



POSGRADOS

Maestría en

SOFTWARE CON MENCIÓN EN DISEÑO DE ARQUITECTURA DE SISTEMAS

RPC-SO-34-NO.778-2021

OPCIÓN DE TITULACIÓN:

PROYECTO DE TITULACIÓN CON COMPONENTES DE INVESTIGACIÓN APLICADA Y/O DE DESARROLLO

TEMA:

COMPARACIÓN DE TÉCNICAS DE SEGMENTACIÓN DE CLIENTES MEDIANTE MÉTRICAS DE VALIDACIÓN INTERNA EN EL CAMPO FINANCIERO CASO DE ESTUDIO EN EMPRESA

AUTOR:

DIEGO JAVIER RUIZ VINTIMILLA

DIRECTOR:

JORGE OSWALDO LOJA CAJAS

CUENCA – ECUADOR

2024

Autor:**Diego Javier Ruiz Vintimilla**

Ingeniero en Sistemas.

Candidato a Magíster en Software con Mención en
Diseño de Arquitectura de Sistemas por la Universidad
Politécnica Salesiana – Sede Cuenca.

diegojavinti@gmail.com

Dirigido por:**Jorge Oswaldo Loja Cajas**

Ingeniero en Sistemas.

Máster Universitario en Análisis y Visualización de
Datos Masivos.

jorgelojam@gmail.com

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra para fines comerciales, sin contar con autorización de los titulares de propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual. Se permite la libre difusión de este texto con fines académicos investigativos por cualquier medio, con la debida notificación a los autores.

DERECHOS RESERVADOS

2024 © Universidad Politécnica Salesiana.

CUENCA – ECUADOR – SUDAMÉRICA

DIEGO JAVIER RUIZ VINTIMILLA

Comparación de técnicas de segmentación de clientes mediante métricas de validación interna en el campo financiero caso de estudio en empresa

Índice general

Índice de Figuras	V
Índice de Tablas	VI
Resumen	VII
Abstract	VII
1. Introducción	1
1.1. Descripción general del problema	2
1.2. Objetivos	2
1.2.1. Objetivo general	2
1.2.2. Objetivos específicos	2
1.3. Contribuciones	3
1.4. Organización del manuscrito	4
2. Antecedentes	5
2.1. Estado del Arte	6
2.1.1. Algoritmos de aprendizaje supervisado para clasificación y regresión	6
2.1.2. Algoritmos de aprendizaje no supervisado para clustering y asociación	7
2.2. Evaluación de la calidad de los algoritmos de clustering	9
2.3. Definiciones Previas	13
2.3.1. K-means	13
2.3.2. Agglomerative Nesting (AGNES)	13
2.3.3. Red Neuronal (RNN)	13
2.3.4. Sum of Squared Within(SSW)	13
2.3.5. Sum of Squared Between (SSB)	14
2.3.6. Elbow method	14

2.3.7. Within-clúster Sum of Squares (WCSS)	14
2.3.8. Dendrogramas	15
2.3.9. Ward	15
2.3.10. Función de Activación	15
2.3.11. Metodología eXtreme Programming (XP)	15
2.4. Librerías de Python	18
2.5. Formulación del Problema	20
2.6. Resumen del capítulo	21
3. Metodología	22
3.1. Diseño del experimento	23
3.1.1. Software	23
3.2. Desarrollo	24
3.2.1. Datos	30
3.2.2. Preprocesamiento de Datos	32
3.2.3. Procesamiento de Datos	34
3.2.4. Métrica de Evaluación	40
3.2.5. Interfaz Gráfica	41
4. Resultado y Discusión	50
4.1. Resultados y Discusión	50
4.1.1. Resultados	50
4.1.2. Discusión	53
5. Conclusiones	54
5.1. Conclusiones	54
5.2. Recomendaciones	56

Índice de Figuras

3.1. Historias de Usuario	25
3.2. Diagrama de Casos de Uso.	26
3.3. Historias de Usuario y Actividades en JIRA	27
3.4. Actividades Sprint 1	27
3.5. Diagrama Entidad Relación.	28
3.6. Diagrama de Clases.	29
3.7. Historias de Usuario Sprint 2	29
3.8. Diagrama de Paquetes.	30
3.9. Variables.	32
3.10. Campos Significativos.	33
3.11. Escalado de Características.	34
3.12. Método del Codo.	36
3.13. K-means.	37
3.14. Dendrograma.	38
3.15. AGNES.	39
3.16. SOM.	41
3.17. Menú Usuario Administrador - Menú Usuario	42
3.18. Vista de Inicio	43
3.19. Vista Cambio de Clave	44
3.20. Vista Gestión de Usuarios	44
3.21. Vista Crear configuración	45
3.22. Vista Procesamiento - Tab - Datos	46
3.23. Vista Procesamiento - Tab - K-MEANS	47
3.24. Vista Procesamiento - Tab - AGNES	48
3.25. Vista Procesamiento - Tab - SOM	49
3.26. Tabla Resultante con Columna Clúster Agregada	49
4.1. Resultado de Evaluación Silueta.	51

Índice de Tablas

3.1. Software.	23
3.2. Librerías.	24
4.1. Resultados de Técnicas de Segmentación.	52
4.2. Resultados de Técnicas de Segmentación (Consulta SQL). . .	52
4.3. Promedios de Técnicas de Segmentación.	52

Resumen

El presente trabajo consiste en la elaboración de un diseño, análisis e implementación de un sistema Web basado en el lenguaje de programación Python con el framework Streamlit bajo la metodología ágil Extreme Programming(XP) refinando la arquitectura todo el tiempo en base a pruebas e iteraciones cortas. Este software permite la evaluación de tres técnicas de segmentación de socios en un conjunto de datos (edad, ingresos, ahorros, etc.) sin referencia a resultados conocidos mediante algoritmos de aprendizaje no supervisados utilizando los siguientes enfoques de agrupación, Agrupación particionaria (K-Means), Agrupación jerárquica (Agglomerative Nesting) y Redes Neuronales (Self Organizing Map). Además, se procede a realizar una evaluación del aglomerado resultante de cada algoritmo por medio de una métrica de evaluación interna llamada Silhouette la cual ayuda a definir cuál o cuáles son las técnicas mejor evaluadas. Finalmente obteniendo como efecto que las Técnicas de K-Means y Agglomerative Nesting en un conjunto de pruebas obtuvieron los valores que mayormente se acercaron a 1 demostrando que los objetos están bien emparejados con sus clústers y mal emparejado con sus clústers vecinos.

Abstract

The present work consists of the development, analysis, and implementation of a Web system based on the Python programming language with the Streamlit framework under the Agile Extreme Programming (XP) methodology, refining the architecture continuously based on tests and short iterations. This software allows the evaluation of three partner segmentation techniques in a dataset (age, income, savings, etc.) without reference to known results using unsupervised learning algorithms. The following clustering approaches are utilized: Partitioning Clustering (K-Means), Hierarchical Clustering (Agglomerative Nesting), and Neural Networks (Self-Organizing Map). Additionally, an evaluation of the resulting clusters of each algorithm is performed using an internal evaluation metric called Silhouette, which helps determine which techniques are better evaluated. Finally, it is observed that the K-Means and Agglomerative Nesting techniques in a set of tests obtained values that mostly approached 1, demonstrating that the objects are well-matched with their clusters and poorly matched with their neighboring clusters.

Capítulo 1

Introducción

Con la cantidad de información almacenada de los clientes/socios acerca de sus características generales, demográficas, sus comportamientos en ahorros y pagos de créditos en las instituciones financieras. Es necesario que las organizaciones transformen esta información en conocimiento con el fin de orientar y enfocar sus recursos hacia grupos focalizados de socios que poseen comportamientos similares.

[Gomathy et al. \[2022\]](#) señala que una buena estrategia de desarrollo empresarial rentable comienza con una buena segmentación de los clientes. En el sector bancario, el proceso de segmentación de clientes se ha convertido en una herramienta útil para conseguir más clientes, pero también para extraer un mayor valor de los existentes.

Puesto que, si bien existen herramientas comerciales como lenguajes de programación que tienen librerías propias o de terceros para desarrollar aplicaciones que faciliten con la implementación de algoritmos para la segmentación de objetos, sería un requisito que un usuario tengan conocimiento en el desarrollo de software o maneje una herramienta específica para el procesamiento de la información. Por lo que se plantea la necesidad de desarrollar una aplicación que sea capaz de interactuar de manera directa con los datos, algoritmos y los resultados que presenten información eficiente, capaz de concretar en la toma de una decisión evitando tener un resultado incierto con algún segmento o producto a analizar.

La solución tendría como finalidad interactuar con el usuario, visualizando información desde el proceso de carga de datos como los resultados obtenidos mediante gráficas interactivas donde se pueda apreciar la segmentación y su relación con cada una de las características que conforman el almacén de datos.

1.1. Descripción general del problema

Puesto que los mercados son cada vez más cambiantes y la competencia por parte de las instituciones en el sector financiero se vuelve más estratégico, en donde muchas de estas han optado por asignar una gran parte de sus recursos económicos al área tecnológica, con el fin de tener resultados mucho más rápidos y efectivos en sus procesos operativos buscando una eficiencia absoluta. La falta de recursos y el desconocimiento de los avances tecnológicos enfocados a ramas de Machine Learning en las Cooperativas de segmentos pequeños (4 y 5) en el Ecuador y países latinoamericanos han limitado el crecimiento de las mismas adquiriendo brechas tecnológicas difíciles de suplir, estancándose en procesos manuales que con la ayuda de herramientas tecnológicas pudiesen ser más eficientes y eficaces con su información.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un aplicativo web mediante herramientas informáticas orientadas al desarrollo WEB para comparar tres técnicas de segmentación de clientes/socios mediante una métrica de validación interna.

1.2.2. Objetivos específicos

- Analizar el estado del arte de 3 técnicas de segmentación de clientes y el tipo de validación (métrica) mediante la revisión sistemática de documentos científicos en aras de sentar las bases conceptuales y metodológicas para el desarrollo de la aplicación.
- Definir los requerimientos de diseño mediante los modelos de entidad relación, clases, casos de uso para la implementación de los diferentes algoritmos de segmentación y validación de los mismos.
- Implementar una aplicación WEB mediante el uso de herramientas de desarrollo WEB en función de los requerimientos de diseño con la finalidad de generar un sistema que permita validar, visualizar los resultados de tres técnicas de segmentación de clientes.

- Determinar que técnica o técnicas de segmentación utilizadas tuvieron un agrupamiento más óptimo en base a una métrica que permita la evaluación de la cohesión de los clústeres.

1.3. Contribuciones

Con este trabajo se realiza una comparativa de la eficiencia de los algoritmos de machine learning (K-means, AGNES, redes neuronales) para el desarrollo, diferenciando los resultados de cada una de ellas en base a su metodología aplicada, así como el resultado evaluado en base a una métrica que indique que técnica se ajusta a un agrupamiento más eficiente. La aplicación Web permitirá obtener los resultados visuales de las técnicas escogidas facilitando el entendimiento, en base a los parámetros seleccionados, con el fin de crear el desarrollo de una estrategia de marketing ya sea de fidelización, búsqueda de un nuevo nicho de mercado o productos que se ajusten a las necesidades de los socios.

En cuanto a las ventajas el aplicativo proporcionará optimización en el tiempo de segmentación de los socios, así como también un ahorro representativo en los recursos asignados. Además, permitirá la visualización de los resultados de manera gráfica facilitando su interpretación.

1.4. Organización del manuscrito

El presente documento está conformado por 3 capítulos los cuales contienen la información del trabajo realizado que se detalla a continuación.

Capítulo 1 contiene una breve introducción sobre lo que va a tratar la tesis, seguido del objetivo general que detalla las actividades a cumplir con los objetivos específicos, finalizando con la contribución que aporta esta tesis para las personas interesadas.

Capítulo 2 En este capítulo consta el análisis de los documentos realizados por diversos investigadores sobre temas de segmentación de clientes y los algoritmos utilizados, así también la evaluación de los clústeres resultantes, siendo un aporte muy valioso a este proyecto. Además, contiene definiciones previas de algunos algoritmos de segmentación que serán utilizados en el presente trabajo, así como también definiciones de métricas utilizadas como parámetros de configuración en los algoritmos utilizados.

Finalizando con el Capítulo 3 el cual contiene una descripción de la metodología de desarrollo empleada, sus diagramas y el software base en el cual fue desarrollado, también se detalla las características de cada elemento que conforman el conjunto de datos a ser empleado. En este capítulo se desarrolla la implementación de las técnicas de segmentación, enfocadas en diferentes tipos de aglomeración y sus respectivas evaluaciones de los resultados obtenidos mediante una métrica de validación interna aplicada en dos conjuntos de datos, verificando mediante pruebas su variabilidad con los mismos conjuntos de datos, pero en repetidas ocasiones. Culminando con la descripción de los resultados obtenidos y especificando como estos han permitido cumplir con los objetivos planteados.

Capítulo 2

Antecedentes

En este capítulo se realiza un análisis de las investigaciones realizadas por los autores sobre la segmentación de los clientes, así como los resultados obtenidos de sus trabajos los mismos que formarán parte del estudio de este documento. Además, contendrá las definiciones de previas de la problemática a realizar.

2.1. Estado del Arte

La segmentación del mercado enfocado en los clientes es un tema de gran interés para una institución, debido a que la segmentación nos permite tener una precisión a la hora de fijar objetivos y recursos cuando queremos promocionar los productos o incluso crear uno nuevo, con el fin de que este se acople a la mayoría de las necesidades que este conjunto de clientes lo requiera.

2.1.1. Algoritmos de aprendizaje supervisado para clasificación y regresión

El autor [Dickson and Ginter, 1987] con el análisis de varios libros concluye opciones estratégicas en donde la primera está determinada por condiciones del mercado siendo un mercado total como un conjunto de submercados o segmentos cada uno con su demanda, y que estos varían de acuerdo a la percepción de cada empresa, la segunda habla de una estrategia de diferenciación del producto que no requiere la existencia del mercado y por último dice que una estrategia es factible cuando la diferenciación del producto ya existe o es complementaria.

[Yang and Yuan, 2007] construye un modelo de clasificación de clientes bancarios basado en la red neuronal de Elman entrenado con el algoritmo PSO en base a los datos de préstamos, luego realiza la comparación con la red neuronal de Elman de forma estándar y puede determinar que la red neuronal PSO-Elman obtiene un resultado más alto en cuanto a robustez en comparación con la estándar en lo que respecta en mantenerse alejado a los riesgos de incumplimiento del cliente. Con las técnicas de dos o más etapas se puede apreciar que los resultados se apegan más a las necesidades planteadas.

[Smith, 2018] introduce el concepto de segmentación en el Marketing indicando que las estrategias serán un enfoque integrado para minimizar los costos totales teniendo prioridad sobre enfoques separados para minimizar los costos de producción y los costos de mercadeo por el otro, y que estas se aplican en respuestas a las condiciones variantes del mercado, también se debe indicar que "la segmentación de mercado consiste en ver un mercado heterogéneo como una serie de mercados homogéneos más pequeños en respuesta a las diferentes preferencias de productos entre segmentos atribuyendo a los deseos" de los consumidores o usuarios de una

satisfacción más precisa de sus diferentes deseos. Por último, se destaca que la explotación de segmentos de mercado proporciona la satisfacción de los consumidores o usuarios, tendiendo a construir una posición de mercado segura y con mayor estabilidad general.

En cuanto a las técnicas de segmentación podemos decir que es el procedimiento que nos permite dividir y agrupar un cierto número de objetos de análisis en base a ciertas características propias del objeto o comportamientos del mismo reflejados en un grupo que nos permita diferenciarlos. Según [Panuš et al., 2016] indica que la segmentación debe contemplar factores de utilidad como la demografía, geografía, comportamiento, valores, etc.

La selección de técnicas de segmentación se ha vuelto más importante debido al hecho de que los desarrollos en la información y tecnologías de la comunicación han tenido un gran avance, especialmente la gestión de bases de datos. Además, la gran disponibilidad de datos y el rendimiento ineficiente de técnicas estadísticas tradicionales (o herramientas de segmentación orientadas a la estadística) en datos tan voluminosos han estimulado a los investigadores a encontrar herramientas de segmentación efectivas para descubrir información útil sobre sus mercados y clientes. [Hiziroglu, 2013]

El autor [Gomathy et al., 2022] indica que una buena estrategia de desarrollo empresarial rentable comienza con una buena segmentación de los clientes. En el sector bancario, el proceso de segmentación de clientes se ha convertido en una herramienta útil para conseguir más clientes, pero también para extraer un mayor valor de los existentes. A continuación, el autor indica que ha utilizado dos métodos: el de clasificación binaria y redes neuronales artificiales (Retropropagación), cuyo objetivo es determinar los clientes afluentes en depósitos y tarjetas de crédito, para lo cual se utilizaron un conjunto de variables (treinta continuas y una categórica). Teniendo buenos resultados en el proceso de segmentación.

2.1.2. Algoritmos de aprendizaje no supervisado para clustering y asociación

El autor [Sundjaja, 2013] indica que en el mercado de la industria bancaria se vuelve cada vez más difícil competir con los mismos productos

razón por lo cual se deben desarrollar estrategias de marketing, para tener una visión del comportamiento del cliente para satisfacer mejor sus necesidades. Con respecto a la segmentación de clientes utilizó la segmentación demográfica y la técnica de minería de datos en la base de datos de la empresa, en donde su enfoque principal es la división de los clientes en 6 segmentos: corporativo, negocios, empleados, amas de casa, estudiantes y otros, en donde concluye que la minería de datos brindó los resultados esperados en la similitud en la que realizan sus transacciones los diferentes segmentos, por ejemplo depósitos en efectivo, transferencias internas, compras, efectivo, otros. En consecuencia, la minería de datos le ayudó a determinar cuál de los canales del banco son utilizados por cada uno de los segmentos, así como la similitud que manejan entre ellos.

El autor [Tsai et al., 2015] indica que en los negocios la fuente de ingresos es la venta de productos por lo que la segmentación de clientes se basa en análisis de los registros históricos para determinar los valores de vida de los clientes, en la industria automotriz el mantenimiento representa el 60 % de las ventas. La retención de los clientes disminuye la necesidad de adquirir nuevos clientes potenciales, es por ese motivo que se debe identificar los clientes valiosos para darles un trato especial. Además, al utilizar un único método de agrupación para la segmentación de clientes genera resultados poco fiables llevando a malas decisiones, por lo que recomienda dos técnicas k-means y maximización de expectativas. Obteniendo los resultados de 4 grupos de clientes diferentes fieles, potenciales, VIP y churn.

[Hosseini and Shabani, 2015] indica que dado el comportamiento y las necesidades cambiantes de los clientes las empresas deben también adoptar decisiones de análisis de acuerdo al factor tiempo. El artículo utilizará el modelo de segmentación RFM como modelo tradicional y el método de agrupación de clústeres K-means en función de cambios en varios periodos de tiempo con el fin de descubrir la tendencia de cambio de valor de cada cliente en los intervalos de tiempo propuestos, para lo cual procedió a dividir los datos extraídos en intervalos de tiempo concluyendo el modelo básico de RFM no considera la transición temporal y dada la falta de detalles sobre los cambios en el comportamiento del cliente entrega resultados poco efectivos.

[Rahim et al., 2021] propone un trabajo basado en Recency Frequency Monetary (RFM). En donde incluyen la actualidad de la última compra (R), la frecuencia de compras (F), y el valor monetario de las compras (M),

indica que este modelo detecta patrones de comportamiento de un cliente y en la investigación enfocará en la recompra de ciertos artículos de un cliente, recalcando un costo mínimo en la recopilación de datos, los mismos que determinarán los patrones de comportamiento para clasificar a los clientes individualmente. Dentro de los experimentos que propone tomamos como interés la segmentación de los clientes por la técnica de agrupamiento K-means en los datos extraídos para segmentación de dos grupos, el primero la frecuencia versus la visualización monetaria y la segunda recencia versus la visualización monetaria, en donde se puede observar que en base de la frecuencia se encuentran valores muy dispersos sin importar el valor y cuando se observa por la recencia se puede apreciar que los valores se agrupan teniendo una clara respuesta visual.

[DR et al., 2022] indica la importancia de incorporar la capacidad de cambiar los proyectos de los mercados enfocado a los fragmentos de clientes identificando los artículos relacionados con cada porción de cliente y medir su interés, para lo cual es fundamental elegir un cálculo apropiado para el conjunto de datos accesible. Este se centrará en características comunes como el género, edad, intereses, y hábitos de gastos utilizando el agrupamiento K-means que es apropiado para problemas de división, la orientación se enfocará en la diferencia de edad, las ganancias anuales fijando relaciones en puntajes de gasto del cliente. Utilizará datos de prueba y aprobación teniendo pronósticos genuinos para los datos de aprobación. Concluyó con éxito el método de segmentación ya que la investigación se enfocó en el agrupamiento.

2.2. Evaluación de la calidad de los algoritmos de clustering

[Zakrzewska and Murlewski, 2005] indica que la capacidad para adquirir nuevos clientes y mantener los existentes es crucial para el sector financiero, la segmentación de clientes mediante la obtención de información sobre patrones desconocidos posee una gran importancia. Además, señala que las técnicas dependen significativamente de las características de los conjuntos de datos, y que los algoritmos deben ser eficientes para grandes volúmenes de datos multidimensionales con ruido.

Propone la utilización de 3 algoritmos DBSCAN (Density Based Spatial clustering of Applications with Noise), K-means y agrupación en dos fases.

Durante las pruebas, examinaron los algoritmos en función del número de dimensiones (atributos), la eficacia en la detección de valores atípicos, la escalabilidad y el comportamiento en caso de datos estandarizados y no estandarizados, también se debe indicar que se eligieron los atributos de datos más común, para el análisis de clientes bancarios como son: edad, ingresos, depósito, crédito, utilidad/pérdida.

Concluyó, primero K-means es muy eficiente para grandes conjuntos de datos multidimensionales, sin embargo, depende en gran medida de la elección del parámetro de entrada k , segundo el algoritmo de clústering bifásico tiene un muy buen rendimiento para datos con ruido y poca cantidad de dimensiones, y por último el algoritmo DBSCAN, la elección incorrecta de los parámetros de entrada, puede dar lugar a una mala calidad de los resultados obtenidos.

Al igual que el autor anterior [2022] considera el análisis histórico de ventas basado en el modelo RFM para después aplicar el enfoque científico de segmentación utilizando el algoritmo K-means indicando los tipos más destacados de segmentación en base a conocimientos cualitativos y cuantitativos. El autor también utiliza RFM recalcando que este modelo no solo brinda el patrón de compra sino también información descriptiva sobre la compra reciente y la ganancia obtenida. Dentro de los pasos de su metodología se debe tomar en cuenta que lo principal está en el procesamiento de los datos al igual que los autores anteriores utilizan este proceso con el fin de tener datos más limpios para ser analizados, también da a conocer que utiliza la validación de la silueta para los conglomerados obtenidos que analiza, que también fueron los conglomerados resultantes. Como conclusión implementa el modelo RFM para conjuntos de datos sintéticos y reales. Además, los conglomerados se evalúan utilizando el algoritmo de agrupamiento Silhouette Analysis for K-Means. Con base en el puntaje de silueta, se pueden analizar las ventas recientes, la frecuencia de ventas y las ventas monetarias y se encuentra una solución óptima.

Existen autores que destacan diferentes criterios en la segmentación interpretándolos de manera diferente y no hay una guía clara de cómo medirlos. Sólo se puede afirmar que la mensurabilidad, la accesibilidad, diferenciabilidad, sustancialidad y accionabilidad son los cinco criterios comunes para una segmentación efectiva. Desde el punto de vista de la agrupación, la homogeneidad puede ser añadida a esta lista. Cabe señalar que uno de los más valiosos es la información del comportamiento de los clientes, características, especialmente compras pasadas de clientes y valor

de los atributos introducidos. [Hiziroglu, 2013]

Las técnicas pueden ser clasificadas en 3 categorías: simple, de dos etapas, y multietapa, esta clasificación está basada en las veces en que el proceso de segmentación trabaja con las variables en el modelo. Dentro de la segmentación de clientes la mayoría de los enfoques son tecnológicos orientados a los métodos que van desde simple estadística inferencial o a la inteligencia artificial. Los tipos de enfoques también pueden ser a-priori y post-hoc.

Los enfoques tradicionales para la segmentación del mercado se basan en características no económicas del cliente mientras otros se basan en la rentabilidad. Según [Øyvind Helgesen, 2006] indica que los enfoques de rentabilidad deben ser vistos como técnicas adicionales enfocados al concepto de marketing que identifica, anticipa y satisface los requisitos de clientes de manera eficiente y rentable.

El autor [Øyvind Helgesen, 2006] indica que es una tarea muy desafiante conocer que técnica funciona mejor que las demás. Teniendo sus ventajas y desventajas y estos dependerán si se utilizan para agrupación o clasificación. La selección de la técnica incorrecta puede causar un impacto negativo, es por eso que se debe tener un enfoque claro si los datos son para a-priori o post-hoc, al igual que las características de los datos y estos pueden basarse en el volumen o en su estructura.

El estudio fue aplicado con 48 artículos dentro de los años de 1986 y 2012 que se implementaron en varias industrias que cubren varios sectores. Siendo el turismo un 19 % del total del estudio, el comercio minorista y el comercio electrónico siguen como las segundas áreas más implementadas. Los estudios muestran que alrededor del 65 % de los estudios utilizaron la neurocomputación como tecnología de computación blanda, además, tres estudios utilizaron tanto la computación neuronal como la evolutiva de manera colaborativa en forma de computación híbrida. Las tecnologías de neurocomputación tenían aplicaciones en todas las industrias. Además, la tecnología de computación aproximada se aplica solo en el área de comercio electrónico, mientras que la aplicación de tecnologías de computación difusa y evolutiva se puede ver en la mitad de las industrias. El método de mapas autoorganizados es la técnica más utilizada en los estudios por el 45 % del total de publicaciones. El 90 % de los estudios fueron realizados dentro de la categoría de modelo simple y el resto dentro del modelo de dos etapas. La homogeneidad se calculó en alrededor del 17 % de estos estudios. La identificabilidad se destaca como los porcentajes más altos entre los criterios.

2.2. EVALUACIÓN DE LA CALIDAD DE LOS ALGORITMOS DE CLUSTERING 12

Como recomendación señaló ocho necesidades industriales futuras asociadas a negocios y finanzas para las cuales las tecnologías Soft Computing (SC) pueden ser muy útiles, entre una de ellas la banca. Dentro del ámbito de la segmentación de clientes se puede decir que, si bien las técnicas basadas en enfoques estadísticos para clasificar a los clientes para formar segmentos han tenido diversos grados de éxito, cabe mencionar que dichos enfoques no son capaces de ejecutar una gran cantidad de datos y no proporcionan una estructura de segmentación flexible como las tecnologías informáticas de software que son capaces de hacerlo.

En conclusión, los avances en los sistemas difusos tales como las investigaciones sobre computación con palabras, aplicaciones de inteligencia artificial distribuida cognitiva y reactiva incluyendo agentes inteligentes, las aplicaciones emergentes de la computación evolutiva incluyendo meta-heurísticas, modelos probabilísticos y computación aproximada, nos conducirán a la construcción de sistemas inteligentes más avanzados, que también pueden ser aplicables a problemas empresariales.

2.3. Definiciones Previas

2.3.1. K-means

"K-means es un algoritmo de clasificación no supervisada (clústerización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática."[unioviedo, 2022]

$$\min_s E(\mu_i) = \min_s \sum_{i=1}^k \sum_{X_j \in S_i} \|X_j - \mu_i\|^2$$

2.3.2. Agglomerative Nesting (AGNES)

AGNES Funciona de forma ascendente. "La agrupación aglomerativa comienza con N grupos, cada uno de los cuales contiene inicialmente una entidad, y luego los dos grupos más similares se fusionan en cada etapa hasta que hay un solo grupo que contiene todos los datos."[Subasi, 2020]

2.3.3. Red Neuronal (RNN)

Las redes neuronales son estructuras que emulan el comportamiento del cerebro, están compuestas por neuronas que forman capas, una capa de entrada que propaga sus valores hasta una o varias ocultas y una capa de salida siendo el resultado del modelo de aprendizaje. Las neuronas se conectan entre si por medio de conexiones (variables) que modifican sus valores de acuerdo al aprendizaje. [Rodrigo, 2021]

2.3.4. Sum of Squared Within(SSW)

"SSW Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \min_s \sum_{i=1}^k \sum_{X \in C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster C_i y m_i el centroide del clúster C_i . "[Guzmán, 2016]

2.3.5. Sum of Squared Between (SSB)

"SSB Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación)

$$SSB = \sum_{j=1}^k n_j dist^2(C_j - \bar{x})$$

Siendo k el número de clústeres, n_j el número de elementos en el clúster j , c_j el centroide del clúster j y \bar{x} es la media del dataset. "[Guzmán, 2016]

"Sum of Squares based Indexes Los índices o medidas basadas en las «sumas de cuadrados» presentadas anteriormente se caracterizan por medir o cuantificar la dispersión de los puntos a nivel inter-clúster e intra-clúster.

$$d * \log \sqrt{\frac{SSW}{dN^2}} + \log k$$

Siendo k el número de clústeres, N el número de datos y d la dimensión de los datos. "[Guzmán, 2016]

2.3.6. Elbow method

El método del codo consiste en elegir un número de grupos para calcular sus varianzas iniciando con sus valores más representativos que son los números de menor valor, hasta sus valores menos representativos, en donde la agregación de cada número de mayor valor ira mostrando un descenso. El numero donde la varianza muestre una caída dramática dará un ángulo en la gráfica, formando un codo, de allí su nombre. [Bholowalia and Kumar, 2014]

2.3.7. Within-clúster Sum of Squares (WCSS)

Consiste en la suma de las desviaciones al cuadrado desde el clúster de la agrupación de cada una de las observaciones.

2.3.8. Dendrogramas

Resume el proceso de aglomeración de los clústeres por medio de una representación gráfica donde los objetos similares se conectan en base a su similitud o disimilitud. [Villardón, 2007]

2.3.9. Ward

Es un método que utiliza la distancia entre agrupaciones con el fin de encontrar la que tenga menos varianza dentro de cada aglomeración, en base a la homogeneidad estadística. [Espinel, 2015]

2.3.10. Función de Activación

Es una función que indica el nivel de activación que alcanzo una neurona frente algún impulso por parte de un valor de entrada. Las funciones de activación más comúnmente utilizadas son: [Matich, 2001]

- Función lineal
- Función sigmoidea
- Función tangente hiperbólica

2.3.11. Metodología eXtreme Programming (XP)

[Letelier and Penadés] define como una metodología especialmente para proyectos con requisitos imprecisos y muy cambiantes que pueden tener un alto riesgo técnico, sus principales características esenciales son:

Historias de usuario es la técnica para especificar los requisitos de software se han estos funcionales o no funcionales con unas breves características que el sistema debe poseer en base a las descripciones del cliente.

Roles XP

- Programador produce el código y ejecuta las pruebas unitarias.
- Cliente escribe las historias de usuarios asignando la prioridad de las mismas, además, realiza las pruebas unitarias.
- Encargado de pruebas, ayuda al cliente a crear las historias de usuario, realiza las pruebas funcionales difundiendo al equipo los resultados.

- Encargado de seguimiento, evalúa los objetivos alcanzados en base a las restricciones de tiempo y recursos presentes proporcionando retroalimentación al equipo en el proceso XP con respecto al grado de aciertos entre las estimaciones realizadas.
- Entrenador, es responsable del proceso o de la tarea global.
- Consultor, es un miembro externo del equipo con conocimiento específico en cierto tema necesario para el proyecto a realizar.
- Gestor, es el coordinador del equipo velando por las condiciones adecuadas, también es el nexo con el cliente.

Proceso XP Consiste en la definición del proceso a implementar por parte del cliente en base a prioridades y restricciones de tiempo, y donde el programador estima el tiempo de la implementación, esto se realiza en todas las iteraciones del proyecto donde existe un aprendizaje mutuo entre cliente y programador. No se debe perder la calidad del producto al ejercer presión en el tiempo de entrega al programador pero a su vez se debe tener el mayor valor de negocio con cada iteración. Sus fases son:

- **Exploración**, en esta fase se plantean las historias de usuario, el equipo de desarrollo se familiariza con las herramientas, se crea un prototipo en base a la arquitectura del sistema.
- **Planificación de la Entrega**, en esta fase se da prioridad a las historias de usuario y los programadores estimaran el tiempo de desarrollo.
- **Iteraciones**, es el conjunto de repeticiones para resolver todas las historias de usuario antes de entregar el producto final.
- **Producción**, en esta fase requiere de pruebas adicionales y revisiones de rendimiento antes de ser colocados en el entorno del cliente.
- **Mantenimiento**, en esta fase se debe mantener estable la primera versión mientras se continúan con las iteraciones para avanzar el proyecto.
- **Muerte del Proyecto**, esta es la fase final donde ya no existen mas historias de usuario por parte del cliente, entonces se genera la documentación final.

Prácticas XP Para que el diseño evolutivo funcione con el fin de disminuir la curva exponencial del costo del cambio a lo largo del proyecto se utilizan las siguientes practicas:

- **El juego de planificación** consisten en la comunicación entre el cliente y los programadores donde se estiman los tiempos de entrega para cada historia de usuario y la prioridad que tiene cada.
- **Entregas pequeñas** consiste en entregar versiones pequeñas del sistema pero con valor para el negocio.
- **Metáfora**, es una historia compartida que indica como debería funcionar el sistema.
- **Diseño simple** consiste en una solución simple y que pueda ser implementada superando todas las pruebas con éxito.
- **Pruebas** la producción de código esta dirigida por las pruebas unitarias, las mismas que están establecidas y son ejecutadas constantemente con cada modificación del código.
- **Refactorización (Refactoring)** es una actividad constante de reestructuración del código con el fin de mejorar su legibilidad, simplificarlo, remover duplicidad para facilitar cambios posteriores.
- **Programación en parejas** se consiguen mejores resultados, se transfieren los conocimientos por consiguiente la tasa de errores disminuye.
- **Propiedad colectiva del código** cualquier programador puede cambiar cualquier parte del código evitando que cualquier programador sea imprescindible.
- **Integración continua** cada pieza de código se integra cuando esta lista.
- **40 horas por semana** se trabajan máximo de 40 horas por semanas.
- **Cliente in-situ** el cliente debe estar siempre disponible para solventar cualquier inquietud al equipo.
- **Estándares de programación XP** enfatiza la comunicación de los programadores a través del código haciendo imprescindible cierto estándares de programación.

2.4. Librerías de Python

SQLAlchemy, es una librería que permite manejar objetos relacionales teniendo muchos más beneficios con el lenguaje SQL.

Pandas es una librería de Python especializada en el manejo y análisis de estructuras de datos.

Numpy es una librería que permite la creación de vectores, matrices, además que posee colecciones de funciones matemáticas.

Hashlib define una interfaz de programación para acceder a diferentes algoritmos criptográficos de hash.

st-pages permite manejar las clases como páginas teniendo acceso a una serie de atributos o características.

Streamlit-extras, es una biblioteca de Python que reúne fragmentos de código Streamlit útiles.

scipy esta librería se compone de herramientas y algoritmos matemáticos.

Matplotlib es una biblioteca para la generación de gráficos en dos dimensiones.

Scikit-learn es una biblioteca para el lenguaje de programación Python y es open source. Posee varios algoritmos para el análisis de información. Además, está diseñada para trabajar con las bibliotecas NumPy, SciPy, etc.

MiniSom es una librería que es una implementación minimalista de SOM y está basada en Numpy en la cual sus datos deben organizarse como una matriz Numpy en donde cada fila corresponde a una observación, o como una lista de listas.

Psycopg2 librería para la conexión con la base de datos PostgreSQL.

oracledb librería para la conexión con la base de datos Oracle

mysqlclient librería para la conexión con la base de datos MySQL

2.5. Formulación del Problema

Este trabajo pretende hacer la comparación de tres algoritmos aplicados en la segmentación de clientes en el sector financiero, específicamente el sector cooperativo, debido a que en las investigaciones realizadas hay poca información de estudios aplicados en este mercado. Se debe tener en cuenta que el concepto de socio abarca una responsabilidad más estrecha con la Institución ya que esta vela por los interés y bienestar de sus integrantes, pero sin dejar al lado la rentabilidad que debe percibir para mantenerse en el mercado financiero. Para lo cual se realizará un análisis mediante los estudios previos descritos en los diferentes sectores industriales y con un mayor énfasis en el sector financiero tomando en cuenta las referencias de variables que consideraron para la segmentación en los documentos o artículos.

Por otra parte, se tendrá en cuenta los resultados de los algoritmos que hayan tenido los mejores resultados en las investigaciones realizadas destacando su mayor efectividad así también con su métrica utilizada. Sin embargo en las investigaciones no se ha encontrado una comparativa con los algoritmos K-means, Agglomerative Nesting (AGNES) y redes neuronales puesto que se desea realizar la comparativa de estos en el desarrollo de una aplicación que permita la carga de los datos y el procesamiento de los mismos, cuyos resultados puedan ser evaluados mediante una métrica de validación interna que se ajuste o brinde mejor resultado de agrupación, se utilizará validación interna ya que no se posee ninguna información externa adicional para su comparación en la mayoría de los escenarios a desarrollar. Además, los resultados serán visualizados en la aplicación para tener una mejor percepción de los mismos.

2.6. Resumen del capítulo

En resumen, la aplicación de varias técnicas de segmentación utilizadas en los diferentes documentos indicados anteriormente demuestra la obtención de resultados beneficiosos para los diferentes tipos de industrias que ha permitido focalizar sus recursos humanos, económicos y por qué no decir tecnológico impulsando en el desarrollo o adquisición de herramientas que permitan realizar este tipo de análisis. Además, indican que la información histórica constituye la clave principal en el éxito de estas herramientas de agrupación de clientes, así también como la capacidad de seleccionar las características de la información cuyo valor sea el más relevante para que permita tener un conjunto de datos que sea significativo y no redundante.

En consecuencia el trabajo a realizar es una herramienta especializada que permita automatizar los procesos de extracción o carga de información, así como limpieza del conjunto de datos finalizando con la aplicación de 3 técnicas de segmentación presentando los resultados de forma visual de tal manera automatizado que únicamente con un par de clics puede tener todo el análisis. Todo esto se realizara con herramientas de software libre evitando costos y limitaciones en el desarrollo así como el uso de la misma, además con tecnología WEB para que pueda ser de fácil acceso por cualquier tipo de navegador adaptable a cualquier dispositivo. También se debe destacar que al se una herramienta especializada esta podrá ser utilizada por personas que no tengan conocimientos técnicos ni especializado en el procesamiento de datos ni uso de herramientas comerciales o gratuitas teniendo un ahorro significativo para la institución.

Capítulo 3

Metodología

En este capítulo se desea abordar el proceso del desarrollo de software en base a la metodología escogida (XP) con el fin de llevar a cabo el análisis de las técnicas de segmentación seleccionadas teniendo un resultado final por cada una de las iteraciones, donde se ha rediseñado y mejorado de manera constante el sistema, en consecuencia, se obtendrá un producto que permita interactuar con el usuario teniendo resultados fáciles de apreciar en tiempos eficaces.

Tabla 3.1: Software.

Software/Recurso	Versión
Windows	11 PRO
Anaconda 3	23.07-2
Python	3.11.5
DB Browser for SQLite	3.12.2
Visual Studio Code	1.87.2
Lucidchart	WEB
Jira	WEB
SQLite	3

3.1. Diseño del experimento

Para el desarrollo del proyecto se ha optado por utilizar la metodología ágil de desarrollo incremental XP para obtener en cada iteración el mayor de los resultados sobre el producto final, con cada requerimiento en una iteración esta debe incluir pruebas y documentación, con la finalidad de ir alineando los objetivos que sean necesarios para evitar correcciones arriesgadas al final del proyecto.

3.1.1. Software

En el presente trabajo se utilizaron los siguientes programas de software que se describen en la tabla 3.1, además se indica que el framework de trabajo en donde fue desarrollado todo el proyecto se llama Streamlit, que es una librería de Python cuyas bondades se enfocan en el desarrollo de aplicaciones web interactivas, visualización de datos, ciencia de datos, etc. acoplándose de forma eficiente a nuestras necesidades de desarrollo.

Streamlit es un framework que agiliza el desarrollo de las aplicaciones web con componentes visuales fáciles de utilizar y con características específicas como extensas para un desarrollo a medida, de esta forma se puede profundizar a un buen nivel el detalle de las acciones que se dé a un componente. Además, se debe tomar en cuenta que estas librerías trabajan en conjunto con otras librerías de Python que nos facilitan el manejo de los datos las definiciones y detalles de cada uno se encuentran en el capítulo 2

En la tabla 3.2 se indica las librerías y sus versiones que se utilizaron en el desarrollo de la presente aplicación, con el fin de evitar cualquier problema

Tabla 3.2: Librerías.

Librerías	Versión
SQLAlchemy	2.0.22
Pandas	2.1.1
Numpy	1.26.0
Hashlib	
Streamlit	1.28.2
st_pages	0.4.5
Streamlit-extras	0.3.5
Matplotlib	3.8.1
scipy	1.11.4
Scikit-learn	1.3.2
MiniSom	2.3.2
psycopg2	2.9.9
oracledb	2.2.0
mysqlclient	2.2.4

en una posible reproducción.

3.2. Desarrollo

Las historias de usuario forman una parte fundamental en el inicio de un proyecto ya que contiene las necesidades de los usuarios finales siendo estos el punto de partida para planificación de las iteraciones, a continuación se indican las historias de usuario en la figura [3.1](#)

El propósito principal de las iteraciones es desarrollar un proyecto de forma incremental, permitiendo al desarrollador realizar correcciones y mejoras del desarrollo anterior en base a lo aprendido. Para lo cual se elaboró un diseño de Casos de Uso simple del sistema, en donde los actores realizarán actividades básicas como el ingreso, actualización, eliminación de un usuario del sistema, la configuración de los parámetros de conexión a la base de datos cuando se tengan las autorizaciones y permisos que mediante una consulta se pueda extraer la información de Core - Financiero de la Institución y la carga de un archivo en formato csv cuando la información sea proporcionada por la institución en base a una estructura de datos requerida como se detalla en la subsección de Datos. Con cada iteración se fueron realizando cambios en el diseño y agregando nuevas funcionalidades



Figura 3.1: Historias de Usuario

al proyecto hasta obtener el producto final como muestra la figura 3.2

En consecuencia, obtenemos dos actores, el primero llamado administrador cuya finalidad es gestionar las operaciones sobre un usuario determinado, así también asignar los permisos de este en el sistema. Hay que tener en cuenta que se está trabajando con información sensible que puede ser mal utilizada, es por ello que es necesario crear confidencialidad para el acceso a la aplicación.

Por el contrario, está el usuario final que hará uso de la aplicación obteniendo los beneficios para lo cual fue creada, ya sea obteniendo los datos desde una base de datos configurada en la aplicación o mediante la carga de un archivo para su respectivo procesamiento.

Para empezar la codificación ingresamos las historias de usuarios y el

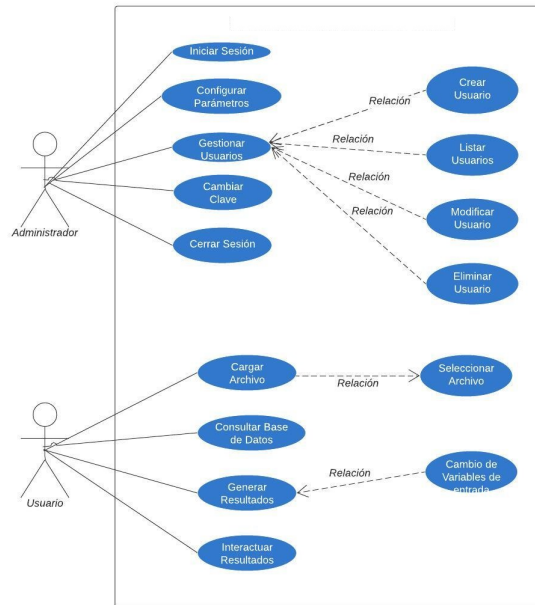


Figura 3.2: Diagrama de Casos de Uso.

conjunto de actividades que la conforman en la herramienta Jira cuya especialidad radica en la ayuda para el desarrollo de proyectos ágiles, de esta manera se obtuvo una mejor organización del trabajo que se realizó. En caso de tener un grupo de desarrolladores también se podrá extender las responsabilidades por cada actividad a cada desarrollador así también el tiempo estimado para cada uno, tal y como recomienda la metodología.

Ahora bien que se tienen las tareas organizadas y siguiendo las prácticas de XP se sabe que los sprints no pueden durar mas de 4 semanas por lo que dividiremos las tareas en 2 sprints en donde el primero contiene las actividades como podemos ver en la figura 3.4, la primera actividad fue la creación de la base de datos, que permitirá la gestión de los usuarios (crear, modificar y eliminar) y la gestión de las configuraciones, la misma que contiene los campos necesarios para una conexión a la base de datos, así también el campo que contiene la consulta (sql) que permitirá extraer la información, como podemos apreciar en la figura 3.5

En este sprint se pretendió tener un primer producto con el cual el usuario pueda interactuar y retro-alimentar el proceso, ofreciendo los requerimientos











Tipo	# Clave	Resumen	Estado	Sprint
▼ 	SCRUM-1	Inicio de sesión	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-5	Creación de base de datos para el proyecto	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-6	Script para la creación de la tabla Usuario	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-7	Incorporar librería hashlib para encriptar la clave de usuario	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-8	Ingreso de usuario en la base de datos	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-9	Función de validación para verificar existencia de usuario	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-10	Vista de inicio de sesión campos usuario y clave	TAREAS POR HAC...	Tablero Sprint 1
▼ 	SCRUM-11	Crear usuarios	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-12	Vista con todos los campos para el ingreso del usuario en la ...	TAREAS POR HAC...	Tablero Sprint 1
	SCRUM-13	Función de ingreso del registro en la base de datos	TAREAS POR HAC...	Tablero Sprint 1

Figura 3.3: Historias de Usuario y Actividades en JIRA



Figura 3.4: Actividades Sprint 1

esenciales del problema, pero que a su vez contenga una solución simple y fácil de implementar. También se debe indicar que no existe un rol de cliente o usuario final específico para el presente trabajo, ya que el usuario es el propio desarrollador, teniendo que plantear los requisitos, así como experimentar con las técnicas propuestas.

Asimismo, cada iteración es una etapa evolutiva que involucró el rediseño

The diagram shows two entity tables. The first table, 'Usuario', has five attributes: 'codigo' (entero), 'usuario' (string), 'clave' (string), 'nombre' (string), and 'rol' (entero). The second table, 'Configuracion', has ten attributes: 'codigo' (entero), 'host' (string), 'puerto' (entero), 'usuario' (string), 'clave' (string), 'nombre_base' (string), 'tipo_base' (entero), 'consulta' (string), and 'estado' (entero).

Usuario	
codigo	entero
usuario	string
clave	string
nombre	string
rol	entero

Configuracion	
codigo	entero
host	string
puerto	entero
usuario	string
clave	string
nombre_base	string
tipo_base	entero
consulta	string
estado	entero

Figura 3.5: Diagrama Entidad Relación.

e implementación del producto, afinando y revisando que los requerimientos cumplan con las necesidades o aspectos primordiales del sistema, con el fin de dar la mejor solución al problema planteado.

Como resultado de las iteraciones se ha obtenido el diagrama de clases que contiene las clases de entidades, clases controladoras y clases de vista, las mismas que en su conjunto conforman la solución final elaborada como se puede observar en la figura [3.6](#)

Teniendo en cuenta que el conjunto de clases entidades fueron creadas con el propósito de aplicar Mapeo Relacional de Objetos (ORM) de la librería SQLAlchemy para obtener el máximo de beneficios y flexibilidad del SQL por medio de la persistencia, ganando tiempo en el desarrollo y simplificando actividades para el desarrollador. Por otro lado, se utilizó el patrón modelo vista controlador en donde las clases controladoras contendrán los métodos que permitirán interactuar con las clases entidades permitiendo hacer el uso de la persistencia al únicamente operar con los objetos en las funciones básicas de crear, actualizar y eliminar, pero sin olvidar los métodos de búsqueda ya sea por objeto como por Id. Para terminar las clases que forman las vistas son las que consumen las funciones de las clases controladoras interactuando directamente con los componentes como formularios, campos de texto, botones, etc., capturando o visualizando la información según la acción solicitada.

El segundo sprint de igual manera se desarrolló en 4 semanas presentando

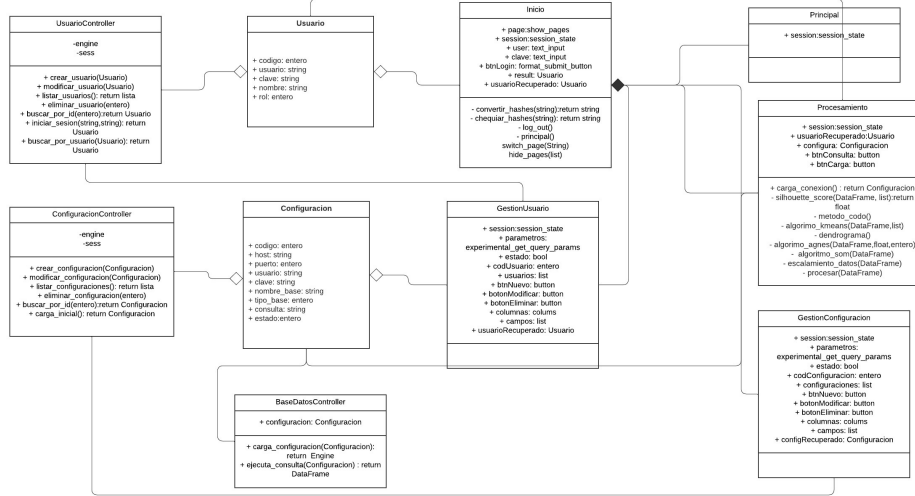


Figura 3.6: Diagrama de Clases.



Figura 3.7: Historias de Usuario Sprint 2

un mayor esfuerzo en la última historia de usuario llamada **procesamiento de datos** la misma que contiene el conjunto de actividades para el tratamiento de los datos así como los resultados finales para ser evaluados por el cliente para finalizar con su aceptación u obtener la retroalimentación respectiva. figura 3.7

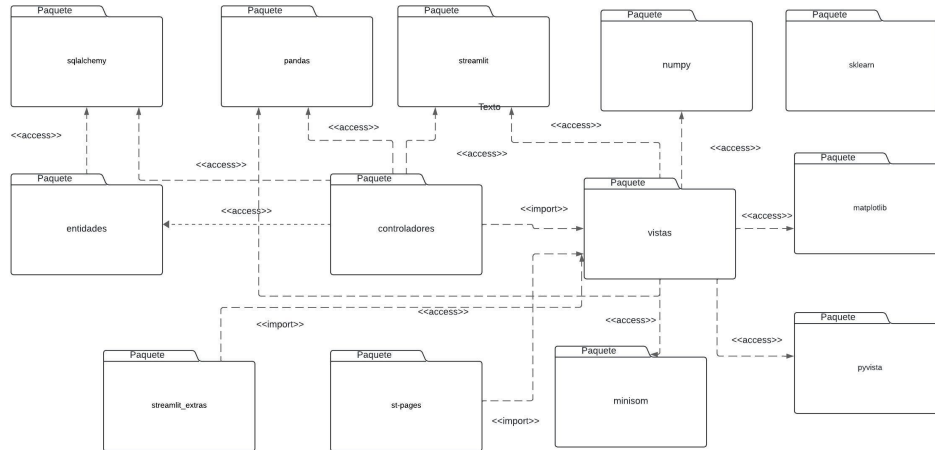


Figura 3.8: Diagrama de Paquetes.

Como un resultado también, se ha visto necesario indicar el respectivo diagrama de paquetes, en donde se puede apreciar la distribución de las clases agrupadas en sus respectivos paquetes, así como las relaciones y dependencias que tienen entre ellos, tanto de paquetes propios del desarrollo como de terceros, como se puede ver en la Figura 3.8 .

3.2.1. Datos

Con respecto a los datos, estos son proporcionados por una Cooperativa de Ahorro y Crédito de la ciudad de Cuenca en la provincia del Azuay país Ecuador con un número de socios activos de 4300, la información brindada es de sus socios con corte al 31 de diciembre del 2023. Es importante conocer las variables que conforman esta información, las mismas se detallan a continuación.

- Socio: muestra la identificación única de cada socio. Etiqueta `cod_socio`
- Edad: es la edad del socio en años, es calculada a partir de la fecha actual menos la fecha de nacimiento. Etiqueta `edad`

- 18 Valor mínimo: es el mínimo valor considerado para ser mayor de edad y que permite formar parte de la Cooperativa como socio
- 78 Valor máximo: considerado el valor máximo debido a que por políticas internas el socio puede ser sujeto a un crédito
- Género: es el sexo biológico registrado en el almacén de datos o consulta. Etiqueta genero
 - 0: masculino
 - 1: femenino
- Instrucción: es el grado de educación que tuvo el socio en adquirir tiempo de vida. Etiqueta cod_instruccion
 - 1: Primaria
 - 2: Secundaria
 - 3: Universitaria
 - 4: Ninguna
 - 5: Tecnología
 - 6: Postgrado
- Estado civil: estado que posee el socio actualmente reconocido en su documento de identificación. Etiqueta estado_civil
 - 1: Soltero
 - 2: Casado
 - 3: Viudo
 - 4: Divorciado
 - 5: Otro
- Parroquia: dato que refiere a la parroquia donde reside, está compuesto por la división política del Estado Ecuatoriano. Etiqueta cod_parroquia
- Cargas Familiares: indica el número de cargas familiares que posee el socio. Etiqueta num_cargas_familiares
- Tiempo trabajo: es el periodo en años que un socio se encuentra en su trabajo. Etiqueta num_tiempo_trabajo

cod_socio	edad	genero	cod_instruccion	estado_civil	num_cargas_familiares	num_tiempo_trabajo	cod_parroquia	val_ingreso_mensual	val_patrimonio	ahorros	polizas	otros_ahorros
14,332	73	1	4	3	0	10	10,170	1,267	40,000	4,623.25	0	0
914	18	0	2	1	0	0	10,170	20	0	10.02	0	0
11,217	33	1	2	1	2	2	10,107	350	350	0	0	0
10,819	73	0	1	2	0	3	11,051	300	68,500	6.74	0	0
10,237	40	1	3	4	1	3	10,170	630.87	2,436.65	18.51	0	0
69	47	0	1	2	0	25	10,170	600	38,000	36.72	0	0
2,814	48	1	3	4	2	7	10,113	2,350	128,693.85	15.12	0	0
443	49	0	1	2	2	10	10,170	600	15,000	3.4	0	0
11,278	38	0	3	1	0	0	10,113	800	0	0	0	0
2,656	58	1	3	2	0	17	10,170	800	255,296.57	57.08	3,151.67	0

Figura 3.9: Variables.

- Ingresos: remuneración percibida del socio de forma mensual. Etiqueta val_ingreso_mensual
- Patrimonio: es el valor que el socio tiene después de haber realizado la operación de activos – pasivos. Etiqueta val_patrimonio
- Ahorros: valor que posee el socio a una fecha de corte. Etiqueta ahorros
- Pólizas: es la suma de valores de las inversiones activas que posee el socio en la institución a una fecha de corte. Etiqueta polizas
- Otros ahorros: es la suma de otros valores de tipo de ahorros que posee el socio en la institución con una fecha de corte. Etiqueta otros_ahorros

3.2.2. Preprocesamiento de Datos

Debido a que la dimensionalidad (características) es un factor importante para el correcto funcionamiento de los algoritmos seleccionados, se optó por realizar la eliminación de características de manera manual fundamentando que las columnas a eliminar son de tipo categórica y no serán muy relevantes en los modelos de machine learning como son el género, instrucción, estado civil, y parroquia ya que un principal enfoque es identificar los socios que tienen dinero en sus diferentes tipos de cuentas de ahorros, solvencia patrimonial e inversiones. Así pues se procede a trabajar con las dimensiones que contengan valores significativos como la Edad, Cargas Familiares, Tiempo trabajo, Ingresos, Patrimonio, Ahorros, Pólizas y Otros

cod_socio	edad	num_cargas_familiares	num_tiempo_trabajo	val_ingreso_mensual	val_patrimonio	ahorros	polizas	otros_ahorros
14,332	73	0	10	1,267	40,000	4,623.25	0	0
914	18	0	0	20	0	10.02	0	0
11,217	33	2	2	350	350	0	0	0
10,819	73	0	3	300	68,500	6.74	0	0
10,237	40	1	3	630.87	2,436.65	18.51	0	0
69	47	0	25	600	38,000	36.72	0	0
2,814	48	2	7	2,350	128,693.85	15.12	0	0
443	49	2	10	600	15,000	3.4	0	0
11,278	38	0	0	800	0	0	0	0
2,656	58	0	17	800	255,296.57	57.08	3,151.67	0

Figura 3.10: Campos Significativos.

ahorros. Para lo cual como un segundo paso después de la carga del archivo o la consulta a la base de datos se procede a eliminar las columnas del conjunto de datos con la función `drop()` de la librería `pandas` donde se especifican las columnas a eliminar quedando como resultado la tabla que se ve en la figura 3.10

Como otra etapa del preprocesamiento se procedió a eliminar todas las filas que poseen valores nulos o vacíos ya que estos pueden causar errores en el procesamiento como en el análisis de predicción de datos teniendo una reducción del conjunto de datos. Esta tarea fue realizada con la función de `Pandas dropna()` que devuelve un nuevo conjunto de datos tras eliminar los valores nulos.

Por otro lado, el conjunto de datos tiene características numéricas que se han medido en diferentes unidades como los años, el patrimonio, etc. por lo general esta desigualdad genera preferencias en los algoritmos, al considerar mayor peso en los valores más altos. Razón por la cual se debe transformar las características numéricas a una misma escala, lo cual se ha realizado mediante el proceso de escalado de características que tiene el paquete de `sklearn - StandardScaler`. Figura 3.11

	edad	num_cargas_familiares	num_tiempo_trabajo	va_ingreso_mensual	va_patrimonio	ahorros	polizas	otros_ahorros
7	0.1718	0.9129	0.0257	-0.2927	-0.425	-0.1472	-0.1872	-0.0561
8	-0.6052	-0.8587	-0.8816	-0.0364	-0.5679	-0.1482	-0.1872	-0.0561
9	0.8074	-0.8587	0.6608	-0.0364	1.8628	-0.1317	0.1927	-0.0561
10	-0.7464	2.6846	1.1145	-0.7413	-0.3774	-0.1472	-0.1872	-0.0561
11	-0.6052	3.5704	-0.6094	-0.8053	-0.0918	-0.1429	-0.1872	-0.0561
12	0.6662	0.9129	-0.5187	0.5724	-0.168	-0.1409	-0.1872	-0.0561
13	-0.0401	1.7987	-0.5187	0.0918	-0.0918	-0.1418	-0.1872	-0.0561
14	0.8781	0.0271	1.4774	-0.2927	0.2366	-0.0684	0.09	-0.0561
15	-1.1702	-0.8587	-0.7909	-0.5642	-0.2186	-0.0773	3.9718	-0.0561
16	0.1718	1.7987	-0.7909	-0.4209	0.289	-0.1468	-0.1872	-0.0561

Figura 3.11: Escalado de Características.

3.2.3. Procesamiento de Datos

El principal objetivo de esta sección consiste en aplicar las técnicas de segmentación escogidas, teniendo en cuenta que nuestro aprendizaje no será supervisado porque únicamente tenemos un conjunto de datos y el objetivo del proyecto es agrupar a los socios en base a su número de características o variables, así como sus valores atípicos.

Como primera opción se escogió la agrupación particionaria, por la característica de que ningún miembro puede formar parte de más de un clúster. Uno de sus principales representantes y el más famoso es K-Means.

K-Means necesita decidir el número de grupos (k) que queremos identificar, este debe ser más de uno si no, no tendría sentido, existen algunas técnicas para estimar el número de clústers, pero la que escogeremos es el método del codo (Elbow method) ya que esta analiza la variabilidad dentro de los clústers, medida como WCSS (Within-Cluster Sum of Squares), cambia a medida que aumenta el número de clústers. El propósito es encontrar un balance en el que el incremento de número de clústers no produce una mejora significativa de la variabilidad (WCSS).

Para desarrollar la técnica del codo y obtener el valor de K se crea una función dentro de la clase procesamiento llamada `metodo_codo()` la que recibe el conjunto de datos como parámetro. A continuación, invocamos a la

función `KMEANS()` de la librería Sklearn en la cual indicamos los siguientes hiperparámetros de entrada:

- `n_clusters = n` (la cantidad de clústeres que queremos).
- `init= k-means++` (esta opción selecciona los clústeres iniciales de forma inteligente y con esto acelera la convergencia).
- `n_init = 10` (el número de veces que el algoritmo se ejecutará con diferentes clústeres iniciales).
- `max_iter = 300` (número de iteraciones que se ejecuta el algoritmo).
- `random_state = 42` (número de aleatorios para la inicialización del centroide).

Para el parámetro de `n_clusters` fijamos el valor de `n` lo que significa que este valor vendrá dado por un rango de 1 a 20. Iteramos esta función por medio de una estructura de repetición en base al rango de `n`, seguido de la función `fit()` de la librería `KMEANS` para que calcule la agrupación cuyo parámetro de entrada será el conjunto de datos, y por último almacenamos en un arreglo(A) los valores de aplicar el atributo `inertia_` que realiza la suma de las distancias al cuadrado de las muestras al centro de su grupo más cercano, mediante `plot` de la clase `pyplot` y de la librería `matplotlib` se realiza la gráfica teniendo como eje X el número de clústeres y Y los valores resultantes(arreglo A). Figura 3.12 extraída de la vista Procesamiento - Tab - K-MEANS

En la gráfica 3.12 como se puede observar se constató que el valor significativo donde descende el número de clústers es el número 6, después de este los valores se suavizan.

Por último en la misma clase se crear una función llamada `algoritmo_kmeans()` cuyos parámetros de entrada son el conjunto de datos y el valor de los clúster(k) que llama al algoritmo con los parámetros indicados anteriormente y con el valor resultante de `k = 6`, se obtiene los resultados de nuestro conjunto de datos, en donde mediante la función `plot` se realiza la gráfica de los ingresos mensuales frente a la edad de los socios y además la ubicación de los centroides como se puede apreciar en la figura 3.13 extraída de la vista Procesamiento - Tab - K-MEANS

Como segunda opción se escogió la agrupación jerárquica enfocada en una subdivisión de esta, llamada agrupación aglomerativa, que consiste en la agrupación ascendente iniciando con dos puntos similares hasta terminar con

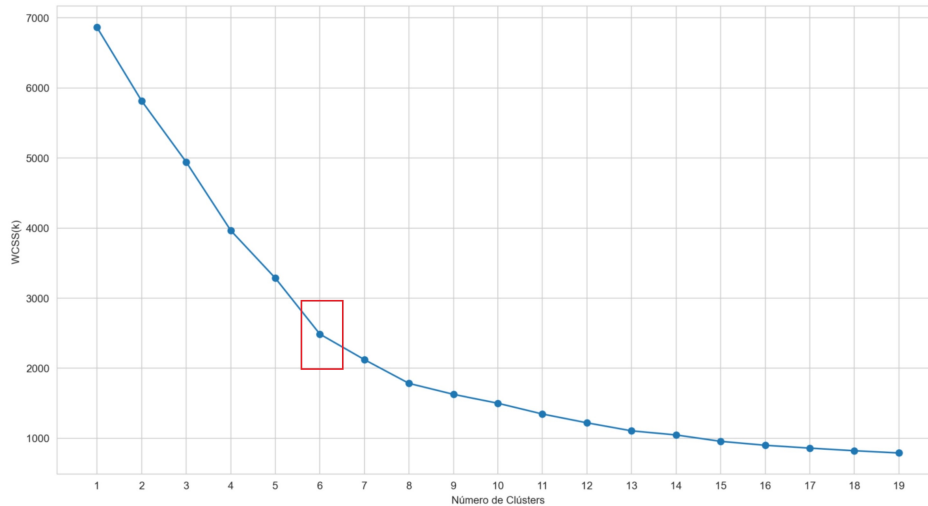


Figura 3.12: Método del Codo.

todos los puntos formando un grupo, el algoritmo de AGNES será utilizado en el presente trabajo.

Los dendrogramas nos darán el número de agrupaciones recomendadas para nuestro problema, este método es tan simple como dibujar una línea paralela de tal forma que intercepte con el mayor número de líneas verticales sin chocar con alguna línea horizontal.

Para realizar esta tarea dentro de la clase procesamiento se crea una nueva función llamada `dendrograma()` que recibe como único parámetro de entrada el conjunto de datos, y se procede a realizar la siguiente instrucción. Invocando a la función `linkage()` de la librería `scipy` ingresamos como parámetros nuestro conjunto de datos, así como segundo parámetro igual a `ward` que es el método para que pueda hacer el análisis de los conglomerados dando como resultado una agrupación, la misma que ingresando como parámetro a la función `dendrogram()` de la librería `scipy` podremos obtener la figura 3.14 extraída de la vista Procesamiento - Tab - AGNES

En la figura 3.14 vemos que la línea horizontal interseca las líneas verticales de mayor distancia marcando 6 grupos diferentes, resultando 6 agrupaciones para el proyecto. Hay que tener en cuenta que esta línea horizontal no debe sobrescribir a ninguna línea de la gráfica.

Continuando con el proceso de desarrollo para este algoritmo procedemos a la creación de la función `algoritmo_agnes()` en la clase procesamiento al

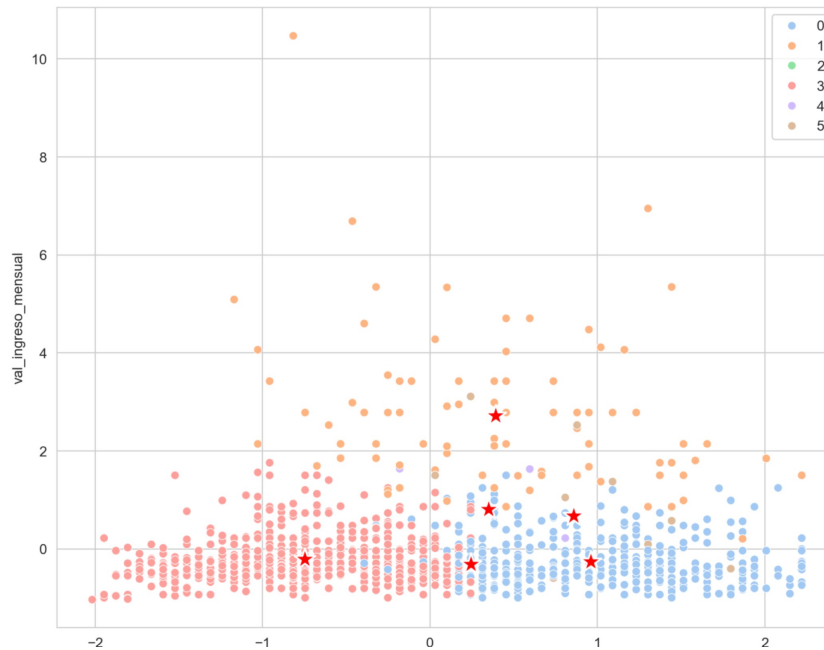


Figura 3.13: K-means.

igual que el anterior también recibirá como parámetro el conjunto de datos. Esta función contendrá la llamada a la función `AgglomerativeClustering()` de la librería `sklearn` con los siguientes hiperparámetros de entrada:

- `n_clusters = 6` (número de clústeres que se van a encontrar).
- `metric= euclidean` (métrica utilizada para calcular el vínculo).
- `linkage = ward` (el criterio de vinculación determina qué distancia utilizar entre conjuntos de observación).

Con el resultado del algoritmo procedemos a obtener la segmentación de cada uno de los objetos con la función `fit_predict()` el cual recibe como parámetro el conjunto de datos. Los resultados de los objetos ya categorizados deben ser agregados adhiriendo una columna a su conjunto de datos original.

Estos resultados fueron graficados de la misma manera que el método `plot` anteriormente descrito, ingresos mensuales frente a la edad de los socios. Cabe recalcar que esta figura 3.15 extraída de la vista `Procesamiento`

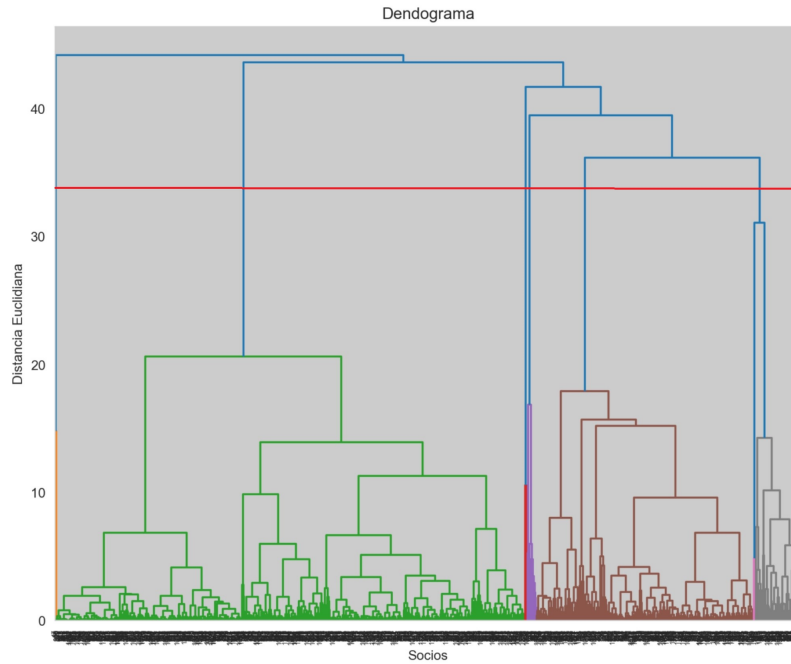


Figura 3.14: Dendrograma.

- Tab - AGNES no posee centroides.

Finalmente, la opción de redes neuronales que se basa en el aprendizaje, simulando a un cerebro humano, e imitan la forma en la que las neuronas biológicas se señalan entre sí. Estas están formadas por una capa de entrada, capas intermedias u ocultas y una capa de salida, todos los nodos se conectan entre sí teniendo un peso y un umbral.

"Mapas Autoorganizados (SOM). se entrena a través de una red neuronal competitiva, una red de retroalimentación de una sola capa que se asemeja a estos mecanismos cerebrales, este está compuesto por una capa de entrada y una capa de salida siendo esta una red bidimensional de $m \times n$ neuronas. En el proceso de aprendizaje las neuronas de la capa de salida compiten entre sí para ser activadas, este proceso competitivo consiste en buscar a la neurona más similar con el patrón de entrada, al ganador se le llama Mejor Unidad Coincidente (BMU).

El criterio ganador se puede medir con la distancia euclidiana que es la medida más común empleada. La BMU comparte con sus vecinos el privilegio

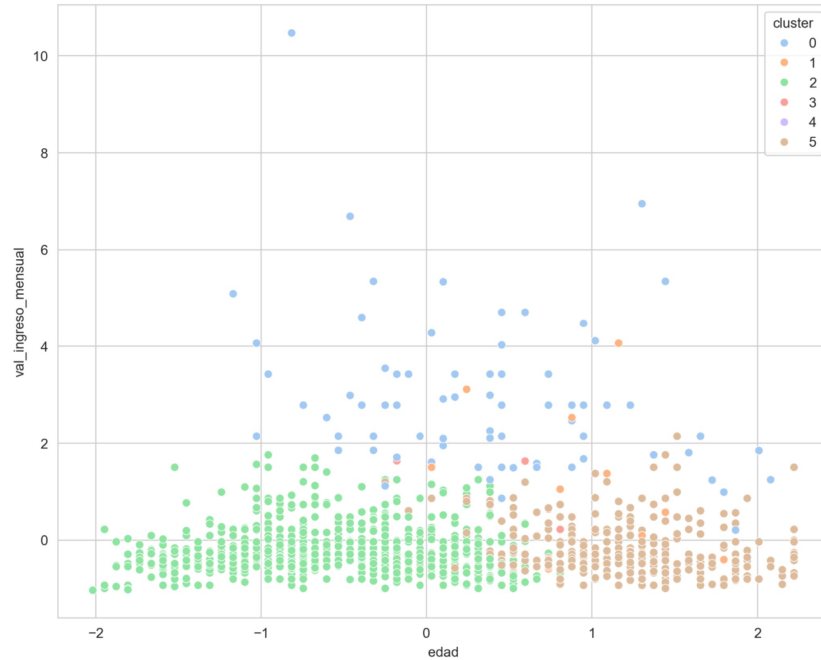


Figura 3.15: AGNES.

de aprender juntos ajustando sus pesos. El efecto del aprendizaje es acercar el peso de las neuronas ganadoras y vecinas al patrón de entrada." [Álvaro José García-Tejedor and Nogales, 2022]

De la misma forma que para los algoritmos anteriores se procederá a crear una función llamada `algoritmo_som()` en la clase `procesamiento` con el parámetro de entrada el conjunto de datos, esta función contendrá el algoritmo MiniSom con sus hiperparámetros de entrada:

- `x_dim` es la dimensión del espacio de entrada.
- `y_dim` es la dimensión del espacio de salida.
- `input_len` es la longitud de un dato de entrada
- `sigma` es el radio de vecindad.
- `learning_rate` es la tasa de aprendizaje.

Teniendo en cuenta las recomendaciones del autor [Vettigli \[2018\]](#) de la librería, quien indica que la aplicación del algoritmo para el caso de agrupamiento, cada dimensión corresponde a un clúster, por lo que se optaron los valores de `x_dim` será 1 y `y_dim` será de 6 para de esta forma equiparar las segmentaciones en base a los algoritmos anteriores, la longitud del dato hace referencia a la dimensionalidad en nuestro caso es 6, el radio de los diferentes vecinos del SOM será 0.5 (Función de Activación), y la tasa de aprendizaje en cada iteración será de 0.5.

Siguiendo con el desarrollo se procede a entrenar la red para lo cual se hizo uso de la librería `MiniSom` y con la función `som.update()` la cual recibirá como parámetros primero un objeto (fila) de nuestro conjunto de datos, como segundo parámetro el valor resultante de la función `som.winner()` con parámetro de entrada el mismo objeto del conjunto de datos y por último el número de iteraciones que se ejecutará el algoritmo para que se equilibren los pesos de los valores de las neuronas. Se procede a realizar esta acción el número de veces que se desee para entrenar la red. Para la aplicación se procedió a realizar un entrenamiento con 1000 iteraciones, tomando de manera aleatoria cada uno de los objetos del conjunto de datos y pasándolos por este proceso.

Para finalizar la función `algoritmo_som()` se procede a pasar el conjunto de datos de entrada por la función interna `som.winner()` almacenando sus valores en un arreglo `(A)` para obtener las neuronas ganadoras para después continuar en la conversión bidimensional por medio de una función de Numpy llamada `ravel_multi_index()` con los valores de entrada `array(A)` y la matriz `(mxn)` entrega los valores de los clústeres para proceder a graficarlos, de igual manera que las gráficas anteriores donde en el eje X esta la edad y en el eje Y el ingreso mensual. Figura 3.16 extraída de la vista Procesamiento - Tab - SOM

3.2.4. Métrica de Evaluación

Ahora bien, ya que se ha obtenido las agrupaciones resultantes por cada una de las técnicas, es necesario saber cuál fue la que tuvo mayor efectividad en la clasificación de sus aglomerados, por lo que es necesario aplicar una métrica que nos permita evaluar nuestro objetivo.

Silhouette es un método de validación interna basada en la cohesión. Siendo una medida que verifica de cuán es la similitud de un objeto con su

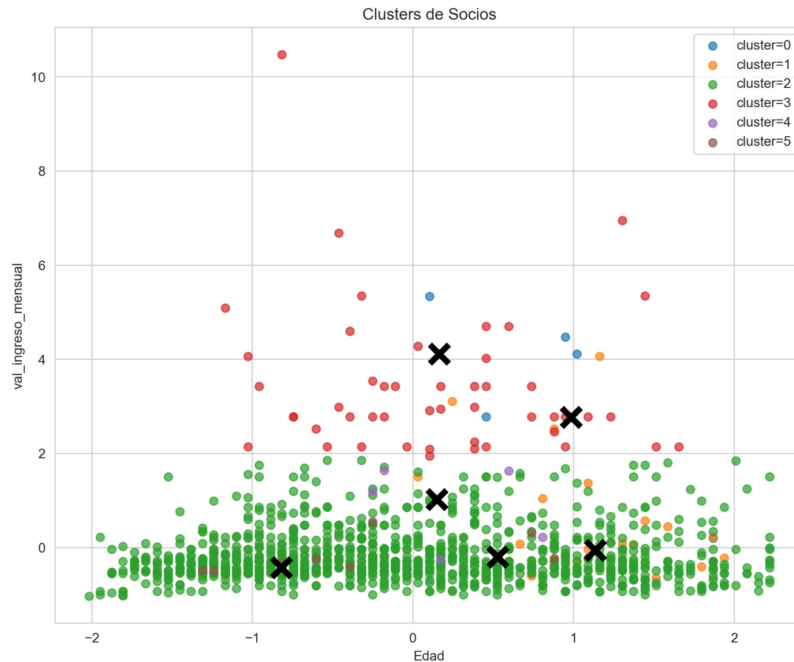


Figura 3.16: SOM.

propio grupo y su disimilitud con los otros grupos. Sus valores van desde -1 a 1 siendo el mayor el resultado de cuán emparejado está el objeto con su propio grupo. [Rousseeuw, 1987]

Puesto que el método de Silueta es una opción que se acopla a las necesidades de valoración para la aplicación se procede a crear la función `silhouette_score()` dentro de la clase `procesamiento` como indica la figura 3.6 extraída de la vista `Procesamiento - Tab - SILUETA SCORE`. Este método recibirá de parámetros, el conjunto de datos y el resultado de los clústeres de las diferentes técnicas de segmentación mostrando en una sección de la vista `procesamiento` los valores resultantes de cada algoritmo aplicado.

3.2.5. Interfaz Gráfica

La interfaz gráfica forma parte fundamental en toda aplicación ya que permite la interacción directa con el usuario, esta debe ser de fácil manejo y contener la mayor cantidad de información que ayude al usuario final a interactuar de forma clara, rápida y precisa por lo que debe describir cual

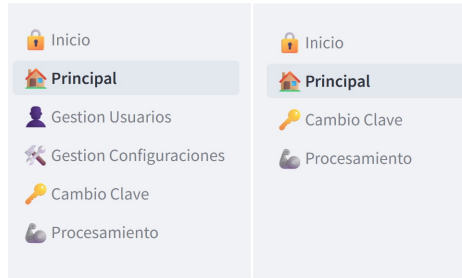


Figura 3.17: Menú Usuario Administrador - Menú Usuario

es la función de cada componente de la vista actual en donde se encuentra el usuario como son menús, campos de texto, botones, etc. Esta simplicidad no debe comprometer ciertos aspectos fundamentales de seguridad y más aún cuando la información que se maneja es de carácter confidencial o sensible, por lo que en el desarrollo de la aplicación se tomaron en cuenta estos aspectos.

Con respecto al control de acceso se utilizó la funcionalidad de Streamlit `session_state` en todas las páginas de la aplicación para verificar su estado activo cuando se desee navegar por las páginas del menú o cuando se quiera cargar directamente la página en un navegador.

Además, se implementaron dos medidas más para controlar el acceso de los usuarios, para empezar una página de inicio de sesión llamada Inicio la cual contiene los campos de texto para el ingreso de usuario y contraseña que una vez verificados crea la sesión actual del usuario para que pueda ser validado en las demás páginas, en segundo lugar, un menú dinámico que en base al rol y privilegios que maneja el usuario se visualizarán las opciones que tiene permitido. Para lo cual se utilizó la librería `st_pages` que posee la funcionalidad de mostrar u ocultar las páginas del menú. Vease la figura [3.17](#)

Con respecto al menú de la aplicación a continuación se detalla cada una de las páginas que lo conforman y a su vez se indica su funcionamiento, así como los métodos primordiales de los que están compuestos.

Inicio es la página cuyos componentes ya fueron indicados al inicio de esta sección, su principal funcionalidad es para la verificación del usuario y su contraseña, así como el inicio de sesión, que una vez activo procede a cambiar su vista desactivando componentes de validación de usuario y activando componentes para la desconexión y cierre de sesión.



Figura 3.18: Vista de Inicio

Principal es la primera página que se muestra una vez que el inicio de sesión fue correcto, contiene únicamente información de cómo funciona la aplicación. Figura 3.18

Cambio Clave la funcionalidad de esta página es para que el usuario pueda cambiar su contraseña con un campo extra para que permita hacer la confirmación del cambio realizado. Figura 3.19

Gestión Usuarios mediante esta página se puede realizar las operaciones básicas de la gestión de usuario como crear, modificar, eliminar y buscar, estos eventos se comunican directamente con la clase UsuarioController del paquete Controladores donde estas funciones interactúan directamente con la base de datos SQLite 3. Visualmente esta página se compone de columnas para formar una tabla que visualiza los registros de los usuarios del sistema, al mismo tiempo que posee botones de modificación y eliminación de cada registro, por ultimo posee un componente de formulario que contiene todos los campos necesarios para crear o modificar un registro en la tabla Usuario de la base de datos. Véase la figura 3.20

Gestión Configuraciones trabaja de igual manera que la descripción

Cambio Clave

Usuario conectado root

Ingrese la clave 0/100

Vuelva a ingrese la misma clave 0/100

Guardar

Figura 3.19: Vista Cambio de Clave

Gestion Usuarios

Nuevo

Codigo	Usuario	Nombre	Rol	Modificar	Eliminar
1	diego	Diego Ruiz	1	Modificar	Eliminar
2	root	Super Usuario	1	Modificar	Eliminar
3	priscila	Priscila Bernal	2	Modificar	Eliminar
4	johanna	Johanna Hurtado	2	Modificar	Eliminar
5	sonia	Sonia Nieves	2	Modificar	Eliminar
6	sandra	Sandra Arevalo	2	Modificar	Eliminar
7	claudio	Claudio Chillogalli	2	Modificar	Eliminar

Figura 3.20: Vista Gestión de Usuarios

del funcionamiento de la página de "Gestión Usuarios", pero los componentes a destacar constituyen los parámetros que permiten la conexión con la base de datos del Core Financiero de la institución al igual que el campo que guarda la consulta SQL por medio de la cual se extraerá la información para crear el conjunto de datos. Esta vista posee los métodos de conexión para 3 bases de datos como son ORACLE, MySQL y PostgreSQL, siendo este último el método de conexión con el que trabaja la Cooperativa, en las que

Ingreso de Configuración

Ingrese el host de la base de datos

Ingrese el nombre de usuario de la base de datos

Ingrese la clave de la base de datos

Ingrese el nombre de la base de datos

Base de Datos

PostgreSQL MySQL Oracle

Ingrese consulta

Estado

Guardar

Figura 3.21: Vista Crear configuración

se hicieron las pruebas y la implementación de la solución. Véase la figura [3.21](#)

Procesamiento, esta página concentra todo el procesamiento de la aplicación de allí su nombre, está conformada por un componen de `st.tabs` que contiene un arreglo con las siguientes opciones:

Procesamiento

Datos K-MEANS AGNES RED NEURONAL (SOM) SILUETA SCORE

Datos

Consultar

Archivo de datos csv

Drag and drop file here
Limit 200MB per file • CSV

Browse files

dataSocio2.csv 52.5KB

cod_socio	edad	genero	cod_instruccion	estado_civil	num_cargas_familiares	num_tiempo_trabajo	cod_parroquia	vaLingreso_mensual	vaLpatrimonio	ahorros	poliz
14,332	73	1	4	3	0	10	10,170	1,267	40,000	4,623.25	
914	18	0	2	1	0	0	10,170	20	0	10.02	
11,217	33	1	2	1	2	2	10,107	350	350	0	

Figura 3.22: Vista Procesamiento - Tab - Datos

- DATOS, esta vista permite la carga del archivo csv, así también la posibilidad de generar la información desde una consulta almacenada que se indicó en la página "Gestión Configuraciones". Después de accionar cualquiera de las 2 opciones se procede a llamar a las funciones para que ejecuten los algoritmos de las técnicas de segmentación. Véase la figura 3.22
- K-MEANS, muestra el resultado de manera gráfica de 2 algoritmos el "Método del Codoz "K-means", además contiene los componentes como un campo de texto para cambiar el número de clústeres y un botón para el reprocesamiento de resultados. Véase la figura 3.23
- AGNES, de igual manera muestra los resultados de manera gráfica del Dendrograma así como el resultado de segmentación del algoritmo AGNES permitiendo modificar el número de clústeres para un reprocesamiento de ser el caso. Véase la figura 3.24
- RED NEURONAL (SOM) en esta vista se realizó un gráfico del resultado del error de cuantificación, así también como el resultado de la segmentación del algoritmo SOM, con los respectivos campos y botones para modificar los valores de la matriz, función de activación (Sigma) y la tasa de aprendizaje. Véase la figura 3.25

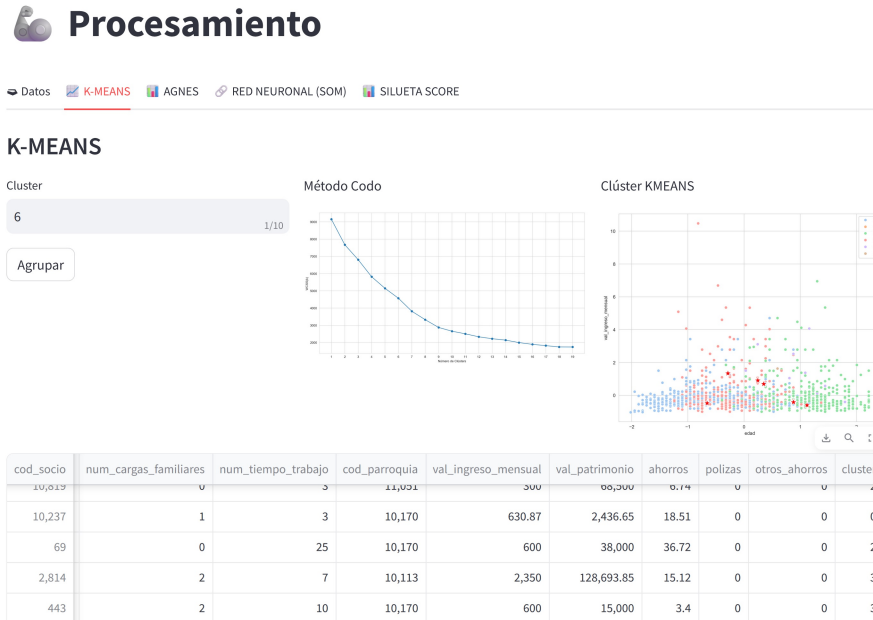


Figura 3.23: Vista Procesamiento - Tab - K-MEANS

- SILUETA SCORE muestra los resultados numéricos del porcentaje evaluado de cada una de las técnicas.

Para finalizar se indica que todas las opciones de la Vista Procesamiento componente tabs que contienen los algoritmos de segmentación poseen las tablas con los resultados finales de la agrupación a la que pertenece cada registro, para proceder a la descarga de ser el caso. De esta forma podrán hacer uso los usuarios finales cumpliendo el objetivo para lo que fue desarrollado la aplicación. Véase la figura 3.26 extraída del sistema vista Procesamiento - Tab - K-MEANS

Procesamiento

Datos K-MEANS **AGNES** RED NEURONAL (SOM) SILUETA SCORE

AGNES

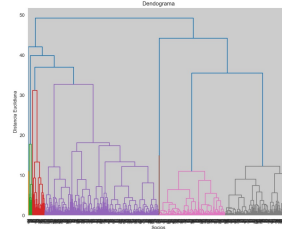
Cluster AGNES

6

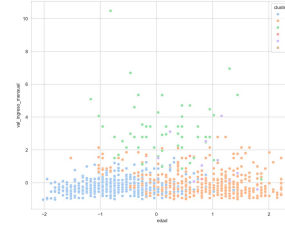
1/10

Agrupar AGNES

Dendrograma



Segmentación AGNES



cod_socio	num_cargas_familiares	num_tiempo_trabajo	cod_parroquia	val_ingreso_mensual	val_patrimonio	ahorros	polizas	otros_ahorros	cluster
14,332	0	10	10,170	1,267	40,000	4,623.25	0	0	1
914	0	0	10,170	20	0	10.02	0	0	0
11,217	2	2	10,107	350	350	0	0	0	0
10,819	0	3	11,051	300	68,500	6.74	0	0	1
10,237	1	3	10,170	630.87	2,436.65	18.51	0	0	0

Figura 3.24: Vista Procesamiento - Tab - AGNES

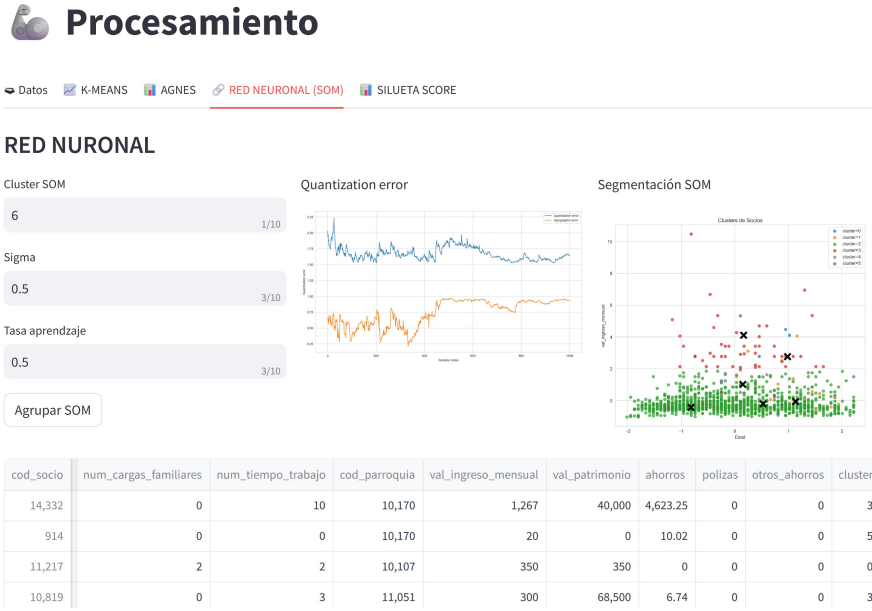


Figura 3.25: Vista Procesamiento - Tab - SOM

-2 -1 0 1

edad

Download as CSV

cod_socio	cccion	estado_civil	num_cargas_familiares	num_tiempo_trabajo	cod_parroquia	val_ingreso_mensual	val_patrimonio	ahorros	polizas	otros_ahorros	cluster
14,332	4	3	0	10	10,170	1,267	40,000	4,623.25	0	0	2
914	2	1	0	0	10,170	20	0	10.02	0	0	0
11,217	2	1	2	2	10,107	350	350	0	0	0	3
10,819	1	2	0	3	11,051	300	68,500	6.74	0	0	2
10,237	3	4	1	3	10,170	630.87	2,436.65	18.51	0	0	0
69	1	2	0	25	10,170	600	38,000	36.72	0	0	2
2,814	3	4	2	7	10,113	2,350	128,693.85	15.12	0	0	3
443	1	2	2	10	10,170	600	15,000	3.4	0	0	3
11,278	3	1	0	0	10,113	800	0	0	0	0	0
2,656	3	2	0	17	10,170	800	255,296.57	57.08	3,151.67	0	2

Figura 3.26: Tabla Resultante con Columna Clúster Agregada

Capítulo 4

Resultado y Discusión

4.1. Resultados y Discusión

4.1.1. Resultados

Con respecto al software desarrollado se puede indicar que el mismo ha resultado exitoso, debido a que se pudo realizar todas las necesidades descritas en las historias de usuario, permitiendo condensar todas las actividades que hacen otros programas en una serie de pasos de personas especializadas a una únicamente función que se acciona desde la carga de información o consulta a la base de datos, realizando la evaluación de cada una de las Técnicas empleadas con resultados visuales de los métodos empleados en cada algoritmo, además, con la bondad de su respectiva descarga para brindar la información al alcance de sus usuarios. También, presenta una interfaz amigable e intuitiva para los usuarios permitiendo una correcta administración para el caso del administrador y una herramienta valiosa en la automatización del análisis y segmentación de los datos para los usuarios finales (marketing) permitiendo apuntar a segmentos de socios de manera focalizada en un tiempo mucho menor.

Hay que destacar que la herramienta Streamlit como las librerías que apoyan a este framework son de potente ayuda ya que permitieron cumplir con los sprints en los tiempos planificados, debido a la facilidad y reacción que posee el framework cuando se realizan cambios y los tiempos que se reflejan los mismos evitando los tiempos muertos de compilación.

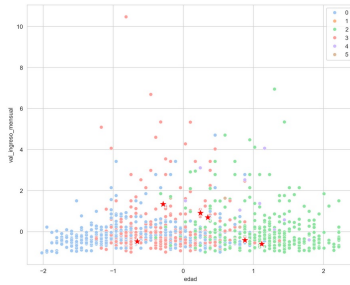
En relación al conjunto de datos conformado de 1144 objetos los cuales fueron cargados mediante archivo en 3 ocasiones para ser procesados por los algoritmos, se puede apreciar que los 2 determinaron la segmentación de

Procesamiento

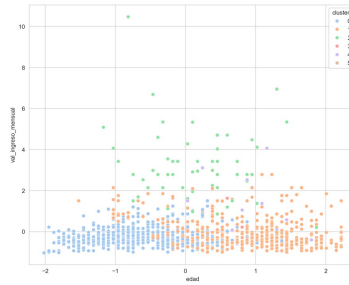
Datos K-MEANS AGNES RED NEURONAL (SOM) SILUETA SCORE

Silueta Score

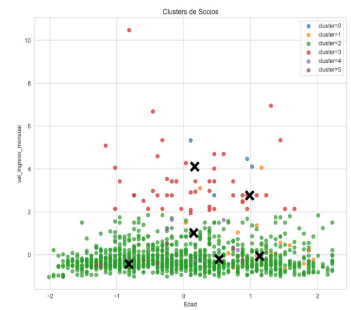
Resultado K-MEANS



Resultado AGNES



Resultado SOM



Algoritmo K-MEANS 0.43636271339072175 %

Algoritmo AGNES 0.28856024807689346 %

Algoritmo SOM 0.24440657503082533 %

Figura 4.1: Resultado de Evaluación Silueta.

los datos en 6 grupos cuyos valores resultantes se repiten para las técnicas de K-Means y AGNES, en cuanto para el algoritmo SOM como se indicó en el desarrollo del proyecto el usuario fija el valor del número de clústeres para el agrupamiento (matriz de $m \times n$) donde podemos indicar que su segmentación se basó en resultados de otros algoritmos para que se pueda comparar de manera equitativa, sus resultados en las pruebas siempre presentaron una variación para la técnica de agrupamiento como se puede ver en la tabla 4.1

A continuación, se realizó la prueba desde una consulta sql de la base de datos de la entidad financiera, entregando un resultado de 3717 registros, la misma se repitió por 3 ocasiones igual al ejercicio anterior y se obtuvieron los siguientes resultados, 6 segmentos en los 3 algoritmos y los valores

Tabla 4.1: Resultados de Técnicas de Segmentación.

Número de prueba	Técnica de Segmentación	Valor Métrica
1	K-Means	0.6165
1	AGNES	0.6150
1	RNN	0.5410
2	K-Means	0.6165
2	AGNES	0.6150
2	RNN	0.3464
3	K-Means	0.6165
3	AGNES	0.6150
3	RNN	0.4911

Tabla 4.2: Resultados de Técnicas de Segmentación (Consulta SQL).

Número de prueba	Técnica de Segmentación	Valor Métrica
1	K-Means	0.8873
1	AGNES	0.8907
1	RNN	0.8527
2	K-Means	0.8873
2	AGNES	0.8907
2	RNN	0.6618
3	K-Means	0.8873
3	AGNES	0.8907
3	RNN	0.8400

Tabla 4.3: Promedios de Técnicas de Segmentación.

Técnica de Segmentación	Promedio
K-Means	0,7519
AGNES	0,75285
RNN	0,622166667

resultantes que se muestran en la tabla 4.2. En consecuencia, se obtuvo el mismo efecto para la Técnica de Redes neuronales(SOM) en la cual los valores fueron distintos.

Con respecto a los resultados en las gráficas observamos que tenemos 6

clústeres, es decir que los socios del conjunto de datos están segmentados en 6 grupos diferentes:

- Clúster 0: socios entre 35 a 57 años de edad con ingresos en un rango de 600 a 9000 dólares y con pólizas entre 8000 a 136000 dólares.
- Clúster 1: socios de edad entre 45 a 56 años con ingresos 600 dolares a 920 dólares y con pólizas entre 0 y 8000 dólares .
- Clúster 2: socios entre 18 a 57 años de edad con ingresos en un rango de 20 a 600 dólares y con pólizas entre 0 y 1500 dólares .
- Clúster 3: socios de edad entre 45 a 57 años de edad con ingresos en un rango de 425 a 600 dólares y con pólizas entre 1500 y 8000 dólares.
- Clúster 4: socios entre 18 a 35 años de edad con ingresos en un rango de 20 a 600 dólares y con pólizas entre 1500 a 8000 dólares.
- Clúster 5: socios de edad entre 45 a 78 años de edad con ingresos en un rango de 20 a 600 dólares y con pólizas entre 0 y 1500 dólares.

4.1.2. Discusión

Se observa que las técnicas de segmentación K-Means y AGNES durante las pruebas repetitivas no varían su resultado por la evaluación de la métrica, indicando una alta fiabilidad en sus respuestas en la calidad de sus segmentaciones, mientras que SOM varía sus resultados en todas las pruebas, volviéndose inestable en sus predicciones por motivo de las variaciones de sus pesos iniciales. En resumen, realizando un cálculo del promedio de valores de los resultados tenemos como consecuencia que AGNES se posiciona con el mejor valor seguido de K-means.

Si bien en los estudios revisados en el capítulo 2 K-Means consta en la mayoría de las conclusiones como el algoritmo con mejor puntuación en la agrupación o aglomeración de los objetos, se debe tomar en cuenta que las características utilizadas en los almacenes de datos son de máximo 3 dimensiones, mientras que para el trabajo actual se utiliza 6 dimensiones, por lo que una de sus desventajas se ve accionada.

Por consiguiente, AGNES tiene una puntuación superior por mínima que sea ya que las relaciones entre las características de cada objeto revelan detalles más finos de sus similitudes.

Capítulo 5

Conclusiones

5.1. Conclusiones

En conclusión las investigaciones previas permitieron apreciar las diferentes aplicaciones de los algoritmos en un sin número de campos, como la industria, el comercio, la medicina, y el sector financiero, a su vez como los autores pudieron evaluar y ver las ventajas y desventajas que presentan cada uno de ellos en los diferentes artículos, hemos tomado los algoritmos más representativos de clasificación para el presente trabajo, enfocándonos en una parte del segmento financiero que no comparte o existen muchas publicaciones abiertas debido a la competitividad que presenta este mercado, pero la aplicación de los algoritmos en otros ejemplos tendría información muy importante.

Seguidamente indicamos que la metodología XP nos ayudó a realizar mejoras significativas por cada iteración según se avanza en el proyecto ya que al ser orientado para el desarrollo por grupos pequeños permite realizar pruebas y rediseño del producto, como también cambios mínimos en los diferentes diagramas convirtiéndose en una guía precisa de cómo está estructurada la aplicación para futuras actualizaciones, mantenibilidad, etc.

Así mismo se indica que el framework Streamlit en su versión 1.28.2 fue de gran ayuda en la visualización de los resultados ya que posee componentes fáciles de usar, y los cambios se ven reflejados al instante, se considera que Streamlit es una herramienta muy versátil a la hora de convertir código Python o scripts de datos en aplicaciones WEB sin tener mucho conocimiento de Front-end teniendo un mayor enfoque en la ciencia de datos. En pocas palabras fue muy fácil desplegar la aplicación WEB

en la intranet de la Cooperativa para que de manera inmediata pueda ser utilizada para el fin que fue creada.

Por otra parte los resultados obtenidos en la comparación de cuál fue la mejor técnica puntuada por la métrica interna(Silueta), se indica que para el conjunto de datos y sus características AGNES hizo una muy buena agrupación de los objetos al igual que k-means, pero se debe tomar en cuenta que estas técnicas poseen ayuda adicional de otros métodos (Dendrograma, Método del codo) donde el usuario determina el número de clústeres a segmentar, por otra parte el algoritmo SOM con su librería MiniSom de igual manera solicita los valores de la cuadrícula o matriz de $m \times n$ para que se ejecute su algoritmo, en donde el usuario debe tener mayor experticia en visualizar el número de clústeres en los que se puedan segmentar los datos.

Por consiguiente, se puede decir que según la dispersión del conjunto de datos habrá técnicas de segmentación que realicen mejor sus propósitos, por lo que se debe tomar en cuenta esto como un factor muy importante al igual que su dimensionalidad (conjunto de características).

Para concluir debemos indicar que la herramienta únicamente puede funcionar con la estructura de datos que se indica en el Capítulo 3 subsección Datos debido a que hace operaciones específicas con este conjunto de datos siendo este el mayor de los limitantes debido a que es una aplicación focalizada para un mercado específico el cual maneja este tipo de información.

5.2. Recomendaciones

En futuros trabajos se puede realizar las evaluaciones de diferentes tipos de métricas internas ya sea por concepto de cohesión, separación, silueta y estadísticas para comparar los resultados de las técnicas de segmentación utilizadas, así como también se pueden incluir nuevos algoritmos de segmentación como por ejemplo los que están basados en densidad, de esta forma agregará un abanico de técnicas a la herramienta.

También se puede aplicar otros tipos de validaciones como son Validación de estabilidad que prueban la solidez y consistencia de los clústeres en diferentes conjuntos de datos o muestras, Validación de la interpretabilidad que examinan el significado y la relevancia de los clústeres para los objetivos de segmentación de datos. Utilizan criterios cualitativos y subjetivos, como el conocimiento del dominio, la lógica empresarial o la retroalimentación de los usuarios.

Para terminar, otro tema en el cual se puede profundizar en el presente trabajo sería la ingeniería de características ya que se analizaría a un nivel más profundo las relaciones que existen entre los datos para descartar características similares o adicionar nuevas. Así también para crear funciones que ayuden a disminuir la tasa de error de un modelo.

Bibliografía

- P. Anitha and M. M. Patil. Rfm model for customer purchase behavior using k-means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34:1785–1792, 5 2022. ISSN 22131248. doi: 10.1016/j.jksuci.2019.12.011.
- P. Bholowalia and A. Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105:975–8887, 2014.
- P. R. Dickson and J. L. Ginter. Market segmentation, product differentiation, and marketing strategy. *Journal of Marketing*, 51:1–10, 4 1987. ISSN 0022-2429. doi: 10.1177/002224298705100201.
- C. G. DR, K. P. K. REDDY, K. SRIKAR, and K. S. SANKAR. Customer segmentation techniques. 2022.
- P. Espinel. Procedimiento para efectuar una clasificación ascendente jerárquica de un conjunto de puntos utilizando el método de ward. *Infociencia*, 9:13–18, 2015. ISSN 1390-339X. doi: 10.24133/INFOCIENCIA.V9I1.977. URL <https://journal.espe.edu.ec/ojs/index.php/Infociencia/article/view/977>.
- C. K. Gomathy, C. K. Dr, K. Gomathy, K. Pavan, K. Reddy, K. Srikar, and S. Siva. Customer segmentation techniques e-file system view project books, articles, and posters for colleagues view project customer segmentation techniques. 2022. URL <https://www.researchgate.net/publication/360032683>.
- E. L. Guzmán. Métricas para la validación de clustering. *Elizabeth León Guzmán*, 2016.

- A. Hızıroglu. Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, 40: 6491–6507, 11 2013. ISSN 0957-4174. doi: 10.1016/J.ESWA.2013.05.052.
- M. Hosseini and M. Shabani. New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics 2015 3:3*, 3:110–121, 10 2015. ISSN 2050-3326. doi: 10.1057/JMA.2015.10. URL <https://link.springer.com/article/10.1057/jma.2015.10>.
- P. Letelier and M. C. Penadés. Metodologías ágiles para el desarrollo de software: extreme programming (xp). URL www.agileuniverse.com.
- D. Matich. Cátedra: Informática aplicada a la ingeniería de procesos-orientación i redes neuronales: Conceptos básicos y aplicaciones, 2001.
- J. Panuš, H. Jonášová, K. Kantorová, M. Doležalová, and K. Hořáčková. Customer segmentation utilization for differentiated approach. *IDT 2016 - Proceedings of the International Conference on Information and Digital Technologies 2016*, pages 227–233, 8 2016. doi: 10.1109/DT.2016.7557178.
- M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain. Rfm-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, 61, 7 2021. ISSN 09696989. doi: 10.1016/j.jretconser.2021.102566.
- J. A. Rodrigo. Redes neuronales con python, 5 2021. URL <https://cienciadedatos.net/documentos/py35-redes-neuronales-python>.
- P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7.
- W. R. Smith. Product differentiation and market segmentation as alternative marketing strategies. <https://doi.org/10.1177/002224295602100102>, 21: 3–8, 11 2018. ISSN 0022-2429. doi: 10.1177/002224295602100102. URL <https://journals.sagepub.com/doi/abs/10.1177/002224295602100102?journalCode=jmxa>.
- A. Subasi. Chapter 7 - clustering examples. In A. Subasi, editor, *Practical Machine Learning for Data Analysis Using Python*, pages 465–511. Academic Press, 2020. ISBN 978-0-12-821379-7. doi: <https://doi.org/10.1016/B978-0-12-821379-7.ch007>.

- 1016/B978-0-12-821379-7.00007-2. URL <https://www.sciencedirect.com/science/article/pii/B9780128213797000072>.
- A. M. Sundjaja. Analysis of customer segmentation in bank xyz using data mining technique. *Article in Asian Journal of Information Technology*, 2013. doi: 10.3923/ajit.2013.39.44. URL <https://www.researchgate.net/publication/256199650>.
- C. F. Tsai, Y. H. Hu, and Y. H. Lu. Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, 32:65–76, 2 2015. ISSN 1468-0394. doi: 10.1111/EXSY.12056. URL <https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.12056https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12056https://onlinelibrary.wiley.com/doi/10.1111/exsy.12056>.
- unioviedo. kmeans, 10 2022. URL https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html.
- G. Vettigli. Minisom: minimalistic and numpy-based implementation of the self organizing map, 2018. URL <https://github.com/JustGlowing/minisom/>.
- J. L. V. Villardón. Introducción al análisis de clúster. *Departamento de Estadística, Universidad de Salamanca*. 22p, 2007.
- G. Yang and X. C. Yuan. Bank customer classification model based on elman neural network optimized by pso. *2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007*, pages 5672–5675, 2007. doi: 10.1109/WICOM.2007.1390.
- D. Zakrzewska and J. Murlewski. Clustering algorithms for bank customer segmentation. *Proceedings - 5th International Conference on Intelligent Systems Design and Applications 2005, ISDA '05*, 2005:197–202, 2005. doi: 10.1109/ISDA.2005.33.
- Álvaro José García-Tejedor and A. Nogales. An open-source python library for self-organizing-maps. *Software Impacts*, 12:100280, 5 2022. ISSN 2665-9638. doi: 10.1016/J.SIMPA.2022.100280.
- Øyvind Helgesen. Customer segments based on customer account profitability. *Journal of Targeting, Measurement and Analysis for Marketing 2006 14:3*, 14:225–237, 6 2006. ISSN 1479-1862. doi: 10.1057/

PALGRAVE.JT.5740183. URL <https://link.springer.com/article/10.1057/palgrave.jt.5740183>.