosques Aleatorios, Regresión Logística, Consensus.

Abstract

The current business environment, characterized by intense competition in globalized markets and the imperative need for innovation to ensure a competitive advantage, has underscored the importance of data analytics as a fundamental tool for optimizing resources, increasing profits, and enhancing performance indicators in companies. One crucial indicator for any company is the minimization of employee attrition, especially in sectors like Business Process Outsourcing (BPOs).

High attrition leads to significant increases in recruitment costs, time required to find replacements, as well as expenses associated with training new employees and the loss of accumulated knowledge by departing staff, among other negative effects. Therefore, it is essential for companies to mitigate this risk by using data analytics and machine learning tools to develop predictive models to identify the likelihood of employee attrition within a specific period.

Various models, including XGBoost, Random Forest, Logistic Regression, and Consensus, were evaluated, comparing metrics such as area under the ROC curve, overall accuracy, sensitivity, and specificity. XGBoost emerged as superior due to its adept predictive capacity, leveraging an ensemble approach with decision trees, regularization techniques to forestall overfitting, hyperparameter optimization for optimal configuration, and scalability for handling large and complex datasets. Additionally, the model's sensitivity was assessed through stress tests on both observations and predictor variables. Since implementation, the model has yielded millions of dollars in the aforementioned savings.

Keywords: Employee Attrition, Machine Learning, XGBoost, Random Forest, Logistic Regression, Consensus.

Índice de cuadros

Índice de figuras

1. Introducción

El entorno competitivo que se establece en la actualidad, con mercados globalizados y necesidades de innovación que permitan obtener una ventaja competitiva frente a los demás integrantes del mercado, ha generado una necesidad inminente en el uso de la analítica de datos para maximizar los recursos y obtener mayores ganancias.

Uno de los objetivos claves de una compañía es minimizar su rotación de personal, ya que una métrica alta de este tipo implica que se pierdan ingresos, al no tener el suficiente personal para satisfacer la demanda operacional de la empresa. Esto hace incurrir a las empresas en mayores costos de tiempos extras para pagar a los colaboradores que están laborando y de esta forma poder cubrir al máximo posible el volumen de trabajo que se recibe todos los días.

También se incrementan los costos de reclutamiento, al tener que buscar nuevas personas para reemplazar las bajas, se aumentan los costos de entrenamiento de las personas para que puedan operar de forma exitosa con los clientes, pérdidas intangibles de fuga de conocimiento al salir personal ya entrenado y con conocimiento adquirido, entre otros efectos negativos.

Por lo tanto, es crucial que las empresas puedan mitigar el riesgo de rotación de personal, con el fin de maximizar su rentabilidad y mejorar su competitividad con respecto a la industria. La innovación en el uso de análisis de datos y herramientas de *machine learning* permiten mitigar este riesgo mediante la creación de modelos estadísticos, como los modelos predictivos, que en este caso permitan definir la probabilidad de salida de los empleados en un periodo determinado. De esta manera, pueden ejecutar un plan de acción con los empleados identificados y disminuir los efectos negativos descritos anteriormente.

En esta investigación se desarrollará un modelo predictor de rotación de personal con el objetivo planteado anteriormente, que permite ayudar a anticipar esas eventuales salidas en cualquier empresa que se requiera. Para la implementación del caso de estudio, se trabajará con los datos de SITEL, una de las empresas líderes a nivel mundial en la industria de *Business Process Outsourcing* (BPO).

Esta empresa tiene presencia en Costa Rica desde el año 2021 al adquirir a la

compañía SYKES, cuyos inicios en el país fueron en 1995. Producto de esta adquisición se cambió el nombre a Foundever, como se conoce hoy en día. A nivel global la empresa tiene más de 170.000 empleados, con operaciones en 45 países distintos, servicios en más de 60 idiomas diferentes, más de 700 clientes distintos e ingresos anuales de aproximadamente $4.3 billones.

En general, los BPO son empresas que se especializan en recibir o realizar llamadas, correos o chats para brindarle apoyo de servicio al cliente o mercadeo para la misma compañía o a un tercero. En el caso de SITEL Costa Rica, la operación se enfoca en proveerle el servicio a clientes de *customer service* y *tech support*.

En esta industria, el departamento de recursos humanos considera métricas importantes y vitales para el desempeño de este tipo de negocio, tales como: atracción de talento, compromiso de los empleados y la rotación de personal. La atracción de talento se refiere a la capacidad de la empresa de poder conseguir a los mejores candidatos para ocupar las plazas vacantes. El compromiso de los empleados consiste en la forma en que las personas de una compañía se sienten con respecto a trabajar con la misma, su posición y la cultura organizacional. La rotación de personal mide la cantidad de empleados que salen de la compañía en un periodo determinado, dividido entre la cantidad de empleados activos al momento del cálculo.

Existen tres tipos de rotación de personal: voluntaria, involuntaria y total. La rotación voluntaria se refiere a aquellas personas que salieron de la empresa por decisión propia, la involuntaria considera aquellas personas que la compañía decidió retirar de su nómina por temas de rendimiento o reducción de la fuerza laboral y la total es la suma de ambos tipos de salidas. Esta investigación se enfocará en la rotación de personal voluntaria.

Según se menciona en varios estudios realizados previamente (Kavyasree y Naresh Kumar, 2020), (Suhasini, 2019) y (NageswaraRao y Swapna, 2019), la industria de los BPO ha presentado desde su inicio una rotación de personal muy alta, debido al alto volumen de trabajo que se recibe todos los días, los pocos descansos para los agentes, la fuerte competencia con múltiples empresas intentando ofrecer mejores beneficios a sus empleados que los demás, salarios relativamente bajos en comparación con el mercado, perfil de las personas que

trabajan en los mismos, pocas oportunidades para crecer dentro de la industria, ambiente laboral, horarios con turnos no habituales, entre otras razones.

En el caso de SITEL, específicamente en la operación en Costa Rica, no es la excepción. Para el año 2023 la empresa estaba conformada por cerca de **6.000** empleados, de los cuales se consideran empleados directos aproximadamente **5.000** personas y que a su vez se dividen en agentes de *customer service* alrededor de **3.800**, de las cuales en un año son bajas aproximadamente **3.000** personas, lo que representa una rotación de personal de **78,95** % anual. Los ciclos de tiempo para reemplazar a una de estas salidas son de más de 6 semanas en promedio y el costo anualizado de perder a cada uno de estos empleados es de $49.000 por persona aproximadamente. Por el volumen de empleados, esto genera millones de dólares de impacto negativo año con año para la compañía, afectando severamente sus estados financieros.

En la Figura 1 se muestra el gráfico de la rotación de personal anualizada de empleados directos de *customer service* en SITEL Costa Rica desde enero 2019 hasta diciembre 2023.

Figura 1: Rotación de Personal Anualizada de SITEL Costa Rica por Mes del 2019 al 2023

Fuente: Elaboración propia con datos de SITEL

Consecuentemente, se trabajará en la elaboración de un modelo que estime la probabilidad de que cada empleado renuncie a la empresa en 3 meses, consi-

derando datos sociodemográficos, económicos y operacionales de los colaboradores de la empresa. De esta forma se podrá brindarle recomendaciones a la compañía sobre la priorización de las personas a las que los planes de retención deberían ser aplicados a la mayor brevedad posible.

El horizonte a 3 meses se seleccionó porque se considera que es un tiempo prudencial para que la empresa identifique a las personas con mayor riesgo de salida, mantenga conversaciones con ellas y ejecute planes de acción para evitar las salidas del personal.

Adicionalmente, se van a incluir análisis fuera de muestra para medir la precisión de los modelos con datos diferentes a los de entrenamiento, y con esto validar los resultados generados a lo largo del tiempo. También se implementará un análisis de sensibilidad, donde se modificarán algunas variables con el objetivo de determinar la robustez del modelo ante cambios que puedan sufrir las variables a lo largo del tiempo.

Al día de hoy se han trabajado en algunos estudios similares publicados en artículos científicos. A continuación se presentarán algunos de ellos. En la investigación elaborada por (Barvey, Kapila, y Pathak, 2018), se elaboró un estudio donde por medio de algunas técnicas de modelación, se predijo la rotación de personal en una compañía particular, para que las personas encargadas pudieran tomar las acciones que correspondieran y evitar la fuga de personal. Para conseguirlo se tomaron distintas variables acomodadas por grupos del medio ambiente, factores financieros, factores externos, relacionados al trabajo, aspectos legales, variables sociodemográficas de las personas, entre otros.

Se han utilizado algunos métodos predictivos de clasificación para entrenar un modelo y poder predecir la rotación de personal de forma voluntaria (Yedida, 2018). En este trabajo también se estudiaron factores demográficos de las personas, financieros, ambiente laboral, potencial de crecimiento, entre otros. Se compararon los distintos modelos que se trabajaron al estudiar el área bajo las curvas ROC de cada uno de ellos. Los modelos de referencia fueron Bayes, Regresión Logística, Redes Neuronales y KNN (*K-Nearest Neighbors*), siendo este último el de mejores resultados.

Por otro lado, (Aulck, 2017) desarrolló modelos para poder predecir la rotación, pero en lugar de que fuera en un ambiente laboral, más bien fue la

rotación estudiantil en las carreras STEM (de las siglas en inglés *Science, Technology, Engineering and Math*). Se estudiaron más de 66.000 estudiantes de la Universidad de Washington, entre 1998 y el 2010. En el estudio también se compararon distintos modelos (ADA-Boosting, Regresión Logística, *XGBoost* y Bosques Aleatorios) por medio del área bajo las curvas ROC de cada uno de ellos. En este caso el mejor modelo fue el de Regresión Logística.

El análisis de la rotación estudiantil e identificación de los cursos que son cuellos de botella para que las personas se gradúen en la CSUN (California State University Northridge) fueron estudiados por (Sajjadi, 2017). Las carreras de estudio fueron Derecho, Administración, Mercadeo, Ingeniería Civil e Ingeniería Eléctrica. Se tomaron las notas académicas de las personas del 2004 al 2014, con más de 9.000 estudiantes involucrados. En este caso se logró determinar el objetivo, basado en las características de los estudiantes, cuáles serían los cursos cuello de botella y el abandono estudiantil por medio de los modelos de KNN y Regresión Logística.

La rotación estudiantil de personas en cursos en línea masivos, con una base de 36.000 estudiantes se investigaron en (Amnueypornsakul, Bhat, y Chinprutthiwong, 2014). Se analizaron aspectos como el número de interacciones con los videos de la página, cantidad de intentos de las evaluaciones, número de visitas a los foros, entre otros. En este caso se entrenó un modelo de SVM (*Support Vector Machine*) para poder clasificar a los estudiantes.

En la investigación de (Alaraj y Abbod, 2016), se abordó la forma de crear un método de consenso entre modelos de Regresión Logística, Redes Neuronales, SVM, Bosques Aleatorios, Árboles de Decisión y Bayes, para poder clasificar a las personas en un *score* de crédito. Este artículo no trabaja propiamente con rotación de personal, sin embargo podría ser de gran utilidad para poder utilizar este método de consenso, ya que en el fondo un *score* de crédito y un predictor de rotación de personal funcionan muy parecidos, al tener como objetivo principal en ambos casos poder evaluar un 1 o 0 con base en diferentes variables para cada observación. Al trabajar con distintas bases de datos de varios países se logró demostrar que en casi todas las oportunidades el área bajo la curva ROC era mayor en el método de consenso.

Los autores (Chourey, Phulre, y Mishra, 2019) trabajaron con la rotación de

personal con datos ficticios, creados por científicos de datos en IBM por medio de los modelos de predicción de Árboles de Decisión, Bosques Aleatorios, ADA-Boosting y *XGBoost*. Después de analizar los 35 atributos distintos de cada observación se terminó definiendo que el modelo de Bosques Aleatorios era el mejor de todos para esta base de datos, al tener una precisión mayor que los demás.

Finalmente, (Shankar, Rajanikanth, Sivaramaraju, y Murthy, 2018) analizaron la rotación de personal en la base de datos de IBM, previamente descrita con 1.470 observaciones y 35 variables, tanto númericas como categóricas, por medio de métodos predictivos de clasificación, tales como Árboles de Decisión, Regresión Logística, SVM, KNN y Bayes. Se creó un algoritmo para determinar quiénes eran los mejores empleados con base en desempeño, salario y educación, para poder definir una prioridad en cuanto a las personas que se iban a prevenir que abandonaran la empresa.

2. Objetivos

2.1. Objetivo General

- Analizar e implementar un modelo predictivo que permita estimar la rotación de personal de empleados directos de *customer service* por medio de una probabilidad asociada en la empresa SITEL Costa Rica en el 2023.

2.2. Objetivos Específicos

- Realizar un análisis exploratorio y de selección de las variables disponibles para determinar las que tienen mejor capacidad predictora de la rotación de personal de empleados directos de *customer service* en la empresa SITEL Costa Rica.
- Ajustar e implementar diferentes modelos predictivos para definir el modelo con mejor ajuste para pronosticar la rotación de personal de empleados directos de *customer service* por medio de una probabilidad asociada en la empresa SITEL Costa Rica.
- Analizar la precisión del modelo seleccionado por medio de una prueba de estimación fuera de muestra y con esto determinar la efectividad del mismo con los datos de las salidas reales en la empresa SITEL Costa Rica y compararlas con sus predicciones respectivas.

3. Marco Teórico

En este capítulo se definirán los principales conceptos relacionados con la investigación, con el objetivo de establecer una guía para comprender el trabajo desarrollado. Estos se consideran como la base del proceso investigativo.

- **Aprendizaje Automático**: También conocido como *machine learning*, es un conjunto de métodos computacionales, algoritmos y modelos para aprender patrones por medio del análisis de la información suministrada, sin ser explícitamente programados. Existen múltiples tipos de aprendizajes derivados de *machine learning*, tales como supervisado, no supervisado, reforzamiento, activo, entre otros. Esto se explica en (Bi, 2019) y (Mohri, Rostamizadeh, y Talwalkar, 2018).
- **Aprendizaje Supervisado**: Rama de *machine learning* donde se desarrolla algún modelo o algoritmo y se entrena a partir de un conjunto de datos que contiene ciertas características (variables) para cada observación y también utiliza etiquetas o variables respuestas. A partir del aprendizaje generado, se pueden producir ciertas predicciones sobre observaciones que no se tengan etiquetas y sus implementaciones más comunes son por medio de problemas de clasificación y regresión. Lo anterior se detalla en (Muhammad y Yan, 2015) y (Mohri y cols., 2018). Todas las metodologías descritas a continuación se consideran aprendizaje supervisado.
- **Árboles de Decisión**: modelo donde se trabaja con nodos (representados gráficamente por cajas y líneas que los conectan) para mostrar distintos escenarios y los efectos de que pasen los mismos hasta que se agoten todos los caminos posibles y se clasifiquen las observaciones. Recibe este nombre por la similitud con las ramas de un árbol, que se van interconectando hasta llegar a las hojas. Este método se describe en (Alaraj y Abbod, 2016), (Chourey y cols., 2019) y (Shankar y cols., 2018).
- **Bosques Aleatorios**: metodología que se compone por un conjunto de Árboles de Decisión donde se entrena el modelo que ejecuta un gran número de escenarios distintos y llega a un consenso en los resultados de las clasificaciones de los árboles ejecutados para tomar una decisión

final. Este método se describe en (Alaraj y Abbod, 2016) y (Chourey y cols., 2019) y también lo implementa (Aulck, 2017).

- **KNN**: algoritmo que clasifica observaciones con base en una distancia que se calcula entre las observaciones a clasificar con los vecindarios formados de observaciones cercanas entre sí. Se dice que al una observación estar más próxima a un grupo en específico es porque tiene características en común con ese grupo y por lo tanto se clasificaría de la misma forma. Este método se describe en (Yedida, 2018) y (Shankar y cols., 2018) y también lo implementa (Sajjadi, 2017).

- **Regresión Logística**: modelo que ajusta las clasificaciones basado en la función logística para variables binarias a predecir. Se obtiene una combinación lineal de coeficientes con los valores de las variables independientes y después se calcula la probabilidad de clasificarse de una u otra forma. Se utiliza para prevenir un sobreajuste. Este método se describe en (Yedida, 2018), (Alaraj y Abbod, 2016) y (Shankar y cols., 2018) y también lo implementan (Aulck, 2017) y (Sajjadi, 2017).

- **SVM**: método cuyo objetivo es encontrar un hiperplano que permita separar a las observaciones por completo en dos grupos que no se traslapen entre sí. Recibe este nombre por los vectores de soporte que definen el hiperplano. Además, se busca que este hiperplano tenga la separación máxima entre ambos grupos clasificados. Este método se describe en (Alaraj y Abbod, 2016) y (Shankar y cols., 2018) y también lo implementa (Amnueypornsakul y cols., 2014).

- **XGBoost**: modelo en el que las clasificaciones se obtienen secuencialmente. Los nuevos predictores aprenden de los errores que cometieron en el pasado, lo que provoca que cada vez se generen menos errores en las predicciones. Además, utiliza técnicas de regularización para prevenir el sobreajuste del modelo, por lo que se adapta en gran forma a nuevos conjuntos de datos. Este método se describe en (Chen y Guestrin, 2016), (Chourey y cols., 2019) y también lo implementa (Aulck, 2017).

- **Consensus**: metodología que permite combinar las predicciones de diferentes metodologías, con el fin de encontrar un modelo con mejor ajuste y predicciones. Al ser un modelo con variable a predecir desbalanceada,

se puede ponderar con base en los resultados de las áreas bajo las curvas ROC como se describe en (Marmion, Hjort, Thuiller, y Luoto, 2009).

- **Backward Stepwise**: procedimiento de regresión múltiple con varios tipos de algoritmos, utilizado para selección de variables, tal y como se describe en (Thayer, 2002). En el caso del *Backward Stepwise* específicamente, se construye un modelo predictivo con el que se consideran todas las variables posibles inicialmente y se van eliminando una por una conforme al menor poder predictivo o que disminuría la capacidad de explicar la varianza de la variable dependiente. Se suele medir el AIC (*Akaike Information Criterion*) para evaluar tanto la bondad de ajuste del modelo como su complejidad. En el caso del *Backward Stepwise* lo que se busca es una minimización de este criterio.

- **Clustering Jerárquico**: método de aprendizaje no supervisado que permite agrupar las observaciones en un dendograma basadas en las características en común de cada una de ellas. Como se detalla en (Wu, Xiong, y Chen, 2009), se consideran *sub-clusters* con los que se compara la distancia de las observaciones a cada uno de ellos para clasificarlas.

- **Análisis Fuera de Muestra**: análisis retrospectivo para evaluar la viabilidad de aplicar los modelos entrenados con nueva información a la entrenada en el modelo. Como se explica en (Olorunnimbe y Viktor, 2023), no basta con determinar la precisión de un modelo al separar la información de la base de datos en *training* y *testing*, sino que se debería de tomar otra base de datos que considere un periodo similar para que, una vez el modelo ya esté entrenado, se pruebe con nueva información de otro periodo de tiempo para garantizar consistencia en los resultados.

- **Matriz de Confusión**: matriz comunmente utilizada al implementar un modelo predictivo, que tiene un tamaño de n x n y ayuda a resumir las predicciones de los modelos y compararlas con los valores reales de cada observación del subgrupo de *testing*. Estas se definen en (Visa, 2011)

- **Precisión y Error Global**: la precisión global va a estar determinada por la suma de las observaciones correctamente clasificadas de cada categoría entre el total de observaciones. Por otro lado, el error global es la métrica calculada al sumar las observaciones incorrectamente clasificadas

de cada categoría entre el total de observaciones. Se describe en (Visa, 2011) y (Coenen, 2012).

A continuación se muestra la fórmula de ambos conceptos, donde VN corresponde a las observaciones clasificadas como negativas correctamente, FP las clasificadas como positivas erróneamente, FN las clasificadas como negativas erróneamente y VP las clasificadas correctamente como positivas:

$$\text{Precisión Global} = \frac{VN + VP}{VN + FP + FN + VP}$$

$$\text{Error Global} = \frac{FN + FP}{VN + FP + FN + VP}$$

- **Sensibilidad**: se refiere a las observaciones clasificadas como verdaderos positivos (en este caso, las predicciones de salidas que realmente fueron salidas) entre la suma de los verdaderos positivos más los falsos negativos (en este caso, personas cuya predicción fue que no iban a salir y en realidad sí terminaron saliendo de la empresa). Se explica en (Coenen, 2012).

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

- **Especificidad**: corresponde a las observaciones clasificadas como verdaderos negativos (en este caso, las predicciones de no salidas que realmente no fueron salidas) entre la suma de los verdaderos negativos más los falsos positivos (en este caso, personas cuya predicción fue que iban a salir y en realidad no terminaron saliendo de la empresa). Se menciona en (Coenen, 2012).

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

- **Probabilidad de Corte**: definida para modelos predictivos, como se detalla en (Habibzadeh y Yadollahie, 2016), es la probabilidad que se define para poder calcular la sensibilidad y especificidad del modelo. Entre más

alta sea la probabilidad de corte a utilizar se incrementa la especificidad y entre más baja sea la probabilidad se incrementa la sensibilidad.

- **Curva ROC**: indicador de calidad y capacidad predictiva de un modelo que representa de forma gráfica la comparación entre la sensibilidad versus la especificidad para cada probabilidad de corte posible de un modelo predictivo. En el eje X se calcula (1 - especificidad) y en el eje Y la sensibilidad. Una vez graficados todos los posibles puntos de cada probabilidad de corte, se calcula el área bajo la curva para determinar la calidad del modelo predictivo, siendo 1 el valor máximo y 0 el mínimo. También se le conoce como (AUC) por sus siglas en inglés *Area Under the Curve*. Puede ser consultado en (Ekelund, 2012).

- **Validación Cruzada**: técnica utilizada en el aprendizaje automático para evaluar el rendimiento de un modelo y estimar su capacidad de generalización a los datos no vistos. Se dividen los datos en subconjuntos de entrenamiento y de prueba en múltiples ocasiones de forma aleatoria en K particiones (también llamadas grupos o pliegues), donde el subconjunto de prueba corresponde a una K-ésima parte del total de los datos y los restantes K-1 grupos son los de entrenamiento. Este proceso se ejecuta K veces para garantizar que cada parte fue tomada como subconjunto de prueba una vez. Esto se detalla en los artículos de (Yates, 2022) y (Ghojogh y Crowley, 2023).

4. Metodología

En este capítulo se describe el funcionamiento de las diferentes metodologías que se aplicaron en la investigación, profundizando en mayor medida en la metodología de *Extreme Gradient Boosting*, al ser el método seleccionado para implementar el modelo predictivo de la rotación de personal en la compañía.

Es importante señalar que para ejecutar este tipo de metodologías se suele dividir la información disponible en dos partes: *training* (datos de entrenamiento para ajuste de los parámetros) y *testing* (datos no vistos por el modelo para determinar si el ajuste es apropiado o existe alguna evidencia de sobreajuste).

4.1. Regresión Logística

Los modelos de Regresión Logística se comenzaron a desarrollar en 1967 para calcular la probabilidad de ocurrencia de un proceso a partir de una función con varias variables, tal como se describe en (Domínguez-Almendros, Benítez-Parejo, y Gonzalez-Ramirez, 2011). El objetivo del modelo es obtener una clasificación de una variable binaria, comúnmente conocida como variable dependiente o respuesta, a partir de diversas variables independientes.

Supóngase que se quiere determinar la probabilidad de que suceda algún evento, dado ciertas variables independientes. Considérese la variable a predecir Y, donde toma un valor de 0 si el evento no ocurre y 1 en el caso que sí ocurra. Además, $\{X_1, X_2, ..., X_n\}$ son las variables independientes que se utilizarán para la predicción.

La variable Y se describe de la siguiente manera:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{1}$$

Por lo que la probabilidad de que el evento Y ocurra se podría calcular de la siguiente manera:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}} \tag{2}$$

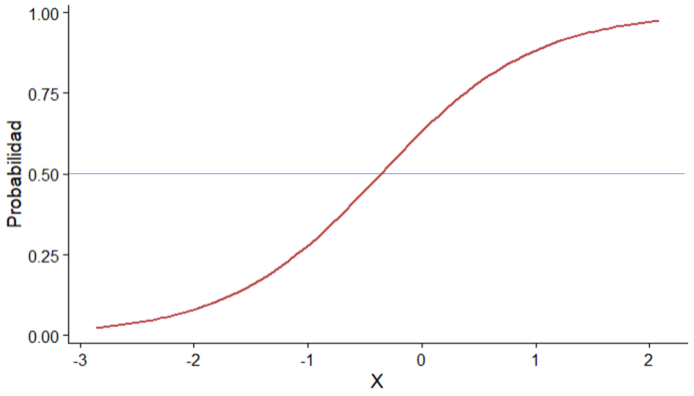
Donde e representa la función exponencial, β_0 es el intercepto y $(\beta_1, \beta_2, ...\beta_n)$ son los coeficientes de las variables independientes $\{X_1, X_2, ..., X_n\}$ respectivamente. Al igual que con la regresión múltiple, entre mayores sean los coeficientes betas del modelo en valor absoluto, estos terminarán con un mayor peso sobre el cálculo de la probabilidad.

Si se utiliza la fórmula de la Ecuación 3, se podría graficar el cociente de las probabilidades por medio de la transformación logística:

$$y = Logit(p) = \frac{p}{1-p} \tag{3}$$

En la Figura 2 se muestra un ejemplo gráfico de una Función Logística, donde la línea de color rojo muestra cómo variaría la probabilidad de ser una categoría 1 o 0 dependiendo del valor de la variable predictora X. Además, se podría establecer una probabilidad de corte que serviría de comparación con la probabilidad obtenida para definir la clasificación final del evento como 1 o 0, en este caso identificada en color azul.

Figura 2: Ejemplo Gráfico de una Función Logística



Fuente: Elaboración propia

4.2. KNN

El algoritmo de KNN es un método de clasificación basado en instancias, es decir que no construye un modelo explícito, sino que almacena todos los datos de entrenamiento y toma decisiones basadas en sus comparaciones, tal como se describe en (Bafandeh y Bolandraftar, 2013) y (Suguna y Thanushkodi, 2010).

KNN se basa en las distancias de los datos de entrenamiento y la observación que se desea clasificar, donde se define la clasificación dependiendo de la etiqueta mayoritaria de los K vecinos más cercanos a la observación.

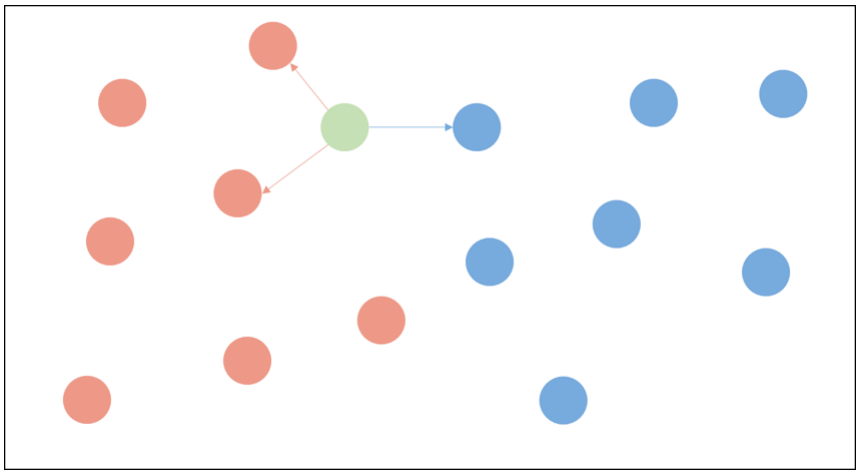
Una vez que se define el número de K vecinos que se van a utilizar, se calcula la distancia mencionada anteriormente. Esta distancia se suele medir con la medida Euclidiana, sin embargo existen otras como *City-Block*, Chebychev, entre otras. La fórmula de la distancia Euclidiana sería la siguiente:

$$D(x, p) = ||x - p|| \tag{4}$$

Donde x representa una muestra desconocida y p las muestras del conjunto de entrenamiento. De esta manera se tomaría la decisión sobre la clasificación con base en la clasificación más frecuente sobre los K vecinos con menor distancia a la observación.

Como se muestra en la Figura 3, si se quisiera clasificar la observación en color rojo o azul con base en los 3 vecinos más cercanos, se hubieran calculado las distancias para determinar cuáles eran las 3 observaciones más cercanas que hubiera tenido y la etiqueta de la mayoría de ellas hubieran definido su clasificación, siendo este caso una clasificación roja.

Figura 3: Ejemplo de Proceso de Clasificación KNN



Fuente: Elaboración propia

4.3. SVM

Support Vector Machine es otro algoritmo de aprendizaje supervisado que puede ser utilizado para clasificación. Como se describe en (Shawe-Taylor y Sun, 2011) y (Gunn, 1997), el objetivo es encontrar un hiperplano óptimo que separe las clases en el espacio de características. Dentro de este espacio se podrían separar los datos de múltiples maneras, sin embargo va a existir solamente una solución que garantice que los márgenes (la distancia entre el plano y el punto más cercano al plano de cada clase) sea la máxima posible.

Considérese un conjunto de datos de entrenamiento etiquetados $\{(x_i, y_i)\}_{i=1}^{n}$, donde $x_i \in \mathbb{R}^d$ es el vector de características y $y_i \in \{-1, +1\}$ son las etiquetas binarias de las categorías de la variable a predecir. Este conjunto de datos se pretende separar por medio del hiperplano:

$$(w \cdot x) + b = 0 \tag{5}$$

Donde w corresponde al vector normal al hiperplano, b es el término de sesgo y x es el vector de características de entrada, donde consideran la restricción de $\underset{x_i}{min}|x \cdot w + b| = 1$.

Por otro lado, la distancia de un punto cualquier al hiperplano se podría calcular como:

$$d(w, b; x) = \frac{|w \cdot x + b|}{||w||} \tag{6}$$

Además, el hiperplano óptimo maximiza los márgenes ρ.

$$\rho(w, b) = \frac{2}{||w||} \tag{7}$$

Con base en la Ecuación 7, en problemas que son linealmente separables, se busca resolver el problema de optimización de la siguiente Ecuación 8.

$$\underset{w,b}{min}\Phi(w) = \frac{||w||^2}{2} \tag{8}$$

Donde se tiene que cumplir la restricción para $i = 1, ..., n$ de la Ecuación 9.

$$y_i(w^T x_i + b) \geq 1 \tag{9}$$

Finalmente, la clasificación de las nuevas observaciones estarían dadas por la posición de la misma con respecto al lado del plano en el que se ubicaría por:

$$f(x) = \text{sign}(w^T x + b) \tag{10}$$

Donde sign$(\cdot)$ es la función signo que asigna +1 si el argumento es positvo o cero y -1 si es negativo.

4.4. Árboles de Decisión

Los árboles de decisión son la base fundamental del método de *Extreme Gradient Boosting* y Bosques Aleatorios. Como mencionan (Song y Lu, 2015) y (Witten y Frank, 2005) en sus publicaciones, los árboles de decisión son una representación gráfica de un método de clasificación por medio de ramas, nodos y hojas. El método considera las características de las distintas variables a considerar y se van formando caminos a través de las ramas y nodos hasta llegar a las clasificaciones.

Estos suelen utilizarse comúnmente para selección de variables, evaluar la importancia relativa de esas variables, manipulación de la información y programar modelos predictivos de futuras observaciones. Por su fácil interpretabilidad, son ampliamente utilizados en diversas áreas para la toma de decisiones.

Sin embargo, tienen algunas desventajas, como que se suele tener cierta propensión al sobreajuste al tener un alto rendimiento con los datos de entrenamiento, por lo que podría no brindar resultados óptimos la hora de generalizar a nuevos datos. Además, se presenta sensibilidad a datos pequeños que pueden generar ruido en el modelo, pueden crear sesgos hacia la clase mayoritaria al trabajar con bases desbalanceadas, entre otros.

A continuación, se detallan algunos conceptos básicos que conforman la estructura de un árbol de decisión:

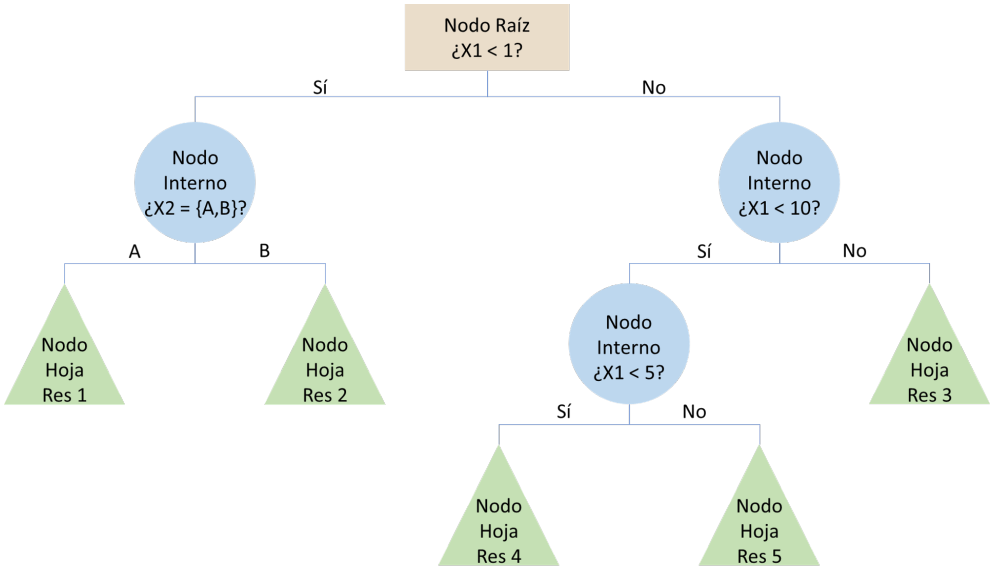
- **Nodos**: Son las diferentes categorizaciones que realizan los árboles para evaluar si continúan por un camino o se detienen en una clasificación. Existen varios tipos de nodos: raíz, internos y hojas.
- **Nodo Raíz**: La raíz representa la primera decisión en la parte superior que va a determinar el origen de todos los siguientes nodos.
- **Nodo Interno**: Los internos corresponden a una de las posibilidades disponibles en ese punto de la estructura del árbol. Estos nodos siempre tienen un nodo padre (parten de algún otro) y nodos hijos (divisiones siguientes que aportan mayor estructura al árbol u hojas).
- **Nodo Hoja**: Finalmente, las hojas son los últimos nodos y contienen los resultados de las clasificaciones según la combinación de eventos previos.
- **Ramas**: Representan los diferentes escenarios de caminos posibles que podría tomar una observación partiendo del nodo raíz hasta llegar alguna de las hojas. En ese recorrido la observación se somete a diferentes pruebas para determinar con base en la información de los nodos internos si debería de considerar un camino u otro.
- **División**: Se consideran las variables relacionadas con la variable objetivo para tener una mejor precisión en las clasificaciones que se apliquen al conjunto de datos establecido. Estas divisiones, también llamadas *splits*, se aplican para generar más nodos desde el nodo raíz hasta que se llega al criterio de detención. Es importante señalar que la estructura de un árbol no necesariamente incluye todas las variables posibles, así como que una misma variable puede servir para hacer varias divisiones en diferentes niveles.
- **Detención**: Es importante que la generación del árbol de decisión se detenga en el momento preciso, ya que sino podría generarse un sobreajuste en los datos al tener un modelo muy complejo que no sirva para predicciones de observaciones futuras. Para prevenir esto se pueden establecer ciertas reglas, como que cada hoja contenga un mínimo de observaciones por hoja, el número mínimo de registros en un nodo antes de dividir y la profundidad del árbol (número de pasos o nodos intermedios entre la raíz y las hojas).

- **Poda**: En algunos casos las reglas de detención no funcionan adecuadamente, por lo que una opción puede ser construir un árbol que tenga una estructura amplia y después se poden las ramas que tengan los nodos que menos información aporten, para evitar el sobreajuste del modelo.

La Figura 4 representa un ejemplo de los conceptos explicados anteriormente de forma gráfica, donde se parte del nodo raíz en color café, se continúa por las ramas hasta los nodos intermedios en celeste y finalmente se llega a los nodos hoja en verde. Se puede observar que el árbol considera tanto variables numéricas (X_1) como categóricas (X_2), donde existen criterios que se evalúan para determinar el camino de las ramas a seguir. En el caso de X_1, se evalúa en el nodo raíz si es menor a 1, donde en caso afirmativo se evalúa en el nodo interno si X_2 corresponde a la categoría A o B para llegar a la clasificación final de los nodos hojas.

En caso que X_1 fuera mayor o igual a 1 entonces se evalúan en los nodos intermedios de la derecha si X_1 es menor a 10 y posteriormente si es menor a 5 para llegar a las clasificaciones finales de los nodos hojas de la derecha del árbol.

Figura 4: Ejemplo de Estructura de Árbol de Decisión



Fuente: Elaboración propia

4.4.1. Método CART

Existen varios tipos de algoritmos para la creación de árboles de decisión, donde uno de los más comunes que se utiliza es el método CART, por sus siglas en inglés *Classification and Regression Tree* y se considera la base de *Extreme Gradient Boosting*. Según indican (Irmanita, Suryani, y Sibaroni, 2021), el método CART se creó en 1984 y dentro de sus principales ventajas están el control de los *outliers*, al quedar separados con la división que se aplica, así como la selección de las variables más importantes y la velocidad computacional que tiene.

Como se describe en (Huang, Peng, Cai, y Chen, 2016) y (Irmanita y cols., 2021), el algoritmo para construir un árbol de decisión con el método CART es:

1. Establecer el nodo raíz sobre el que se iniciará la estructura del árbol y asignar la regla para la división de este nodo basado en el índice de impureza de Gini.
2. Si todas las observaciones corresponden al mismo subconjunto o es solamente una sola observación, el nodo raíz será al mismo tiempo el único nodo hoja.
3. Para crear las ramas, se verifica que se cumpla con el criterio de la división al considerar las variables posibles del conjunto de datos para la división de los subconjuntos, con el objetivo de minimizar la impureza de los nodos determinada por el índice de Gini para todas las variables posibles.
4. Se continúa de forma recursiva este proceso con cada uno de los subconjuntos que se vayan obteniendo, hasta que no sea posible continuar adicionando más ramas.
5. Finalmente, se podan las ramas para garantizar que no exista un sobreajuste del modelo.

El índice de Gini que funciona como criterio al momento de dividir un nodo en una nueva rama, midiendo la impureza del conjunto de los datos de *training* por medio de la siguiente fórmula descrita en (Irmanita y cols., 2021):

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \tag{11}$$

Donde D representa al conjunto de datos, P_i corresponde a la probabilidad de que una observación pertenezca a una clase determinada y se realiza la suma sobre las m clases posibles. Se conoce como impureza porque esta mide qué tan puros son los nodos con base en la división por una característica determinada y permite aislar lo más posible que cada nodo quede solamente con una categoría de la variable predictora posible.

En otras palabras, entre mayor sea el número de observaciones en un nodo solamente de una de las categorías posible, más pura será la divisón y el valor de la fórmula será más cercano a cero. Con base en este resultado de cada una de las características de las variables, se puede calcular el índice de Gini con la fórmula:

$$Gini_A(D) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2) \tag{12}$$

Donde $Gini_A(D)$ es el índice de Gini de una característica y D_1, D_2 son las particiones de los subconjuntos de observaciones basados en esa característica; es decir, la cantidad de observaciones que cumplen la característica para obtener un promedio ponderado por la cantidad de observaciones.

Finalmente, para la poda de las ramas se utilizan los métodos descritos en (Ravi y Serra, 2017) y (Quinlan, 1987). Lo primero es establecer una función de costo asociada a cada nodo interno del árbol. La fórmula está dada por:

$$R_\alpha(T) = R(T) + \alpha \cdot L(T) \tag{13}$$

donde R(T) es la sumatoria de todos los errores generado con los datos de *training* para cada uno de los nodos hojas, L(T) son la cantidad de nodos hojas del árbol y α es un parámetro que se suele obtener por medio de una validación sobre los datos de *testing* para determinar el mejor valor.

El incorporar L(T) sirve para poder compensar el hecho que un árbol con más hojas pueda tener menor cantidad de errores en los nodos hojas y con esto

un R(T) menor, pero esto no es necesariamente positivo, porque precisamente podría deberse a un sobreajuste del modelo. Además, el α entre mayor sea, mayor será el impacto para podar el árbol al castigar más los árboles conforme más nodos hojas tengan.

4.5. Bosques Aleatorios

El método de Bosques Aleatorios fue desarrollado en el 2001 por Leo Breiman para resolver problemas de predicción basado en el método de *bagging* y en árboles de decisión con el método CART, tal como mencionan los autores (Han, Gui, Xu, y Lacidogna, 2019) y (Xu, 2013).

Bagging es un método de ensamble utilizado para prevenir el sobreajuste de los modelos, al reducir la varianza del modelo promediando múltiples modelos entrenados con diferentes subconjuntos del conjunto de datos original.

Supóngase que se tiene un conjunto de datos de *training* $\mathbf{C} = \{C_1, ..., C_n\}$, de forma independiente e idénticamente distribuidos mediante el remuestreo aleatorio del conjunto de datos de *training* original. De estos se toman M muestras aleatorias con reemplazo de $\mathbf{C}$, $B_1, ..., B_M$. Para cada subconjunto de B_m se ajusta un modelo base, donde $m = 1, ..., M$.

El número de muestras de cada nuevo conjunto de datos es el mismo que los datos originales y dado que el muestreo se aplica con reemplazo, es posible que algunas observaciones sean seleccionadas múltiples veces en el mismo conjunto de *training*, mientras algunos podrían no aparecer del todo.

Para el caso específico de los bosques aleatorios, el modelo base serían árboles de decisión CART para cada B_m con todas las consideraciones descritas en las secciones anteriores. La aplicación del *bagging* evita que los árboles generados sean los mismos al tener un remuestreo con reemplazo, previniendo de esta manera el sobreajuste del modelo.

Cada uno de estos árboles de decisión también consideran solamente un subconjunto de todas las variables predictoras posibles a considerar y estas variables predictoras se van alternando de forma aleatoria en las diferentes estructuras de cada árbol. El autor Breiman sugiere utilizar un tercio del total de las variables para calibrar cada uno de ellos.

Una vez que ya se tiene la estructura de todos los árboles definidas, las predicciones de las observaciones del conjunto de *testing* del modelo de Bosques Aleatorios son basadas en cuál fue la categoría que obtuvo mayoría de clasificaciones en cada uno de los árboles de manera individual, tal como se muestra en la Ecuación 14.

$$argmax_g\{\sum_{m=1}^{M} L[\hat{f}_m^*(x_0) = g]\} \tag{14}$$

Donde $\hat{f}_m^*(x_0)$ sea la predicción de C_0 del m-ésimo árbol que depende del predictor x_0.

4.6. Extreme Gradient Boosting

Como se mencionó anteriormente, corresponde a un algoritmo de *machine learning*, que aplica una serie de árboles de decisión sobre los cuales se van aprendiendo los errores conforme se realizan las clasificaciones de los árboles anteriores.

De esta manera, termina siendo un método que comete cada vez menos errores de clasificación al obtenerse secuencialmente, aprendiendo del pasado y es justo lo que lo cataloga como uno de los métodos más novedosos, precisos y mayormente recomendado por consultores y validadores de modelos predictivos.

Justamente esta es la principal diferencia con Bosques Aleatorios, ya que como se estudió en la sección anterior, utiliza *bagging* para combinar múltiples árboles de decisión independientes en lugar de corregir secuencialmente los errores del árbol anterior. Además, *XGBoost* considera una función de pérdida utilizando el gradiente descendente y un término de regularización para reducir el sobreajuste, como se analizará en esta sección.

El método de *Extreme Gradient Boosting*, también conocido como *XGBoost*, fue desarrollado originalmente por parte de Tianqu Chen y Carlos Guestrin, ambos de la Universidad de Washington en el año 2016. La publicación fue consultada en (Chen y Guestrin, 2016) y desarrolla todo el fundamento matemático que se estará detallando en esta sección.

Este método es una versión mejorada del *Gradient Boosting Decision Tree*, ya que ese solamente utiliza la primera derivada para la optimización, mientras que *Extreme Gradient Boosting* utiliza la primera y segunda derivada para este fin. Otra diferencia de *XGBoost* es la utilización de un término para regular la función objetivo y con esto evitar un sobreajuste del árbol, como se verá a continuación.

Supóngase que se tiene un conjunto de datos $D = (x_i, y_i)$, con n observaciones y m variables, donde ($|D| = n$, $x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$). Los x_i corresponden a las variables independientes que se van a utilizar para predecir a la variable objetivo y_i. Cada una de las predicciones del modelo serán definidas como $\hat{y}_i$. El método empieza aplicando un ensamble de árboles de decisión con K funciones aditivas para predecir el resultado.

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{15}$$

donde $F = \{f(x) = w_{q(x)}\}$ ($w \in \mathbb{R}^T$, $q : \mathbb{R}^m \to T$) es el espacio de los árboles de decisión con regresión. Según definen los autores, este espacio q representa la estructura de cada árbol que mapea un ejemplo al índice de la hoja correspondiente. T es el número de hojas (nodos terminales) en el árbol. Cada f_k corresponde a una estructura de un árbol independiente q y hoja con pesos w.

El modelo considera las reglas de los árboles (dadas por q) para la clasificación de cada hoja y para realizar el cálculo de la predicción final suma las evaluaciones de las hojas correspondientes (dadas por w). Esto difiere de los árboles de decisión comunes, ya que cada árbol de regresión contiene una evaluación continua en cada hoja, donde se usa w_i para representar el valor de la i-ésima hoja.

El objetivo de *XGBoost* es poder aprender de cada uno de los k árboles que se examinen, mediante la minimización de la siguiente función objetivo regularizada:

$$L(\phi) = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{16}$$

En la ecuación (16), l es una función de pérdida diferenciable y convexa que mide la diferencia entre la variable objetivo y_i y la predicción $\hat{y}_i$. El segundo término de la ecuación, Ω penaliza la complejidad del modelo (i.e., las funciones de los árboles de regresión) y se puede calcular de la siguiente manera:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{17}$$

En este caso, Ω es el término de regularización que caracteriza al método de *XGBoost*, pues ayuda a suavizar los pesos del aprendizaje final utilizando γ como parámetro que limita la complejidad del modelo, junto con la cantidad de nodos hoja T y λ como el coeficiente de penalización para los pesos de las hojas (w) que suele ser una constante, controlando entre las dos la regularización de los árboles y de esta forma evitar un sobreajuste del modelo.

Entre mayores sean los γ y λ, más grande va a ser el castigo que aporte Ω a la función objetivo, provocando que el valor óptimo w esté cada vez más cercano a 0. Además, γ es el parámetro que controla una posible división de los nodos, ya que si la función de costo del nodo después de separado resulta menor a γ, entonces no se realizará la separación y viceversa. λ es el peso para regularizar la función. Intuitivamente, el objetivo regularizado va a tender a seleccionar un modelo empleando funciones simples y con altas capacidades predictivas.

Además, (Chen y Guestrin, 2016) indican que el modelo de ensamble de árboles de la ecuación (16) incluye como parámetros algunas funciones que no pueden ser optimizadas usando métodos tradicionales en un espacio Euclidiano. En lugar de eso, el modelo es entrenado de forma aditiva.

Al momento de entrenar el modelo, se incluyen nuevos árboles para ajustar los residuos de las rondas previas, por lo que si un modelo tiene t árboles, se define:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{18}$$

Formalmente, sea $\hat{y}_i^{(t)}$ la predicción de la i-ésima instancia en la t-ésima iteración, donde se tendrá que adicionar f_t para minimizar la siguiente función objetivo:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{19}$$

Esto significa que se adiciona la f_t que tiene mayor mejora al modelo con base en (16). Esto sucede en el punto en el que el valor de f_t resulta en la pérdida total mínima posible, porque las diferencias entre y_i y $\hat{y}_i^{(t-1)} + f_t(x_i)$ son las menores posibles.

Se puede usar la aproximación de Taylor de segundo orden para optimizar la función objetivo anterior:

$$L^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{20}$$

donde $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ y $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ son el primer y segundo gradiente estocástico de la función de pérdida. Si se remueven los términos constantes para obtener la siguiente función objetivo simplificada en el paso t se tendría:

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [(g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{21}$$

Estos gradientes g_i y h_i son los parámetros sobre los cuales depende optimización de la función objetivo.

Ahora, se define $I_j = \{i | q(x_i) = j\}$ como el conjunto de instancias de la hoja j. Se podría sustituir el Ω de la ecuación (21), con base en la ecuación (17) de la siguiente manera:

$$\begin{aligned} \tilde{L}^{(t)} &= \sum_{i=1}^{n} [(g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i)] + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \end{aligned} \tag{22}$$

Para una estructura fija $q(x)$, se puede obtener el peso óptimo w_j^* de la hoja j

por medio de:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{23}$$

Considerando este peso óptimo w_j^*, se podría calcular el valor óptimo:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{24}$$

Como indican los autores (Chen y Guestrin, 2016), esta ecuación (24) se puede utilizar como función para medir la calidad de la estructura del árbol $q(x)$. Esta evaluación funciona de forma similar que la nota de impureza al evaluar árboles de decisión, excepto que está derivada de un rango mucho más amplio de funciones objetivos.

Una vez optimizada la función objetivo $\tilde{L}^{(t)}$ de la ecuación (22), se puede encontrar la estructura óptima del árbol de decisión que se debería de aplicar. El valor de la función objetivo se podría comprender como el índice de una puntuación de la información ganada y que por su estructura entre menor sea el valor mejor.

XGBoost utiliza un algoritmo que considera inicialmente una única hoja y de forma iterativa va generando ramas al árbol, creando de esta manera más nodos intermedios y nodos hojas. Esto se debe a que resulta en la mayoría de las ocasiones imposible enumerar todas las posibles estructuras q.

El proceso mencionado asume que I_L y I_R son las instancias de los conjuntos de los nodos izquierdo y derecho después de la división respectivamente. Sea $I = I_L \cup I_R$, entonces la reducción de la pérdida después de la división estaría dada por:

$$L_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{25}$$

Esta fórmula se utiliza para poder evaluar las posibles divisiones de candidatos que se apliquen. Se estaría comparando la ganancia de información aportada

por las hojas de la izquierda y la derecha contra el nodo original que estaba considerando. Como se mencionó anteriormente, el resultado de este *split* debería de ser mayor a 0, ya que si no el valor de la ganancia estaría siendo inferior a γ y no podría ocurrir el *split*.

4.6.1. Extreme Gradient Boosting para Clasificación Binaria

Ahora bien, para resolver problemas de clasificación binaria con *XGBoost* es necesario realizar algunas transformaciones a las fórmulas planteadas anteriormente. Como se detalla en (Qin, 2021) y (Wang, Deng, y S., 2020), la función de pérdida sería de la siguiente manera:

$$L(y_i, \hat{y}_i) = -\sum_{i=1}^{N} [y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{26}$$

Donde y_i es el valor real de la observación y $\hat{y}_i$ la probabilidad de ocurrencia de la predicción del modelo. De esta forma se tiene una función de pérdida considerando las posibilidades de cada i y que es diferenciable:

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}^{(t-1)}} = -(y_i - \hat{y}_i)$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}^{(t-1)}} = \hat{y}_i(1 - \hat{y}_i)$$

Una vez calculados los nuevos g_i y h_i se continúa con el mismo proceso descrito de la ecuación (23) para definir w^* en adelante, ya que la única diferencia entre *Extreme Gradient Boosting* para regresión y para clasificación es la función de pérdida descrita anteriormente.

En resumen, para problemas de regresión, el método de *Extreme Gradient Boosting* utiliza árboles de decisión para encontrar el valor f_t que minimiza la función objetivo de la ecuación (19). Esta ecuación utiliza una función de pérdida y un término de regularización para evitar un sobreajuste de los árboles. En el caso de problemas de clasificación binaria funciona de forma similar, pero utilizando la función de pérdida de la ecuación (26).

Una vez que se encuentra el peso óptimo w_j^*, se utiliza para calcular el valor de la similitud de la ecuación (24). Finalmente, el método evalúa si al hacer una división de ramas adicionales se incrementa la precisión en la clasificación cumpliendo los requisitos establecidos por el término de regularización.

La Figura 5 representa un ejemplo del algoritmo de *XGBoost* basado en la investigación de (Muneeb y Henschel, 2021). Considérese el Cuadro 1 para el ejemplo de la Figura 5.

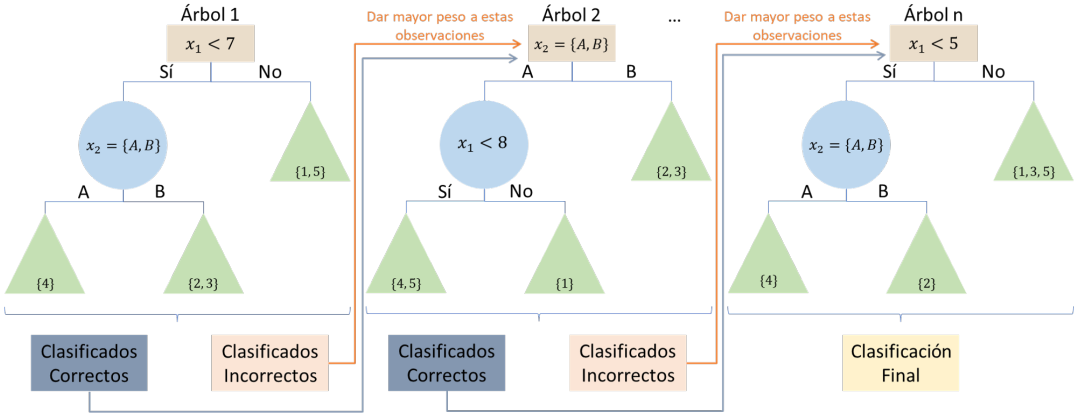
Cuadro 1: Datos para Ejemplo de Estructura de XGBoost

Observación	x_1	x_2
1	10	A
2	3	B
3	5	B
4	4	A
5	7	A

Fuente: Elaboración propia

En este ejemplo se muestra cómo se utilizan n árboles de decisión, donde las observaciones podrían llegar a ser clasificadas correcta o incorrectamente. Después de la primera clasificación que se haga del Árbol 1, se le dará un mayor peso a las observaciones que estuvieron incorrectas, para que el modelo intente clasificarlas correctamente en el Árbol 2 y así sucesivamente, hasta llegar al n-ésimo árbol, que brinda la clasificación final.

Figura 5: Ejemplo de Estructura de XGBoost



Fuente: Elaboración propia

5. Análisis Descriptivo y Selección de Variables

La calibración del modelo estará dirigida a determinar la probabilidad de que los empleados directos de *customer service* de SITEL Costa Rica abandonen la compañía en un horizonte de tres meses, a partir del estudio de variables socioeconómicas, demográficas, laborales y personales de los empleados.

Se va a trabajar con una base de datos con información de la empresa en SQL. Estos datos se exportarán al lenguaje de programación *R* para el manejo de los mismos, así como la programación y calibración de los modelos predictivos a desarrollar.

5.1. Definición de Variables

La base de datos consta de aproximadamente 3.200 empleados directos de *customer service* y está compuesta por las siguientes variables:

1. **Código Año**: Variable numérica que contiene el código identificador del año al cual corresponde toda la información de la observación. El nombre registrado en la base de datos es *RT_Year*.

2. **Código Mes**: Variable numérica que contiene el código identificador del mes al cual corresponde toda la información de la observación. El nombre registrado en la base de datos es *RT_Month*.

3. **Número de Empleado**: Variable numérica que contiene el número de empleado único de cada persona. El nombre registrado en la base de datos es *Employee_ID*.

4. **Nombre**: Variable de texto que contiene el nombre de las personas. El nombre registrado en la base de datos es *Name*.

5. **Edad**: Variable numérica que contiene la edad de las personas. El nombre registrado en la base de datos es *Age*.

6. **Género**: Variable categórica que contiene el género de las personas. El nombre registrado en la base de datos es *Gender*.

7. **Estado Civil**: Variable categórica que contiene el estado civil de las personas. El nombre registrado en la base de datos es *Marital_Status*.

8. **Nacionalidad**: Variable categórica que contiene la nacionalidad de las personas. El nombre registrado en la base de datos es *Nationality*.

9. **Provincia**: Variable categórica que contiene la provincia de residencia de las personas. El nombre registrado en la base de datos es *Province*.

10. **Sucursal**: Variable categórica que contiene el nombre de la sucursal en la que están registradas las personas. El nombre registrado en la base de datos es *WorkLocation*.

11. **Teletrabajo**: Variable categórica que contiene la información de si las personas tienen teletrabajo o no. El nombre registrado en la base de datos es *FlagLocation*.

12. **Vertical**: Variable categórica que contiene la clasificación del tipo de cliente para el que trabajan las personas. El nombre registrado en la base de datos es *Vertical*.

13. **Cliente**: Variable categórica que contiene el nombre del cliente para el que trabajan las personas. El nombre registrado en la base de datos es *Client*.

14. **Posición**: Variable categórica que contiene el nombre del puesto de las personas. El nombre registrado en la base de datos es *Position*.

15. **Tipo Trabajador**: Variable categórica que contiene la distinción de si la persona es agente o no. El nombre registrado en la base de datos es *ClassName*.

16. **Fecha Contratación**: Variable de fecha que contiene la fecha en la que se contrató a la persona. El nombre registrado en la base de datos es *Hire_Date*.

17. **Fecha Salida**: Variable de fecha que contiene la fecha de salida de la persona, en caso que sea un empleado inactivo. El nombre registrado en la base de datos es *Exit_Date*.

18. **Antigüedad**: Variable numérica que contiene el tiempo laborado en la compañía calculado en días. El nombre registrado en la base de datos es *Seniority*.

19. **Salario**: Variable numérica que contiene el salario total que recibe la persona por mes. El nombre registrado en la base de datos es *SalaryTotal*.

20. **Nota Evaluación**: Variable numérica que contiene la nota de evaluación que reciben las personas en sus evaluaciones de desempeño. El nombre registrado en la base de datos es *PA_Score*.

21. **Acciones Disciplinarias**: Variable numérica que contiene el número total de acciones disciplinarias que han recibido las personas. El nombre registrado en la base de datos es *DisciplinaryAction*.

22. **Reingreso**: Variable numérica que contiene el número de veces que las personas han trabajado en la compañía. El nombre registrado en la base

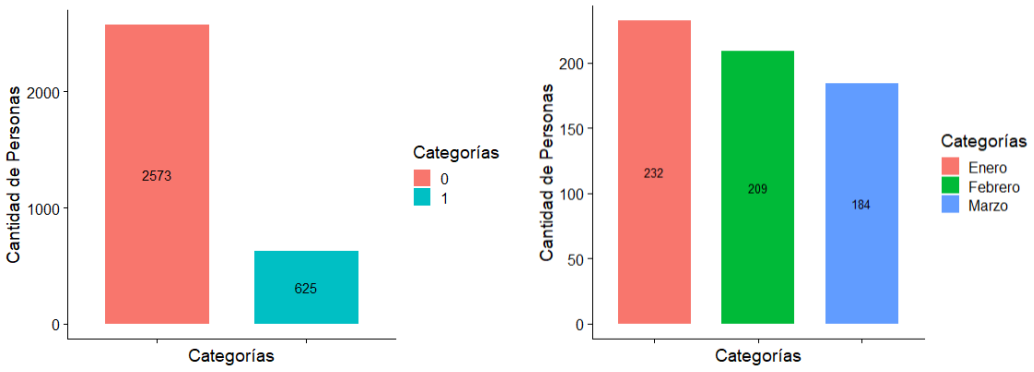
de datos es *ReIngreso*.

23. **Target**: Variable categórica que contiene un indicador 1 si el empleado salió de la compañía en un horizonte de 3 meses y 0 si no. El nombre registrado en la base de datos es *Target*. Esta variable sería la objetivo a predecir por medio del modelo de estudio, para poder determinar la probabilidad de que las personas salgan de la empresa en los próximos 3 meses de estudio.

5.2. Análisis Descriptivo y Selección

De la variable *Target* se observa que un 80,5 % corresponde a las personas que no salieron durante los primeros 3 meses del 2023, mientras que un 19,5 % sí salieron en ese periodo. Esto lo vuelve un problema ligeramente desbalanceado, al tener un mayor porcentaje de no salidas versus salidas. Sin embargo, es esperable este comportamiento al tratarse de una métrica de rotación de personal de 3 meses en una compañía de esta industria. Además, un 37,1 % de esas salidas correspondieron a enero, 33,4 % a febrero y 29,5 % a marzo.

Figura 6: Distribución Total y por Mes de Variable Target



(a) Distribución Total Target

(b) Distribución por Mes Target

Fuente: Elaboración propia con datos de SITEL

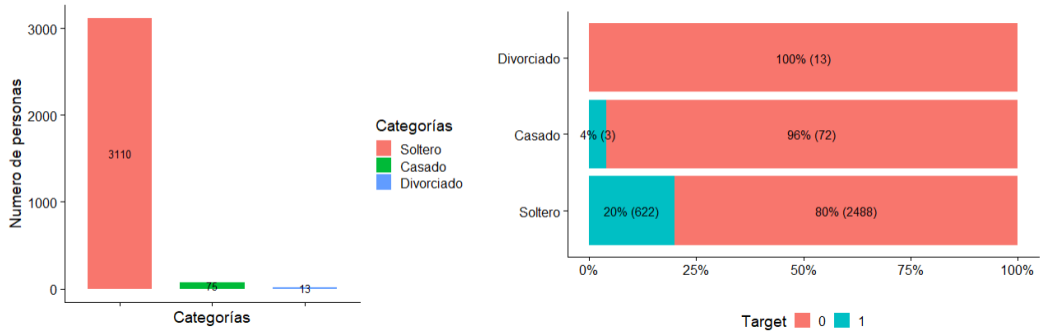
Después de utilizar el método de Stepwise con dirección *backward* descrito en (Thayer, 2002) aplicado a la Regresión Logística, se determinó que las variables

con mayor capacidad predictiva son: Estado Civil, Provincia, Reingreso, Salario, Antigüedad, Sucursal, Teletrabajo y Cliente con un AIC final de 2.825,09.

A continuación se presenta el comportamiento del conjunto de variables con mayor capacidad predictiva y que le aportan mayor información al negocio, por medio de gráficos y un breve análisis cuantitativo. Las demás variables a considerar en la modelación se muestran en la sección de Anexos.

Como se muestra en la Figura 7, un 97,3 % las personas del estudio son solteras, seguidas por 2,3 % casadas y 0,4 % divorciadas. Adicionalmente, se muestra que tiene un muy buen poder predictivo al tener diferentes proporciones en sus distribuciones de salidas versus no salidas en cada una de las categorías.

Figura 7: Distribución y Densidad de Variable Estado Civil



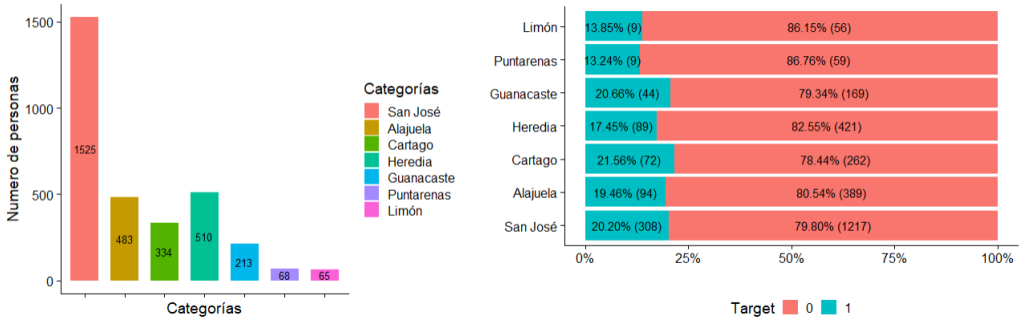
(a) Distribución Estado Civil

(b) Densidad Estado Civil

Fuente: Elaboración propia con datos de SITEL

Por otro lado, la distribución por Provincias está compuesta por un 47,7 % de personas que viven en San José, 15,1 % en Alajuela, 10,5 % en Cartago, 15,9 % en Heredia, 6,7 % en Guanacaste, 2,1 % en Puntarenas y 2,0 % en Limón. Al igual que con la variable anterior, se muestra que también tienen diferencias entre sus categorías.

Figura 8: Distribución y Densidad de Variable Provincia

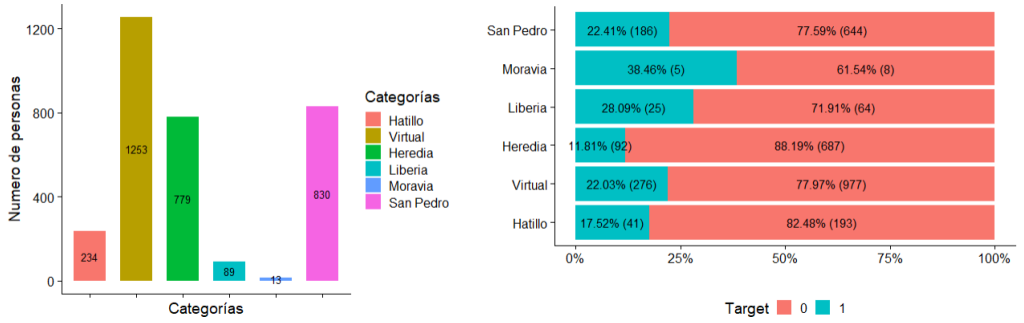


(a) Distribución Provincia (b) Densidad Provincia

Fuente: Elaboración propia con datos de SITEL

La distribución por Sucursal muestra un 7,3 % de personas que están asignados a la sucursal de Hatillo, 39,2 % de forma virtual, 24,3 % en Heredia, 2,8 % en Liberia, 0,4 % en Moravia y 26 % en San Pedro. La densidad por categoría evidencia diferencias muy marcadas entre las composiciones de las mismas, considerándose como muy buena predictora.

Figura 9: Distribución y Densidad de Variable Sucursal



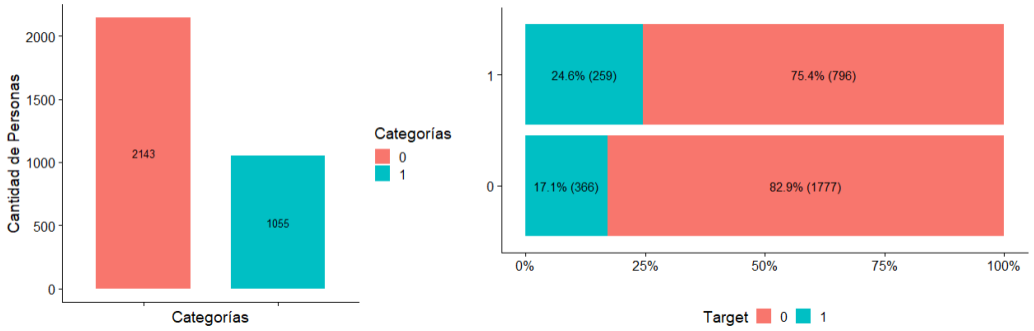
(a) Distribución Sucursal (b) Densidad Sucursal

Fuente: Elaboración propia con datos de SITEL

Un 67 % de personas están clasificadas como que no poseen teletrabajo, mientras que el restante 33 % sí trabajan forma virtual. La densidad por categoría también presenta diferencias entre ambas categorías al agrupar por salidas versus no salidas, por lo que se confirma como una variable predictora muy buena. Resulta interesante que se ve un mayor porcentaje de salidas en los que

sí tienen teletrabajo, por lo que se estudiará su impacto posteriormente en las siguientes secciones al momento de analizar los perfiles que se elaboren.

Figura 10: Distribución y Densidad de Variable Teletrabajo

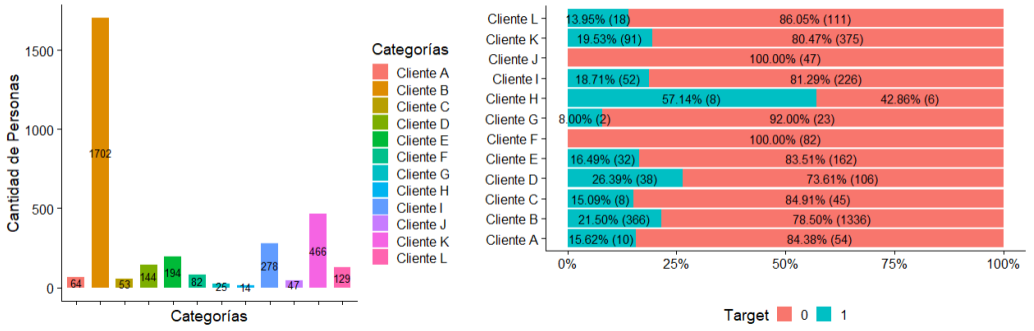


(a) Distribución Teletrabajo (b) Densidad Teletrabajo

Fuente: Elaboración propia con datos de SITEL

Los clientes se clasificaron de forma anónima, donde el más importante de todos (Cliente B) agrupa al 53,2 % de los empleados directos de *customer service*, seguido por el Cliente K con 14,6 %, Cliente I con 8,7 %, Cliente E con 6,1 % y los demás clientes que agrupan menos de 5,0 % cada uno. En su densidad se puede apreciar que también existen diferencias muy marcadas en la composición de las salidas dependiendo del cliente que sea, por lo que es una excelente variable a considerar para el modelo.

Figura 11: Distribución y Densidad de Variable Cliente



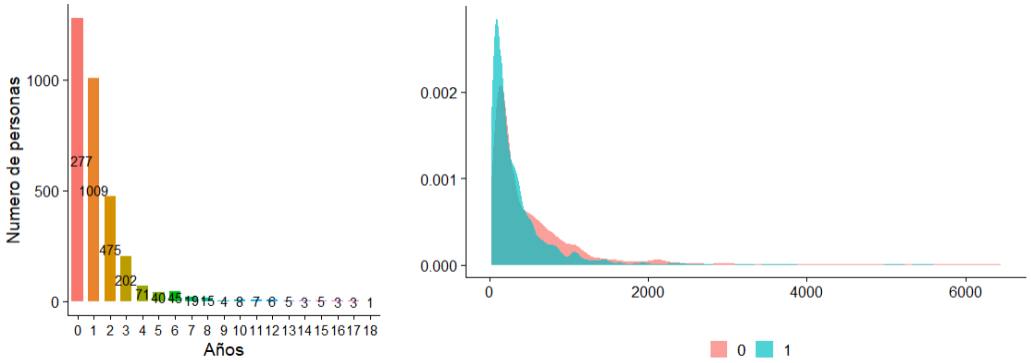
(a) Distribución Cliente (b) Densidad Cliente

Fuente: Elaboración propia con datos de SITEL

La antigüedad al medirse por la cantidad de días laborados de cada persona

en la empresa se considera una variable númerica, que para mejor interpretabilidad del gráfico descriptivo se agruparon a las personas por años cumplidos de laborar. El 39,9 % de los colaboradores tiene menos de un año de estar en la empresa, seguido por el 31,6 % de los que tienen menos de dos años, 14,9 % de los que tienen menos de tres años y el restante representa un 13,6 %. Al ser una variable numérica, su distribución de densidad se representa diferente a las anteriores. Su interpretación sería que las personas que acumulan menos de 365 días suelen salir de la empresa con mucho mayor frecuencia que las personas que no lo hicieron. Sin embargo, después del primer año esas distribuciones se invierten, por lo que se infiere que son mucho más personas que no se van de la empresa que las que sí se van después de ese periodo de tiempo. Es por esto que es una variable que se puede utilizar para la clasificación de las salidas.

Figura 12: Distribución y Densidad de Variable Antigüedad

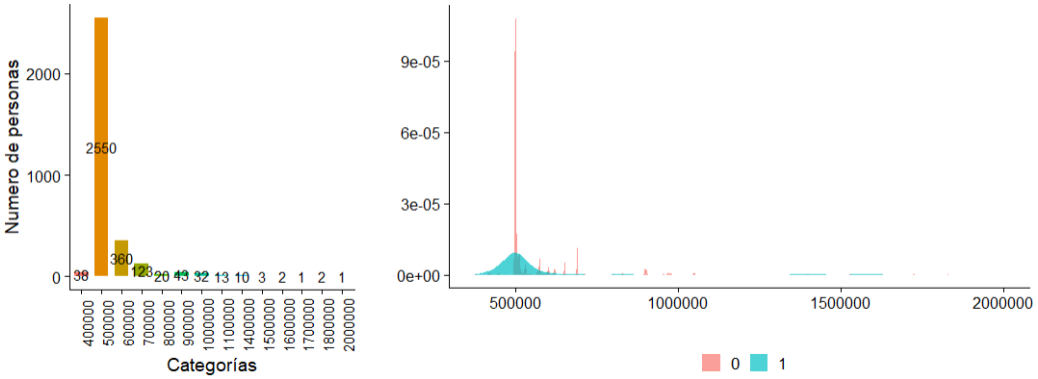


(a) Distribución Antigüedad (b) Densidad Antigüedad

Fuente: Elaboración propia con datos de SITEL

En cuanto a los salarios de los empleados directos de *customer service* se ubican en un 91,0 % de las ocasiones entre 500.000 y 600.000 colones por mes. De igual forma, su distribución de densidad representa que las personas que se fueron de la empresa están dentro de este intervalo y que por encima de este no hay salidas identificadas.

Figura 13: Distribución y Densidad de Variable Salario

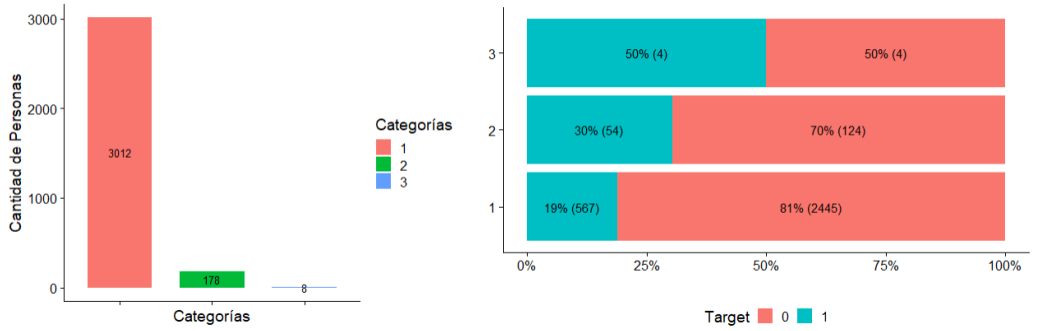


(a) Distribución Salario (b) Densidad Salario

Fuente: Elaboración propia con datos de SITEL

Por último, la variable de Reingreso muestra que un 94,2 % de las personas han trabajado solamente una vez en la compañía, mientras que un 5,5 % han trabajado en 2 periodos distintos y un 0,3 % en 3 periodos. Además, en su distribución de densidad se muestra que es mucho más probable que entre más veces una persona ya hubiera salido y regresado a la empresa anteriormente, más probable es que vuelva a salir una vez más.

Figura 14: Distribución y Densidad de Variable Reingreso



(a) Distribución Reingreso (b) Densidad Reingreso

Fuente: Elaboración propia con datos de SITEL

6. Modelación y Comparación de Resultados

Para pronosticar la rotación de personal de los empleados directos de *customer service* por medio de una probabilidad de salida de la compañía, se ajustaron e implementaron modelos predictivos con diferentes métodos como Regresión Logística, KNN, Árboles de Decisión, Bosques Aleatorios, SVM, *XGBoost* y Consensus.

Como se mencionó en el capítulo anterior, con el análisis descriptivo y la aplicación del método *Backward Stepwise* sobre la Regresión Logística se determinó que las variables con mayor capacidad predictiva son: Estado Civil, Provincia, Reingreso, Salario, Antigüedad, Sucursal, Teletrabajo y Cliente, por lo que serán incluidas en el modelaje, tomando en cuenta otras variables que podrían ser consideradas como valiosas por la empresa, dado el conocimiento en rotación.

El método *Backward Stepwise*, además de identificar las variables predictoras clave, ofrece una estrategia sistemática para simplificar modelos complejos sin sacrificar significativamente la capacidad explicativa. Este enfoque iterativo permite eliminar variables redundantes o con menor contribución, lo que resulta en modelos más parsimoniosos y fáciles de interpretar. Este equilibrio es crucial para generar resultados que sean tanto estadísticamente sólidos como accionables desde una perspectiva de negocio, especialmente en la gestión de rotación de personal.

Adicionalmente se segmentó la población de estudio en dos grupos posterior al uso de *Backward Stepwise*: *training* (datos de entrenamiento para ajuste de los parámetros) y *testing* (datos no vistos por el modelo para determinar si el ajuste es apropiado o existe alguna evidencia de sobreajuste), donde el 75 % de los datos estarán en el primer grupo y el 25 % restante en el segundo grupo.

6.1. Extreme Gradient Boosting

A continuación se detalla el proceso de modelaje considerando ajuste y calibración del método *XGBoost*.

6.1.1. Definición de Parámetros de XGBoost

Como se describirá más adelante, se decidió utilizar *XGBoost*, al presentar mejores resultados que los otros modelos considerados. Según se explica en (Yuan, 2023), el paquete de *R* que implementa la función *xgb.train* se llama *xgboost*. Para esta función se usan diversos parámetros, tales como:

- **Data**: Datos de *training* para entrenar el modelo. Esta función solamente acepta información en formato xgb.DMatrix, por lo que es necesario realizar transformaciones a los datos.
- **Nrounds**: Número máximo de iteraciones. En este caso número de árboles, que estará ejecutando la función para entrenar al modelo.
- **Watchlist**: Lista con los datos de *training* y *testing*. Ambos en formato de xgb.DMatrix.
- **Print_every_n**: Número asignado para definir cada cuantas iteraciones se van a imprimir los mensajes con la información del estado del entrenamiento del modelo. En este caso se seleccionó que se imprimieran los mensajes cada 10 iteraciones.
- **Early_stop_round**: Parámetro de paradas que determina que si el modelo no mejora sus resultados después de n iteraciones consecutivas, terminando prematuramente sin necesidad de recorrer por completo las iteraciones definidas en *nrounds*. Se definió que se detuviera después de 10 iteraciones consecutivas de estabilidad en los resultados.
- **Eval_metric**: Medida para especificar la métrica de evaluación que se utilizará para monitorear el rendimiento del modelo en su etapa de entrenamiento. En el caso de problemas de clasificación binaria se puede utilizar igual a *'error'*, que mide la proporción de errores de clasificación de los árboles generados y se calcula como el número de predicciones incorrectas sobre el número total de observaciones.
- **Maximize**: Ayuda a determinar si la función objetivo debe maximizarse (TRUE) o minimizarse (FALSE). Dado que se estaba evaluando el error, se consideró un problema de minimización.
- **Params**: Lista de parámetros que se determinan el comportamiento del

algoritmo *XGBoost*. A continuación se detalla cada uno de ellos:

- **Booster**: Identifica el modelo base que se entrenará. La opción de *gbtree* se utiliza para modelos que entrenen árboles de decisión para clasificación.
- **Objective**: Determina la función objetivo que se va a optimizar con el modelo. Dependiendo del parámetro que se utilice podría utilizarse para problemas de regresión, clasificación binaria, clasificación multiclase, entre otros.
- **Eta**: Tasa de aprendizaje que controla la contribución de cada nuevo árbol al modelo. Funciona como un regulador para prevenir un sobreajuste del modelo. Varía entre 0 y 1, donde un menor valor evita en mayor medida el sobreajuste, pero se vuelve computacionalmente más lento de ejecutar.
- **Max_Depth**: Profundidad máxima de cada árbol en la fase de entrenamiento; es decir, el número máximo de niveles de nodos desde la ráiz hasta las hojas. Entre mayor sea el valor que se utilice, más complejo y más patrones podrá aprender el modelo, pero tendría un mayor riesgo de sobreajuste.
- **Gamma**: Controla la reducción mínima de la pérdida requerida para dividir un nodo del árbol. La partición del nodo solamente se ejecutará si la reducción de la pérdida es mayor al valor del gamma, por lo que entre mayor sea su valor, más conservador será el algoritmo y se tendrán árboles con una menor profundidad.
- **Colsample_bytree**: Determina la proporción de variables que se deben muestrear para construir los árboles de las iteraciones. De esta manera se puede usar la aleatoriedad para la construcción de los árboles y prevenir el sobreajuste del modelo.
- **Min_child_weight**: Muestras mínimas que se necesitan por cada nodo hoja del árbol. Esto significa que si al aplicar una división de un nodo y terminan menos observaciones clasificadas que el valor del parámetro, entonces no se aplicará la división.
- **Subsample**: Proporción de observaciones que se van a muestrear

para construir cada árbol de las iteraciones del total de observaciones posibles de la base de *training*. Al igual que con Colsample_bytree, permite darle aleatoriedad al modelo para evitar un sobreajuste con la construcción de los árboles.

6.1.2. Calibración de Parámetros

Para definir cúales eran los mejores valores de los parámetros a utilizar, se aplicaron validaciones cruzadas formando 10 particiones distintas de la muestra de calibración de forma aleatoria, donde se tomaron 9 de esos grupos como datos de *training* y el otro grupo se consideró como *testing*. Estos grupos se fueron alternando durante el proceso hasta que cada grupo fue considerado como *testing* exactamente una vez.

En cada una de estas iteraciones se midieron los indicadores de las precisiones y área bajo la curva ROC para comparar posteriormente cúales obtuvieron mejores resultados. Finalmente, el proceso completo se repitió 5 veces con diferentes combinaciones de parámetros para definir cúales eran los óptimos con base en los indicadores seleccionados, abarcando todas las permutaciones posibles.

Además, esta optimización de parámetros se aplicó sobre los datos *training* en 4 grupos diferentes de datos considerando: variables predictoras completas, solamente las variables que fueron seleccionadas como mejores predictoras, eliminando de las mejores predictoras la de mayor capacidad de predicción (Cliente) y también eliminando de las mejores predictoras la de menor capacidad de predicción (Estado Civil).

A continuación se presentan en el Cuadro 2 las distintas combinaciones de variables para cada grupo de estudio, ordenadas de mayor a menor poder predictivo según los resultados del *Backward Stepwise* con respecto al AIC y los p valores de la regresión logística:

Cuadro 2: Combinación de Variables Contempladas para Cada Modelo

Modelo Completo	Modelo Selección	Modelo Selección Sin Mejor	Modelo Selección Sin Peor
Cliente	Cliente	Teletrabajo	Cliente
Teletrabajo	Teletrabajo	Sucursal	Teletrabajo
Sucursal	Sucursal	Antigüedad	Sucursal
Antigüedad	Antigüedad	Salario	Antigüedad
Salario	Salario	Reingreso	Salario
Reingreso	Reingreso	Provincia	Reingreso
Provincia	Provincia	Estado Civil	Provincia
Estado Civil	Estado Civil		
Edad			
Género			
Acciones Disciplinarias			
Nota Evaluación			

Fuente: Elaboración propia

Para todos los 4 escenarios se consideraron esta lista de opciones para definir la optimización de los parámetros:

Cuadro 3: Posibles Valores de Parámetros a Optimizar XGBoost

Nrounds	50	75	100	150
Max_Depth	3	6	9	
Eta	0,3	0,1	0,01	
Gamma	0	1		
Colsample_bytree	0,8	1		
Min_child_weight	1	5	10	
Subsample	0,8	1		

Fuente: Elaboración propia

Estas selecciones de las opciones para los parámetros se basaron en considerar los atributos por defecto que se señalan en (Yuan, 2023), así como algunas modificaciones de los mismos para medir la sensibilidad sobre los resultados finales del modelo. Se consideró que un Max_Depth superior a 9 podría generar una mayor probabilidad de sobreajuste sobre los datos de *training*, al igual que se recomienda tener valores de Eta bajos, en caso que los números máximos de iteraciones sean altos, aunque esto resulte computacionalmente más lento. Gamma considera las opciones que se adaptan según la profundidad del árbol, junto con la cantidad mínima de observaciones permitidas en cada nodo hoja de los árboles.

Después de realizar todas las combinaciones de parámetros posibles y probarlos

por medio distintas validaciones cruzadas, se definieron que los valores del Top 3 de los modelos con mejores indicadores para cada uno de los datos de *training* distintos son los siguientes:

Cuadro 4: Parámetros de los Mejores Top 3 Modelos con base en el Área Bajo la Curva

Modelo	Nrounds	Max_Depth	Eta	Gamma	Colsample_bytree	Min_child_weight	Subsample
XGB Completo 1	150	9	0,1	1	1,0	10	1,0
XGB Completo 2	150	9	0,1	1	0,8	10	0,8
XGB Completo 3	150	9	0,1	0	1,0	10	1,0
XGB Selección 1	75	6	0,3	1	0,8	1	1
XGB Selección 2	150	6	0,3	1	0,8	1	1
XGB Selección 3	75	6	0,1	1	0,8	1	0,8
XGB Selección Sin Mejor 1	50	3	0,3	1	0,8	1	0,8
XGB Selección Sin Mejor 2	150	9	0,01	0	0,8	1	0,8
XGB Selección Sin Mejor 3	75	9	0,01	1	0,8	1	0,8
XGB Selección Sin Peor 1	75	6	0,1	1	1,0	1	0,8
XGB Selección Sin Peor 2	150	3	0,3	1	0,8	1	0,8
XGB Selección Sin Peor 3	100	6	0,1	0	0,8	1	0,8

Fuente: Elaboración propia

6.1.3. Ejecución de los Modelos y sus Predicciones

Para obtener las probabilidades de salida de los modelos desarrollados con los parámetros de la sección anterior, se utilizó la función *predict* sobre los datos de *testing*, indicando en el parámetro *type* la opción "prob". Después, se determinó una probabilidad de corte para cada modelo a partir de un balance idóneo en las precisiones por categoría obtenidas de las matrices de confusión.

Posteriormente, se compararon las probabilidades obtenidas a partir del modelo con la probabilidad de corte establecida para definir la clasificación de si cada una de las observaciones de *testing* se considerarían como salidas o no. En caso que la probabilidad de una observación fuera mayor o igual al umbral de la probabilidad de corte sería tomada como salida y viceversa.

Así que estuvieron todas las clasificaciones de *testing*, se compararon las clasificaciones con los datos reales de si fueron salidas o no, por medio de matrices de confusión. A partir de estas matrices de confusión se pudo calcular la precisión y error global, especificidad y sensibilidad de cada modelo.

Finalmente, se analizaron las distribuciones de las curvas de densidad por cada categoría de salidas y no salidas, se graficaron las curvas ROC y se calcularon

su área bajo la curva para la comparación entre los modelos y definir los que se iban a comparar contra los otros tipos de modelos posteriormente.

6.1.4. Resultados Validación Cruzada

Con base en estos parámetros, se utilizaron 20 validaciones cruzadas con 10 grupos para garantizar que los resultados son consistentes, independientemente de la muestra que se utilice. El proceso para los cálculos de probabilidades y clasificaciones de las observaciones en cada validación cruzada es el mismo que el descrito en la sección anterior.

Los resultados promedios de las métricas de cada uno de los modelos fueron los siguientes:

Cuadro 5: Resumen de los Principales Indicadores para Cada Modelo de XGBoost

Modelo	AUC	Precisión Global	Error Global	Especificidad	Sensibilidad	Probabilidad de Corte
XGB Completo 1	0,8351	0,7415	0,2585	0,7326	0,7785	0,135
XGB Completo 2	0,8353	0,7281	0,2719	0,7129	0,7911	0,130
XGB Completo 3	0,8353	0,7343	0,2657	0,7224	0,7840	0,130
XGB Selección 1	0,7960	0,6932	0,3068	0,6864	0,7225	0,170
XGB Selección 2	0,7978	0,6930	0,3070	0,6866	0,7208	0,170
XGB Selección 3	0,7978	0,6811	0,3189	0,6680	0,7368	0,170
XGB Selección Sin Mejor 1	0,7504	0,6298	0,3702	0,6054	0,7316	0,170
XGB Selección Sin Mejor 2	0,7461	0,6518	0,3482	0,6351	0,7218	0,255
XGB Selección Sin Mejor 3	0,7461	0,6535	0,3465	0,6413	0,7054	0,335
XGB Selección Sin Peor 1	0,7808	0,6752	0,3248	0,6587	0,7448	0,165
XGB Selección Sin Peor 2	0,7814	0,6724	0,3276	0,6590	0,7297	0,160
XGB Selección Sin Peor 3	0,7814	0,6939	0,3061	0,6868	0,7256	0,170

Fuente: Elaboración propia

A partir de los resultados del Cuadro 5, se puede observar que los que utilizaron mayor cantidad de variables tuvieron mejores resultados, con áreas bajo la curva, precisión global, especificidad y sensibilidad. Esto puede deberse a que los modelos con menor cantidad de variables sacrifican en parte estos indicadores con tal de conseguir mayor parsimoniosidad, entendido como la preferencia de modelos menos complejos y que explican los datos de manera similar.

En promedio, los modelos que utilizan solamente las mejores variables predictoras tienen áreas bajo la curva 380bps por debajo de las de los modelos con

todas las variables, así como 455bps menos de precisión global, 423bps menos de especificidad y 578bps menos de sensibilidad.

Otra diferencia es que las probabilidades de corte de los modelos con todas las variables son aproximadamente 0,13 en promedio, mientras que de los modelos que seleccionan variables son 0,17. Esto significa que los modelos que incluyen todas las variables tienen la distribución de las densidades por categorías más separadas una de otra y agrupan las personas activas en los primeros percentiles, a diferencia de los modelos con variables seleccionadas. Por esta misma razón el área bajo la curva resulta mayor, ya que se consideran modelos con una mayor capacidad predictiva, al ajustarse más rápido que los otros.

Además, se puede observar la sensibilidad de excluir la variable Cliente en el tercer grupo que corresponde a la variable predictora más influyente. En promedio, tiene una reducción en las áreas bajo la curva de aproximadamente 877bps con respecto a los mejores modelos y 496bps con respecto a los modelos con las mejores variables seleccionadas. La precisión de las salidas disminuyó 650bps y 72bps respectivamente.

También se muestra que el efecto de remover la variable predictora menos influyente de las mejores variables seleccionadas es menor, puesto que el área bajo la curva se redujo en promedio 541bps con respecto a los mejores modelos y 160bps con respecto a los modelos con las mejores variables seleccionadas. La precisión de las salidas disminuyó 512bps y mejoró 66bps respectivamente.

A partir de los resultados anteriores, se decidió analizar con mayor profundidad los modelos de XGB Completo 2 y XGB Selección 3 para compararlos contra las demás tipos de metodologías que se desarrollaron. Estos serán identificados como *XGBoost* de ahora en adelante, tomando el primero para las secciones con las variables predictoras completas y el segundo para las secciones con las variables predictoras seleccionadas respectivamente. Esta selección se hizo con base tanto en el área bajo la curva, como de las precisiones globales y por categoría.

6.2. Comparación de Metodologías Alternativas

Como se mencionó anteriormente, los mejores modelos seleccionados de *XGBoost* en la sección anterior se compararán contra los de Regresión Logística, KNN, Árboles de Decisión, Bosques Aleatorios, SVM y Consensus. Esto con el fin de poner a competir metodologías y brindarle a la compañía el mejor modelo posible.

Con todos estos modelos se siguieron los mismos pasos que con *XGBoost*, donde una vez que se desarrollaron, se utilizó la función *predict* para obtener las probabilidades, se clasificaron las observaciones con base en ellas, se calcularon las matrices de confusión para determinar las precisiones y finalmente se calcularon las áreas bajo la curva.

La descripción de las funciones para la implementación de cada modelo en el lenguaje de programación *R* es la siguiente:

- **Regresión Logística**: La función *glm* permite establecer una regresión lineal o logística entre las variables predictoras y la variable respuesta. Se consideraron los conjuntos de variables completas y mejores predictoras para cada caso, con los datos de *training* y familia binomial para que se definiera una regresión logística.
- **KNN**: La función *train.knn* crea un modelo de KNN, donde se conforman grupos de vencindarios con base en ciertas características y su distancia con cada vecindario. Se usaron los conjuntos de variables predictoras, los datos de *training*, un máximo de vecindarios que fuera igual a la raíz cuadrada del número de observaciones a utilizar y se probaron diferentes *kernels* (rectangular, triangular, *optimal*, *epanechnikov*, *inv* y *triweight*) y se eligió triangular, al ser el que representaba mejores resultados.
- **Árboles de Decisión**: La función *rpart* se utiliza para ajustar modelos de árboles de decisión basados en criterios de partición, como la reducción de la varianza para regresión o la ganancia de información para clasificación. Esta función se aplicó tanto para el conjunto con todas las variables, como el conjunto con las mejores variables predictoras.
- **Bosques Aleatorios**: La función *train.randomForest* implementa el algoritmo de bosques aleatorios, que ejecuta múltiples árboles de decisión

con los que se toma un consenso para obtener predicciones más robustas y precisas. En este caso se tomaron en cuenta las variables en totalidad, así como la selección de las mejores variables respectivamente y con el parámetro *importance* con valor de TRUE. Este parámetro mide la importancia de las variables predictoras para el rendimiento del modelo.

- **SVM**: La función *svm* permite elaborar modelos de *Support Vector Machines*, donde las predicciones dependen del hiperplano que se genera. Al igual que con los demás, se usaron ambos conjuntos de variables predictoras, distintos tipos de *kernels* (linear, radial, polynomial y sigmoid), siendo el último el que obtuvo mejores resultados. El último parámetro fue *probability* que tuviera un valor de TRUE, para que el modelo estimara las probabilidades de pertenencia a cada clase para todas las observaciones.

- **Consensus**: Por último, se generó el modelo de Consensus, el cual pretende combinar las predicciones de diferentes metodologías. Como se muestra en la Ecuación 27, se tomaron las probabilidades de cada una de los métodos ($P(x_i)$) y se ponderaron según la magnitud del área bajo la curva ROC de cada uno (Pj_i), dando un mayor peso a las probabilidades de los métodos que tenían mayores valores, porque se consideran mejores modelos.

$$P(x)_{Consensus} = \sum_{i}^{n} P(x_i) \cdot Pj_i \tag{27}$$

$$Pj_i = \frac{AUC_i}{\sum_i^n AUC_i}$$

Una vez que se tuvieron los valores de probabilidades ponderados obtenidos por ($P(x)_{Consensus}$), se compararon contra la probabilidad de corte del Consensus para definir las predicciones del modelo de cada observación. Finalmente, se calcularon la matriz de confusión, precisiones y el área bajo la curva ROC para la comparación con los demás modelos.

Este modelo de Consensus se aplicó considerando los resultados de todas las metodologías y también sin el modelo de SVM (que fue el del

ponderador más bajo por los resultados del área bajo la curva ROC). Los modelos con ponderaciones más altas fueron *XGBoost* y Bosques Aleatorios, respectivamente.

Al igual que con los de *XGBoost*, se aplicaron 20 validaciones cruzadas con 10 pliegues, para garantizar que los resultados son consistentes, independientemente de la muestra que se utilice, separando la base en 75 % de *training* y 25 % *testing*. Además, se compararon los modelos tanto en su versión con todas las variables, así como con las variables seleccionadas como mejores predictoras.

6.2.1. Modelos con Variables Completas

A continuación se muestran los resultados de los promedios de las validaciones cruzadas para cada uno de los modelos:

Cuadro 6: Resumen Principales Indicadores Modelos con Variables Completas

Método	AUC	Precisión Global	Error Global	Especificidad	Sensibilidad	F1 Score	Probabilidad de Corte
XGBoost	0,8434	0,7259	0,2741	0,7104	0,7911	0,5436	0,130
Bosques Aleatorios	0,8286	0,7434	0,2566	0,7397	0,7595	0,5329	0,160
Consensus	0,8346	0,7474	0,2526	0,7463	0,7534	0,5471	0,180
Árbol de Decisión	0,7558	0,4635	0,5365	0,3596	0,8914	0,4322	0,125
Regresión Logística	0,7207	0,6522	0,3478	0,6560	0,6383	0,4672	0,215
KNN	0,7130	0,6519	0,3481	0,6513	0,6556	0,4623	0,220
SVM	0,6515	0,5773	0,4227	0,5661	0,6329	0,3917	0,200

Fuente: Elaboración propia

Con base en el Cuadro 6, se puede concluir que los modelos que tienen mejores resultados con base en el área bajo la curva, precisión global, especificidad y sensibilidad son *XGBoost*, Bosques Aleatorios y Consensus.

En el caso de *XGBoost* tiene un área bajo la curva 794bps mayor que el promedio, 148bps que Bosques Aleatorios y 88bps que Consensus. Además, su precisión global es 742bps superior al promedio, 176bps inferior a Bosques Aleatorios y 216bps al de Consensus. Esto se explica porque *XGBoost* tiene una mayor precisión en cuanto a Sensibilidad, que es 594bps mayor que el promedio, 316bps que Bosques Aleatorios y 377bps que Consensus. Por otro lado *XGBoost* tiene una menor precisión en cuanto a Especificidad, que es 777bps mayor que el promedio, pero 293bps inferior que Bosques Aleatorios y 359bps que Consensus.

La Sensibilidad podría considerarse el principal indicador de las precisiones, ya que determina el porcentaje de acierto sobre las salidas. Al ser una base desbalanceada con más personas que no salieron de la compañía que las que sí lo hicieron, la Especificidad va a ponderar en mayor medida la Precisión Global que la Sensibilidad.

Además, en el Cuadro 6 tanto en el modelo de *XGBoost* como Bosques Aleatorios se usaron de las menores probabilidades de corte. Esto evidencia lo explicado anteriormente, de que ambos modelos logran una mejor distinción de sus curvas de densidades, así como que logran esta distinción más rápido que los demás métodos. Es por esta razón que tienen una mayor área bajo la curva ROC.

Finalmente, podría asegurarse que, con base en las validaciones cruzadas, *XGBoost* es el modelo con mejores resultados, seguido por Bosques Aleatorios y Consensus. Esto fortalece la idea que esta metodología tiene ventajas sobre los demás, al combinar modelos más débiles, como los árboles de decisión, para desarrollar un modelo más preciso, previene un sobreajuste por medio del término de regularización, permite optimización de hiperparámetros, selección de variables, manejo de datos desbalanceados y por esto se obtienen mejores resultados.

6.2.2. Modelos con Variables Predictoras Seleccionadas

A continuación se muestran los resultados de los promedios de las validaciones cruzadas de los modelos con variables seleccionadas:

Cuadro 7: Resumen Principales Indicadores Modelos con Variables Seleccionadas

Método	AUC	Precisión Global	Error Global	Especificidad	Sensibilidad	F1 Score	Probabilidad de Corte
XGBoost	0,8000	0,6817	0,3183	0,6686	0,7373	0,5127	0,170
Consensus	0,7766	0,6425	0,3575	0,6184	0,7434	0,4831	0,150
Bosques Aleatorios	0,7372	0,6515	0,3485	0,6537	0,6437	0,4362	0,025
KNN	0,7349	0,6275	0,3725	0,6095	0,7028	0,4382	0,175
Regresión Logística	0,7212	0,6483	0,3517	0,6467	0,6559	0,4672	0,215
SVM	0,6721	0,5968	0,4032	0,5960	0,6163	0,4017	0,200
Árbol de Desición	0,6492	0,4491	0,5509	0,3693	0,7830	0,3483	0,150

Fuente: Elaboración propia

En el Cuadro 7 se muestra una situación similar que con el Cuadro 6, donde *XGBoost*, Consensus y Bosques Aleatorios son los mejores modelos a partir de los resultados del área bajo la curva, precisión global, especificidad y sensibilidad.

XGBoost tiene un área bajo la curva 727bps mayor que el promedio, 234bps que Consensus y 234bps que Bosques Aleatorios. En este caso su precisión global es superior tanto al promedio por 678bps, Consensus por 392bps y Bosques Aleatorios por 302bps. La Sensibilidad registrada de *XGBoost* también supera al promedio con una diferencia de 740bps y Bosques Aleatorios de 937bps, inferior por 60bps con respecto a Consensus. Sin embargo, utilizando solamente las mejores variables predictoras *XGBoost* sí es superior en Especificidad a todos, mejorando en 740bps con el promedio, 501bps con Consensus y 149bps con Bosques Aleatorios.

Cabe destacar que todas las metodologías empleadas tuvieron una desmejora al reducir el número de variables predictoras, por lo que podría considerarse que es el costo de conseguir modelos más parsimoniosos. Sin embargo, tanto en el caso de los modelos usando todas las variables como solamente las mejores variables predictoras, el que obtuvo mejores resultados fue *XGBoost*. De hecho, *XGBoost* fue el único que registro un promedio de áreas bajo la curva superior a 0.80 con ambos grupos de datos.

Por lo tanto, podría asegurarse que, con base en las validaciones cruzadas, resulta *XGBoost* la metodología con resultado más satisfactorios en términos de ajuste y precisión, seguido por Consensus y Bosques Aleatorios, a la hora de utilizar solamente las mejores variables predictoras.

6.3. Modelos Finales Seleccionados

Esta sección pretende comparar los modelos finalistas de *XGBoost* con mayor profundidad, ya que se comprobó previamente que obtuvieron mejores resultados que las otras metodologías.

6.3.1. Resultados Modelos Finales Seleccionados

A continuación se presentan las matrices de confusión de los modelo de *XGBoost* utilizando todas las variables predictoras posibles (Completo) con una probabilidad de corte de 0.13 y la selección de las mejores predictoras (Selección) con una probabilidad de corte de 0.17 respectivamente:

Cuadro 8: Matrices de Confusión Modelos XGBoost

(a) Variables Predictoras Completas

Matriz de Confusión	Predicción No Salida	Predicción Salida
Clase No Salida	455	179
Clase Salida	43	122

(b) Variables Predictoras Seleccionadas

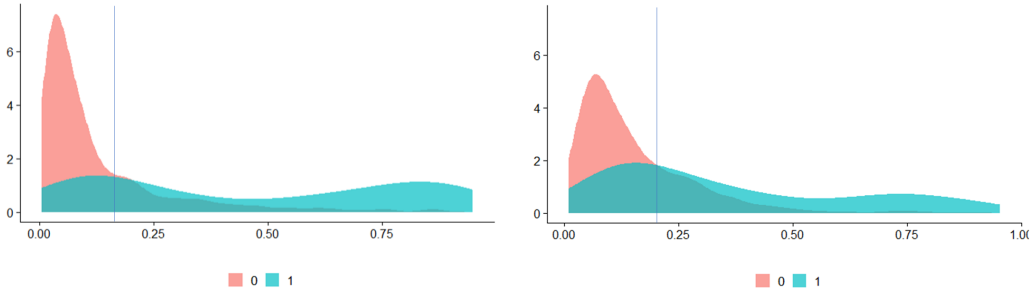
Matriz de Confusión	Predicción No Salida	Predicción Salida
Clase No Salida	444	190
Clase Salida	49	116

Fuente: Elaboración propia

De esta manera, se puede comprobar numéricamente que el modelo que contempla todas las variables tiene una mayor precisión al clasificar tanto las salidas como las no salidas. Estas probabilidades de corte definidas pueden llegar a cambiar según la capacidad instalada de la compañía para poder atender a todas las personas que el modelo pronostique como una posible salida.

Las distribuciones de densidad de cada categoría están representadas la Figura 15, siendo 1 las salidas y 0 las demás:

Figura 15: Distribuciones de Densidad por Categoría Modelos XGBoost



(a) Variables Predictoras Completas (b) Variables Predictoras Seleccionadas

Fuente: Elaboración propia

Antes de la probabilidad de corte, marcada con una línea azul vertical, las distribuciones de densidad de las personas que no salieron de la compañía son mayores a las de las densidades de las salidas. Sin embargo, después de las

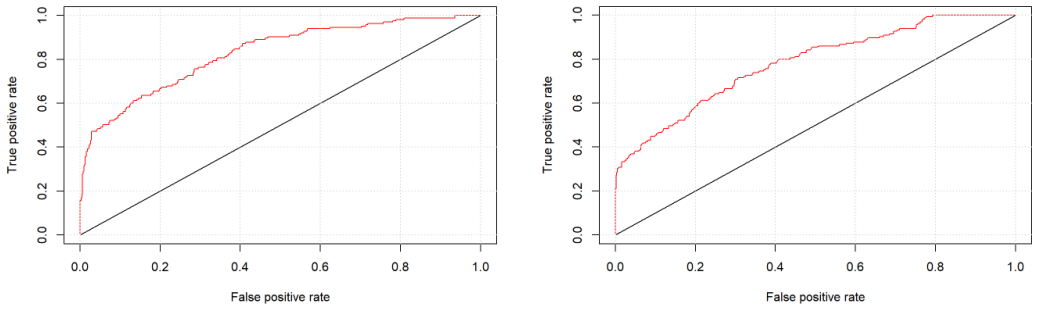
probabilidades de corte las distribuciones de densidad de la categoría salidas superan en todo momento a las no salidas.

Esto es algo positivo, ya que entre mayor sea la separación entre ambas densidades, el modelo puede discriminar mejor las observaciones a predecir, porque contienen características distintas entre sí. Además, se corre un menor riesgo de que el modelo pronostique de forma incorrecta si una observación pertenece a una clase o la otra, ya que si ambas densidades estuvieran sobrepuestas, existiría la posibilidad de que la clasificación sea cualquiera de las dos categorías para esa probabilidad en específico.

Es por esta razón que la sensibilidad y especificidad se incrementan al clasificar de mejor manera, por lo que se consideran modelos más robustos y se facilita la selección de una probabilidad de corte, al considerar que a partir de cierto punto se puede garantizar en mayor medida que bajo una probabilidad asociada sean consideradas de una u otra forma.

En la Figura 16 se presentan los gráficos de las curvas ROC resultantes de ambos modelos:

Figura 16: Curvas ROC Modelos XGBoost



(a) Variables Predictoras Completas (b) Variables Predictoras Seleccionadas

Fuente: Elaboración propia

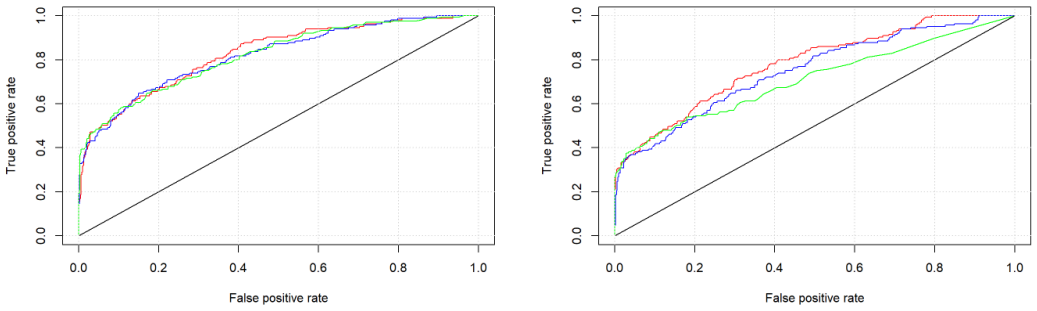
La idea con estas curvas es que se logre abarcar la mayor cantidad de área bajo la curva posible, porque significa que el modelo tiene una mayor capacidad discriminativa, al distinguir entre las salidas y no salidas. De esta manera se consideran modelos más robustos que garantizan mantener un buen rendimiento con las nuevas observaciones que se incluyan en el futuro. Sin embargo, el punto más importante es que se puede asegurar que existe una menor

dependencia de la probabilidad de corte utilizada al analizarse la capacidad predictiva del modelo independientemente del umbral.

En el caso del modelo que considera todas las variables predictoras posibles registró un área bajo la curva ROC de 0.8434, mientras que el modelo con variables seleccionadas fue de 0.8000. Ambos resultados se consideran que son buenos resultados para un modelo predictivo.

Finalmente, en la Figura 17 se comparan las curvas ROC de los modelos de *XGBoost* en color rojo, con los de Consenso en color azul y Bosques Aleatorios en color verde. En ambas gráficas la curva de *XGBoost* supera a las demás y por esta razón es que su área bajo la curva es mayor, como se mostró en la sección anterior.

Figura 17: Curvas ROC Mejores 3 Metodologías



(a) Variables Predictoras Completas (b) Variables Predictoras Seleccionadas

Fuente: Elaboración propia

6.4. Comparación de Perfiles según Rendimiento del Modelo

En esta sección se pretende determinar la calidad y capacidad de los modelos calibrados, a partir de estudiar los perfiles de las personas que son clasificadas de manera óptima con base en las predicciones de salidas a 3 meses de los modelos de *XGBoost*, así como identificar aquellos perfiles para los cuales el modelo presenta desafíos en su clasificación. Para esto se aplicó Clustering Jerárquico para conformar los diferentes grupos y también se incluyó una variable adicional que determina si la persona se fue de la empresa en algún momento, aunque no fuera en el intervalo de los 3 meses del estudio.

6.4.1. Modelo XGBoost con Variables Predictoras Completas

Con base en el dendograma de la Figura 18 y el Cuadro 9, se observa que el clustering conformó 4 grupos distintos, siendo el más grande el Grupo 1 con 110 personas, donde se acertó en un 99 % que las personas iban a salir en los siguientes 3 meses de la compañía, mientras que el Grupo 2 contiene 20 personas que un 20 % fueron salidas, el Grupo 3 68 personas que un 13 % fueron salidas y el Grupo 4 incluye 103 personas de las cuales ninguna se acertó que fueran a salir en los siguientes 3 meses, ya que todas permanecieron dentro de la empresa durante ese periodo.

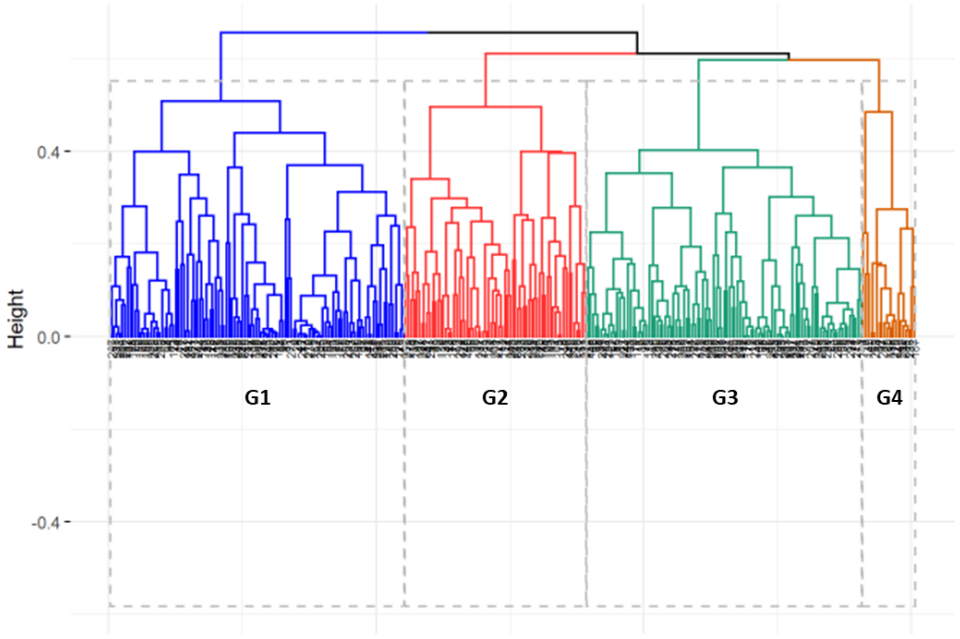
Por lo anterior, se considera que los mayores grupos de interés son el 1 y el 4, ya que tienen mayor y menor cantidad de aciertos de salidas de la población de estudio respectivamente.

Estos grupos de mayor interés tienen los mayores promedios de acciones disciplinarias, así como antigüedad intermedia en comparación con el Grupo 2 y Grupo 3, edad cercana a los 24 años, mayores notas de evaluación que con respecto a los Grupos 2 y 3, menor número de reingreso y menor salario que los otros. Además, tanto el Grupo 1 como el Grupo 4 consideran personas de los mismos clientes (un 90 % de los clientes del Grupo 1 son el 100 % de los clientes del Grupo 4).

Las diferencias más marcadas entre los Grupos 1 y 4 son que el Grupo 1 estaba conformado en un 54 % por hombres, mientras que el Grupo 4 estaba conformado en un 52 % por mujeres. Sin embargo, la mayor diferencia es que el Grupo 1 son las personas con menor salario de los 4 grupos y que tienen mayor porcentaje de teletrabajo disponible, excepto por el Grupo 3.

Por lo que se concluye que para un perfil de personas con un menor salario, mejores notas de desempeño y mayor porcentaje de teletrabajo, el modelo tiene un mejor rendimiento predictivo a la hora de clasificar a las personas como un riesgo de salida; es decir, que fueran clasificadas como salidas y efectivamente terminaron saliendo durante los siguientes 3 meses.

Figura 18: Dendograma de Clasificaciones Salidas con Variables Predictoras Completas



Fuente: Elaboración propia

Cuadro 9: Resumen del Clustering Jerárquico de las Personas Clasificadas como Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	110	20	68	103
Acciones Disciplinarias	$\overline{X} = 0{,}75$	$\overline{X} = 0{,}25$	$\overline{X} = 0{,}21$	$\overline{X} = 0{,}81$
Antigüedad	$\overline{X} = 313$	$\overline{X} = 480$	$\overline{X} = 256$	$\overline{X} = 306$
Cliente	Cliente B: 76 % Cliente K: 12 % Cliente D: 4 % Otros: 10 %	Cliente D: 85 % Cliente H: 15 %	Cliente I: 37 % Cliente K: 31 % Cliente L: 18 % Otros: 14 %	Cliente B: 97 % Cliente D: 2 % Cliente K: 1 %
Edad	$\overline{X} = 24{,}0$	$\overline{X} = 23{,}0$	$\overline{X} = 26{,}3$	$\overline{X} = 23{,}7$
Estado Civil	Solteros: 100 %	Solteros: 100 %	Solteros: 100 %	Solteros: 99 % Casados: 1 %
Género	M: 54 % F: 46 %	M: 50 % F: 50 %	M: 62 % F: 38 %	M: 48 % F: 52 %
Nota Evaluación	$\overline{X} = 3{,}15$	$\overline{X} = 2{,}92$	$\overline{X} = 2{,}42$	$\overline{X} = 3{,}32$
Provincia	San José: 54 % Alajuela: 14 % Heredia: 13 % Otros: 20 %	San José: 80 % Heredia: 10 % Alajuela: 5 % Cartago: 5 %	San José: 40 % Alajuela: 23 % Cartago: 15 % Otros: 22 %	San José: 40 % Heredia: 20 % Guanacaste: 15 % Otros: 25 %
Reingreso	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}10$	$\overline{X} = 1{,}20$	$\overline{X} = 1{,}03$
Salario	$\overline{X} = 491.976$	$\overline{X} = 635.000$	$\overline{X} = 532.827$	$\overline{X} = 499.205$
Sucursal	Virtual: 43 % San Pedro: 39 % Heredia: 9 % Otros: 9 %	Hatillo: 85 % Virtual: 10 % Heredia: 5 %	Virtual: 100 %	San Pedro: 49 % Heredia: 32 % Liberia: 13 % Otros: 6 %
Teletrabajo	Sí: 39 % No: 61 %	Sí: 10 % No: 90 %	Sí: 100 %	No: 100 %
Salidas Reales 3 Meses	99 %	20 %	13 %	0 %
Salidas Reales Indefinido	100 %	60 %	34 %	29 %

Fuente: Elaboración propia

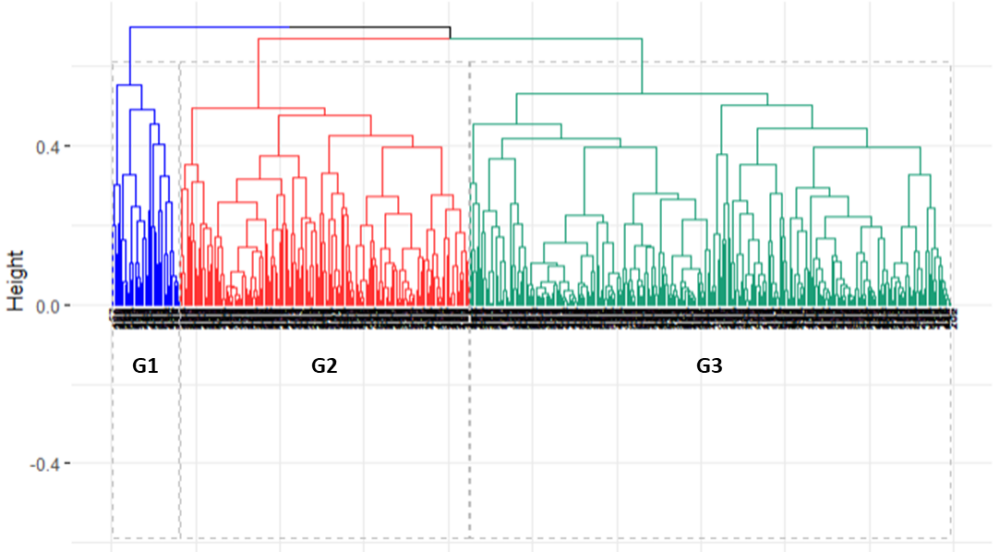
Para el caso del dendograma de la Figura 19 y el Cuadro 10, el clustering jerárquico conformó 3 grupos distintos, siendo el más grande el Grupo 1 con 286 personas, donde se acertó en un 99 % que las personas no iban a salir en los siguientes 3 meses de la compañía, mientras que el Grupo 2 está conformado por 172 personas que un 91 % no fueron salidas y el Grupo 3 40 personas que un 37 % no fueron salidas en los siguientes 3 meses.

Por lo anterior, se considera que los Grupos 1 y 2 tuvieron rendimientos del modelo bastante acertados, mientras que el perfil de las personas del Grupo 3 tuvo peores resultados.

Al aplicar un ejercicio similar al del dendograma y cuadro anterior se puede notar que las personas del Grupo 3 son casi en su totalidad hombres con 85 %, siendo el grupo con menor nota de evaluación con 2,08, mayor número de personas viviendo en San José con 65 %, mayor número de Reingreso 1,15, mayor salario en promedio y menor disponibilidad de teletrabajo con 15 %.

Por lo que se concluye que para un perfil de personas hombres, con un mayor salario, peores notas de desempeño, menor porcentaje de teletrabajo y mayor número de reingresos, el modelo tiene un menor rendimiento predictivo a la hora de clasificar a las personas como un riesgo de salida; es decir, que fueran clasificadas como no salidas y sin embargo terminaron saliendo durante los siguientes 3 meses. Esto hace concordancia con las conclusiones de las clasificaciones del clustering anterior.

Figura 19: Dendograma de Clasificaciones No Salidas con Variables Predictoras Completas



Fuente: Elaboración propia

Cuadro 10: Resumen del Clustering Jerárquico de las Personas Clasificadas como No Salidas

Variable	Grupo 1	Grupo 2	Grupo 3
Observaciones	286	172	40
Acciones Disciplinarias	$\overline{X}$ = 1,96	$\overline{X}$ = 2,75	$\overline{X}$ = 2,08
Antigüedad	$\overline{X}$ = 719	$\overline{X}$ = 496	$\overline{X}$ = 609
Cliente	Cliente B: 82 % Cliente F: 6 % Cliente J: 5 % Otros: 7 %	Cliente K: 28 % Cliente I: 24 % Cliente E: 20 % Otros: 28 %	Cliente B: 55 % Cliente H: 13 % Cliente K: 12 % Otros: 20 %
Edad	$\overline{X}$ = 27,9	$\overline{X}$ = 28,3	$\overline{X}$ = 27,5
Estado Civil	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 96 % Casados: 3 % Divorciados: 1 %	Solteros: 95 % Casados: 5 %
Género	M: 61 % F: 39 %	M: 58 % F: 42 %	M: 85 % F: 15 %
Nota Evaluación	$\overline{X}$ = 3,50	$\overline{X}$ = 2,75	$\overline{X}$ = 2,08
Provincia	San José: 46 % Heredia: 22 % Alajuela: 14 % Otros: 18 %	San José: 53 % Alajuela: 15 % Heredia: 11 % Cartago: 21 %	San José: 65 % Alajuela: 13 % Heredia: 12 % Cartago: 10 %
Reingreso	$\overline{X}$ = 1,02	$\overline{X}$ = 1,02	$\overline{X}$ = 1,15
Salario	$\overline{X}$ = 552.167	$\overline{X}$ = 552.451	$\overline{X}$ = 650.162
Sucursal	Heredia: 49 % San Pedro: 38 % Hatillo: 11 % Otros: 2 %	Virtual: 100 %	San Pedro: 38 % Virtual: 30 % Heredia: 17 % Hatillo: 15 %
Teletrabajo	No: 100 %	Sí: 84 % No: 16 %	Sí: 15 % No: 85 %
Salidas Reales 3 Meses	1 %	9 %	63 %
Salidas Reales Indefinido	20 %	19 %	98 %

Fuente: Elaboración propia

Los dendogramas y tablas con la información de los clusterings, desde el punto de vista de personas bien clasificadas versus mal clasificadas con variables completas, pueden ser consultados en la sección de anexos en las Figuras 35, 19 y los Cuadros 18 y 19.

6.4.2. Modelo XGBoost con Variables Predictoras Seleccionadas

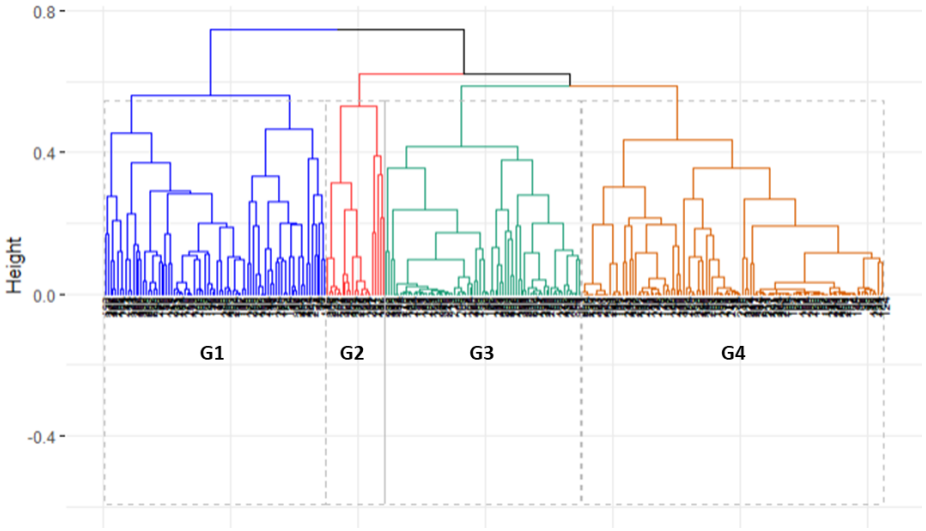
Al considerar solamente a las variables mejores predictoras se obtuvo el dendograma de la Figura 20 y el Cuadro 11, conformando 4 grupos distintos. El primero de ellos contempla 77 personas, de las cuales 97 % salieron de la empresa en los siguientes 3 meses del estudio, el segundo 23 personas con 30 % de salidas, el tercero 87 personas con 34 % de salidas y el cuarto 119 con 3 % de salidas.

Aunque solamente el Grupo 1 tuvo un alto rendimiento del modelo en cuanto a su clasificación de salidas, se puede notar que en los Grupos 2 y 3 las personas terminaron saliendo de la empresa en un 74 % y 55 % respectivamente, por lo que, aunque no fueron un riesgo en los 3 meses siguientes, sí se detectó una eventual salida en su mayoría.

Las personas de los Grupos 1 y 4 son las que tienen una antigüedad intermedia en comparación con los Grupos 2 y 3, siendo conformados en su totalidad por los clientes B e I, menores números de reingreso y salarios. La principal diferencia entre el Grupo 1 y el Grupo 4 es que el primero incorpora personas con posibilidad de teletrabajo y que no incorpora personas de afuera de la Gran Área Metropoitana (GAM).

Por lo que se concluye que para un perfil de personas con un menor salario, que viven en la GAM y con posibilidad de teletrabajo, el modelo tiene un menor rendimiento predictivo a la hora de clasificar a las personas como un riesgo de salida; es decir, que fueran clasificadas como salidas y efectivamente terminaron saliendo durante los siguientes 3 meses.

Figura 20: Dendograma de Clasificaciones Salidas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Cuadro 11: Resumen del Clustering Jerárquico de las Personas Clasificadas como Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	77	23	87	119
Antigüedad	$\overline{X} = 374$	$\overline{X} = 458$	$\overline{X} = 213$	$\overline{X} = 303$
Cliente	Cliente B: 97 % Cliente I: 3 %	Cliente D: 78 % Cliente H: 13 % Cliente K: 9 %	Cliente K: 42 % Cliente I: 32 % Cliente L: 9 % Otros: 17 %	Cliente B: 99 % Cliente I: 1 %
Estado Civil	Solteros: 100 %	Solteros: 100 %	Solteros: 100 %	Solteros: 99 % Casados: 1 %
Provincia	San José: 56 % Heredia: 20 % Alajuela: 14 % Otros: 10 %	San José: 79 % Alajuela: 13 % Cartago: 4 % Heredia: 4 %	San José: 40 % Alajuela: 24 % Cartago: 16 % Otros: 20 %	San José: 48 % Heredia: 16 % Guanacaste: 15 % Otros: 21 %
Reingreso	$\overline{X} = 1{,}04$	$\overline{X} = 1{,}09$	$\overline{X} = 1{,}17$	$\overline{X} = 1{,}03$
Salario	$\overline{X} = 486.395$	$\overline{X} = 610.870$	$\overline{X} = 559.901$	$\overline{X} = 498.544$
Sucursal	San Pedro: 52 % Virtual: 34 % Heredia: 13 % Otros: 1 %	Hatillo: 78 % Moravia: 18 % Heredia: 4 %	Virtual: 100 %	San Pedro: 64 % Heredia: 19 % Liberia: 14 % Otros: 3 %
Teletrabajo	Sí: 27 % No: 73 %	No: 100 %	Sí: 100 %	No: 100 %
Salidas Reales 3 Meses	97 %	30 %	34 %	3 %
Salidas Reales Indefinido	100 %	74 %	55 %	24 %

Fuente: Elaboración propia

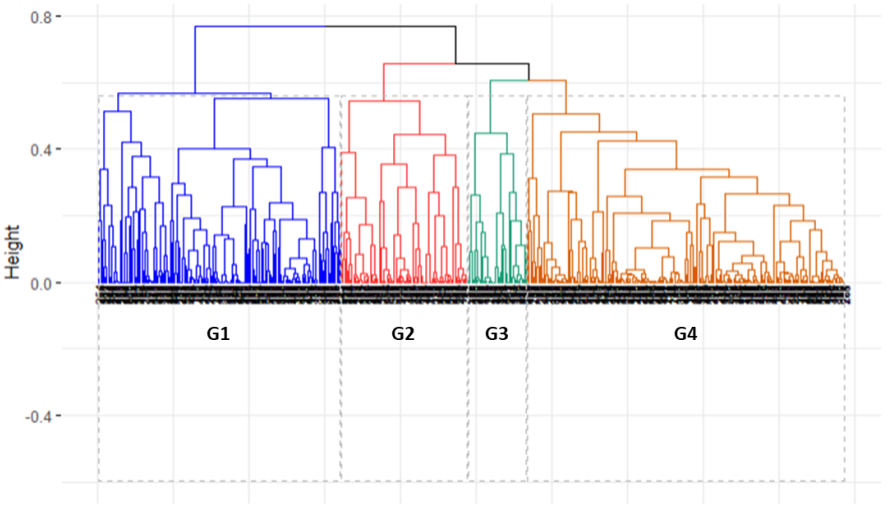
Para el caso del dendograma de la Figura 21 y el Cuadro 12, el clustering jerárquico conformó 4 grupos distintos, siendo el más grande el Grupo 1 con 210 personas, donde se acertó que ninguna de las personas iban a salir en los siguientes 3 meses de la compañía, mientras que el Grupo 2 contiene 84 personas que un 70 % no fueron salidas, Grupo 3 39 personas con 90 % de no salidas y el Grupo 4 160 personas que un 87 % no fueron salidas en los siguientes 3 meses.

Por lo anterior, se considera que los Grupos 1, 3 y 4 tuvieron rendimientos del modelo bastante acertados, mientras que el perfil de las personas del Grupo 2 tuvo peores resultados.

Las personas del Grupo 2 pertenecen en un 97 % al Cliente B, son las que tienen menor antigüedad laboral, menor promedio de reingreso, menor salario y no tienen teletrabajo. Esto vuelve a confirmar que personas con estas características son más propensas a salir de la empresa que las de los otros grupos, como se indicó en los análisis anteriores.

Cabe destacar que el rendimiento del modelo de *XGBoost* fue bastante satisfactorio a lo largo de todos los grupos, ya que el perfil con mayor desacierto tuvo solamente un 30 % de personas clasificadas como que no se iban a ir y realmente se fueron y los otros grupos con un 13 % o menos, lo que es la combinación de clasificación versus resultado más crítica de predecir correctamente.

Figura 21: Dendograma de Clasificaciones No Salidas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Cuadro 12: Resumen del Clustering Jerárquico de las Personas Clasificadas como No Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	210	84	39	160
Antigüedad	$\overline{X} = 774$	$\overline{X} = 452$	$\overline{X} = 523$	$\overline{X} = 461$
Cliente	Cliente B: 78 % Cliente K: 14 % Cliente J: 7 % Otros: 1 %	Cliente B: 97 % Cliente C: 1 % Cliente J: 1 % Cliente K: 1 %	Cliente D: 54 % Cliente F: 43 % Cliente E: 3 %	Cliente E: 24 % Cliente I: 24 % Cliente L: 19 % Otros: 33 %
Estado Civil	Solteros: 94 % Casados: 5 % Divorciados: 1 %	Solteros: 96 % Casados: 4 %	Solteros: 100 %	Solteros: 96 % Casados: 3 % Divorciados: 1 %
Provincia	San José: 44 % Heredia: 28 % Alajuela: 17 % Otros: 11 %	San José: 39 % Alajuela: 24 % Cartago: 16 % Otros: 21 %	San José: 69 % Alajuela: 21 % Cartago: 5 % Heredia: 5 %	San José: 54 % Alajuela: 14 % Guanacaste: 12 % Otros: 20 %
Reingreso	$\overline{X} = 1{,}02$	$\overline{X} = 1{,}01$	$\overline{X} = 1{,}05$	$\overline{X} = 1{,}06$
Salario	$\overline{X} = 534.216$	$\overline{X} = 516.892$	$\overline{X} = 688.718$	$\overline{X} = 571.317$
Sucursal	Heredia: 57 % San Pedro: 29 % Virtual: 14 %	San Pedro: 49 % Heredia: 43 % Liberia: 6 % Virtual: 2 %	Hatillo: 100 %	Virtual: 99 % Hatillo: 1 %
Teletrabajo	No: 100 %	No: 100 %	No: 100 %	Sí: 96 % No: 4 %
Salidas Reales 3 Meses	0 %	30 %	10 %	13 %
Salidas Reales Indefinido	0 %	87 %	36 %	28 %

Fuente: Elaboración propia

De igual manera, los dendogramas y tablas con la información de los clusterings desde el punto de vista de personas bien clasificadas versus mal clasificadas con variables seleccionadas pueden ser consultados en la sección de anexos en las Figuras 37, 38 y los Cuadros 20 y 21.

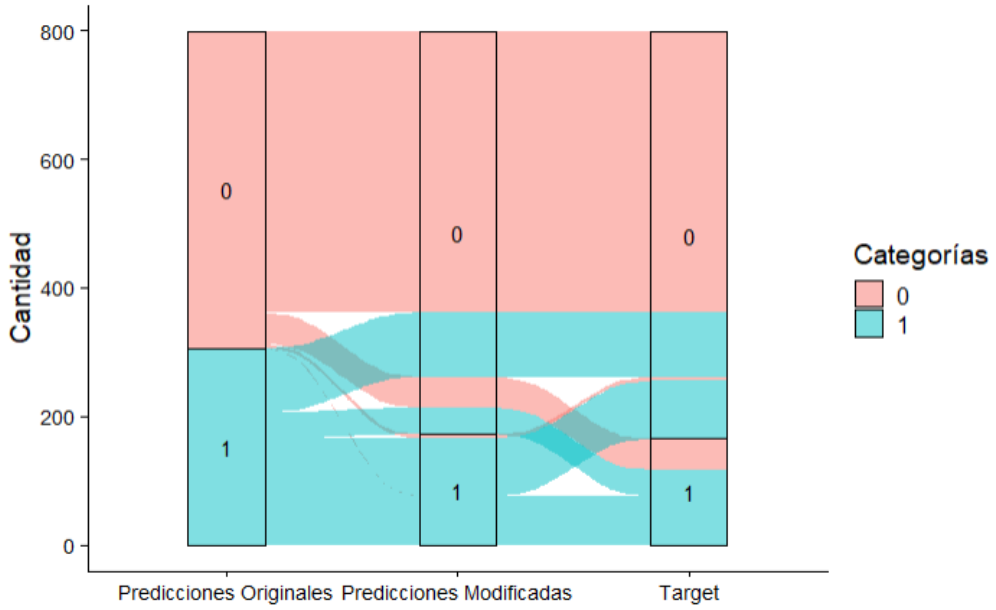
6.5. Análisis de Sensibilidad

Para medir el nivel de precisión del modelo ante circunstancias distintas, se aplicó un análisis de sensibilidad por medio de estresar la variable del salario. Se estudiaron los cambios de las predicciones al modificar en un 5 %, 10 % y 20 % adicional el salario de los empleados para definir perfiles más afectados y medir el impacto de las precisiones con el fin de determinar que ante un cambio en las políticas de compensación de la empresa, se debería de recalibrar el modelo para no perder poder predictivo del mismo.

6.5.1. Variación de la Variable Salario

Con un incremento en el salario del 5 %, se observa en la Figura 22 que no hubo mayores cambios en las predicciones de las personas que se habían clasificado originalmente como no salidas de forma incorrecta, ya que casi todas siguen estando en esa misma categoría en el modelo con los datos estresados. Estas son el grupo que originalmente eran un 0, siguieron siendo 0 en las nuevas predicciones, pero en realidad eran un 1 en la variable *Target*.

Figura 22: Predicciones al Modificar un 5 % el Salario



Fuente: Elaboración propia

También se puede identificar que hay un grupo de personas que se clasificaron correctamente como salidas por el modelo inicialmente, que al modificar los valores del salario, se clasificaron como que no representaban un riesgo de salida, cuando en realidad sí terminaron abandonando a la empresa. Este es el grupo de 41 personas que se considera más crítico, ya que originalmente se habían clasificado como salidas correctamente, pero al modificar su salario se clasificaron de forma incorrecta. Se pueden identificar en la Figura 22 como el grupo que originalmente eran un 1, pasaron a ser un 0 y tenían un valor de 1 en la variable *Target*.

Estos son colaboradores que un 93 % viven en la GAM (44 % en San José), un 66 % de los casos no tenían posibilidad de teletrabajo, 48 % asiste a la sede de San Pedro, con un promedio de antigüedad de 346 días y un salario original promedio de 519.000 colones aproximadamente.

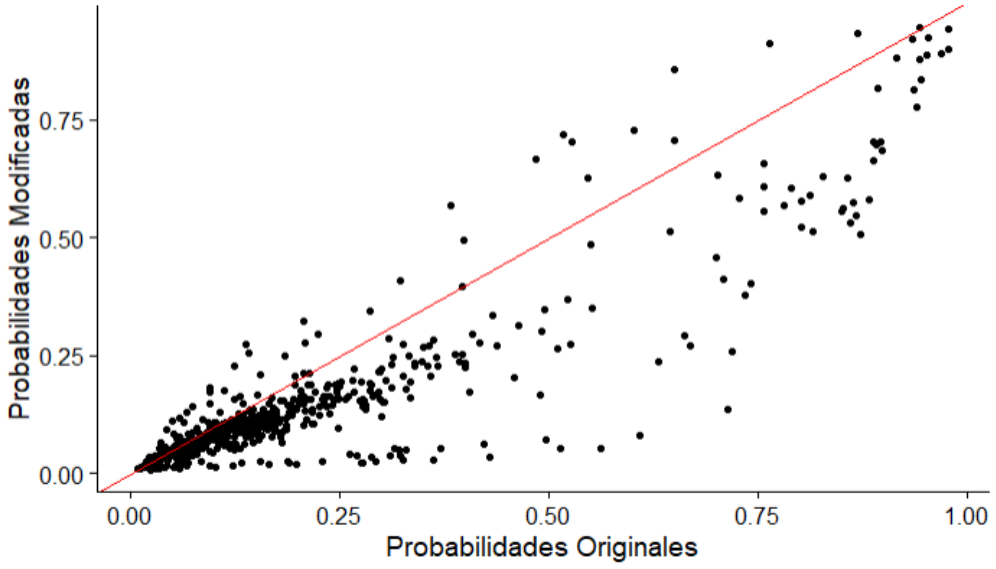
Sin embargo, se puede apreciar que el mayor cambio que hubo en las predicciones fue el grupo que originalmente se había clasificado como salidas de forma incorrecta, que en las nuevas predicciones del modelo se corrigió la predicción y ahora se clasificaron como no salidas de forma correcta. Ante un incremento en el salario con el perfil de estas 100 personas, se podrían utilizar los recursos de

una mejor manera, atendiendo a personas que sí representaban un riesgo real de salir de la empresa. Este grupo se puede identificar en la Figura 22 como las personas que originalmente tenían una predicción de 1, pasaron a tener una predicción de 0 y realmente eran 0 en la variable *Target*.

Un 92 % de estos empleados viven en la GAM (47 % en San José), 43 % trabaja en las oficinas de San Pedro, 65 % no tienen teletrabajo, tienen una antigüedad laboral de 262 días en promedio y un salario original promedio de 512.000 colones aproximadamente.

Además de analizar los cambios en las predicciones, también se analizaron los cambios en las probabilidades de las observaciones. En la Figura 23 se muestran todas las personas del estudio representadas por un punto que se interseca con el valor de la probabilidad original y la probabilidad modificada por el ajuste del 5 % del salario. También, se muestra una línea roja que representa la ecuación de la recta de la identidad.

Figura 23: Gráfico Dispersión de Probabilidades al Modificar un 5 % el Salario



Fuente: Elaboración propia

Si las probabilidades no hubieran tenido cambios en ambos escenarios, los puntos de las observaciones estarían todas encima de la identidad, ya que la probabilidad original sería la misma que la modificada. Por otro lado, si los puntos se ubican por debajo de la identidad, esto significa que la probabilidad

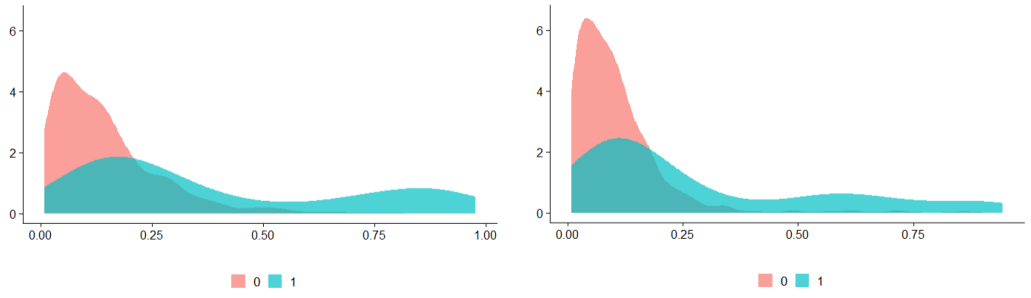
disminuyó con el cambio del salario y viceversa.

Es esperable que la mayoría de las personas disminuyan su probabilidad de salida ante un incremento salarial, como sucedió con el 89 % de las observaciones. Estas son un 98 % solteras, donde un 90 % viven en la GAM (51 % en San José), 30 % trabaja en las oficinas de San Pedro, 68 % no tienen teletrabajo, antigüedad laboral de 413 días en promedio y un salario promedio de 537.000 colones aproximadamente.

Por otro lado, el 11 % de la población que sufrió un incremento en la probabilidad de salida con el incremento salarial del 5 % eran personas 91 % solteras, donde un 93 % viven en la GAM (35 % en San José), 39 % trabaja en las oficinas de San Pedro, 64 % no tienen teletrabajo, antigüedad laboral de 1.294 días en promedio y un salario promedio de 584.000 colones aproximadamente.

Este comportamiento de las probabilidades de salida también se puede apreciar por medio de las distribuciones de densidad como se muestra en la Figura 24. Tanto la curva de la distribución de las salidas, como la de las no salidas se movieron hacia la izquierda, reflejando el efecto de la disminución en promedio de las probabilidades. Además, esto a su vez provoca que la probabilidad de corte también se debería disminuir al desplazarse hacia la izquierda, donde debería de pasar de 17 % a 11 % con base en las nuevas distribuciones.

Figura 24: Distribuciones de Densidad por Categoría



(a) Distribuciones con Salario Original (b) Distribuciones con Salario Modificado

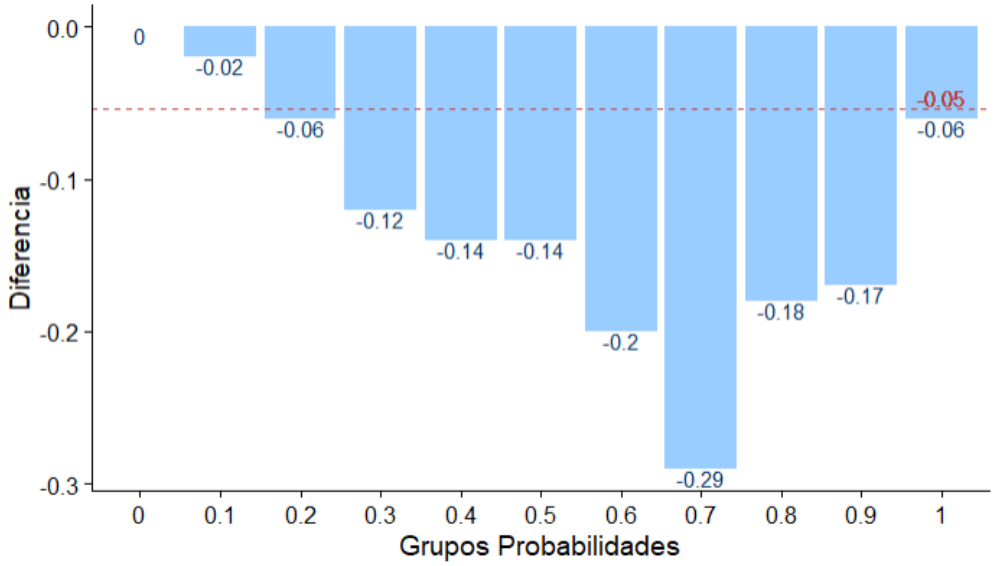
Fuente: Elaboración propia

El impacto de estas diferencias también se puede medir al agrupar por categorías, como se muestra en la Figura 25. En promedio, las probabilidades de salida se disminuyeron 5 %, como se marca en la línea punteada roja. Se evi-

dencia cómo conforme las probabilidades originalmente eran mayores, tuvieron una reducción más grande en su probabilidad con el ajuste salarial.

En general, las personas que tenían una probabilidad de salida original entre 60 % y 90 % fueron las que tuvieron un mayor cambio en su cálculo, específicamente la categoría original de los 70 %, donde tuvieron una reducción cercana a los 29 % de su probabilidad de salida.

Figura 25: Análisis de Sensibilidad al Modificar un 5 % el Salario



Fuente: Elaboración propia

Por lo tanto, se concluye que el salario representa una variable numérica sensible para el modelo, donde un incremento del 5 % del salario disminuyó en promedio un 5 % las probabilidades de salida. Además, las personas más sensibles a estos cambios son personas solteras, con una antigüedad menor y un salario original menor.

Estos efectos descritos al variar 5 % el salario se vieron amplificados en su impacto sobre las predicciones y probabilidades generadas a partir del modelo al variar el salario en un 10 % y 20 % adicional. Pueden ser consultados en la sección de Anexos.

Entre más se varíe el salario, mayor es el número de personas cuya predicción se deja de considerar un riesgo de salida en las Figuras 39 y 43. Además, entre mayor sea el incremento salarial, las probabilidades se disminuyen en mayor

medida, reflejando observaciones cada vez más por debajo de la recta de la identidad en las Figuras 40 y 44 y menor población con mayor probabilidad que la que se tenía originalmente.

También en las Figuras 41 y 45 se muestran que las distribuciones de densidad por categoría se desplazan en mayor medida hacia la izquierda, reflejando el impacto en la disminución de las probabilidades y la necesidad de reestablecer nuevas probabilidades de corte menores.

Finalmente, estos datos también son congruentes con los mostrados en las Figuras 42 y 46, donde se aprecia que entre mayor fuera el salario, la probabilidad disminuyó en promedio 7 % y 11 % respectivamente, manteniendo el mismo comportamiento en la distribución por categoría analizada para el incremento salarial del 5 %, donde las probabilidades mayores tuvieron un mayor impacto negativo que las menores.

7. Análisis Fuera de Muestra

Como se determinó en la sección anterior, las metodologías que presentaron mejores resultados de forma consistente con la validación cruzada fueron los modelos de *XGBoost*, junto con Bosques Aleatorios y Consensus. Por esta razón, en conjunto con la compañía, se aprobó su implementación y uso para validar por medio de una estimación fuera de muestra la efectividad de los mismos con observaciones fuera del periodo de entrenamiento.

Con base en los resultados del capítulo anterior, se decidió aplicar en la empresa el modelo de *XGBoost* con variables seleccionadas, ya que se considera un modelo más parsimonioso, resulta un modelo de menor complejidad, más eficiente computacionalmente, reduce el riesgo de sobreajuste y es menos susceptible a errores de multicolinealidad. Sin embargo, al igual se comparó en este capítulo con las demás metodologías y con los modelos que incluyen las variables completas, con el fin de medir su rendimiento con respecto a los demás.

Para esto se consideraron a las personas activas en abril 2023 y se esperó para observar su comportamiento de rotación de las salidas de abril, mayo y junio 2023 para definir la variable objetivo del periodo. Posteriormente se repitió el mismo proceso con los datos de los meses siguientes del año, es decir, personas activas en mayo 2023 y las salidas de mayo, junio y julio y así sucesivamente, hasta las personas activas de noviembre 2023. Dada la temporalidad, se tiene un total de 8 meses de prueba de los modelos.

7.1. Comparación de Resultados de las Mejores Metodologías

Al igual que en el capítulo anterior, la comparación de los rendimientos de los modelos se medirá a través del área bajo la curva, precisión global, error global, especificidad y sensibilidad. Esta comparación se hará entre los valores generados en la estimación fuera de muestra y con los resultados de los modelos originales. Es importante señalar que se estableció una meta de que el promedio del área bajo la curva ROC para los meses de estudio del modelo seleccionado

no disminuyeran más de 10 % en comparación a los de entrenamiento.

7.1.1. Modelos con Todas las Variables

A continuación se muestran los resultados de la estimación fuera de muestra realizada en abril 2023 para cada modelo:

Cuadro 13: Resumen Principales Indicadores Modelos con Variables Completas

Método	AUC	Precisión Global	Error Global	Especificidad	Sensibilidad	Probabilidad de Corte
XGBoost	0,7801	0,7167	0,2833	0,7159	0,7200	0,130
Bosques Aleatorios	0,7360	0,6666	0,3334	0,6635	0,6804	0,160
Consensus	0,7572	0,6916	0,3084	0,6894	0,7018	0,180

Fuente: Elaboración propia

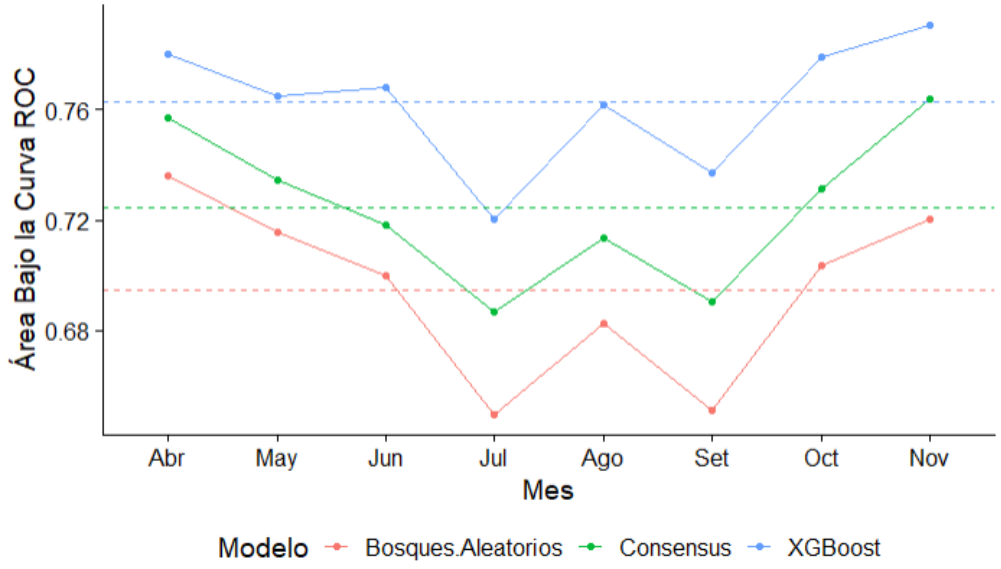
Con base en el Cuadro 13, el modelo de *XGBoost* sigue siendo el que genera los mejores resultados de los 3 modelos, seguido por Consensus y por último el de Bosques Aleatorios. Este comportamiento es esperable, ya que como se describe en (Zamani, 2019) y (Sahin, 2020), los modelos de *XGBoost* tienden a ser superiores a los de Bosques Aleatorios por el factor del control para prevenir sobreajustes por el término de regularización explicado en capítulos anteriores.

Para la estimación fuera de muestra, el modelo de *XGBoost* obtuvo 223bps más área bajo la curva ROC que el promedio, 250bps mayor precisión global, 263bps mayor Especificidad y 193bps mayor Sensibilidad. Este último indicador es el de mayor relevancia, ya que como se mencionó anteriormente, es el crítico al momento de determinar el porcentaje de acierto de las salidas en el periodo.

Al comparar los resultados de *XGBoost* contra el modelo de entrenamiento, se tuvo una reducción de 485bps del área bajo la curva ROC, 55bps menos precisión global, 17bps menos Especificidad y 194bps menos Sensibilidad.

En la Figura 26 se muestran los resultados de las áreas bajo la curva ROC de cada modelo para cada mes en análisis del 2023. Las líneas punteadas horizontales representan el promedio de cada modelo respectivamente.

Figura 26: Análisis Fuera de Muestra con Variables Predictoras Completas



Fuente: Elaboración propia

Es esperable que los valores de un modelo al momento de aplicar una estimación fuera de muestra suelan disminuir un poco, ya que existen estacionalidades, condiciones macroeconómicas y políticas de la empresa que van cambiando conforme pasa el tiempo y pueden llegar a afectar la eventual salida o no de una persona y que el modelo no responda al 100 % sobre estos cambios.

El promedio del área bajo la curva ROC del modelo de *XGBoost* de los meses de estudio de la estimación fuera de muestra fue de 0,7628, siendo 281bps menor que el mayor registro (noviembre) y 658bps menor en comparación al modelo calibrado.

Los meses que afectaron negativamente al promedio de la estimación fuera de muestra fueron los comprendidos entre julio y setiembre. Particularmente en estos meses se anunció por parte de la empresa que existirían aumentos salariales que beneficiarían a los empleados, por lo que podrían haberse generado algunas predicciones que en condiciones normales se hubieran podido ir, pero con la comunicación generada no se concretaron.

También hubo algunos cambios en las necesidades de personal por parte de algunos clientes, principalmente en esos meses, donde al reubicar a las personas con otros clientes algunos decidieron renunciar, al no adaptarse a sus nuevas

funciones. Esto también afectó los resultados, al contemplar algunas personas como que no eran un riesgo de salida y al final salieron por esas circunstancias.

Además, al estar implementando un conjunto de acciones sobre las personas que se levantó una alerta de un potencial riesgo de salida, se tiende a distorsionar los resultados del modelo, al evitar que las personas clasificadas como salidas no terminaran concretándose. Esto porque se hacen planes de retención sobre las personas con mayor probabilidad de salida, por lo que se impacta directamente las clasificaciones de este tipo de colaboradores por parte del modelo.

7.1.2. Modelos con Variables Seleccionadas

A continuación se muestran los resultados de la estimación fuera de muestra realizado en abril 2023 para cada modelo:

Cuadro 14: Resumen Principales Indicadores Modelos con Variables Seleccionadas

Método	AUC	Precisión Global	Error Global	Especificidad	Sensibilidad	Probabilidad de Corte
XGBoost	0,7574	0,6708	0,3292	0,6628	0,7068	0,170
Bosques Aleatorios	0,6832	0,6334	0,3666	0,6314	0,6424	0,025
Consensus	0,7367	0,6741	0,3259	0,6727	0,6804	0,150

Fuente: Elaboración propia

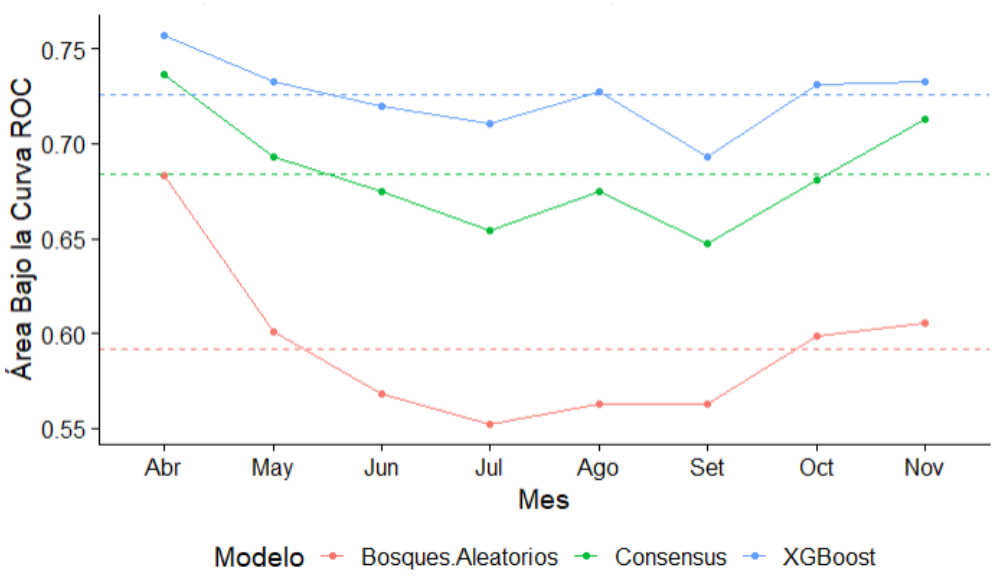
Como se muestra en el Cuadro 14, con el modelo de *XGBoost* se obtuvieron los mejores resultados de todos los métodos aplicados en la estimación fuera de muestra. El segundo mejor modelo fue el de Consensus y el tercero el de Bosques Aleatorios.

Si se compara contra el promedio de los 3 modelos, *XGBoost* registró un área bajo la curva ROC superior por 316bps, 114bps mayor precisión global, 71bps mayor Especificidad y 303bps mayor Sensibilidad. Por otro lado, el rendimiento de la estimación fuera de muestra contra el modelo de *XGBoost* original muestra una caída de 231bps en el área bajo la curva ROC, 301bps de precisión global, 375bps de Especificidad, pero una mejora de 37bps de Sensibilidad.

En la Figura 27 se pueden apreciar los resultados de las áreas bajo la curva

ROC de cada modelo para cada mes en análisis del 2023. Las líneas punteadas horizontales representan el promedio de cada modelo respectivamente.

Figura 27: Análisis Fuera de Muestra con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Tal como los modelos que utilizan todas las variables predictoras, los resultados se vieron más impactados en el rango de los meses de julio a setiembre con respecto a su promedio. Sin embargo, la diferencia entre el promedio de los valores registrados en la estimación fuera de muestra contra los valores de los modelos entrenados de cada metodología, tuvieron un menor impacto negativo en el caso de los modelos más parsimoniosos.

Además, los resultados son menos variables en el caso de los modelos de *XGBoost* con variables predictoras seleccionadas versus los que contemplaban todas las variables, pero sucede lo contrario con los modelos de Consensus y Bosques Aleatorios.

A pesar de bajar en los indicadores, estos se mantienen en un rango tolerable. Se observa en los resultados de (Praphutikul y Limpiyakorn, 2023) y (Loterman, 2014) que se puede tener una reducción de hasta 1.800bps al realizar una estimación fuera de muestra, debido a variantes que se pueden dar en el entorno.

Al registrar un promedio de 0.7256 en el área bajo la curva ROC del modelo

de *XGBoost*, se confirma que pasó la meta previamente establecida de que el promedio no disminuyera más de 10 % con respecto al modelo seleccionado en el análisis fuera de muestra, ya que tuvo una reducción de 7.57 % con respecto al modelo entrenado.

La disminución del modelo de *XGBoost* se justifica porque la misma rotación de personal en la industria hace que las variables cambien de comportamiento, así como la efectividad de los planes de retención aplicados según las probabilidades de salida. Finalmente, se espera tener datos de 12 periodos, por lo que se necesita de 15 meses de observaciones para poder determinar si se ocupan recalibrar los parámetros.

7.2. Comparación de Perfiles según sus Clasificaciones

Al igual que en el capítulo anterior, se compararon los perfiles de las personas que son clasificadas de manera óptima con base en las predicciones de salidas a 3 meses de los modelos de *XGBoost* y los perfiles para los cuales el modelo presenta desafíos en su clasificación por medio de un Clustering Jerárquico. Estos perfiles se aplicaron sobre los datos de la estimación fuera de muestra de abril 2023 y que contemplan las salidas de abril, mayo y junio 2023 para el modelo seleccionado.

7.2.1. Modelo XGBoost con Variables Predictoras Seleccionadas

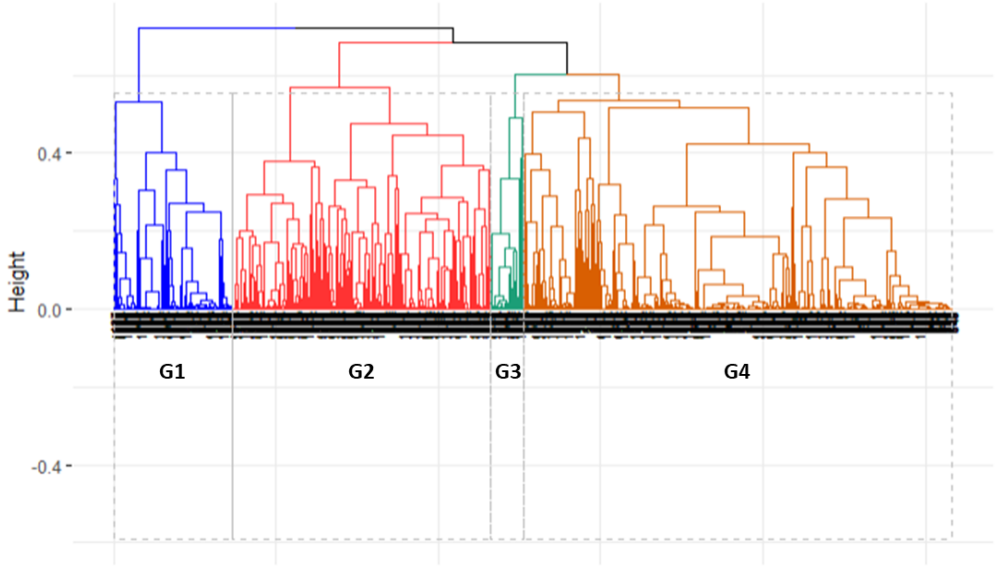
La Figura 28 y el Cuadro 15 contemplan 4 grupos distintos, siendo el Grupo 1 el más grande con 617 personas y 36 % de salidas en los siguientes 3 meses, Grupo 2 171 con 0 % de salidas, Grupo 3 371 personas con 45 % de salidas y Grupo 4 48 personas con 95 % de salidas.

Los 2 grupos más pequeños son los que tuvieron mayor y menor precisión de las salidas durante el periodo. El Grupo 2 es el que tiene el mayor número de antigüedad, con 360 días en promedio, mientras que el Grupo 4 el menor número, con 207 días. El número de Reingresos es menor en el Grupo 2 con 1,06 y en el Grupo 4 se presenta el mayor número, con 1,20. Ambos salarios promedios están más cercanos al del Grupo 1, que es el de menor salario, versus el Grupo 3, que es el de mayor salario. El Grupo 2 es el que tiene mayor número

de personas fuera de la GAM, mientras que el Grupo 4 tiene mayor número de personas trabajando en la sucursal de San Pedro.

Por lo tanto, se concluye que el perfil de una persona con mayor antigüedad, menor número de reingreso y trabajando fuera de la GAM genera que el modelo tenga una menor capacidad predictiva para acertar que la persona sea clasificada como una salida y, efectivamente, termine saliendo de la empresa en un lapso de 3 meses.

Figura 28: Dendograma de Clasificaciones Salidas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Cuadro 15: Resumen del Clustering Jerárquico de las Personas Clasificadas como Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	617	171	371	48
Antigüedad	$\overline{X} = 250$	$\overline{X} = 360$	$\overline{X} = 209$	$\overline{X} = 207$
Cliente	Cliente B: 88 % Cliente D: 10 % Cliente K: 1 % Otros: 1 %	Cliente B: 72 % Cliente D: 16 % Cliente A: 3 % Otros: 9 %	Cliente K: 33 % Cliente I: 22 % Cliente L: 21 % Otros: 24 %	Cliente B: 84 % Cliente K: 14 % Cliente A: 2 %
Estado Civil	Solteros: 99 % Casados: 1 %	Solteros: 98 % Casados: 1 % Divorciados: 1 %	Solteros: 99 % Casados: 1 %	Solteros: 100 %
Provincia	San José: 58 % Cartago: 15 % Alajuela: 10 % Otros: 17 %	Heredia: 35 % San José: 34 % Guanacaste: 17 % Otros: 14 %	San José: 44 % Alajuela: 19 % Heredia: 14 % Otros: 23 %	Cartago: 65 % Heredia: 14 % Puntarenas: 8 % Otros: 13 %
Reingreso	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}12$	$\overline{X} = 1{,}20$
Salario	$\overline{X} = 497.006$	$\overline{X} = 505.737$	$\overline{X} = 535.565$	$\overline{X} = 500.784$
Sucursal	San Pedro: 68 % Heredia: 13 % Hatillo: 10 % Otros: 9 %	Heredia: 57 % Liberia: 22 % Hatillo: 19 % Otros: 2 %	Virtual: 100 %	San Pedro: 88 % Moravia: 8 % Virtual: 4 %
Teletrabajo	Sí: 1 % No: 99 %	Sí: 1 % No: 99 %	Sí: 99 % No: 1 %	Sí: 4 % No: 96 %
Salidas Reales 3 Meses	36 %	0 %	45 %	95 %
Salidas Reales Indefinido	69 %	0 %	61 %	98 %

Fuente: Elaboración propia

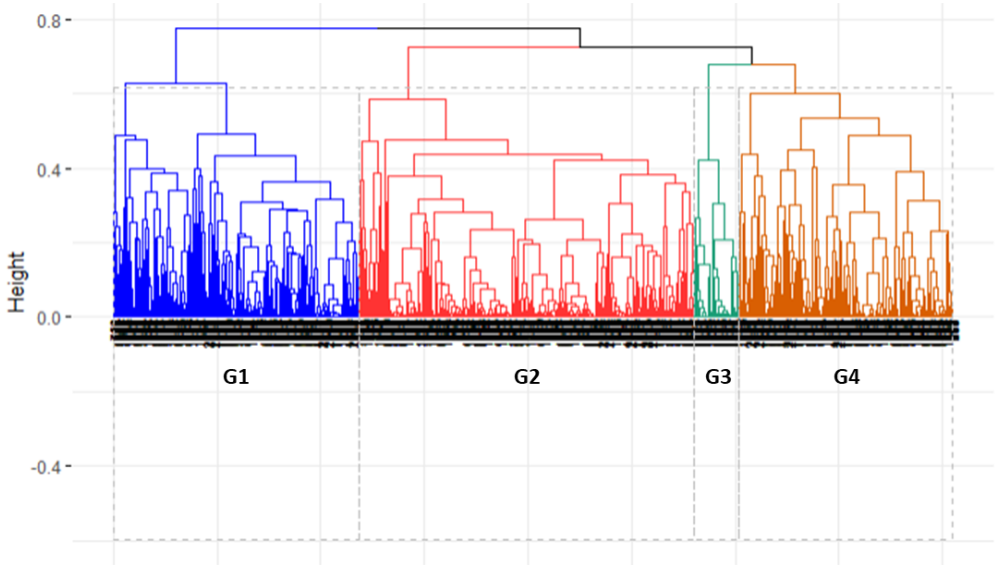
En el dendograma de la Figura 29 y el Cuadro 16 se representan 4 grupos distintos, donde el Grupo 1 agrupa 841 personas y un acierto de no salida durante el periodo de estudio del 100 %, el Grupo 2 618 personas, de las cuales un 93 % no fueron salidas, Grupo 3 537 personas con 77 % de no salidas y el Grupo 4 114 personas con 97 % que no fueron salidas en los siguientes 3 meses.

Por lo anterior, se considera que los Grupos 1, 2 y 4 tuvieron rendimientos del modelo bastante acertados, mientras que el perfil de las personas del Grupo 3 representó el mayor número de salidas no detectadas de todos.

La principal diferencia del Grupo 2 con respecto a los demás es que las personas tienen el menor salario promedio de 523.604 colones y son las personas con menor experiencia de los Clientes B, K y D, al tener en promedio 518 días de antigüedad versus 920 días de las personas del Grupo 1 que considera a los mismos clientes.

Al igual que con el modelo de *XGBoost* de entrenamiento, los resultados fueron bastante satisfactorios a lo largo de todos los grupos, ya que el perfil con mayor desacierto tuvo solamente un 23 % de personas clasificadas como que no se iban a ir y realmente se fueron y los otros grupos con un 7 % o menos.

Figura 29: Dendograma de Clasificaciones No Salidas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Cuadro 16: Resumen del Clustering Jerárquico de las Personas Clasificadas como No Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	841	618	537	114
Antigüedad	$\overline{X} = 920$	$\overline{X} = 560$	$\overline{X} = 518$	$\overline{X} = 249$
Cliente	Cliente B: 72 % Cliente K: 18 % Cliente D: 4 % Otros: 6 %	Cliente I: 34 % Cliente E: 21 % Cliente K: 17 % Otros: 28 %	Cliente B: 71 % Cliente K: 17 % Cliente D: 6 % Otros: 6 %	Cliente F: 99 % Cliente E: 1 %
Estado Civil	Solteros: 93 % Casados: 6 % Divorciados: 1 %	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 99 % Casados: 1 %	Solteros: 100 %
Provincia	San José: 43 % Heredia: 21 % Alajuela: 18 % Otros: 18 %	San José: 47 % Alajuela: 15 % Guanacaste: 10 % Otros: 27 %	San José: 41 % Heredia: 20 % Alajuela: 18 % Otros: 21 %	San José: 67 % Cartago: 15 % Heredia: 12 % Alajuela: 6 %
Reingreso	$\overline{X} = 1{,}03$	$\overline{X} = 1{,}08$	$\overline{X} = 1{,}05$	$\overline{X} = 1{,}07$
Salario	$\overline{X} = 537.732$	$\overline{X} = 572.577$	$\overline{X} = 523.604$	$\overline{X} = 945.413$
Sucursal	Heredia: 44 % San Pedro: 33 % Virtual: 17 % Otros: 6 %	Virtual: 99 % San Pedro: 1 %	Heredia: 41 % San Pedro: 34 % Virtual: 14 % Otros: 11 %	Hatillo: 99 % San Pedro: 1 %
Teletrabajo	No: 100 %	Sí: 99 % No: 1 %	No: 100 %	No: 100 %
Salidas Reales 3 Meses	0 %	7 %	23 %	3 %
Salidas Reales Indefinido	1 %	35 %	98 %	39 %

Fuente: Elaboración propia

7.3. Impacto del Modelo a la Empresa

El impacto generado a la empresa derivado del modelo ha sido positivo, tanto desde el punto de vista de retención de sus colaboradores, como económico. Desde que se ha aplicado el modelo se ha evitado la salida de 676 personas a lo largo del año, que al considerar la no generación de ingresos por no tener el personal requerido (aproximadamente $3.006.000), compensados por los costos del salario (aproximadamente $1.178.000), sumado a los costos de contratación (aproximadamente $366.000) y de entrenamiento por su reemplazo (aproximadamente $1.862.000) representan más de $4 millones en ahorros para la empresa.

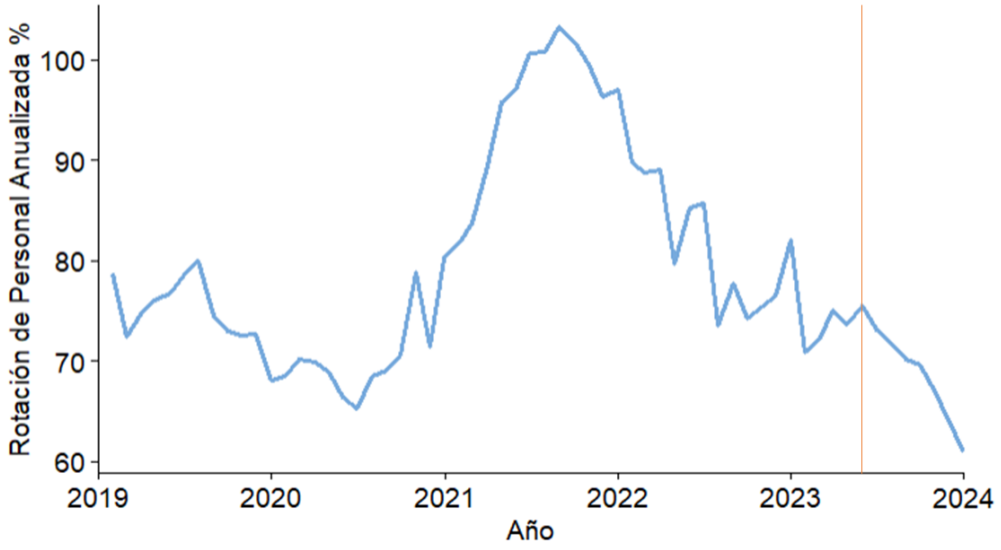
Esto a su vez no cuantifica el intangible del aprendizaje adquirido por estas personas, que ya tienen experiencia para asumir sus roles y, en caso de haber

salido de la compañía, se hubiera perdido, al colocar a otra persona nueva en esa posición.

Estas potenciales salidas de los colaboradores se evitaron al generar las alertas tempranas a los departamentos de Operaciones y Recursos Humanos, donde cada departamento trabajó con distintos planes de acción individualizados a partir de la identificación de los colaboradores clasificados como riesgos potenciales de abandono de la empresa. Sin la generación de estas alertas tempranas, la organización no tendría la capacidad de identificar todos los riesgos potenciales y actuar a tiempo en la mayoría de las ocasiones para evitar el abandono de las personas.

En la Figura 30 se muestra el comportamiento de la rotación de personal de la empresa desde enero 2019 hasta diciembre 2023. Después de la pandemia, se comenzó a incrementar el volumen de salidas de la empresa a un ritmo más acelerado que el crecimiento de las personas activas, hasta el tercer trimestre del año 2021, donde la empresa comenzó a contratar de forma acelerada por crecimiento de varios clientes, hasta inicios del 2023.

Figura 30: Rotación de Personal Anualizada de SITEL Costa Rica por Mes del 2019 al 2023



Fuente: Elaboración propia con datos de SITEL

Al existir una base mayor de personas activas, hace que cada salida tenga un impacto menor en el cálculo de la rotación de personal. En el 2023, por decisio-

nes estratégicas de negocio, se decidió que las salidas no se iban a reemplazar al mismo ritmo que se generaban, lo que provocó que cada salida hoy en día tenga un mayor impacto sobre la rotación.

A pesar de esta situación, según la Figura 30, desde el momento en que se comenzó a ejecutar el modelo predictor de esta investigación (marcado con la línea vertical naranja), la rotación anualizada ha llegado a niveles históricos entre los agentes directos, que no se observan desde hace más de 5 años. Esto se consiguió mediante los esfuerzos de detección temprana de posibles salidas por parte del modelo, junto con los planes de acción ejecutados por parte de la empresa para retener a las personas.

8. Conclusiones y Recomendaciones

Como principales conclusiones se tienen los siguientes puntos:

- Se desarrolló una metodología que consiste en un modelo predictivo que permite predecir la rotación de personal, no solamente en la empresa SITEL, sino en cualquier empresa que se requiera. Al ser una metodología que puede ser aplicada a cualquier tipo de empresa, algunas variables se podrían modificar según las especificaciones y disponibilidad de información, sin embargo la estructura de la selección de las mejores variables predictoras, comparación de los diferentes modelos predictivos implementados y calibración de parámetros para ejecución del modelo seleccionado es independiente de la empresa donde se aplique.
- La metodología de *XGBoost* resultó superior a los otros algoritmos desarrollados en la investigación realizada, dado que tiene una alta capacidad predictiva y por eso suele ser de los métodos con resultados más satisfactorios, como se comprobó en la bibliografía, y este trabajo no es la excepción. Esto se debe a que combina múltiples métodos, como los árboles de decisión, implementa un término de regularización para prevenir sobreajuste del modelo, permite la optimización de los hiperparámetros para una configuración óptima y su potencialización de las predicciones le permite manejar datos de gran tamaño y complejidad. Sin embargo, no puede asegurarse que esta metodología vaya a ser superior a los demás, sin importar el contexto ni el conjunto de datos que se utilice.
- A partir de la modelación se consiguió identificar un perfil de colaboradores más propensos a salir de la empresa y que el modelo los categorice de forma correcta. Este perfil está conformado por un empleado que tiene un menor salario, mayores notas de desempeño, mayor posibilidad de teletrabajo y menor número de reingresos a la empresa.
- La aplicación del modelo en la empresa es simple y altamente efectiva, ya que cada mes las personas encargadas de los colaboradores de cada cliente reciben una actualización de las probabilidades de salida para los siguientes 3 meses de los colaborades de su departamento, donde se priorizan según su probabilidad. Al implementar el modelo se ha generado

un impacto anualizado de más de $4.000.000 en el país en ahorros al tener una detección temprana de las posibles salidas de los colaboradores, al considerar los impactos financieros descritos anteriormente.

- Es utópico pensar que se van a evitar salidas voluntarias en la empresa a partir del modelo, ya que las razones de salida pueden ser múltiples. El objetivo del modelo es generar una alerta temprana de las eventuales salidas a partir de los pronósticos que se originen. Una vez generados estos pronósticos, se necesita de la implementación de un plan de acción para identificar qué podría ser lo que esté aquejando a la persona y con esto fomentar un ambiente laboral satisfactorio que evite la eventual salida del colaborador.
- Algunas recomendaciones de acciones a implementar para evitar la salida de los empleados serían las siguientes:
    - Brindar la posibilidad de horarios flexibles, para que puedan continuar con sus estudios sin que los horarios del trabajo sean un impedimento en la medida de lo posible.
    - Verificar la posibilidad para que las personas puedan trabajar en alguna sucursal cercana a su hogar o lugar de estudio, para disminuir el impacto negativo que les pueden generar los traslados en su vida cotidiana.
    - Establecer sistemas de compensación variable, como bonos más atractivos, para incentivar que puedan conseguir un mejor rendimiento en sus funciones y al mismo tiempo aumentar su poder adquisitivo
    - Agendar sesiones con los supervisores, mínimo de forma semanal, que no se basen solamente en su desempeño en la cuenta, sino también que se pueda identificar si hay algún tema ajeno al trabajo que pueda estar provocando una insatisfacción.
    - Incrementar las ceremonias de reconocimiento de los mejores agentes de cada cliente, para mejorar la motivación de las personas.
    - Ofrecer la posibilidad de cambiar de cliente o a otras áreas después de cierto tiempo, para no tener un trabajo rutinario, si así lo desean las personas después de cierto tiempo. Esto aumentaría la

percepción de oportunidades de crecimiento en las personas.

- Se recomienda ejecutar un modelo para determinar la probabilidad de salida de los candidatos a contratación, para que los reclutadores puedan tener un sistema de detección temprana de posibles salidas y, en caso de contratarse a la persona, alerten a los supervisores de las cuentas para tener un seguimiento más cercano de esas personas durante las primeras semanas de entrenamiento.
- Finalmente, con los impactos positivos tanto a nivel de detección temprana de posibles salidas, como financieros derivados del modelo, se recomienda a la empresa desarrollar un modelo para predecir las cuentas de la vertical de tecnología, así como ejecutar modelos similares en otros países, ya que la alta rotación de personal no es un tema solamente de Costa Rica, sino que de todos los países dentro de la empresa.

9. Referencias

Alaraj, M., y Abbod, M. (2016, 7). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89-105. Descargado de https://www.sciencedirect.com/science/article/abs/pii/S0950705116300569 doi: https://doi.org/10.1016/j.knosys.2016.04.013

Amnueypornsakul, B., Bhat, S., y Chinprutthiwong, P. (2014, 10). Predicting attrition along the way: The uiuc model. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 55-59. Descargado de https://paperswithcode.com/paper/predicting-attrition-along-the-way-the-uiuc

Aulck, L. e. a. (2017, 8). Stem-ming the tide: Predicting stem attrition using student transcript data. Descargado de https://paperswithcode.com/paper/stem-ming-the-tide-predicting-stem-attrition

Bafandeh, S., y Bolandraftar, M. (2013, 9). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *Int. Journal of Engineering Research and Applications*, *3*(5), 605-610. Descargado de https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf

Barvey, A., Kapila, J., y Pathak, K. (2018, 7). Proactive intervention to downtrend employee attrition using artificial intelligence techniques. Descargado de https://paperswithcode.com/paper/proactive-intervention-to-downtrend-employee

Bi, Q. e. a. (2019, 10). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222-2239. Descargado de https://doi.org/10.1093/aje/kwz189 doi: 10.1093/aje/kwz189

Chen, T., y Guestrin, C. (2016, 8). Xgboost: A scalable tree boosting system. , 785–794. Descargado de https://dl.acm.org/doi/pdf/10.1145/2939672.2939785 doi: https://doi.org/10.1145/2939672.2939785

Chourey, A., Phulre, S., y Mishra, S. (2019, 12). Employee attrition prediction using various machine learning techniques. *The International journal of analytical and experimental modal analysis*, *11*, 233-239. Descargado de http://www.ijaema.com/gallery/23-december-2932.pdf

Coenen, F. (2012, 5). On the use of confusion matrixes. Des-

cargado de https://pyrobots.csc.liv.ac.uk/~frans/Notes/confusionMatrices2012-5-12.pdf

Domínguez-Almendros, S., Benítez-Parejo, N., y Gonzalez-Ramirez, A. (2011). Logistic regression models. *Allergologia et Immunopathologia*, *39*(5), 295-305. Descargado de https://www.sciencedirect.com/science/article/pii/S0301054611002011 doi: https://doi.org/10.1016/j.aller.2011.05.002

Ekelund, S. (2012, 3). Roc curves—what are they and how are they used? *Point of Care: The Journal of Near-Patient Testing Technology*, *11*(1), 16-21. Descargado de https://journals.lww.com/poctjournal/fulltext/2012/03000/ROC_Curves_What_are_They_and_How_are_They_Used_.6.pdf doi: 10.1097/POC.0b013e318246a642

Ghojogh, B., y Crowley, M. (2023). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. Descargado de https://arxiv.org/pdf/1905.12787.pdf doi: https://doi.org/10.48550/arXiv.1905.12787

Gunn, S. (1997, 11). Support vector machines for classification and regression. Descargado de https://www.svms.org/tutorials/Gunn1997.pdf

Habibzadeh, H. P., F., y Yadollahie, M. (2016, 9). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica*, *26*(3), 297-307. Descargado de https://hrcak.srce.hr/file/247319 doi: https://doi.org/10.11613/BM.2016.034

Han, Q., Gui, C., Xu, J., y Lacidogna, G. (2019). A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm. *Construction and Building Materials*, *226*, 734-742. Descargado de https://www.sciencedirect.com/science/article/pii/S0950061819319890 doi: https://doi.org/10.1016/j.conbuildmat.2019.07.315

Huang, N., Peng, H., Cai, G., y Chen, J. (2016). Power quality disturbances feature selection and recognition using optimal multi-resolution fast s-transform and cart algorithm. *Energies*, *9*(11). Descargado de https://www.mdpi.com/1996-1073/9/11/927 doi: 10.3390/en9110927

Irmanita, R., Suryani, S., y Sibaroni, Y. (2021). Classification of malaria complication using cart (classification and regression tree) and naïve bayes. *Resti Journal*, *5*(1), 10-16. Descargado de http://jurnal.iaii.or.id/

index.php/RESTI/article/view/2770 doi: https://doi.org/10.29207/resti.v5i1.2770

Kavyasree, B., y Naresh Kumar, T. (2020, 8). A study on employee attrition and retention strategies in bpo industry. *Mukt Shabd Journal*, *9*(8), 1289-1292. Descargado de http://shabdbooks.com/gallery/166-aug2020.pdf

Loterman, G. e. a. (2014, 3). A proposed framework for backtesting loss given default models. *Journal of Risk Model Validation*. Descargado de https://eprints.soton.ac.uk/386766/1/A_proposed_framework.pdf doi: 10.21314/JRMV.2014.117

Marmion, M., Hjort, J., Thuiller, W., y Luoto, M. (2009, 3). Statistical consensus methods for improving predictive geomorphology maps. *Computers Geosciences*, *35*(3), 615-625. Descargado de https://www.sciencedirect.com/science/article/pii/S0098300408001350 doi: https://doi.org/10.1016/j.cageo.2008.02.024

Mohri, M., Rostamizadeh, A., y Talwalkar, A. (2018). Foundations of machine learning. Descargado de https://cs.nyu.edu/~mohri/mlbook/

Muhammad, I., y Yan, Z. (2015). Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, *5*(3). Descargado de https://ictactjournals.in/paper/IJSC_Paper_4_946-952.pdf doi: https://doi.org/10.21917/ijsc.2015.0133

Muneeb, M., y Henschel, A. (2021, 4). Eye-color and type-2 diabetes phenotype prediction from genotype data using deep learning methods. *BMC Bioinformatics*. Descargado de https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04077-9 doi: https://doi.org/10.1186/s12859-021-04077-9

NageswaraRao, S., y Swapna, M. (2019, 6). A study on "employee attrition" with reference to indian bpo industry: Factors and measures. *International Journal of Innovative Studies in Sociology and Humanities (IJISSH)*, *4*(6), 115-119. Descargado de https://ijissh.org/storage/Volume4/Issue3/IJISSH-040315.pdf

Olorunnimbe, K., y Viktor, H. (2023, 3). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artifcial Intelligence Review*, *56*, 2057-2109. Descargado de https://link.springer.com/article/10.1007/s10462-022-10226-0 doi: https://

doi.org/10.1007/s10462-022-10226-0

Praphutikul, T., y Limpiyakorn, Y. (2023, 5). Xgboost for smart portfolio management based on multi factor stock selection. Descargado de https://dl.acm.org/doi/pdf/10.1145/3605423.3605442 doi: https://doi.org/10.1145/3605423.3605442

Qin, C. e. a. (2021, 3). Xgboost optimized by adaptive particle swarm optimization for credit scoring. *Mathematical Problems in Engineering, 2021*. Descargado de https://www.hindawi.com/journals/mpe/2021/6655510/ doi: https://doi.org/10.1155/2021/6655510

Quinlan, J. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221-234. Descargado de https://www.sciencedirect.com/science/article/pii/S0020737387800536 doi: https://doi.org/10.1016/S0020-7373(87)80053-6

Ravi, K. B., y Serra, J. (2017, 7). Cost-complexity pruning of random forests. Descargado de http://jurnal.iaii.or.id/index.php/RESTI/article/view/2770 doi: https://doi.org/10.48550/arXiv.1703.05430

Sahin, E. (2020, 6). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. *Springer Nature*. Descargado de https://link.springer.com/article/10.1007/s42452-020-3060-1#Sec12 doi: https://doi.org/10.1007/s42452-020-3060-1

Sajjadi, S. e. a. (2017, 5). Finding bottlenecks: Predicting student attrition with unsupervised classifier. Descargado de https://paperswithcode.com/paper/finding-bottlenecks-predicting-student

Shankar, R. S., Rajanikanth, J., Sivaramaraju, V., y Murthy, K. (2018, 8). Prediction of employee attrition using datamining. En *2018 ieee international conference on system, computation, automation and networking (icscan)* (p. 335-342). Descargado de https://www.researchgate.net/profile/Shiva-Reddy-7/publication/329136719_PREDICTION_OF_EMPLOYEE_ATTRITION_USING_DATAMINING/links/6095745c92851c490fc35cdb/PREDICTION-OF-EMPLOYEE-ATTRITION-USING-DATAMINING.pdf doi: 10.1109/ICSCAN.2018.8541242

Shawe-Taylor, J., y Sun, S. (2011). A review of optimization methodologies

in support vector machines. *Neurocomputing*, *74*(17), 3609-3618. Descargado de https://www.sciencedirect.com/science/article/pii/S0925231211004371 doi: https://doi.org/10.1016/j.neucom.2011.06.026

Song, Y., y Lu, Y. (2015, 4). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*, 130-135. Descargado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/ doi: https://doi.org/10.11919/j.issn.1002-0829.215044

Suguna, N., y Thanushkodi, K. (2010, 7). An improved k-nearest neighbor classification using genetic algorithm. *IJCSI International Journal of Computer Science Issues*, *7*(2), 18-21. Descargado de https://www.researchgate.net/profile/Amir-Mosavi-3/publication/46093676_Domain_Driven_Data_Mining_-_Application_to_Business/links/00b7d51c0a0025b857000000/Domain-Driven-Data-Mining-Application-to-Business.pdf#page=32

Suhasini, L. (2019, 2). Attrition and retention of a human resource management, the huge challenge in indian bpo companies. *International Journal of Research in Engineering*, *9*, 486-491. Descargado de http://indusedu.org/pdfs/IJREISS/IJREISS_2719_41436.pdf

Thayer, J. (2002, 4). Stepwise regression as an exploratory data analysis procedure.. Descargado de https://files.eric.ed.gov/fulltext/ED464932.pdf

Visa, S. e. a. (2011, 4). Confusion matrix-based feature selection. *Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011,*, 120-127. Descargado de https://ceur-ws.org/Vol-710/paper37.pdf

Wang, C., Deng, C., y S., W. (2020). Imbalance-xgboost: leveraging weighted and focal losses for binary label-imbalanced classification with xgboost. *Pattern Recognition Letters*, *136*, 190-197. Descargado de https://www.sciencedirect.com/science/article/pii/S0167865520302129 doi: https://doi.org/10.1016/j.patrec.2020.05.035

Witten, I., y Frank, E. (2005). Data mining: Practical machine learning tools and techniques. Descargado de https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf

Wu, J., Xiong, H., y Chen, J. (2009, 6). Towards understanding hierarchical clustering: A data distribution perspective. *Neurocomputing*, *72*(10), 2319-2330. Descargado de https://www.sciencedirect.com/science/article/pii/S0925231208005663 (Lattice Computing and Natural Computing (JCIS 2007) / Neural Networks in Intelligent Systems Designn (ISDA 2007)) doi: https://doi.org/10.1016/j.neucom.2008.12.011

Xu, R. (2013). Improvements to random forest methodology. Descargado de https://www.proquest.com/openview/2e57452e00a58cd98f45a22c5d375891/1?pq-origsite=gscholar&cbl=18750

Yates, L. e. a. (2022, 11). Cross validation for model selection: A reviewwith examples from ecology. Descargado de https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ecm.1557 doi: https://doi.org/10.1002/ecm.1557

Yedida, R. e. a. (2018, 6). Employee attrition prediction. Descargado de https://paperswithcode.com/paper/employee-attrition-prediction

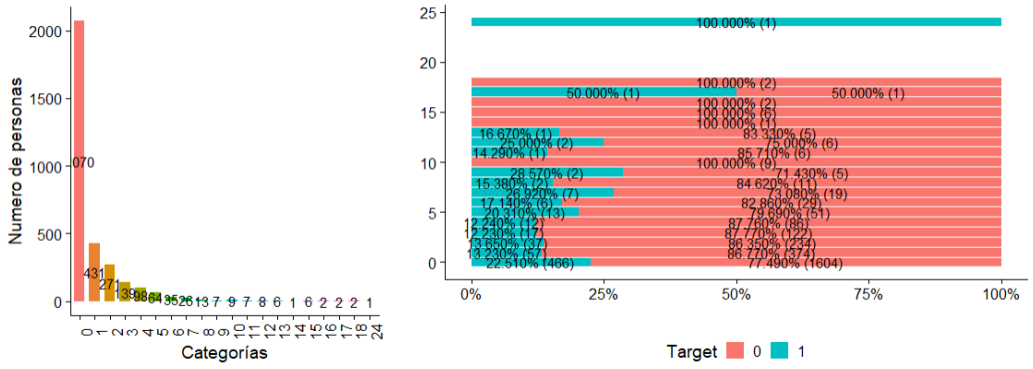
Yuan, J. (2023, 3). Package 'xgboost'. *CRAN*, *1.7.5.1*. Descargado de https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

Zamani, M. e. a. (2019). Pm2.5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7). Descargado de https://www.mdpi.com/2073-4433/10/7/373 doi: 10.3390/atmos10070373

10. Anexos

10.1. Variables Complementarias de las Mejores Predictoras

Figura 31: Distribución y Densidad de Variable Acciones Disciplinarias

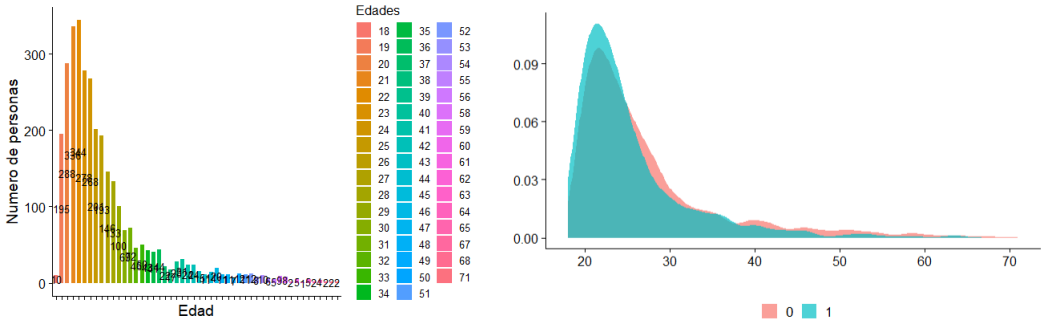


(a) Distribución Acciones Disciplinarias

(b) Densidad Acciones Disciplinarias

Fuente: Elaboración propia con datos de SITEL

Figura 32: Distribución y Densidad de Variable Edad

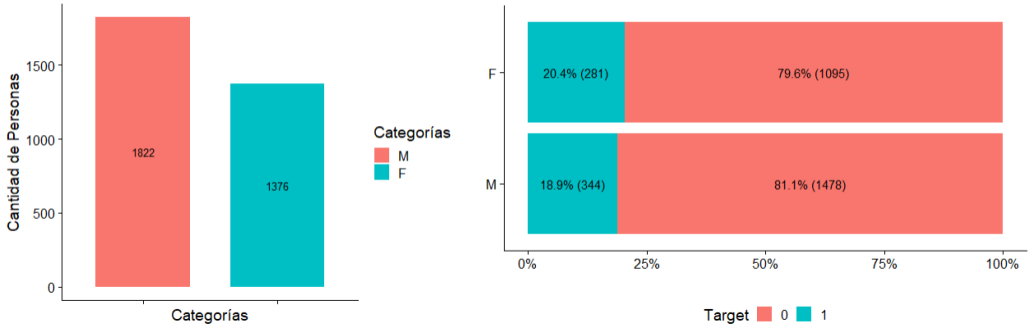


(a) Distribución Edad

(b) Densidad Edad

Fuente: Elaboración propia con datos de SITEL

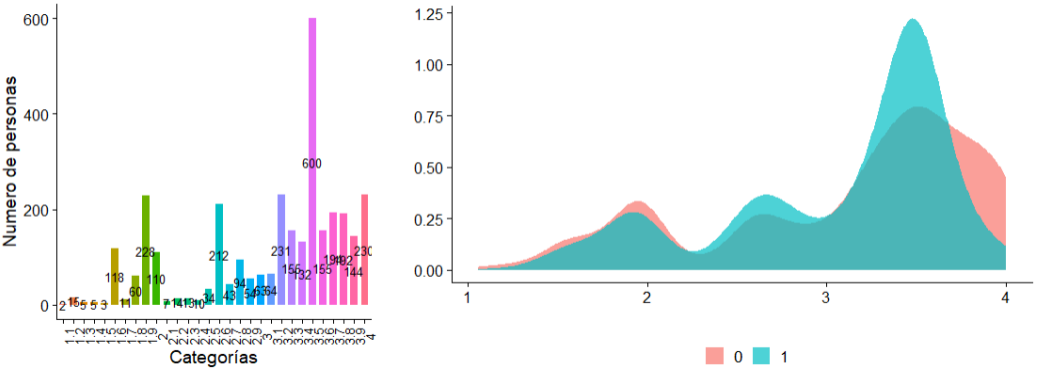
Figura 33: Distribución y Densidad de Variable Género



(a) Distribución Género (b) Densidad Género

Fuente: Elaboración propia con datos de SITEL

Figura 34: Distribución y Densidad de Variable Nota Evaluación



(a) Distribución Nota Evaluación (b) Densidad Nota Evaluación

Fuente: Elaboración propia con datos de SITEL

10.2. Resultados de la aplicación del método *Backward Stepwise*

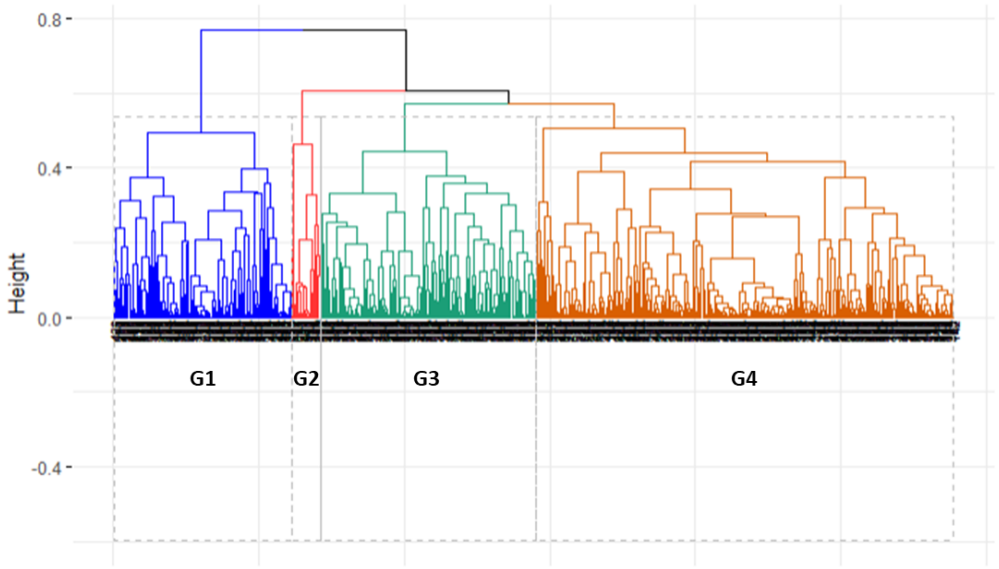
Cuadro 17: Resumen de los Resultados del Método *Backward Stepwise*

	Df	*Deviance*	AIC
none		2767.1	2825.1
Estado Civil	2	2773.4	2827.4
Provincia	6	2781.5	2827.5
Reingreso	1	2777.5	2833.5
Salario	1	2788.7	2844.7
Antigüedad	1	2788.8	2844.8
Sucursal	5	2811.6	2859.6
Teletrabajo	1	2830.6	2886.6
Cliente	11	2988.6	3024.6

Fuente: Elaboración propia

10.3. Modelación Dendogramas Variables Predictoras Completas

Figura 35: Dendograma de Clasificaciones Correctas con Variables Predictoras Completas



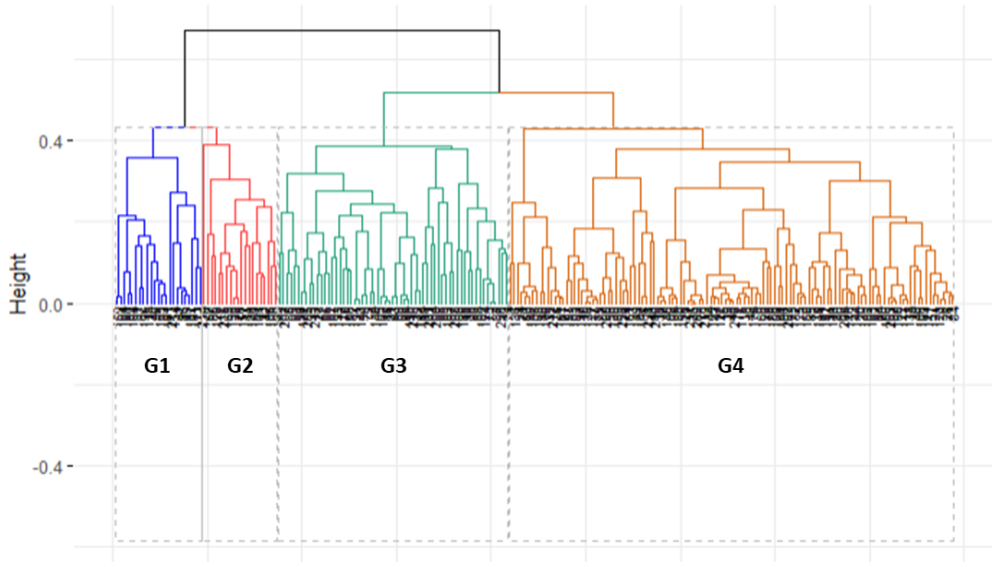
Fuente: Elaboración propia

Cuadro 18: Resumen del Clustering Jerárquico de las Personas Clasificadas Correctamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	122	287	148	20
Acciones Disciplinarias	$\overline{X} = 0{,}68$	$\overline{X} = 1{,}94$	$\overline{X} = 0{,}60$	$\overline{X} = 0{,}65$
Antigüedad	$\overline{X} = 310$	$\overline{X} = 713$	$\overline{X} = 530$	$\overline{X} = 389$
Cliente	Cliente B: 67 % Cliente K: 10 % Cliente I: 9 % Otros: 14 %	Cliente B: 82 % Cliente F: 6 % Cliente J: 5 % Otros: 7 %	Cliente I: 26 % Cliente K: 24 % Cliente E: 22 % Otros: 28 %	Cliente K: 85 % Cliente H: 10 % Cliente C: 5 %
Edad	$\overline{X} = 24{,}3$	$\overline{X} = 27{,}9$	$\overline{X} = 28{,}7$	$\overline{X} = 25{,}0$
Estado Civil	Solteros: 100 %	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 100 %
Género	M: 53 % F: 47 %	M: 63 % F: 37 %	M: 62 % F: 38 %	M: 30 % F: 70 %
Nota Evaluación	$\overline{X} = 3{,}11$	$\overline{X} = 3{,}50$	$\overline{X} = 2{,}80$	$\overline{X} = 2{,}14$
Provincia	San José: 51 % Alajuela: 15 % Heredia: 14 % Otros: 20 %	San José: 48 % Heredia: 22 % Alajuela: 13 % Otros: 17 %	San José: 55 % Alajuela: 14 % Heredia: 12 % Otros: 19 %	San José: 35 % Alajuela: 25 % Cartago: 15 % Otros: 25 %
Reingreso	$\overline{X} = 1{,}07$	$\overline{X} = 1{,}02$	$\overline{X} = 1{,}02$	$\overline{X} = 1{,}15$
Salario	$\overline{X} = 512.957$	$\overline{X} = 552.449$	$\overline{X} = 559.274$	$\overline{X} = 596.000$
Sucursal	Virtual: 46 % San Pedro: 35 % Heredia: 9 % Otros: 10 %	Heredia: 49 % San Pedro: 38 % Hatillo: 11 % Liberia: 2 %	Virtual: 100 %	Virtual: 90 % Hatillo: 5 % Moravia: 5 %
Teletrabajo	Sí: 43 % No: 57 %	No: 100 %	Sí: 88 % No: 12 %	Sí: 5 % No: 95 %
Salidas Reales 3 Meses	100 %	0 %	0 %	0 %
Salidas Reales Indefinido	100 %	20 %	13 %	45 %

Fuente: Elaboración propia

Figura 36: Dendograma de las Clasificaciones Incorrectas con Variables Predictoras Completas



Fuente: Elaboración propia

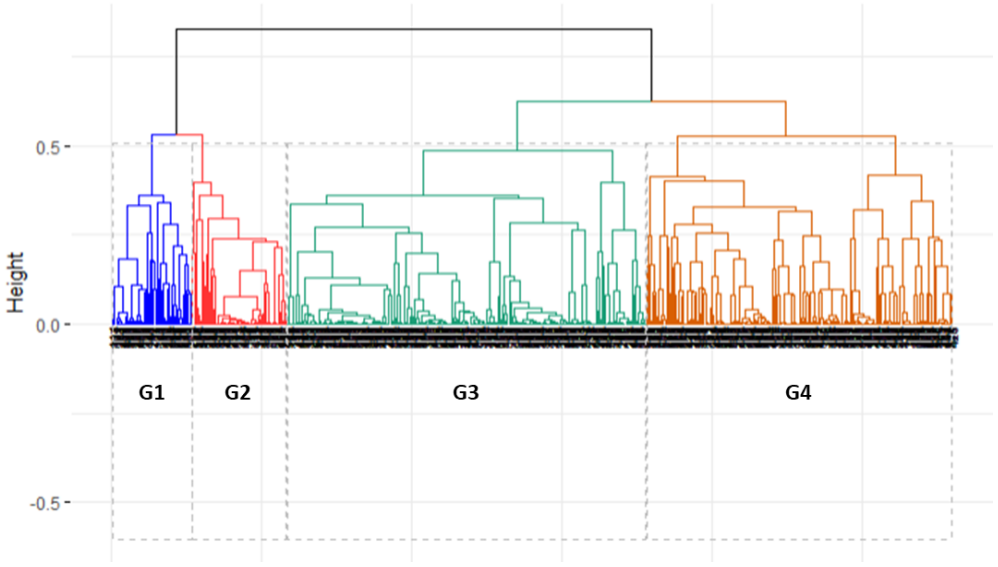
Cuadro 19: Resumen del Clustering Jerárquico de las Personas Clasificadas Incorrectamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	118	61	20	23
Acciones Disciplinarias	$\overline{X} = 0{,}75$	$\overline{X} = 0{,}21$	$\overline{X} = 0{,}40$	$\overline{X} = 2{,}74$
Antigüedad	$\overline{X} = 310$	$\overline{X} = 305$	$\overline{X} = 428$	$\overline{X} = 695$
Cliente	Cliente B: 85 % Cliente D: 14 % Cliente K: 1 %	Cliente K: 36 % Cliente I: 26 % Cliente L: 20 % Otros: 18 %	Cliente A: 30 % Cliente I: 15 % Cliente K: 15 % Otros: 40 %	Cliente B: 87 % Cliente D: 9 % Cliente H: 4 %
Edad	$\overline{X} = 23{,}6$	$\overline{X} = 24{,}7$	$\overline{X} = 28{,}7$	$\overline{X} = 27{,}4$
Estado Civil	Solteros: 99 % Casados: 1 %	Solteros: 100 %	Solteros: 95 % Casados: 5 %	Solteros: 100 %
Género	M: 48 % F: 52 %	M: 64 % F: 36 %	M: 75 % F: 25 %	M: 65 % F: 35 %
Nota Evaluación	$\overline{X} = 3{,}27$	$\overline{X} = 2{,}41$	$\overline{X} = 2{,}93$	$\overline{X} = 3{,}38$
Provincia	San José: 46 % Heredia: 18 % Guanacaste: 12 % Otros: 24 %	San José: 43 % Alajuela: 21 % Heredia: 15 % Otros: 21 %	San José: 55 % Alajuela: 15 % Cartago: 15 % Otros: 15 %	San José: 57 % Alajuela: 17 % Cartago: 13 % Heredia: 13 %
Reingreso	$\overline{X} = 1{,}04$	$\overline{X} = 1{,}20$	$\overline{X} = 1{,}10$	$\overline{X} = 1{,}04$
Salario	$\overline{X} = 499.306$	$\overline{X} = 540.474$	$\overline{X} = 650.370$	$\overline{X} = 551.957$
Sucursal	San Pedro: 43 % Heredia: 28 % Hatillo: 15 % Otros: 14 %	Virtual: 100 %	Virtual: 100 %	San Pedro: 65 % Heredia: 22 % Hatillo: 9 % Moravia: 4 %
Teletrabajo	No: 100 %	Sí: 100 %	Sí: 100 %	No: 100 %
Salidas Reales 3 Meses	0 %	0 %	100 %	100 %
Salidas Reales Indefinido	31 %	26 %	100 %	100 %

Fuente: Elaboración propia

10.4. Modelación Dendogramas Variables Predictoras Seleccionadas

Figura 37: Dendograma de Clasificaciones Correctas con Variables Predictoras Seleccionadas



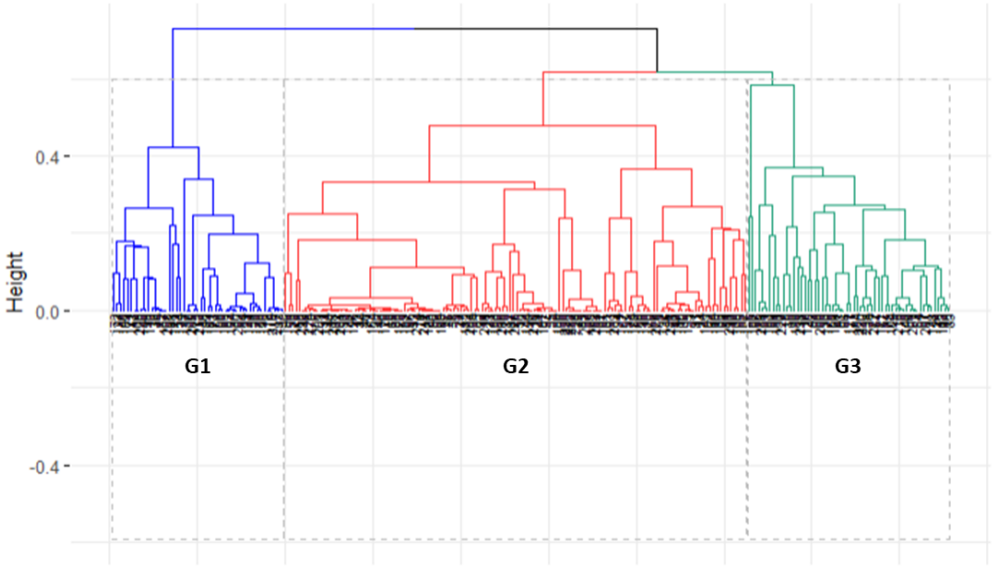
Fuente: Elaboración propia

Cuadro 20: Resumen del Clustering Jerárquico de las Personas Clasificadas Correctamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	53	240	204	63
Antigüedad	$\overline{X} = 281$	$\overline{X} = 803$	$\overline{X} = 487$	$\overline{X} = 400$
Cliente	Cliente B: 45 % Cliente K: 21 % Cliente I: 17 % Otros: 17 %	Cliente B: 92 % Cliente J: 6 % Cliente K: 1 % Cliente G: 1 %	Cliente K: 24 % Cliente E: 18 % Cliente I: 17 % Otros: 41 %	Cliente B: 86 % Cliente D: 6 % Cliente H: 3 % Otros: 5 %
Estado Civil	Solteros: 100 %	Solteros: 94 % Casados: 5 % Divorciados: 1 %	Solteros: 96 % Casados: 3 % Divorciados: 1 %	Solteros: 100 %
Provincia	San José: 40 % Alajuela: 24 % Heredia: 15 % Otros: 21 %	San José: 41 % Heredia: 26 % Alajuela: 17 % Otros: 16 %	San José: 54 % Alajuela: 13 % Heredia: 12 % Otros: 21 %	San José: 57 % Heredia: 14 % Alajuela: 11 % Otros: 18 %
Reingreso	$\overline{X} = 1{,}15$	$\overline{X} = 1{,}01$	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}05$
Salario	$\overline{X} = 552.257$	$\overline{X} = 532.977$	$\overline{X} = 585.341$	$\overline{X} = 512.276$
Sucursal	Virtual: 98 % San Pedro: 2 %	Heredia: 62 % San Pedro: 35 % Liberia: 2 % Otros: 1 %	Virtual: 83 % Hatillo: 17 %	San Pedro: 64 % Heredia: 14 % Liberia: 8 % Otros: 14 %
Teletrabajo	Sí: 96 % No: 4 %	No: 100 %	Sí: 66 % No: 34 %	No: 100 %
Salidas Reales 3 Meses	100 %	0 %	0 %	100 %
Salidas Reales Indefinido	100 %	21 %	16 %	100 %

Fuente: Elaboración propia

Figura 38: Dendograma de las Clasificaciones Incorrectas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

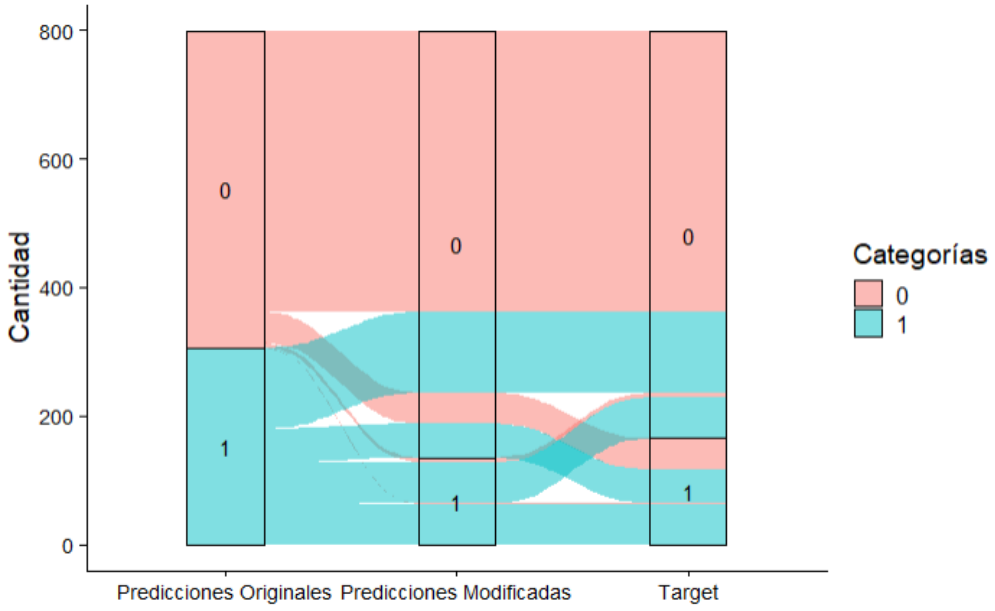
Cuadro 21: Resumen del Clustering Jerárquico de las Personas Clasificadas Incorrectamente

Variable	Grupo 1	Grupo 2	Grupo 3
Observaciones	132	58	49
Antigüedad	$\overline{X} = 307$	$\overline{X} = 229$	$\overline{X} = 454$
Cliente	Cliente B: 88 % Cliente D: 10 % Cliente A: 1 % Cliente K: 1 %	Cliente K: 43 % Cliente I: 33 % Cliente L: 10 % Otros: 14 %	Cliente B: 49 % Cliente A: 16 % Cliente D: 8 % Otros: 27 %
Estado Civil	Solteros: 99 % Casados: 1 %	Solteros: 100 %	Solteros: 98 % Casados: 2 %
Provincia	San José: 54 % Heredia: 16 % Cartago: 11 % Otros: 19 %	San José: 43 % Alajuela: 24 % Cartago: 17 % Otros: 16 %	San José: 61 % Cartago: 12 % Alajuela: 10 % Otros: 17 %
Reingreso	$\overline{X} = 1{,}04$	$\overline{X} = 1{,}14$	$\overline{X} = 1{,}02$
Salario	$\overline{X} = 497.703$	$\overline{X} = 553.652$	$\overline{X} = 541.859$
Sucursal	San Pedro: 58 % Heredia: 19 % Hatillo: 11 % Otros: 12 %	Virtual: 98 % Moravia: 2 %	Virtual: 43 % San Pedro: 35 % Heredia: 14 % Hatillo: 8 %
Teletrabajo	No: 100 %	Sí: 100 %	Sí: 41 % No: 59 %
Salidas Reales 3 Meses	0 %	0 %	100 %
Salidas Reales Indefinido	27 %	33 %	100 %

Fuente: Elaboración propia

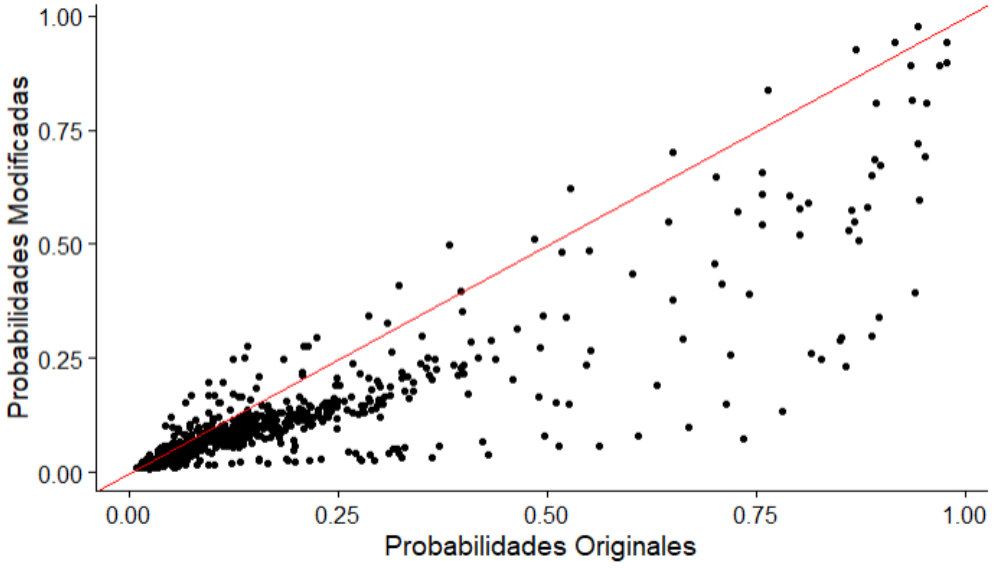
10.5. Análisis Sensibilidad Variación del Salario en un 10 %

Figura 39: Predicciones al Modificar un 10 % el Salario



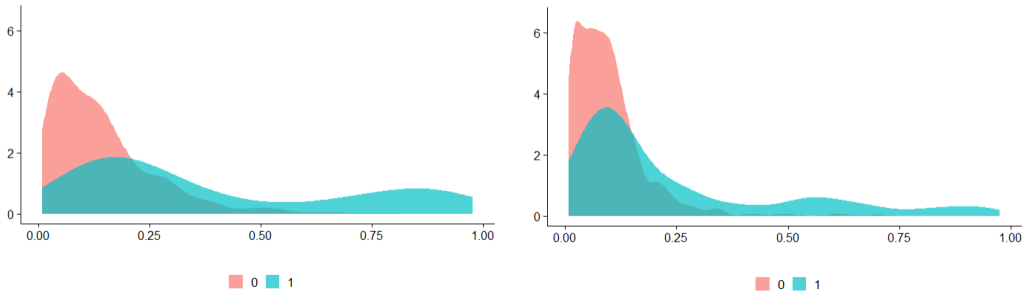
Fuente: Elaboración propia

Figura 40: Gráfico Dispersión de Probabilidades al Modificar un 10 % el Salario



Fuente: Elaboración propia

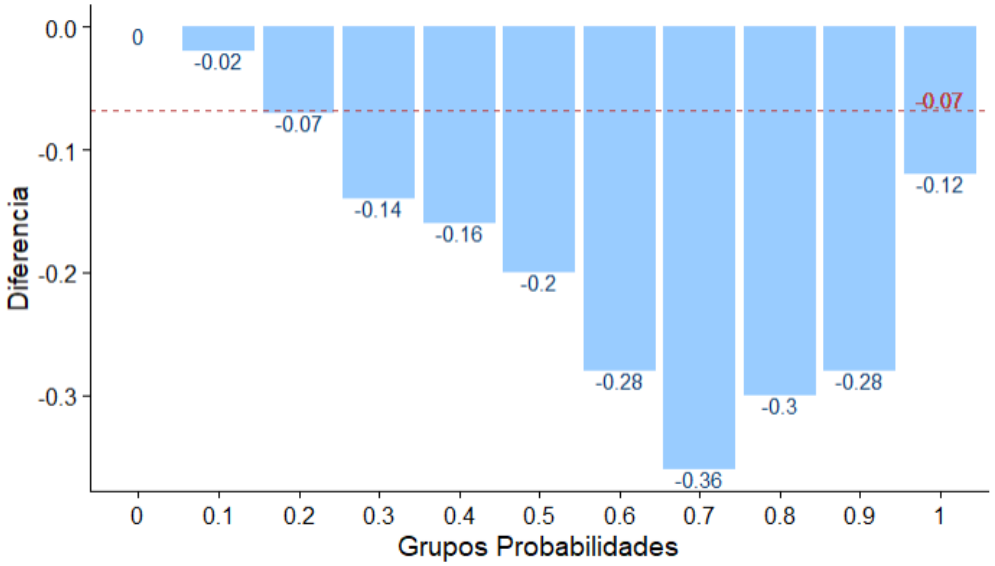
Figura 41: Distribuciones de Densidad por Categoría al Modificar un 10 % el Salario



(a) Distribuciones con Salario Original (b) Distribuciones con Salario Modificado

Fuente: Elaboración propia

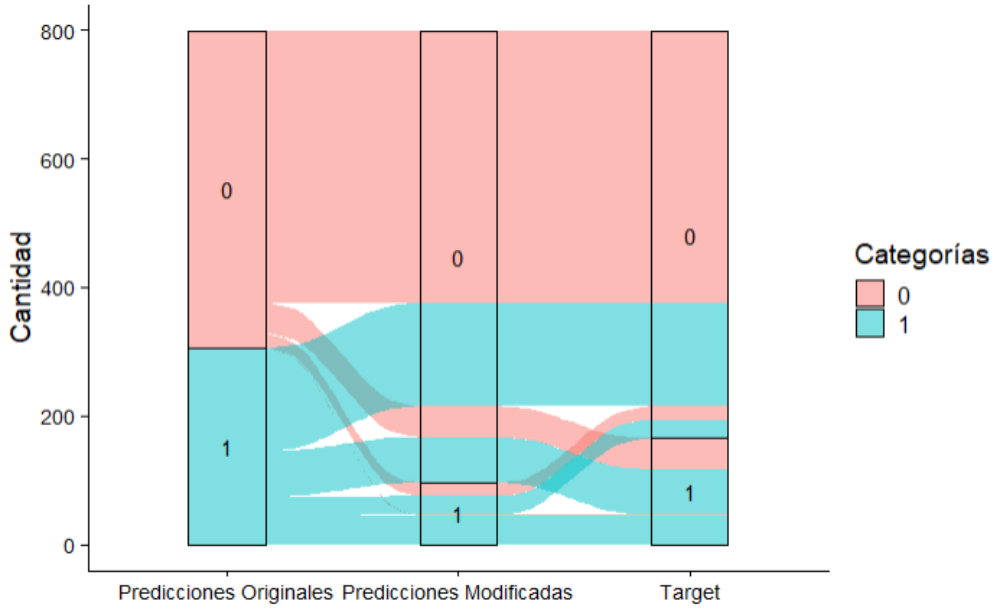
Figura 42: Análisis de Sensibilidad al Modificar un 10 % el Salario



Fuente: Elaboración propia

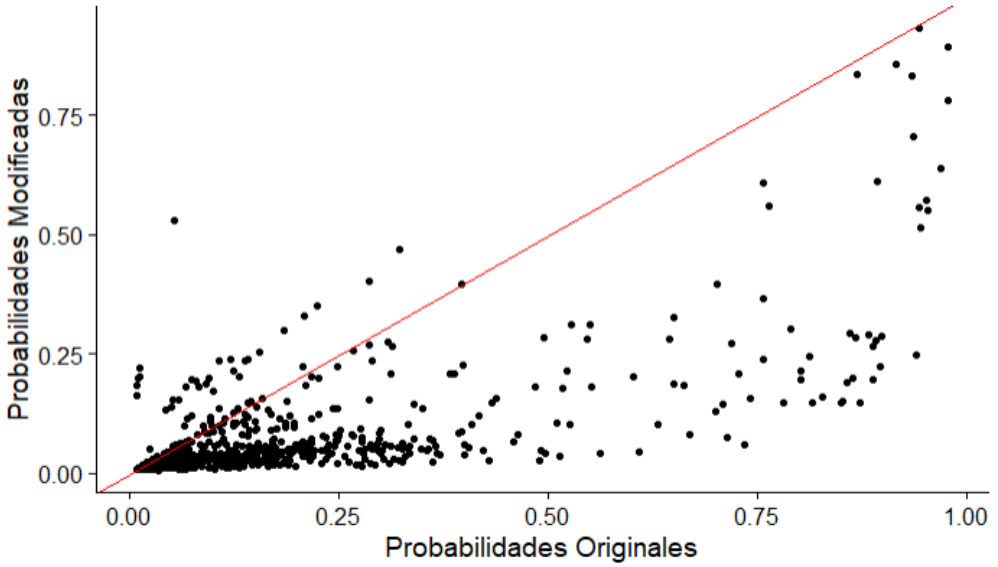
10.6. Análisis Sensibilidad Variación del Salario en un 20 %

Figura 43: Predicciones al Modificar un 20 % el Salario



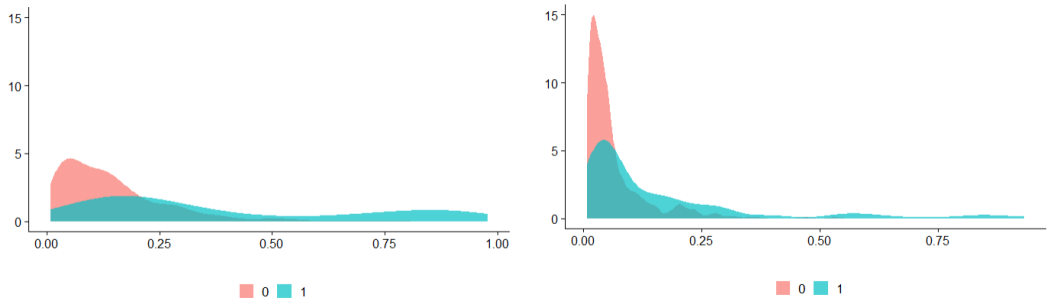
Fuente: Elaboración propia

Figura 44: Gráfico Dispersión de Probabilidades al Modificar un 20 % el Salario



Fuente: Elaboración propia

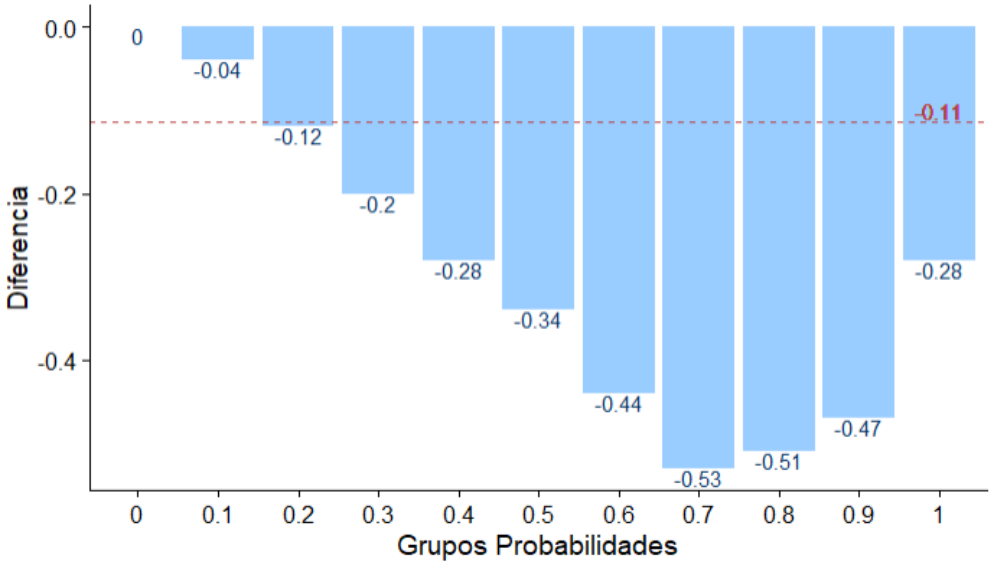
Figura 45: Distribuciones de Densidad por Categoría al Modificar un 20 % el Salario



(a) Distribuciones con Salario Original (b) Distribuciones con Salario Modificado

Fuente: Elaboración propia

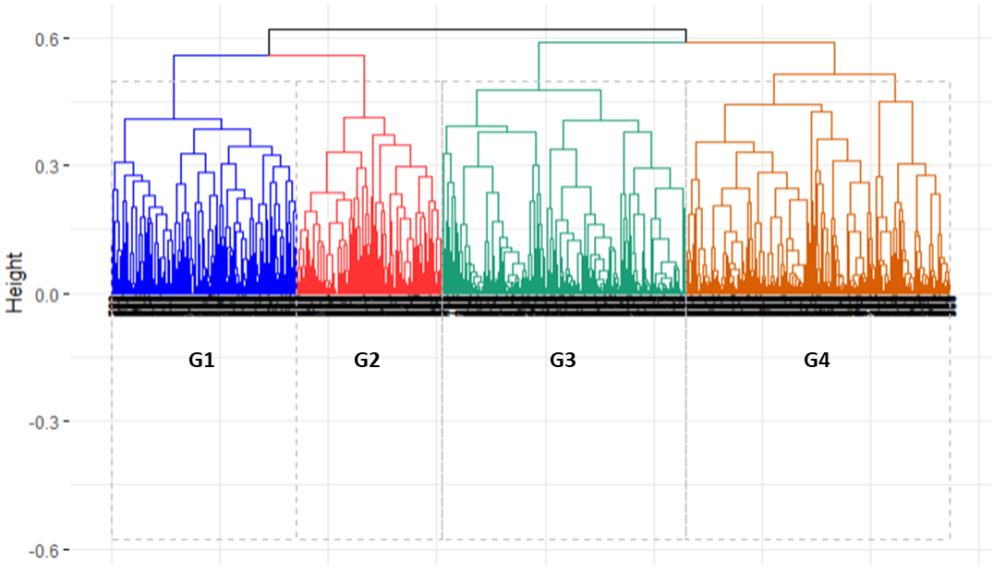
Figura 46: Análisis de Sensibilidad al Modificar un 20 % el Salario



Fuente: Elaboración propia

10.7. Estimación Fuera de Muestra Dendogramas Variables Predictoras Completas

Figura 47: Dendograma de Clasificaciones Salidas con Variables Predictoras Completas



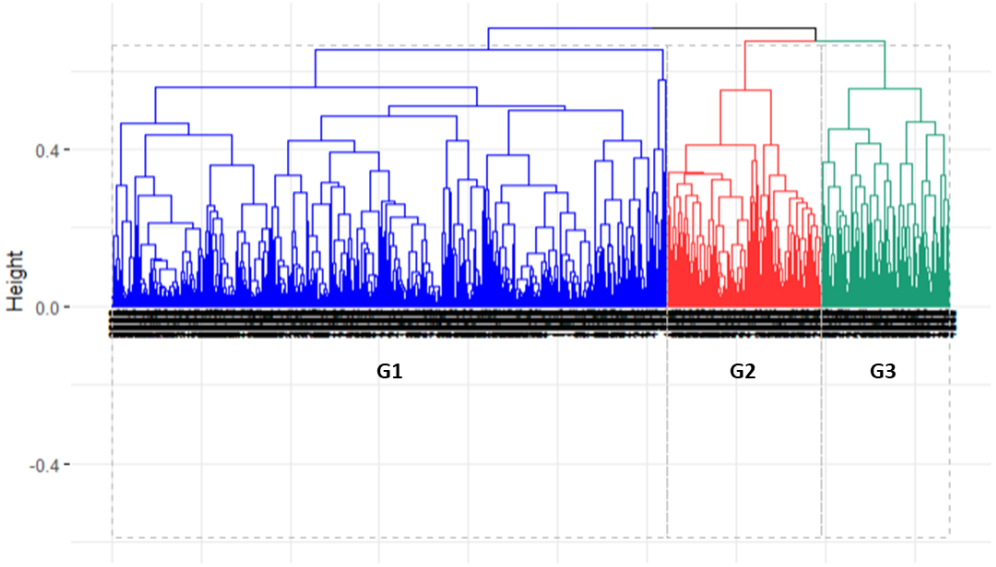
Fuente: Elaboración propia

Cuadro 22: Resumen del Clustering Jerárquico de las Personas Clasificadas como Salidas

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	232	296	422	392
Acciones Disciplinarias	$\overline{X} = 0{,}20$	$\overline{X} = 0{,}32$	$\overline{X} = 1{,}07$	$\overline{X} = 1{,}07$
Antigüedad	$\overline{X} = 156$	$\overline{X} = 197$	$\overline{X} = 249$	$\overline{X} = 240$
Cliente	Cliente K: 39 % Cliente I: 35 % Cliente L: 14 % Otros: 12 %	Cliente L: 28 % Cliente K: 22 % Cliente I: 20 % Otros: 30 %	Cliente B: 59 % Cliente D: 32 % Cliente K: 4 % Otros: 5 %	Cliente B: 91 % Cliente K: 6 % Cliente D: 1 % Otros: 2 %
Edad	$\overline{X} = 26{,}3$	$\overline{X} = 26{,}1$	$\overline{X} = 24{,}3$	$\overline{X} = 23{,}4$
Estado Civil	Solteros: 100 %	Solteros: 100 %	Solteros: 99 % Casados: 1 %	Solteros: 100 %
Género	M: 49 % F: 51 %	M: 70 % F: 30 %	M: 44 % F: 56 %	M: 45 % F: 55 %
Nota Evaluación	$\overline{X} = 3{,}17$	$\overline{X} = 3{,}01$	$\overline{X} = 3{,}21$	$\overline{X} = 3{,}36$
Provincia	San José: 42 % Alajuela: 22 % Cartago: 12 % Otros: 24 %	San José: 35 % Alajuela: 19 % Heredia: 17 % Otros: 29 %	San José: 53 % Heredia: 19 % Alajuela: 10 % Otros: 18 %	San José: 47 % Cartago: 15 % Heredia: 15 % Otros: 23 %
Reingreso	$\overline{X} = 1{,}14$	$\overline{X} = 1{,}18$	$\overline{X} = 1{,}07$	$\overline{X} = 1{,}07$
Salario	$\overline{X} = 526.374$	$\overline{X} = 523.328$	$\overline{X} = 501.174$	$\overline{X} = 495.218$
Sucursal	Virtual: 99 % San Pedro: 1 %	Virtual: 99 % Heredia: 1 %	Hatillo: 35 % Heredia: 27 % San Pedro: 26 % Otros: 12 %	San Pedro: 68 % Heredia: 17 % Virtual: 11 % Otros: 4 %
Teletrabajo	Sí: 100 %	Sí: 97 % No: 3 %	Sí: 1 % No: 99 %	Sí: 1 % No: 99 %
Salidas Reales 3 Meses	81 %	0 %	8 %	55 %
Salidas Reales Indefinido	100 %	31 %	31 %	100 %

Fuente: Elaboración propia

Figura 48: Dendograma de Clasificaciones No Salidas con Variables Predictoras Completas



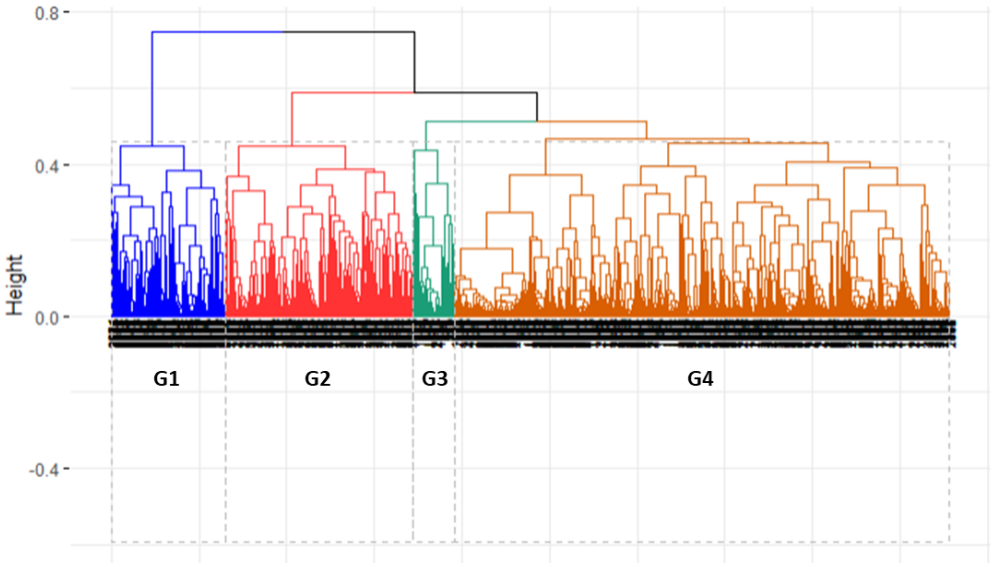
Fuente: Elaboración propia

Cuadro 23: Resumen del Clustering Jerárquico de las Personas Clasificadas como No Salidas

Variable	Grupo 1	Grupo 2	Grupo 3
Observaciones	1.306	365	302
Acciones Disciplinarias	$\overline{X} = 1{,}39$	$\overline{X} = 0{,}62$	$\overline{X} = 0{,}57$
Antigüedad	$\overline{X} = 691$	$\overline{X} = 575$	$\overline{X} = 435$
Cliente	Cliente B: 79 % Cliente K: 10 % Cliente F: 4 % Otros: 7 %	Cliente I: 30 % Cliente E: 25 % Cliente K: 17 % Otros: 28 %	Cliente K: 30 % Cliente I: 16 % Cliente B: 14 % Otros: 40 %
Edad	$\overline{X} = 27{,}5$	$\overline{X} = 29{,}5$	$\overline{X} = 28{,}7$
Estado Civil	Solteros: 96 % Casados: 3 % Divorciados: 1 %	Solteros: 94 % Casados: 5 % Divorciados: 1 %	Solteros: 99 % Casados: 1 %
Género	M: 57 % F: 43 %	M: 53 % F: 47 %	M: 67 % F: 33 %
Nota Evaluación	$\overline{X} = 3{,}55$	$\overline{X} = 3{,}40$	$\overline{X} = 3{,}28$
Provincia	San José: 44 % Heredia: 19 % Alajuela: 16 % Otros: 21 %	San José: 46 % Alajuela: 15 % Guanacaste: 12 % Otros: 27 %	San José: 57 % Alajuela: 14 % Heredia: 12 % Otros: 17 %
Reingreso	$\overline{X} = 1{,}04$	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}07$
Salario	$\overline{X} = 546.105$	$\overline{X} = 579.846$	$\overline{X} = 590.860$
Sucursal	San Pedro: 43 % Heredia: 36 % Virtual: 10 % Otros: 11 %	Virtual: 100 %	Virtual: 61 % Hatillo: 20 % Heredia: 15 % Otros: 4 %
Teletrabajo	Sí: 1 % No: 99 %	Sí: 100 %	Sí: 50 % No: 50 %
Salidas Reales 3 Meses	5 %	1 %	38 %
Salidas Reales Indefinido	32 %	9 %	100 %

Fuente: Elaboración propia

Figura 49: Dendograma de Clasificaciones Correctas con Variables Predictoras Completas



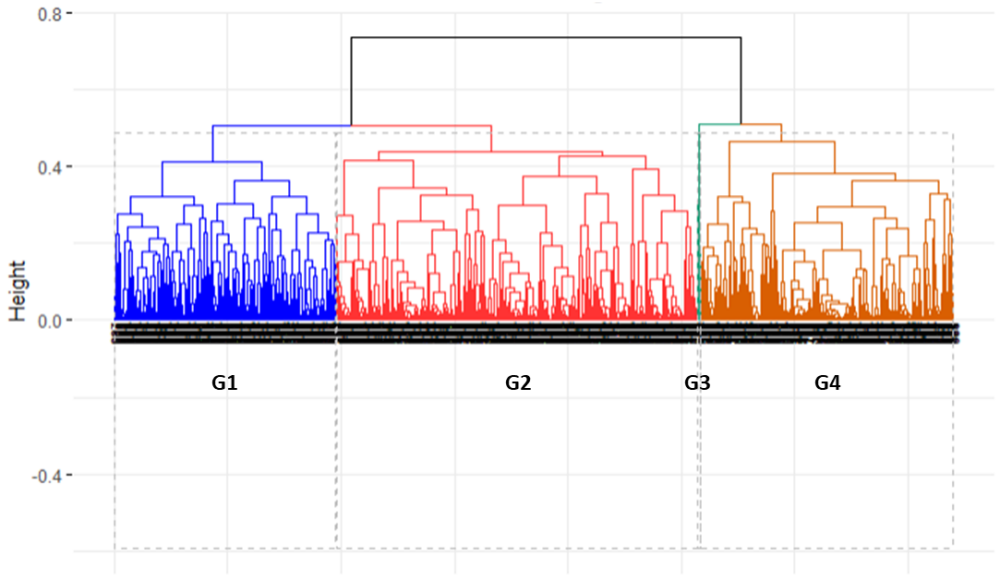
Fuente: Elaboración propia

Cuadro 24: Resumen del Clustering Jerárquico de las Personas Clasificadas Correctamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	103	466	1.227	428
Acciones Disciplinarias	$\overline{X} = 1{,}89$	$\overline{X} = 0{,}68$	$\overline{X} = 1{,}29$	$\overline{X} = 0{,}34$
Antigüedad	$\overline{X} = 1{,}951$	$\overline{X} = 565$	$\overline{X} = 577$	$\overline{X} = 189$
Cliente	Cliente B: 93 % Cliente K: 3 % Cliente C: 3 % Cliente I: 1 %	Cliente I: 31 % Cliente E: 24 % Cliente K: 18 % Otros: 27 %	Cliente B: 71 % Cliente K: 13 % Cliente F: 7 % Otros: 9 %	Cliente B: 46 % Cliente K: 19 % Cliente I: 17 % Otros: 18 %
Edad	$\overline{X} = 37{,}9$	$\overline{X} = 29{,}6$	$\overline{X} = 27{,}0$	$\overline{X} = 24{,}8$
Estado Civil	Solteros: 52 % Casados: 38 % Divorciados: 10 %	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 98 % Casados: 1 % Divorciados: 1 %	Solteros: 100 %
Género	M: 26 % F: 74 %	M: 59 % F: 41 %	M: 60 % F: 40 %	M: 48 % F: 52 %
Nota Evaluación	$\overline{X} = 3{,}66$	$\overline{X} = 3{,}39$	$\overline{X} = 3{,}52$	$\overline{X} = 1{,}11$
Provincia	Heredia: 67 % San José: 17 % Alajuela: 13 % Otros: 3 %	San José: 50 % Alajuela: 15 % Guanacaste: 11 % Otros: 24 %	San José: 47 % Alajuela: 16 % Heredia: 15 % Otros: 22 %	San José: 47 % Alajuela: 16 % Heredia: 14 % Otros: 23 %
Reingreso	$\overline{X} = 1{,}03$	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}04$	$\overline{X} = 1{,}11$
Salario	$\overline{X} = 538.684$	$\overline{X} = 580.882$	$\overline{X} = 557.911$	$\overline{X} = 509.649$
Sucursal	Heredia: 87 % San Pedro: 9 % Virtual: 3 % Liberia: 1 %	Virtual: 100 %	San Pedro: 39 % Heredia: 33 % Virtual: 13 % Otros: 15 %	Virtual: 47 % San Pedro: 31 % Heredia: 12 % Otros: 10 %
Teletrabajo	Sí: 1 % No: 99 %	Sí: 100 %	Sí: 1 % No: 99 %	Sí: 43 % No: 57 %
Salidas Reales 3 Meses	0 %	0 %	0 %	100 %
Salidas Reales Indefinido	2 %	26 %	34 %	100 %

Fuente: Elaboración propia

Figura 50: Dendograma de las Clasificaciones Incorrectas con Variables Predictoras Completas



Fuente: Elaboración propia

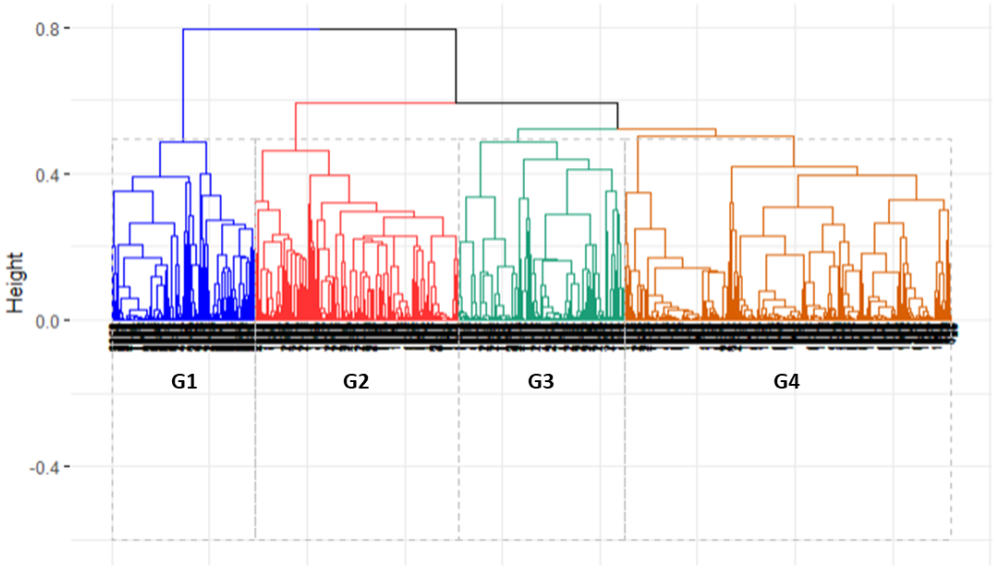
Cuadro 25: Resumen del Clustering Jerárquico de las Personas Clasificadas Incorrectamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	5	173	347	566
Acciones Disciplinarias	$\overline{X} = 1{,}75$	$\overline{X} = 0{,}77$	$\overline{X} = 0{,}28$	$\overline{X} = 0{,}77$
Antigüedad	$\overline{X} = 1{,}964$	$\overline{X} = 442$	$\overline{X} = 191$	$\overline{X} = 260$
Cliente	Cliente B: 75 % Cliente C: 25 %	Cliente B: 57 % Cliente K: 18 % Cliente I: 7 % Otros: 18 %	Cliente L: 29 % Cliente K: 24 % Cliente I: 19 % Otros: 28 %	Cliente B: 73 % Cliente D: 18 % Cliente K: 5 % Otros: 4 %
Edad	$\overline{X} = 44{,}8$	$\overline{X} = 26{,}4$	$\overline{X} = 25{,}8$	$\overline{X} = 24{,}1$
Estado Civil	Casados: 100 %	Solteros: 99 % Casados: 1 %	Solteros: 100 %	Solteros: 99 % Casados: 1 %
Género	F: 100 %	M: 57 % F: 43 %	M: 61 % F: 39 %	M: 47 % F: 53 %
Nota Evaluación	$\overline{X} = 3{,}77$	$\overline{X} = 3{,}39$	$\overline{X} = 3{,}03$	$\overline{X} = 3{,}25$
Provincia	Alajuela: 50 % San José: 25 % Heredia: 25 %	San José: 51 % Alajuela: 13 % Heredia: 12 % Otros: 24 %	San José: 35 % Alajuela: 21 % Heredia: 16 % Otros: 28 %	San José: 51 % Heredia: 18 % Cartago: 11 % Otros: 20 %
Reingreso	$\overline{X} = 1{,}00$	$\overline{X} = 1{,}08$	$\overline{X} = 1{,}18$	$\overline{X} = 1{,}06$
Salario	$\overline{X} = 579.641$	$\overline{X} = 532.372$	$\overline{X} = 522.674$	$\overline{X} = 499.304$
Sucursal	Heredia: 50 % Virtual: 25 % San Pedro: 25 %	San Pedro: 37 % Virtual: 33 % Heredia: 20 % Otros: 10 %	Virtual: 100 %	San Pedro: 42 % Heredia: 24 % Hatillo: 20 % Otros: 14 %
Teletrabajo	Sí: 25 % No: 75 %	Sí: 27 % No: 73 %	Sí: 97 % No: 3 %	Sí: 1 % No: 99 %
Salidas Reales 3 Meses	100 %	100 %	0 %	0 %
Salidas Reales Indefinido	100 %	100 %	39 %	47 %

Fuente: Elaboración propia

10.8. Estimación Fuera de Muestra Dendogramas Variables Predictoras Seleccionadas

Figura 51: Dendograma de Clasificaciones Correctas con Variables Predictoras Seleccionadas



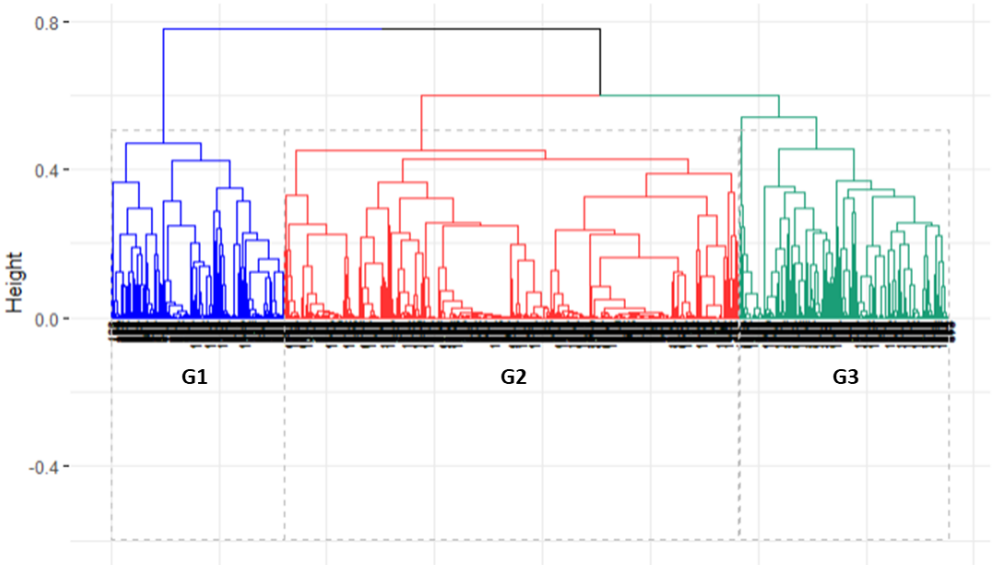
Fuente: Elaboración propia

Cuadro 26: Resumen del Clustering Jerárquico de las Personas Clasificadas Correctamente

Variable	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Observaciones	911	569	460	436
Antigüedad	$\overline{X} = 929$	$\overline{X} = 558$	$\overline{X} = 408$	$\overline{X} = 183$
Cliente	Cliente B: 95 % Cliente J: 3 % Cliente L: 1 % Otros: 1 %	Cliente I: 34 % Cliente E: 21 % Cliente K: 16 % Otros: 29 %	Cliente K: 49 % Cliente F: 24 % Cliente D: 12 % Otros: 15 %	Cliente B: 51 % Cliente K: 19 % Cliente I: 14 % Otros: 16 %
Estado Civil	Solteros: 93 % Casados: 5 % Divorciados: 2 %	Solteros: 95 % Casados: 4 % Divorciados: 1 %	Solteros: 99 % Casados: 1 %	Solteros: 100 %
Provincia	San José: 39 % Heredia: 25 % Alajuela: 20 % Otros: 16 %	San José: 47 % Alajuela: 15 % Guanacaste: 11 % Otros: 27 %	San José: 54 % Alajuela: 11 % Heredia: 10 % Otros: 25 %	San José: 51 % Heredia: 14 % Cartago: 14 % Otros: 21 %
Reingreso	$\overline{X} = 1{,}02$	$\overline{X} = 1{,}09$	$\overline{X} = 1{,}05$	$\overline{X} = 1{,}10$
Salario	$\overline{X} = 538.871$	$\overline{X} = 577.217$	$\overline{X} = 625.590$	$\overline{X} = 511.839$
Sucursal	Heredia: 59 % San Pedro: 40 % Liberia: 1 %	Virtual: 100 %	Virtual: 45 % Hatillo: 37 % San Pedro: 8 % Otros: 10 %	Virtual: 42 % San Pedro: 37 % Heredia: 9 % Otros: 12 %
Teletrabajo	No: 100 %	Sí: 100 %	No: 100 %	Sí: 39 % No: 61 %
Salidas Reales 3 Meses	0 %	0 %	0 %	100 %
Salidas Reales Indefinido	30 %	26 %	30 %	100 %

Fuente: Elaboración propia

Figura 52: Dendograma de las Clasificaciones Incorrectas con Variables Predictoras Seleccionadas



Fuente: Elaboración propia

Cuadro 27: Resumen del Clustering Jerárquico de las Personas Clasificadas Incorrectamente

Variable	Grupo 1	Grupo 2	Grupo 3
Observaciones	527	243	170
Antigüedad	$\overline{X} = 300$	$\overline{X} = 231$	$\overline{X} = 522$
Cliente	Cliente B: 90 % Cliente D: 6 % Cliente K: 1 % Otros: 3 %	Cliente L: 25 % Cliente K: 23 % Cliente E: 13 % Otros: 39 %	Cliente B: 50 % Cliente K: 19 % Cliente I: 10 % Otros: 21 %
Estado Civil	Solteros: 98 % Casados: 1 % Divorciados: 1 %	Solteros: 99 % Casados: 1 %	Solteros: 98 % Casados: 2 %
Provincia	San José: 47 % Heredia: 17 % Cartago: 15 % Otros: 21 %	San José: 48 % Alajuela: 17 % Heredia: 13 % Otros: 22 %	San José: 46 % Heredia: 17 % Alajuela: 15 % Otros: 22 %
Reingreso	$\overline{X} = 1{,}06$	$\overline{X} = 1{,}11$	$\overline{X} = 1{,}08$
Salario	$\overline{X} = 500.095$	$\overline{X} = 530.274$	$\overline{X} = 533.796$
Sucursal	San Pedro: 57 % Heredia: 25 % Liberia: 9 % Otros: 9 %	Virtual: 88 % Hatillo: 11 % Heredia: 1 %	Virtual: 38 % San Pedro: 29 % Heredia: 26 % Otros: 7 %
Teletrabajo	No: 100 %	Sí: 88 % No: 12 %	Sí: 30 % No: 70 %
Salidas Reales 3 Meses	0 %	0 %	100 %
Salidas Reales Indefinido	44 %	35 %	100 %

Fuente: Elaboración propia

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Sergio Ramírez Rodríguez, con cédula de identidad 1-1546-0809, en mi condición de autor del TFG titulado PREDICCIÓN DE LA ROTACIÓN DE PERSONAL EN LA EMPRESA SITEL COSTA RICA

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

Sergio Ramírez

SERGIO VINICIO RAMIREZ RODRIGUEZ (FIRMA)

Firmado digitalmente por SERGIO VINICIO RAMIREZ RODRIGUEZ (FIRMA)
Fecha: 2025.03.26 12:02:10 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.