**APLICACIÓN DE UN MÉTODO DE CLASIFICACIÓN SUPERVISADO A UN PROBLEMA DE SELECCIÓN DE CLIENTES MANEJANTES DE COMPETENCIA PARA BAVARIA S.A.**

*Seminario de Analítica y Ciencia de datos*

*Cristian Bolaños, Mateo Cabarcas*

**Resumen Descriptivo del Proyecto**

Este proyecto surge bajo la oportunidad de seleccionar correctamente los clientes para ejecutar la estrategia de participación de mercado de Bavaria. El objetivo principal es diseñar e implementar un algoritmo de clasificación que permita a través de históricos de los clientes clasificarlos en una base de competencia y poder disminuir la incertidumbre con la que se toma esta decisión en la actualidad. Los resultados esperados son por lo menos ganar un punto de participación de mercado para Bavaria en la regional ANDES.

**Marco Teórico:**

1. **Revisión de literatura**

En primer lugar, es preciso recalcar que el objetivo principal del proyecto es clasificar la base de clientes de Bavaria para la regional ANDES en clientes “base foco” o “no base foco”. Los clientes de la base foco son aquellos en los que posteriormente se aplicarán distintas estrategias de Marketshare como planes promocionales o de implementaciones, destinadas a que el volumen de venta de productos de Bavaria sea mayor que el de la competencia, dando como resultado un incremento en la participación de mercado. Luego de una revisión bibliográfica exhaustiva, se identifica que en general la técnica que se busca aplicar en el proyecto se puede asemejar a la que se emplea en marketing para la predicción de abandono. En muchos casos como lo mencionan (Zadoo, Jagtap, Khule, Kedari, & Khedkar, 2022) esta es una tarea de clasificación binaria y es semejante al caso de estudio ya que el algoritmo debe seleccionar los clientes, basados en ciertas variables como disminución de ROS[[1]](#footnote-2) de cierto SKU, reporte de competencia y demás y arrojar una respuesta binaria, de si debe pertenecer o no a la base foco por su propensión a adquirir productos de la competencia.

Un buen acercamiento a la tarea de predicción de abandono y su importancia en el ámbito competitivo empresarial es hecha por (Agudelo, 2023). En su trabajo se destaca que “la capacidad de clasificar con precisión a los clientes le permite a las instituciones y a las empresas ofrecer servicios personalizados, opciones de crédito ajustados a su historial y campañas de marketing dirigidas”. Por lo tanto, como lo menciona el autor y aterrizándolo al objeto de estudio de este proyecto, el desarrollo de técnicas sólidas y sofisticadas para la clasificación de clientes, basadas en la información de historiales y tendencias de compra, ubicación geográfica y demás, podrían resultar muy útil para que Bavaria pueda tomar decisiones informadas, mitigue los riesgos y optimice los recursos al ser empleados en los clientes que generaran un mayor impacto en la variable objetivo (Marketshare).

Adicionalmente, la segmentación de clientes comúnmente mencionada como una importante herramienta en la literatura de CRM[[2]](#footnote-3), maximiza la satisfacción del cliente y, por lo tanto, mejora significativamente las ganancias de la empresa (Ozan, 2018). En este estudio Ozan menciona que, si bien el enfoque más común es etiquetar a los clientes con dos etiquetas diferentes, como "estándar" y "premium". También es posible segmentar los datos en un mayor número de grupos o definir las características de cada grupo dependiendo del caso de estudio. Además, menciona las ventajas de aplicar dichos métodos de selección sobre grandes volúmenes de datos: “cuando el número de clientes es alto, la operación es manual y no siempre es posible detectar clientes que se pueden segmentar en una categoría u otra”. Así, resulta muy útil la aplicación de métodos de aprendizaje automático que alcanzan a brindar una eficiencia en la clasificación de casi el 89.43% (Ozan, 2018). Este enfoque resolvería dos de los problemas planteados en este estudio: Disminución en la incertidumbre de la selección de los clientes base foco y optimización de los tiempos de gestión de la base foco mes a mes.

Finalmente, la idea del uso de aprendizaje automático en la resolución de problemas como el planteado en este proyecto es reforzada por (Wang, 2024) cuando menciona que los métodos tradicionales de segmentación de clientes, que enfrentan limitaciones a la hora de explorar la información latente, han impulsado un cambio hacia tecnologías avanzadas. Estas tecnologías, sin embargo, muestran una falta de explicabilidad en el proceso de aprendizaje, lo que crea desafíos para que los usuarios comprendan y confíen en los procedimientos de toma de decisiones. Los estudios futuros pueden mejorar la comprensión del proceso de optimización del modelo desde la perspectiva de la interpretabilidad. Al mismo tiempo, el entrenamiento del modelo depende de un volumen sustancial de datos, lo que impone un nuevo desafío para el presente proyecto y se puede configurar como un riesgo adicional. No obstante, se puede concluir a partir de la revisión bibliográfica que los métodos de aprendizaje supervisado a pesar de sus requerimientos se configuran como los ideales para resolver el problema de selección de clientes para la base foco en la regional ANDES para BAVARIA, básicamente por su buena efectividad en la clasificación (disminución de la incertidumbre), reducción de operatividad al tratarse volúmenes de datos grandes (aumento de productividad) y su buen respaldo bibliográfico (ampliamente usados en resolución de problemas de clasificación de usuarios, clientes, empresas, etc).

1. **Modelos y métodos ya existentes**

En esta sección se analizan los distintos modelos de aprendizaje supervisado que se emplean en los recursos bibliográficos revisados para la solución de problemas de clasificación y regresión como el de este estudio. Entre las técnicas más utilizadas se encuentran la regresión logística, los árboles de decisión, Random Forest y AdaBoost. Cada uno de estos algoritmos ofrece ventajas y desventajas específicas que los hacen más o menos apropiados según la naturaleza del conjunto de datos y los objetivos del análisis. A lo largo de este marco, exploraremos cada uno de estos métodos, sus características, y cómo pueden contribuir a obtener mejores resultados en la modelización de los datos de este estudio.

**La regresión logística**

La regresión logística es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación, especialmente en situaciones donde la variable dependiente es binaria. Aunque su nombre sugiere "regresión", se aplica comúnmente a la clasificación (Agudelo, 2023). Sin embargo, su eficacia disminuye en escenarios donde la relación entre las variables es altamente no lineal. Además, los valores atípicos en los datos de entrenamiento pueden tener un impacto significativo en el modelo, lo que requiere métodos complejos de preprocesamiento para mitigarlos (Wang, 2024). Debido a su eficiencia en problemas de clasificación binaria, la regresión logística es una opción muy útil para implementar en el modelo de este estudio, ya que la variable de respuesta es una decisión, esto permite interpretar la salida como la probabilidad de pertenecer a una clase. Según el análisis de Zhiyue Wang, este algoritmo ofrece una alta precisión en los resultados. Adicionalmente, según el estudio realizado por (Ozan, 2018) donde se aplican tres métodos de aprendizaje automático (NEM, LiRM y LoRM)[[3]](#footnote-4) para resolver un problema de segmentación de clientes de una empresa utilizando la información de pago de los mismos, con los parámetros óptimos encontrados al usar LoRM, se crea un límite de decisión que imita la estrategia de segmentación de clientes de la empresa con un alto porcentaje de eficiencia del 89,43 %.

**Árbol de decisión**

Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, utilizado tanto para tareas de clasificación como de regresión. Su estructura es jerárquica y consta de un nodo raíz, ramas, nodos internos y nodos hoja (IBM, 2024). El objetivo de un árbol de decisión es construir un modelo que prediga la clase objetivo de una instancia de prueba desconocida considerando varias características de entrada (Wang, 2024)

La principal ventaja de los árboles de decisión es su capacidad para modelar relaciones complejas de manera intuitiva. Sin embargo, es importante que los árboles de decisión no agreguen complejidad innecesaria, ya que la explicación más simple suele ser la mejor. Para reducir la complejidad y evitar el sobreajuste, se emplea comúnmente un proceso llamado poda. Este proceso elimina ramas que se dividen en características con poca importancia, lo que ayuda a simplificar el modelo (IBM, 2024).

Los árboles de decisión ofrecen varias ventajas, entre las que destaca su capacidad para modelar relaciones no lineales complejas en los datos sin requerir suposiciones lineales. (Wang, 2024) Esta versatilidad los hace adecuados para la data de este estudio y valioso para clasificar instancias en una amplia gama de conjuntos de datos con muchas variables.

**Random Forest**

Es un algoritmo de aprendizaje supervisado que utiliza múltiples árboles de decisión para realizar predicciones. Se trata de una técnica de conjunto (ensemble) que combina las predicciones de varios modelos base con el fin de mejorar tanto la precisión como la robustez del modelo. Este enfoque implica construir una gran cantidad de árboles de decisión durante la fase de entrenamiento y combinar sus predicciones en la fase de prueba, lo que resulta en una precisión general superior. Entre los métodos de clasificación, Random Forest destaca por su alta precisión y menor sensibilidad a los valores atípicos. El promedio de las predicciones de los árboles tiende a reducir el efecto del ruido, lo que hace que el modelo sea más resistente en presencia de conjuntos de datos imperfectos o ruidosos (Wang, 2024). Además, la idea de emplear árboles de decisión para resolver problemas de clasificación como el de este estudio, se ve reforzada por (Agudelo, 2023) quien, en su estudio de clasificación de clientes para predecir la pérdida de usuarios por parte de un banco, logra una buena precisión y recall[[4]](#footnote-5) para en la clasificación de clientes desertores usando Random Forest. Esto indica que, en general, el modelo tiene la capacidad de predecir tanto clientes que abandonan como aquellos que se quedan, minimizando los falsos positivos y falsos negativos.

**Algoritmo Boosting Adaptativo**

Fue el primer algoritmo de boosting adaptativo en el campo de los métodos de aprendizaje automático. A través de un entrenamiento iterativo de clasificadores débiles o de base, asigna mayor importancia a los datos que han sido mal clasificados en iteraciones anteriores, lo que permite obtener un nuevo clasificador más preciso. Este enfoque adaptativo contribuye a mejorar la precisión del algoritmo (Barbona & Celina , 2018)

Sin embargo, AdaBoost puede ser sensible a datos ruidosos o atípicos, lo que podría provocar un sobreajuste si este ruido se interpreta erróneamente como un patrón significativo. Además, para alcanzar un rendimiento óptimo, AdaBoost requiere una cantidad de datos suficientemente grande (Wang, 2024)

La fortaleza de AdaBoost radica en combinar múltiples aprendices débiles en un aprendiz poderoso, utilizando su diversidad para trabajar con diferentes subconjuntos de los datos, mejorando así la precisión general. Esta capacidad para adaptarse y centrarse en las instancias difíciles de clasificar, junto con su enfoque iterativo, hace que AdaBoost sea una herramienta valiosa en el arsenal del aprendizaje automático. Sin embargo, es crucial tener en cuenta su sensibilidad a los datos ruidosos y la necesidad de un conjunto de datos adecuado para maximizar su eficacia.

(Barbona & Celina , 2018) en su estudio de clasificación evalúan el desempeño de este algoritmo aplicado a los métodos de clasificación Regresión Logística y SMO (Sequential minimal optimization). Los resultados finales de los distintos clasificadores considerados se evalúan mediante porcentajes de mala clasificación. Se observa que al aplicar AdaBoost teniendo en cuenta como algoritmo de base el método de Regresión Logística no se presenta una reducción en el porcentaje de mala clasificación. En cambio, para el caso del método SMO como algoritmo de base, el porcentaje de mala clasificación bajó un 8,67%.

Finalmente, luego de analizar los distintos recursos bibliográficos y los métodos existentes para la resolución del problema de clasificación de clientes base foco de Bavaria, se procede a describir su originalidad con respecto a los trabajos revisados en la bibliografía.

1. **Diferenciación del Proyecto**

El actual proyecto aborda la tarea de predicción de abandono desde una perspectiva diferente, entendiendo que el problema en cuestión permite ser resuelto usando el mismo principio, pero de cara al B2B[[5]](#footnote-6) y no consumidores finales. Luego de una revisión exhaustiva no se encuentra algo similar en la literatura y puede sentar un precedente en la forma en que se analiza la propensión hacia la competencia en un problema de negocio como el de Bavaria. Esto debido a que no se tienen en cuenta variables demográficas de los negocios sino históricos de venta, indicadores como volumen y cobertura, ROS e información recolectada mediante encuestas de la fuerza comercial. Adicionalmente, se pretenden analizar varios algoritmos de clasificación con el fin de encontrar el idóneo, algo que se replica en numerosos trabajos, pero puede ser muy útil y disruptivo si se tiene en cuenta que en la compañía ninguna Regional emplea métodos de aprendizaje supervisado para la toma de decisiones en sus bases de competencia.

# Referencias

Agudelo, H. (2023). *ANÁLISIS DE RETENCIÓN DE CLIENTES EN INSTITUCIONES BANCARIAS.* Medellín: Universidad de Antioquia.

Barbona, I., & Celina , B. (2018). *Aplicación del algoritmo Boosting Adaptativo (ADABOOST) a un problema de clasificación automática de textos.* Rosario: Universidad Nacional de Rosario.

IBM. (8 de 10 de 2024). *IBM*. Obtenido de ¿Qué es un árbol de decisión?: https://www.ibm.com/es-es/topics/decision-trees

Ozan, S. (2018). A Case Study on Customer Segmentation by using Machine Learning Methods. *International Conference on Artificial Intelligence and Data Processing (IDAP). doi:10.1109/idap.2018.8620892*.

Wang, Z. (2024). Customer Segmentation Based on Machine Learning Methods. *Highlights in Science, Engineering and Technology*, 131.

Zadoo, A., Jagtap, T., Khule, N., Kedari, A., & Khedkar, S. (2022). A review on Churn Prediction and Customer Segmentation using Machine Learning. *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*.

1. Facturación mensual de cajas de cierto SKU por cliente [↑](#footnote-ref-2)
2. Gestión de relaciones con los clientes, que es un sistema para gestionar todas las interacciones de la empresa con los clientes actuales y potenciales. [↑](#footnote-ref-3)
3. Método de ecuaciones normales (NEM), método de regresión lineal multivariada (LiRM) y método de regresión logística (LoRM) [↑](#footnote-ref-4)
4. Métricas para efectividad del modelo de clasificación: Recall nos da información sobre el rendimiento de un clasificador con respecto a falsos negativos (cuántos fallaron), mientras que la precisión nos proporciona información sobre su rendimiento con respecto a los falsos positivos (cuántos capturados) [↑](#footnote-ref-5)
5. Abreviación de business to business o “de empresa para empresa” en español. Por lo tanto, una venta B2B es aquella se realiza de una empresa a otra [↑](#footnote-ref-6)