



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

**Prototipo de un Sistema de Recomendación para la Oferta de  
Convenios de los Empleados de la Compañía Suramericana**

**Madelin Uribe Lujan**

Universidad Nacional de Colombia  
Facultad de Minas, Maestría en Ingeniería - Analítica  
Medellín, Colombia  
2022

# **Prototipo de un Sistema de Recomendación para la Oferta de Convenios de los Empleados de la Compañía Suramericana**

**Madelin Uribe Lujan**

Trabajo Final de Maestría presentado como requisito parcial para optar al título de:  
**Magister en Ingeniería - Analítica**

Director:  
Juan David Velásquez Henao, MSc, PhD

Universidad Nacional de Colombia  
Facultad de Minas, Maestría en Ingeniería - Analítica  
Medellín, Colombia  
2022

## **Resumen**

En este trabajo final de maestría se presenta el desarrollo de un prototipo para la oferta de convenios de los empleados de las compañías que conforman el Grupo Suramericana, el cual inició con la comprensión del proceso, la recopilación de la información, describiendo las variables usadas y a partir de estas se realizó un análisis exploratorio con las principales características demográficas y socio-demográficas seleccionadas que representan la población objeto de este trabajo; se caracterizó la población usando las técnicas de clusterización, partiendo de la construcción de la base de datos con la respectiva transformación de las variables y se construyó el modelo haciendo uso del algoritmo K-MODES. Como resultado se describen las características de cada clúster obtenido, se corrió el algoritmo Árbol de Decisión para encontrar aquellas variables que explicaban el modelo de clusterización; a continuación se evaluó para cada clúster la usabilidad de los convenios por tipología, donde se calculó su porcentaje de uso y en base a esto se definió una métrica acorde a cada clúster, se usó el algoritmo Árbol de Decisión para evaluar si una persona usaría o no un tipo de convenio; finalmente se construyó un tablero que recogiera los resultados del prototipo del sistema de recomendación, cerrando este trabajo con los resultados y conclusiones.

**Palabras clave:** Sistema recomendación, K-MODES, Árbol de decisión, Clúster, Beneficios, Gestión Humana

## **Prototype of a Recommendation System for the Offer of Agreements of the Employees of the Suramericana Company**

### **Abstract**

This final master's thesis presents the development of a prototype for the offer of agreements of the employees of the companies that make up the Grupo Suramericana, which began with the understanding of the process, the collection of information, describing the variables used and from these an exploratory analysis was carried out with the main selected demographic and socio-demographic characteristics that represent the population object of this work; The population was characterized using clustering techniques, starting from the construction of the database with the respective transformation of the variables and the model was built using the K-MODES algorithm. As a result, the characteristics of each

cluster obtained are described, the Decision Tree algorithm was run to find those variables that explained the clustering model; Next, the usability of the agreements by typology was evaluated for each cluster, where their percentage of use was calculated and based on this, a metric was defined according to each cluster, the Decision Tree algorithm was used to evaluate whether a person would use or not a type of agreement; Finally, a board was built that collected the results of the recommendation system prototype, closing this work with the results and conclusions.

**Keywords:** recommendation system, K-MODES, Decision Tree, cluster, benefits, human management

## Tabla de Contenido

1	Introducción .....	8
1.1	Definición del Problema Real.....	9
1.2	Definición del Problema de Analítica .....	10
1.3	Estado del Arte / Definición de Conceptos Básicos.....	11
1.4	Información Disponible .....	12
1.5	Modelos o Metodología a Utilizar para Desarrollar el Modelo .....	12
1.6	Hipótesis.....	13
1.7	Objetivos .....	13
1.7.1	Objetivo General.....	13
1.7.2	Objetivos Específicos .....	13
1.8	Mapa del Documento.....	14
2	Recopilación de la Información .....	15
3	Análisis Exploratorio .....	17
4	Caracterización de los Empleados.....	22
4.1	Construcción de la Base de Datos para el Modelo de Clusterización .....	22
4.2	Construcción del Modelo .....	22
4.3	Variables que Explican el Modelo de Clusterización .....	28
5	Análisis de la Utilización de los Convenios y Metas Propuestas .....	31
5.1	Análisis del Uso de los Convenios .....	31
5.2	Definición de las Metas de Uso.....	32
5.3	Pronóstico de Uso de un Tipo de Convenio .....	34
6	Dashboard de Resultados .....	36
7	Conclusiones .....	39
7.1	Cumplimiento de Objetivos .....	39
7.2	Principales Conclusiones y Trabajo Futuro .....	39
8	Bibliografía.....	41

## Índice de Figuras

Figura 1. Total de empleados por estado civil y rango de edad. ....	17
Figura 2. Porcentaje de empleados que indican tener hijos y mascotas. ....	18
Figura 3. Total de empleados por estrato de vivienda y medio de transporte.....	19
Figura 4. Total de empleados por actividades artísticas y culturales que asisten y deporte que practican. ....	20
Figura 5. Total de empleados por en lo que les interesa recibir formación y que esperan que la compañía les ofrezca. ....	21
Figura 6. Total de empleados por rango de edad en los diferentes clústers.....	23
Figura 7. Total de empleados por estado civil en los diferentes clústers. ....	24
Figura 8. Total de empleados por estrato socioeconómico en los diferentes clústers. ....	25
Figura 9. Total de empleados por medio de transporte usado para desplazamientos casa – trabajo en los diferentes clústers.....	25
Figura 10. Total de empleados por los temas en los que le interesa recibir formación en los diferentes clústers. ....	26
Figura 11. Total de empleados por lo que esperan que la compañía les ofrezca en los diferentes clústers.....	26
Figura 12. Total de empleados por actividades artísticas y culturales que realizan en los diferentes clústers.....	27
Figura 13. Total de empleados por deportes que practican en los diferentes clústers.....	27
Figura 14. Árbol de Decisión.....	30
Figura 15. Árbol de Decisión tipo texto. ....	30
Figura 16. comparación usabilidad de los convenios por tipología y clúster.....	32
Figura 17. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Cambios y Permanencias. ....	33
Figura 18. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Millennials. ....	33
Figura 19. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Familias Nucleares. ....	34
Figura 20. Dashboard, análisis exploratorio de las variables.....	36
Figura 21. Dashboard, caracterización variables demográficas. ....	37
Figura 22. Dashboard, caracterización de las variables socio-demográficas. ....	37
Figura 23. Dashboard, usabilidad por tipo de convenio y meta propuesta. ....	38

## Índice Tablas

Tabla 1. Variables que componen el conjunto de datos.....	15
Tabla 2. Total y porcentaje de empleados que tienen hijos en los diferentes clústers.....	24
Tabla 3. Total y porcentaje de empleados que tienen mascotas en los diferentes clústers. .....	24
Tabla 4. Variables que explican el modelo de clusterización.....	28
Tabla 5. Resultados de los modelos .....	35

## 1 Introducción

En este documento se propone el desarrollo de un prototipo para la oferta de convenios que hace parte de la categoría Compensación Motivacional del Plan de Beneficios, plan diseñado por la Dirección de Compensación y Beneficios de la Vicepresidencia de Talento Humano, para los empleados de las compañías que conforman el Grupo Suramericana y sus familias en cada una de las regiones de Colombia.

Las relaciones comerciales que tiene establecidas Suramericana con las diferentes empresas, busca brindar al empleado productos y servicios con tarifas especiales de acuerdo a la región en la que se encuentran ubicados; convenios que se han habilitado pensando en el bienestar de cada uno; sin embargo, la compañía no está logrando llegar a los diferentes públicos que espera, se percibe que no todos los colaboradores conocen este tipo de beneficios, ya que la oferta que se está entregando actualmente no los está impactando lo suficiente, debido a que actualmente cada empleado debe entrar al sitio de beneficios y buscar que servicios pueden ajustarse a su necesidad actual. Desde el proceso no se tiene conocimiento del nivel de satisfacción que tienen las personas respecto a los beneficios que se les ofrece en la actualidad a través de estos convenios.

Es necesario contar con un sistema de recomendación donde se pueda diseñar sugerencias personalizadas de los convenios, sistema que filtre la información de cada uno de estos servicios que la compañía dispone, teniendo en cuenta variables demográficas, generación de grupos afines y preferencias identificadas en cada uno de ellos.

El prototipo del sistema de recomendación, le permitirá a la compañía tener opciones para presentar de forma personalizada a los empleados los convenios a los cuales ellos pueden acceder, mediante recomendaciones que se ajusten a las características y particularidades actuales de cada uno de los empleados; este tipo de recomendaciones permitirá aumentar la usabilidad, la oportunidad en la comunicación de los beneficios que se tienen capturando a su vez información de la percepción del empleado frente a los convenios brindados.

El objetivo de este documento es presentar una propuesta para desarrollar un Prototipo de un Sistema de Recomendación para la Oferta de Convenios de los Empleados de las compañías que conforman el Grupo Suramericana.

Para desarrollar esta propuesta se utilizará la Metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), un modelo flexible que está compuesto por 6 fases que necesariamente no deben seguir una misma secuencia, esta puede variar y ajustarse de acuerdo con la necesidad del proyecto que se esté desarrollando. Para conocer cada una de las fases a continuación se dará un contexto general de cada una de ellas.

La fase Comprensión del Negocio incorpora la compilación de la información de la empresa, se determinan los objetivos comerciales y de minería de datos y se evalúa la situación teniendo en cuenta principalmente los riesgos, costos y beneficios para la generación del plan del proyecto evaluando técnicas y herramientas a usar; en la fase de Comprensión de los Datos, se recopilan los datos iniciales para una posterior descripción y exploración de los mismos verificando su calidad; la fase Preparación de Datos es donde se realiza

principalmente el proceso técnico de selección y limpieza de los datos con el fin de definir los conjuntos de datos a utilizar minimizando los errores existentes y a su vez ampliar el número de atributos logrando integración entre ellos; otra de las fases es Modelado donde se realiza la Selección de Técnicas de modelado, generación y evaluación de los modelos; a continuación se encuentra la fase de Evaluación donde se realiza todo el proceso de revisión y evaluación de resultados determinando los pasos a seguir y finalmente esta la fase de Distribución donde se aplica el modelo, se planifica su control y mantenimiento , cerrando con la creación y presentación del informe final.

Una gran ventaja que se puede obtener de esta metodología es que en cada una de las fases se pueden generar preguntas claves que permiten evaluar si el proceso realizado se hizo de manera correcta y así pasar a la siguiente fase teniendo claridad y seguridad de que las actividades ejecutadas se realizaron correctamente.

Este trabajo está organizado de la siguiente forma: en la Sección 1.1 se presenta la definición del problema real en la que se explica la situación actual del sistema de beneficios, especialmente la oferta de convenios que actualmente se encuentra disponible para los empleados; en la Sección 1.2 se define el problema de analítica ; en la Sección 1.3 se desarrolla el estado del arte donde se definen conceptos y se presentan modelos similares que han sido desarrollados con éxito, presentando que modelo fue escogido y sus respectivas conclusiones; en la Sección 1.4 se indica la información disponible para el análisis y cómo será presentada; en la Sección 1.5 se expone la metodología a utilizar y que se abordará en cada una de las fases dispuestas; la Sección 1.6 presenta la hipótesis; en la Sección 1.7 se plantea el objetivo general y los objetivos específicos que se desarrollaron en este trabajo y en la Sección 1.8 se indica cómo se desarrolló el presente trabajo.

## 1.1 Definición del Problema Real

La compañía Suramericana cuenta con un amplio portafolio de beneficios para el bienestar de los empleados, el cual se encuentra definido dentro de cuatro categorías que son:

- Nuestro compromiso contigo: la cual comprende los pagos a los que tiene derecho el empleado;
- Beneficios y prestaciones más allá del salario: aquí se podrá encontrar los auxilios económicos y prestaciones extralegales con las que cuenta el empleado por la compañía que ha sido contratado;
- Los resultados los alcanzamos juntos: es la compensación variable que depende de los resultados de la compañía; y finalmente está la categoría
- Bienestar más allá de lo económico: donde están las diferentes formas de compensación que recibe el colaborador, las cuales contribuyen a su bienestar y con ello a la calidad de vida de las personas que lo rodean, dentro de esta última categoría también llamada “Compensación Motivacional” se encuentran los convenios empresariales que son acuerdos comerciales que permiten que los

colaboradores y sus familias puedan acceder a productos y servicios con tarifas especiales.

Estos convenios se encuentran a nivel nacional con más de 100 empresas y fueron definidos pensando en la necesidad de los empleados, pero estos beneficios no están llegando a los diferentes públicos que se espera, su usabilidad es muy baja debido a que muchos de los empleados no se enteran o la forma en que son ofrecidos no es la ideal.

La información de los convenios esta publicada en el sitio de beneficios, clasificados por ubicación y categoría, pero estos filtros no son suficientes para delimitar e identificar la necesidad actual de cada empleado y que cada uno pueda darse cuenta del producto o servicio que dispone la compañía para su bienestar y el de su familia, de manera que estos puedan ofrecerse en el momento de vida de cada uno de ellos, es decir que de acuerdo a la etapa de vida en la que se encuentren puedan enterarse de los beneficios a los que pueden acceder sin tener que ir a buscar cada uno de acuerdo a su preferencia. Lo anterior permite preguntar, ¿qué tan oportuna está siendo la comunicación de los beneficios?, ¿qué tan satisfechos están los empleados con la oferta de convenios actual? Y ¿Cómo llegar a cada empleado de manera personalizada?

Muchas veces las organizaciones se quedan cortas a la hora de brindarle y asegurarle al empleado un sistema de incentivos que le permita ver sus objetivos individuales y que estos estén alineados con los de la compañía, por lo que se vuelve necesario buscar nuevas alternativas para conocer las necesidades de cada persona y poder llegar a cada una de ellas.

Si existiera un sistema de recomendación que identificara la necesidad actual y las preferencias de cada empleado, de acuerdo a sus costumbres, núcleo familiar y demás características, podría llegarse a cada uno de ellos con una oferta personalizada que de interés a su necesidad actual, permitiendo a su vez capturar información de lo que cada uno espera que la compañía le ofreciera.

## 1.2 Definición del Problema de Analítica

El prototipo del sistema de recomendación de la oferta de convenios, le permitirá a la compañía presentar opciones a los empleados de los productos y servicios a los cuales pueden acceder con tarifas preferenciales, que se ajusten a las características y particularidades actuales de cada uno; esta recomendación personalizada permitirá, mejorar la comunicación de los beneficios a los que pueden acceder, aumentar la usabilidad ya que podrán enterarse oportunamente y a su vez se podrá evaluar la usabilidad de los empleados frente a los convenios brindados.

### 1.3 Estado del Arte / Definición de Conceptos Básicos

En problemas de selección y recomendación de alternativas basándose en información disponible a continuación se enuncian varios trabajos e investigaciones basados en este tipo de problemas:

En [5], se evalúan varias técnicas para mejorar el proceso de los estudiantes a la hora de escoger las mejores asignaturas para ellos en el proceso de matrícula, lo cual se vuelve una tarea compleja; en este trabajo se ha descubierto que las técnicas de Data Mining como las basadas en reglas y las basadas en árboles no muestran adecuadamente la relevancia que los atributos de los docentes y las asignaturas tienen a la hora de predecir el rendimiento académico. Se hace uso de modelos colaborativos, basados en contenido y realizando también un enfoque híbrido, buscando el enfoque que garantice eficiencia y precisión, a lo que se concluye que las recomendaciones obtenidas mediante el enfoque colaborativo reflejan adecuadamente la relevancia de los atributos anteriormente mencionados con una tasa de precisión del 84%.

En [6], se elabora un sistema basado en filtrado colaborativo que busca dar solución a la pérdida de tiempo a la hora de buscar objetos de aprendizaje en el gran volumen de información que se tiene, en el desarrollo de este trabajo se utilizó una adaptación de los K-Vecinos con fundamentado en la percepción de usabilidad y utilidad que tiene el usuario sobre los objetos de aprendizaje, siguiendo las siguientes fases:

- Obtener los objetos votados por el usuario
- Definir vecindario
- Definir correlación del vecindario
- Delimitar el vecindario
- Definir las recomendaciones
- Definir los pesos de las recomendaciones
- Seleccionar las n recomendaciones con el mayor peso

Se concluye que el SR propuesto permite formular recomendaciones de tipo top-N, en el cual los métodos basados en la vecindad se pueden utilizar de acuerdo con los valores de similitud. El SR está basado en el filtrado colaborativo y ha demostrado que es fácil de implementar, el cual podría servir como referencia y modelo para otros repositorios cuyos fines sean similares.

En [7], se buscó determinar cuáles son los productos que generan mayor interés en los usuarios, con el fin de que estaban experimentando una pérdida en la participación del mercado frente a la competencia constante. En la elaboración del SR se evaluaron métodos basados en popularidad, contenido y finalmente un híbrido (Factorización matricial con contenido); se usaron las métricas de AUC y Rango recíproco medio (MRR), para determinar el modelo que otorgara mejores rankings. Se concluye que, entre los Sistemas de Recomendación planteados, sobresalió el modelo híbrido con contenidos en cuanto a su desempeño respecto a los métodos basados en memoria o contenido.

En [8], se desarrolla un Sistema de Recomendación Híbrido para mejorar la retención de los clientes actuales, buscando incrementar su satisfacción; en este trabajo se utiliza la información disponible de cada cliente en cuanto a datos demográficos, transaccionales, hábitos de consumo y comportamiento con productos, con el fin de entrenar modelos de Inteligencia Artificial como el sistema de recomendación por medio de ratings para predecir el comportamiento de los clientes, y un modelo de cruce de productos para predecir la probabilidad de los clientes actuales a aceptar ofertas sobre los productos que aún no tiene. Finalmente, se compara el desempeño de metodologías de ensamblaje de árboles de decisión como son: Random Forest, Gradient Boosting Models (GBM). Al final, los modelos desarrollados demuestran una gran capacidad predictiva en cuanto al comportamiento de los clientes y probabilidad de cruce de los productos del banco.

#### 1.4 Información Disponible

La información de los empleados puede ser extraída de los sistemas de información donde son gestionados los procesos de Talento Humano, que son: SuccessFactors, Base de Datos Nómina (Nomus), Teradata; otro medio para obtener información son las encuestas realizadas donde los resultados se encuentran en archivos de Excel. La información es reservada, no se van a colocar datos que sean sensibles y los datos de los empleados no quedarán a nombre propio por lo que su identificación será encriptada.

#### 1.5 Modelos o Metodología a Utilizar para Desarrollar el Modelo

Para el desarrollo del prototipo se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual proporciona una descripción del ciclo de vida de un proyecto estándar de análisis de datos, cubriendo las fases del proyecto, las tareas y relaciones entre ellas; las fases que componen la metodología son 6 las cuales se enuncian a continuación: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación y Distribución.

Para el desarrollo del prototipo no será necesario partir de la comprensión del negocio ya que de acuerdo a la experiencia en la compañía y el proceso dentro del cual se va desarrollar, se tiene un amplio conocimiento de él, por lo que se iniciará con la comprensión de los datos donde se hará una recolección y un estudio para familiarizarse con los mismos, generar información relevante y descubrir las primeras señales dentro de ellos. En la preparación de los datos se hará todo el proceso de limpieza y transformación para la posterior selección de las características y la construcción del conjunto final de los datos con los que se trabajará el modelado; en la fase de modelado se procederá con la definición y aplicación de los modelos que se consideren más pertinentes para el prototipo; pasando a la fase de evaluación donde se tienen los modelos con los mejores valores y de los que se realizará una evaluación profunda con el fin de revisar si están cumpliendo con los objetivos de negocio y la minería de datos; para finalizar está la etapa de despliegue que en este caso será la presentación del informe del prototipo de la oferta de convenios.

Se utilizó el lenguaje Python para realizar el análisis exploratorio de la información, definición y evaluación del modelo seleccionado para el prototipo del sistema de recomendación.

## 1.6 Hipótesis

Es posible desarrollar un prototipo de un sistema de recomendación utilizando la metodología tal que permita ofertar las diferentes opciones de los convenios a los empleados de las diferentes compañías Suramericana, con el fin de mejorar la comunicación de los beneficios de manera que sean visibles a cada uno y así se logre aumentar el uso y conocimiento de los mismos; se puede contar con la generación de ofertas que estén definidas bajo la caracterización demográfica y sociodemográfica que desde el área Talento Humano se le realiza a los colaboradores de acuerdo a los procesos que de cara al Talento se desarrollan.

También es posible lograr percibir a partir del uso, si la oferta actual de convenios contiene los beneficios esperados por los empleados, evaluar si las alianzas actuales con las entidades que se tienen estos convenios están generando el valor y bienestar que como compañía se pretende brindar con este tipo de recursos.

## 1.7 Objetivos

### 1.7.1 Objetivo General

Desarrollar un prototipo de un sistema de recomendación utilizando la metodología de desarrollo de proyectos de análisis de datos CRISP-DM, tal que permita la presentación de la oferta de convenios a los empleados del Grupo Suramericana, llegando a cada uno de ellos de forma personalizada y entregándoles alternativas que vayan de la mano con su interés actual.

### 1.7.2 Objetivos Específicos

Los objetivos específicos de esta propuesta son los siguientes:

- Recopilar la información de los empleados a través de los sistemas de gestión de información y encuestas aplicadas por el área para la evaluación de las características a trabajar.
- Realizar el análisis exploratorio de la información disponible con el fin de entender el problema planteado.
- Caracterizar mediante técnicas de Clustering la población de los empleados de las compañías del Grupo Suramericana.
- Analizar para cada tipo de clúster la utilización de los convenios actuales.
- Proponer metas de utilización de los convenios de acuerdo con el clúster.

## 1.8 Mapa del Documento

El resto de este trabajo está organizado como sigue: En el Capítulo 2 se presenta el proceso de recopilación de la información, las diferentes fuentes de donde se extrajo la información y la descripción de las variables usadas; en el Capítulo 3 se realizó el análisis exploratorio donde se describió los empleados a partir de las principales características demográficas y socio-demográficas seleccionadas; en el Capítulo 4 se caracteriza los empleados usando las técnicas de clusterización, partiendo de la construcción de la base de datos con la respectiva transformación de las variables y se construyó el modelo haciendo uso del algoritmo K-MODES. Como resultado se describen las características de cada clúster obtenido, se corrió el algoritmo Árbol de Decisión para encontrar aquellas variables que explicaban el modelo de clusterización; en el Capítulo 5 se evaluó para cada clúster la usabilidad de los convenios por tipología, donde se calculó su porcentaje de uso y en base a esto se definió una métrica acorde a cada clúster, se usó el algoritmo Árbol de Decisión para evaluar si una persona usaría o no un tipo de convenio; finalmente en el Capítulo 6 se construyó un tablero que recogiera los resultados del prototipo del sistema de recomendación, cerrando este proyecto en el Capítulo 7 con los resultados, conclusiones y trabajo futuro.

## 2 Recopilación de la Información

Las bases de datos utilizadas en este proyecto fueron extraídas de tres fuentes de información que almacenan el resultado de la gestión de los procesos de Talento Humano en la organización, los cuales contienen la información demográfica y socio-demográfica de los empleados. Estos sistemas de información son:

- **Teradata:** Es un repositorio de datos donde es almacenada la información proveniente de los sistemas transaccionales, allí se encuentran almacenados todos los datos demográficos y laborales de los empleados.
- **Cosmos DB:** Servicio de base de datos NoSQL donde es almacenada la información socio-demográfica, que es alimentada por los empleados a través de una encuesta realizada desde la herramienta Qualtrics.
- **Reportes proveedor:** reportes en Excel enviados por cada proveedor con el que se tiene convenio, los cuales contienen la usabilidad de cada convenio.

La base de datos que se recopiló para este trabajo consta de 11.130 registros y 30 variables demográficas y socio-demográficas del año 2021, las cuales se listan en la Tabla 1.

Tabla 1. Variables que componen el conjunto de datos.

Nombre Variable	Categorías
Clase Vinculación	Aprendiz, Empleado, Asesor Dependiente
Regional	Antioquia, Eje cafetero, Norte, Centro, Occidente
Compañía	Ayudas Diagnósticas, CGR, EPS, IPS, Seguros de Vida, Seguros Generales, Otra
Nivel cargo	Aprendiz, Practicante, Auxiliar, Analista, Asesor, Coordinador, Gerente, Vicepresidente
Tipo contrato	Fijo, Indefinido, Practicante, Relación Aprendizaje
Sexo	Femenino, Masculino
Discapacidad	Si, No
Alguna discapacidad grupo familiar	Si, No
Tiene Hijos	Si, No
Planea Hijos	Si, No, Tal vez
Planea más hijos	Si, No, Tal vez
Requieres llevar hijo cuiden	Si, No
Tiene mascotas	Si, No
Zona Vivienda	Rural, Urbana
Tipo vivienda	Apartamento, Casa, Pensión/Residencia
Tipo pago vivienda	Cuota/arriendo total, Nada, Parcial
Estrato vivienda	1,2,3,4,5,6

Trabajas desde casa	Si, No
Estado civil	Soltero, Casado, Divorciado, Viudo
Idioma nativo	Español, Frances, Inglés, Italiano, Portugués, Otro
Medio transporte casa trabajo	Bicicleta, Bus/Colectivo, Caminando, Carro, Moto, Sistema integrado de transporte público (Metro, Transmilenio, Mio, Tranvía, etc), Taxi/Plataformas colaborativas de transporte
Situaciones que te afectan	Desempleo, Deudas que superan los ingresos, Enfermedad de algún familiar, Muerte reciente de seres cercanos, Problemas de pareja, Ninguna, Otro
Edad	Edad de cada uno de los empleados
Qué actividades sociales te apasionan	Cuidado de los animales, cuidado del medio ambiente, Voluntariado, Otra, Ninguna
Actividades artísticas y culturales	Bailar, Canto, Cine, Escuchar música, Fotografía, Gastronomía, Leer/Club lectura, Manualidades y artesanías, Moda - maquillaje
Deporte que practica	Baile, Caminar, Ciclismo, Fútbol, Gimnasio, Natación, Trotar, Otro
Frecuencia con la que practica el deporte	1 o 2 veces al mes, 1 vez por semana, 2 veces por semana, 3 o más veces por semana, Todos los días, Ninguna
Recibir formación	Cuidado del medio ambiente, Cuidado parental / crianza de hijos, Emprendimiento, Enfermedades crónicas, Estilos de vida saludable, Expresión oral y corporal, Idiomas, Manejo de conflictos, Manejo del duelo, Relaciones familiares, Salud financiera, Salud mental, Otro
Esperan de la Compañía	Actividades deportivas y recreativas, Beneficios para ti y tu familia, Compensación económica, Formación académica, Programas de idiomas, Préstamos de vivienda, Tiempo libre, Otro
Reconocimientos	Bonos para actividades de recreación, cultura y deporte, Compensación económica, Participar en iniciativas o proyectos, Pasantías en otros países, Patrocinios para el desarrollo personal, Tiempo libre, Viajes

### 3 Análisis Exploratorio

Para el análisis exploratorio fueron combinadas las diferentes fuentes de información a partir del identificador único definido en cada una de ellas, obteniendo como resultado un único dataframe con la totalidad de las variables. La información recolectada permitió describir los empleados a partir de las principales características demográficas y socio-demográficas las cuales son:

- Estado civil
- Rango de edad
- Tiene hijos
- Tiene mascotas
- Estrato de vivienda
- Medio de Transporte
- En que desean recibir formación
- Que esperan de la compañía
- Actividades artísticas y culturales
- Deporte que practican

En la Figura 1 se observa la distribución de los empleados de acuerdo con su estado civil y rango de edad; la mayoría son solteros, seguido de los casados y también se tiene una alta proporción de los que viven en unión libre.

En cambio, para la variable rango de edad, la población está concentrada entre los 26 y 51 años edad, teniendo un mayor número de empleados entre los 36 y 51 años; se observó también que más del 50% de los empleados ya se encuentran en edad superior a los 26 años.

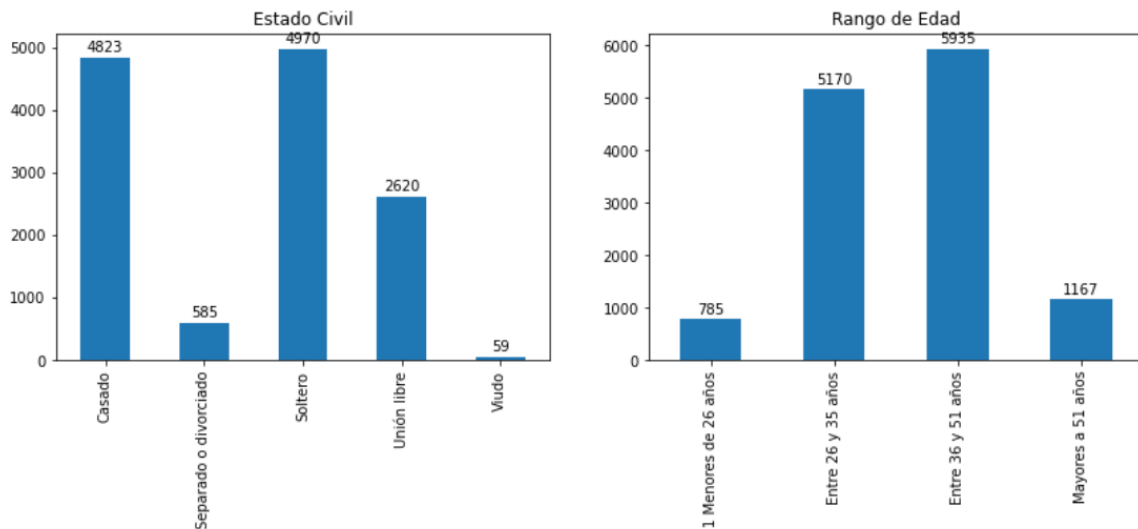


Figura 1. Total de empleados por estado civil y rango de edad.

En la Figura 2, se encuentran las variables que indican si tiene hijos o si tienen mascotas, las cuales tienen distribuciones muy similares ya que cerca del 50% responden no en ambas características.



Figura 2. Porcentaje de empleados que indican tener hijos y mascotas.

Para la variable estrato de vivienda (Figura3) se encontró que los empleados viven principalmente en los estratos 2 ,3 y 4, con una mayor concentración en el estrato 3. Cerca del 50% utilizan transporte propio para el desplazamiento a su lugar de trabajo. Aunque la mayoría indicó caminar como uno de los deportes que practican (ver Figura 4) una proporción muy pequeña indica que se desplaza caminando a su lugar de trabajo. De forma muy similar, se observa que de los empleados que practican ciclismo muy pocos se desplazan en este medio de transporte.

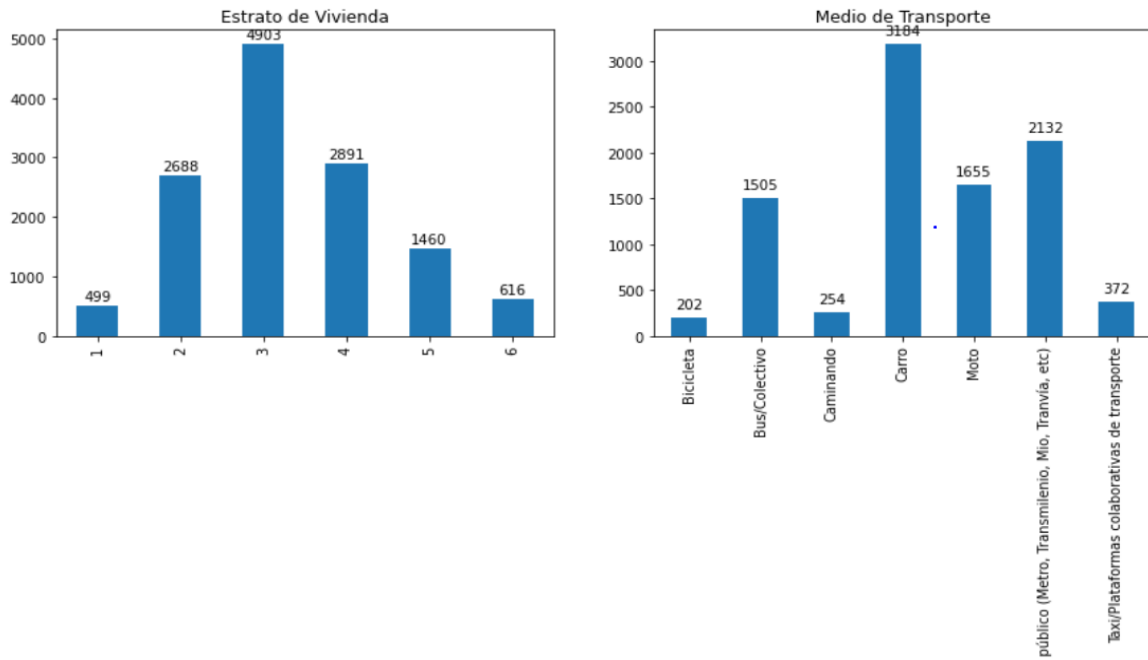


Figura 3. Total de empleados por estrato de vivienda y medio de transporte.

Según la Figura 4, en las Actividades Artísticas y Culturales que los empleados realizan, se observó una inclinación por los viajes, seguido del baile, encontrando proporciones similares para las demás actividades seleccionadas.

Para la variable Actividades Deportivas se encontró que a parte de la caminata muchos prefieren el gimnasio.

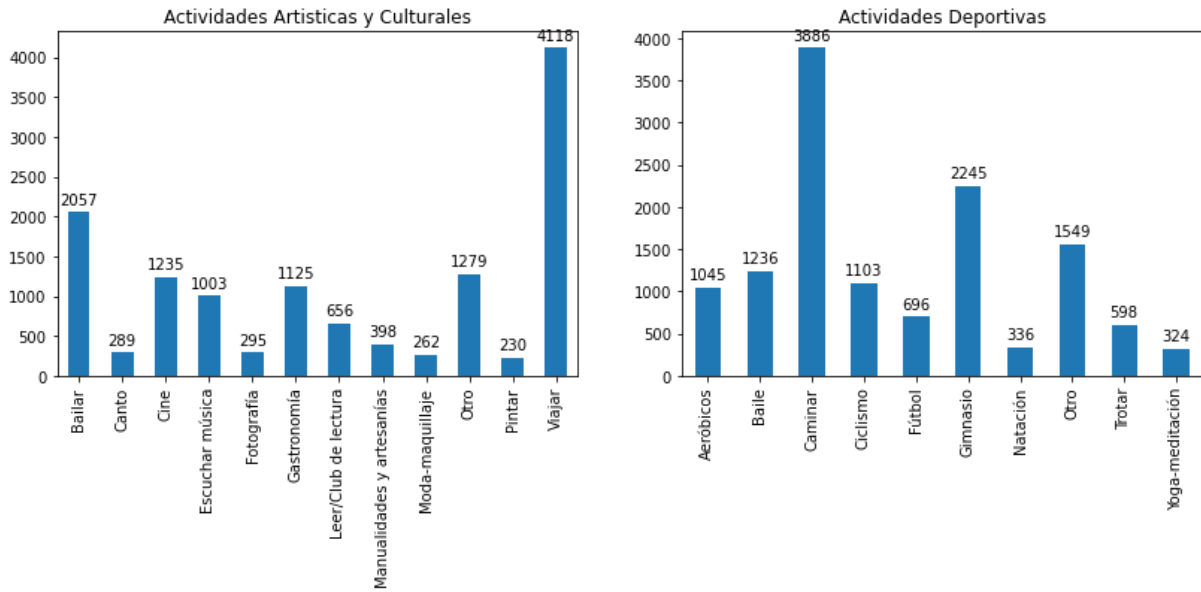


Figura 4. Total de empleados por actividades artísticas y culturales que asisten y deporte que practican.

Finalmente, en la Figura 5, se observó que los empleados prefieren recibir formación en los temas que se enuncian a continuación en orden de mayor a menor frecuencia:

- Emprendimiento.
- Estilos de vida saludable.
- Salud financiera.
- Idiomas.

Para la variable Esperan de la Compañía se encontró una inclinación por los programas educativos.

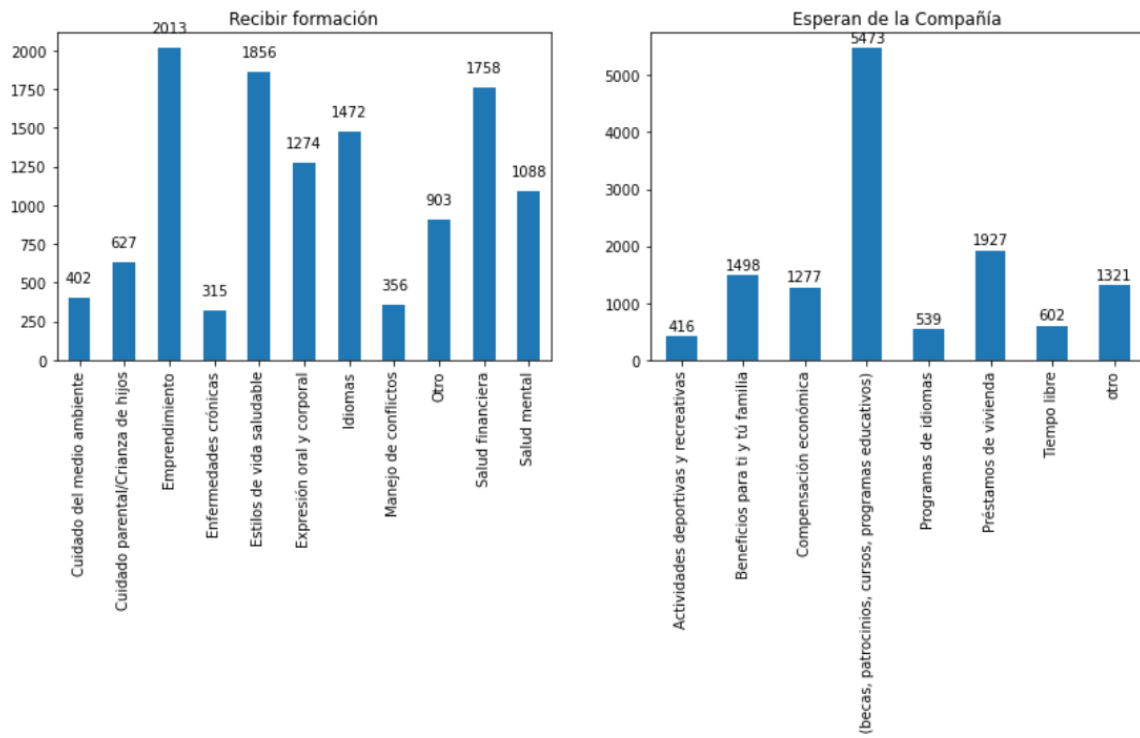


Figura 5. Total de empleados por en lo que les interesa recibir formación y que esperan que la compañía les ofrezca.

## 4 Caracterización de los Empleados

### 4.1 Construcción de la Base de Datos para el Modelo de Clusterización

En el proceso de construcción de los datos utilizados para la segmentación de los empleados, se encontró que la data recopilada está compuesta principalmente por variables categóricas, las cuales corresponden al 97 % del total de variables; estas fueron transformadas a valores numéricos utilizando la técnica one-hot-encoder. El otro 3% son variables continuas las cuales se transformaron a variables categóricas, agrupando los datos en rangos categóricos, obteniendo así un dataframe compuesto 100% de variables numéricas.

### 4.2 Construcción del Modelo

Para el proceso de clusterización se utilizó el algoritmo K-MODES, comúnmente utilizado para trabajar con el tipo de variables recopiladas en este trabajo. Al correr el algoritmo de clusterización se obtuvo que el número recomendado de clústers es 3, por lo que se construyó el modelo con este resultado. Se analizaron variables como:

- Rango de edad
- Estado civil
- Estrato socioeconómico
- Tiene hijos
- Tiene Mascotas
- Medio de transporte
- Intereses en formación
- Actividades artísticas y culturales
- Oferta de la compañía

En las figuras 6 a la 13, se comparan las características para los diferentes clústers obtenidos. Su análisis permitió encontrar que:

- El Clúster Cambios y Permanencias, contiene una proporción similar de empleados en los rangos de edad entre 26 y 51 años; la mayoría son solteros y el 65% tienen hijos. Viven principalmente en los estratos 2 y 3. El 69% tienen mascotas en sus hogares. Los medios de transporte público y moto son los más usados para desplazarse a su lugar de trabajo. Prefieren recibir formación en temas de emprendimiento y en estilos de vida saludable; esperan de la compañía principalmente formación y préstamos de vivienda. Según la descripción de este clúster se podría concluir que se compone principalmente de aquellos empleados que son padres solteros, quienes aparte de interesarse por actividades que propicien espacios en familia, buscan un crecimiento económico y profesional.
- En el Clúster de los Millennials, la mayoría se encuentran en el rango de edad de los 26 y 35 años. El estado civil principal es soltero. Solo el 11% de ellos tiene hijos. Viven principalmente en los estratos socioeconómicos 3 y 4. Menos del 50% tiene mascotas. La mayoría tienen medio transporte propio y prefieren recibir formación

en temas de salud financiera y emprendimiento. Esperan que la compañía les ofrezca formación y préstamos de vivienda. Este clúster se puede considerar que lo conforman aquellos empleados más jóvenes, con mayor interés en el crecimiento económico, formación profesional y viajar, más que en constituir un hogar con hijos.

- En el Clúster Familias Nucleares están los empleados en el rango de edad de los 36 y 51 años. La mayoría son casados. El 87% tiene hijos. Viven principalmente en los estratos socioeconómicos 3 y 4. Cerca del 50% tienen mascotas. Una alta proporción posee transporte propio para sus desplazamientos al lugar de trabajo. Prefieren recibir formación en temas como emprendimiento y estilos de vida saludable y esperan que la compañía les ofrezca formación y beneficios para ellos y sus familias. Según lo descrito se podría decir que en este clúster se encuentran aquellos empleados con mayor estabilidad económica, hogares tradicionales compuestos por esposos e hijos, y sus preferencias se enfocan en propiciar espacios familiares.

Aunque se encontró un comportamiento similar en los tres clústers para las características, actividades artísticas y culturales, actividades deportivas; otras como estado civil, rango de edad, estrato de vivienda, muestran una segmentación clara de la población.

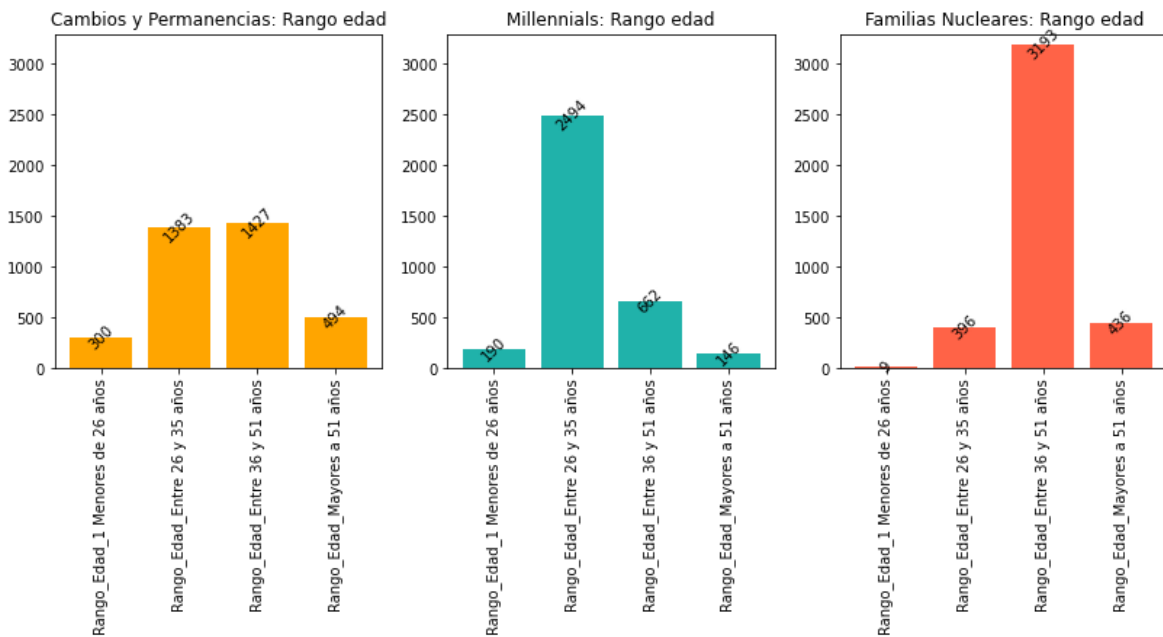


Figura 6. Total de empleados por rango de edad en los diferentes clústers.

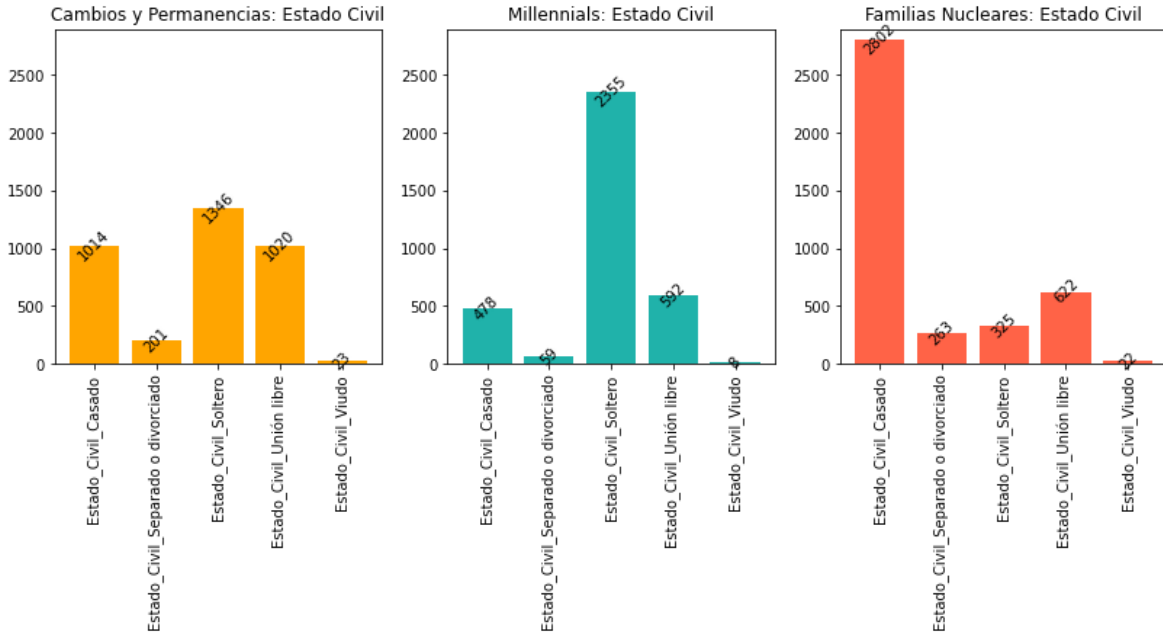


Figura 7. Total de empleados por estado civil en los diferentes clústers.

Tabla 2. Total y porcentaje de empleados que tienen hijos en los diferentes clústers.

Cambios y Permanencias			Millennials			Familias Nucleares		
Característica	Total	%	Característica	Total	%	Característica	Total	%
Tiene_Hijos	2356	65.37%	Tiene_Hijos	416	11.91%	Tiene_Hijos	3538	87.70%

Tabla 3. Total y porcentaje de empleados que tienen mascotas en los diferentes clústers.

Cambios y Permanencias			Millennials			Familias Nucleares		
Característica	Total	%	Característica	Total	%	Característica	Total	%
Tiene_Mascotas	2475	68.7%	Tiene_Mascotas	1575	45.1%	Tiene_Mascotas	1883	46.7%

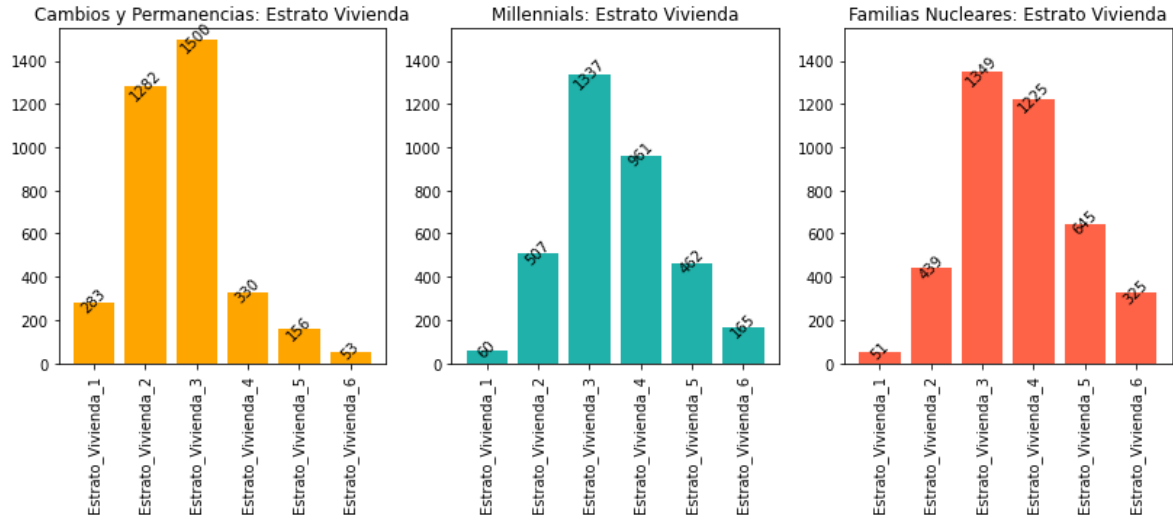


Figura 8. Total de empleados por estrato socioeconómico en los diferentes clústers.

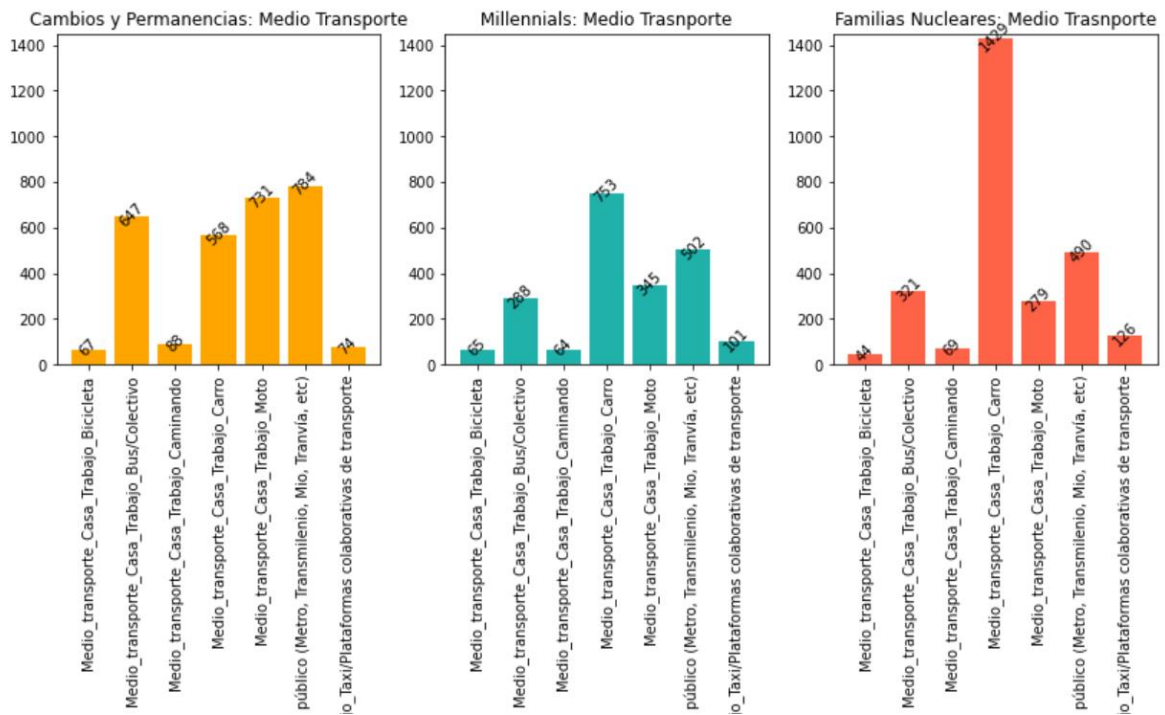


Figura 9. Total de empleados por medio de transporte usado para desplazamientos casa – trabajo en los diferentes clústers.

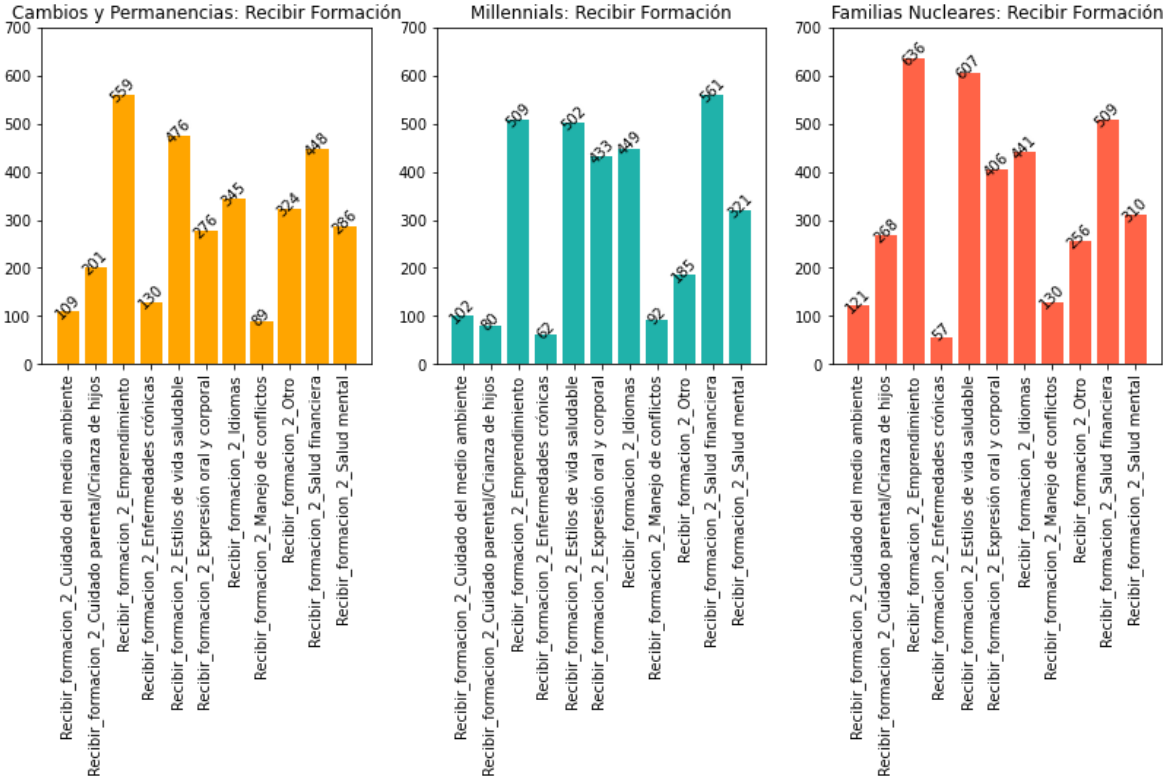


Figura 10. Total de empleados por los temas en los que le interesa recibir formación en los diferentes clústers.

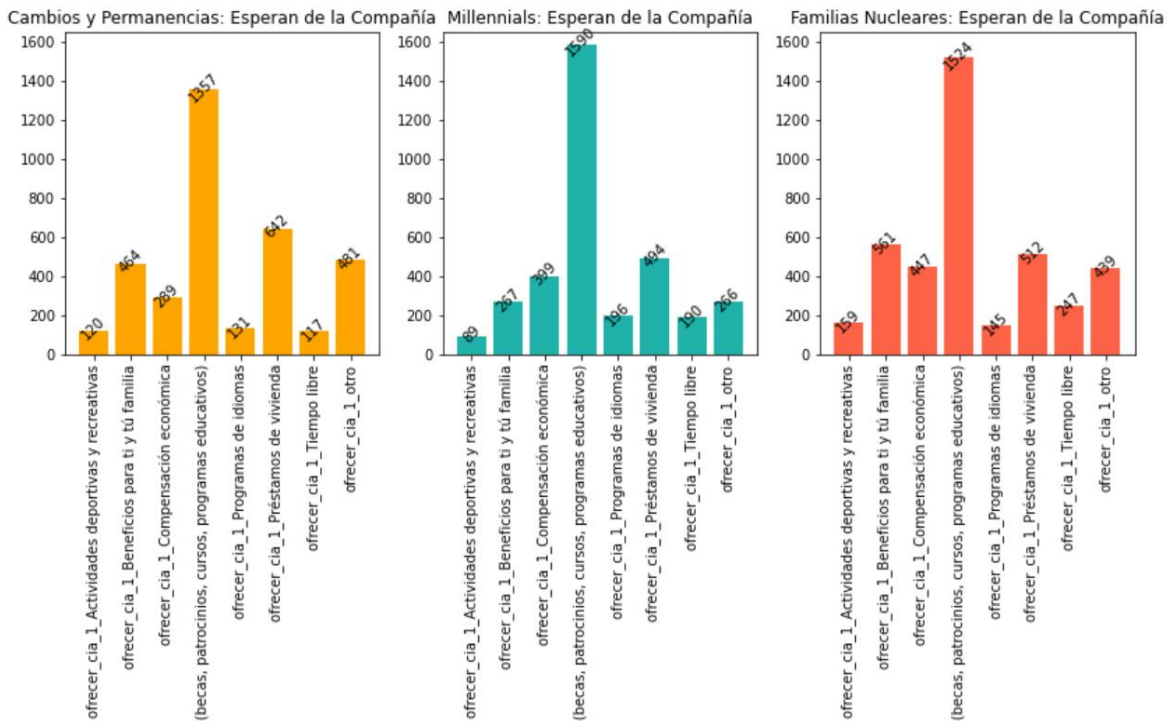


Figura 11. Total de empleados por lo que esperan que la compañía les ofrezca en los diferentes clústers.

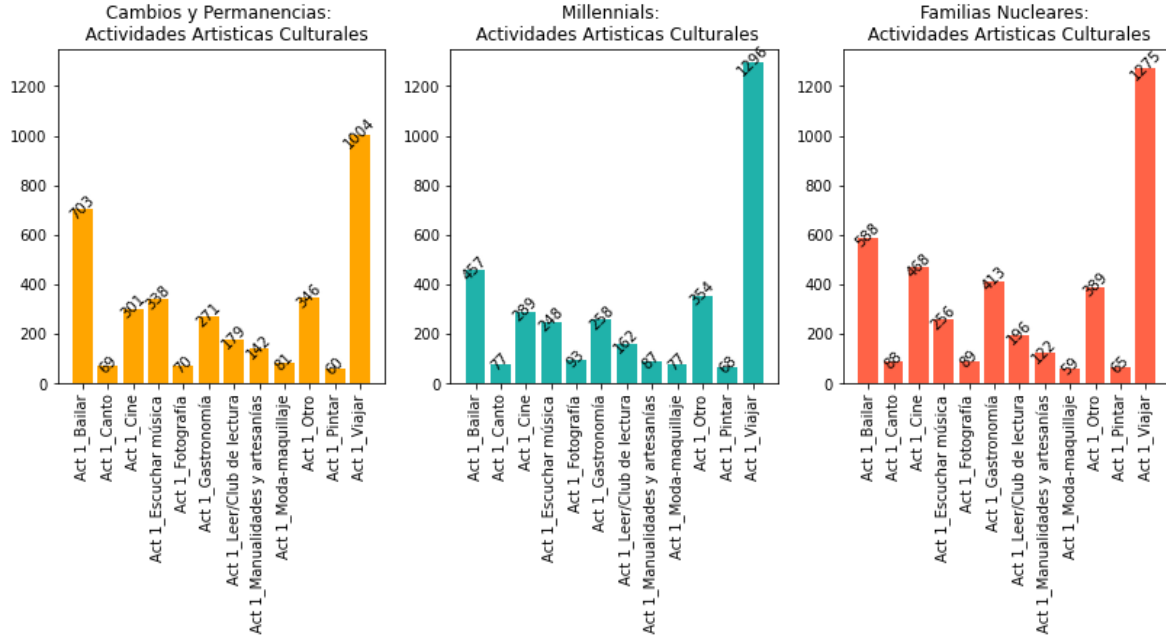


Figura 12. Total de empleados por actividades artísticas y culturales que realizan en los diferentes clústers.

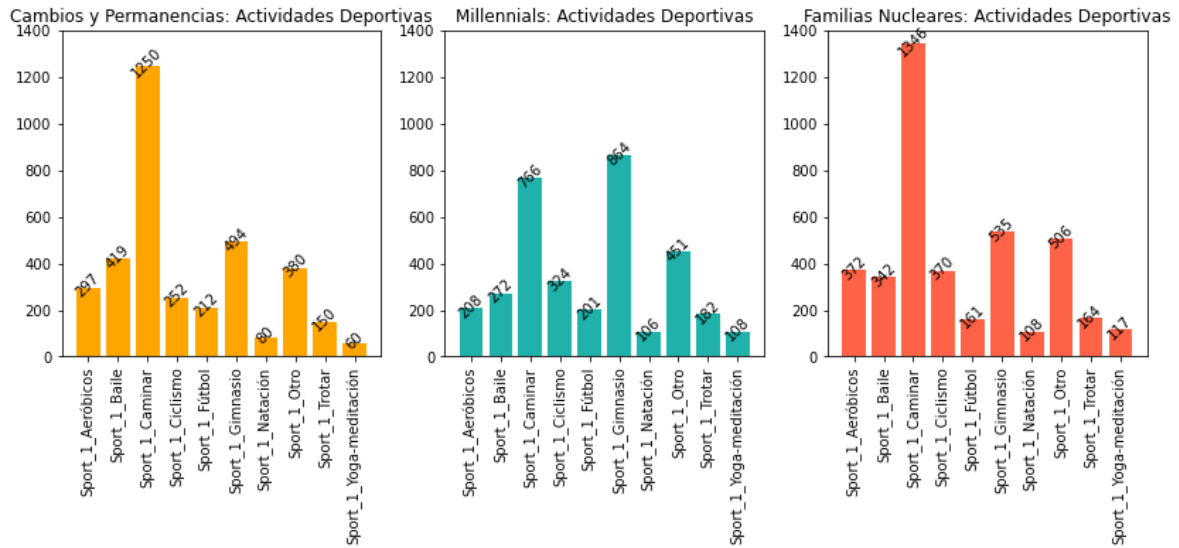


Figura 13. Total de empleados por deportes que practican en los diferentes clústers.

### 4.3 Variables que Explican el Modelo de Clusterización

En el proceso de clusterización que se construyó anteriormente se usaron las variables demográficas y socio-demográficas de los empleados, por lo que se procedió a evaluar cuáles de ellas explican mejor el modelo, para esto se usó el dataframe completo sin separarlo por clústers, donde las características de los empleados se tomaron como variables explicativas y el clúster donde quedaron clasificados los empleados se tomó como la variable de salida.

Para la estimación de la importancia de cada atributo se usó un Árbol de Decisión de profundidad 3, con el objetivo de encontrar las combinaciones de características que permiten clasificar a los empleados en el clúster correspondiente. Al correr el Árbol de Decisión, se obtuvo como resultado que las variables que se muestran en la Tabla 4. son las más significativas para la clasificación de los empleados en cada uno de los clústers.

Tabla 4. Variables que explican el modelo de clusterización.

	predictor	importancia
2	Tiene_Hijos	0.380878
26	Nivel_Cargo_AUXILIAR	0.318376
47	Tipo_Vivienda_Apartamento	0.182410
48	Tipo_Vivienda_Casa	0.075285
220	Rango_Edad_Entre 36 y 51 años	0.043051

En las figuras 14 y 15, se obtuvo las diferentes combinaciones de las características que clasifican los empleados en cada clúster, en este caso el Clúster0 corresponde al clúster de Cambios y Permanencias, el Clúster1 corresponde al clúster de los Millennials y el Clúster2 al clúster de Familias Nucleares.

Para llegar a los diferentes clústers se interpretó 8 caminos de decisión.

**Camino 1:** Si Tiene Hijos es  $\leq 0.50$  y Tipo de Vivienda Apartamento es  $\leq 0.50$  y Nivel Cargo Auxiliar es  $\leq 0.50$ , entonces lo más probable es que pertenezca al clúster de los Millennials.

**Camino 2:** Si Tiene Hijos es  $\leq 0.50$  y Tipo de Vivienda Apartamento es  $\leq 0.50$  y Nivel Cargo Auxiliar es  $> 0.50$ , entonces lo más probable es que pertenezca al clúster Cambios y Permanencias.

**Camino 3:** si Tiene Hijos es  $\leq 0.50$  y Tipo de Vivienda Apartamento es  $\leq 0.50$  y Rango de Edad entre 36 y 51  $\leq 0.50$ , entonces lo más probable es que pertenezca al clúster de los Millennials.

**Camino 4:** si Tiene Hijos es  $\leq 0.50$  y Tipo de Vivienda Apartamento es  $\leq 0.50$  y Rango de Edad entre 36 y 51  $> 0.50$ , entonces lo más probable es que pertenezca al clúster de los Millennials.

**Camino 5:** si Tiene Hijos es  $> 0.50$  y Nivel Cargo Auxiliar es  $\leq 0.50$  y Tipo Vivienda Casa  $\leq 0.50$ , entonces lo más probable es que pertenezca al clúster Familias Nucleares.

**Camino 6:** si Tiene Hijos es  $> 0.50$  y Nivel Cargo Auxiliar es  $\leq 0.50$  y Tipo Vivienda Casa  $> 0.50$ , entonces lo más probable es que pertenezca al clúster Familias Nucleares.

**Camino 7:** si Tiene Hijos es  $> 0.50$  y Nivel Cargo Auxiliar es  $> 0.50$  y Tipo Vivienda Apartamento  $\leq 0.50$ , entonces lo más probable es que pertenezca al clúster Cambios y Permanencias.

**Camino 8:** si Tiene Hijos es  $> 0.50$  y Nivel Cargo Auxiliar es  $> 0.50$  y Tipo Vivienda Apartamento  $> 0.50$ , entonces lo más probable es que pertenezca al clúster Familias Nucleares.

Al compararse los resultados aquí obtenidos con los de la clusterización realizada mediante el algoritmo K-MODES, se encontró que las características sugeridas en el Árbol de Decisión explican el proceso de clusterización.

Por ejemplo, en la caracterización que se realizó a través de los resultados de la clusterización se describió que, en el clúster de los Millennials se indicó que allí se encuentran los empleados más jóvenes y que solo el 11% de ellos tienen hijos.

Al interpretar el árbol, se encontró que 3 de los caminos de decisión indican la probabilidad de pertenecer al clúster de los Millennials, por lo que se evaluó el camino con menor índice de Gini el cual es el 3 y se obtuvo que: si Tiene Hijos es  $\leq 0.50$  y Rango de Edad entre 36 y 51  $\leq 0.50$ , describe la probabilidad de pertenecer a este clúster, resultado que coincide con la descripción anterior. Para el clúster Cambios y Permanencias y el clúster Familias Nucleares se observó el mismo comportamiento lo que permite concluir que las variables que mejor explican el modelo de clusterización son Tiene Hijos, Nivel Cargo Auxiliar, Tipo Vivienda Apartamento, Tipo Vivienda Casa y Rango de Edad 36 y 51 años.

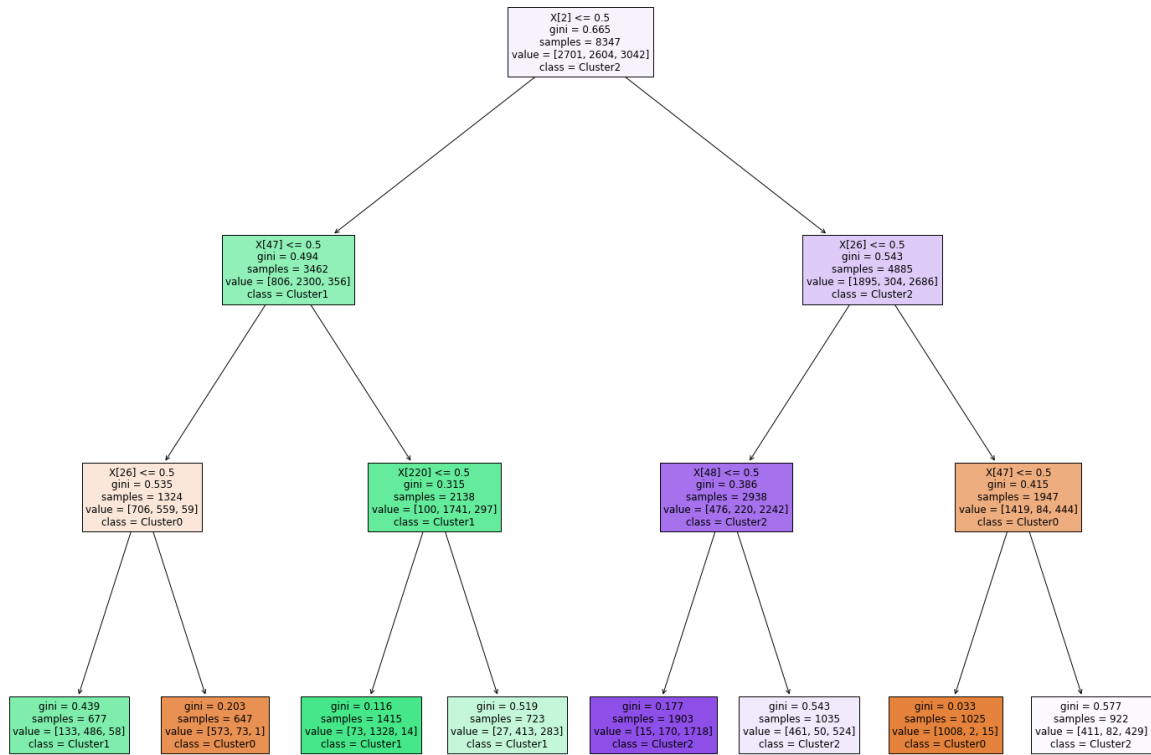


Figura 14. Árbol de Decisión.

```

--- feature_2 <= 0.50
|--- feature_47 <= 0.50
| |--- feature_26 <= 0.50
| | |--- class: 1
| | |--- feature_26 > 0.50
| | | |--- class: 0
| |--- feature_47 > 0.50
| | |--- feature_220 <= 0.50
| | | |--- class: 1
| | |--- feature_220 > 0.50
| | | |--- class: 1
|--- feature_2 > 0.50
|--- feature_26 <= 0.50
| |--- feature_48 <= 0.50
| | |--- class: 2
| |--- feature_48 > 0.50
| | |--- class: 2
|--- feature_26 > 0.50
| |--- feature_47 <= 0.50
| | |--- class: 0
| |--- feature_47 > 0.50
| | |--- class: 2
  
```

Figura 15. Árbol de Decisión tipo texto.

## 5 Análisis de la Utilización de los Convenios y Metas Propuestas

### 5.1 Análisis del Uso de los Convenios

Una vez caracterizada la población se evaluó para cada clúster la usabilidad de los convenios por tipología, donde se calculó su porcentaje de uso y en base a este se definió una métrica acorde a cada clúster para la propuesta de las recomendaciones.

Las tipologías que agrupan los diferentes productos y servicios que se ofrecen a través de los diferentes convenios se describen a continuación:

**Ahorro y Finanzas:** consiste en productos y servicios que contribuyen en el mejoramiento de la calidad de vida del empleado desde su economía, ofreciéndoles un amplio portafolio de ahorro, crédito, auxilios y beneficios.

**Salud:** servicios de atención médica y domiciliaria, ayudas diagnósticas, óptica, vacunación y odontología con tarifas preferenciales para los colaboradores y sus familias (empleado, cónyuge, compañero(a) permanente, hijos, padres, hermanos, abuelos y suegros).

**Acompañamiento del Duelo:** Tarifas especiales en servicios exequiales, acompañamiento del duelo para los colaboradores.

**Deporte:** natación, hidro aeróbicos, aeróbicos, multifuncional, gimnasia en general para los colaboradores con tarifas especiales.

**Cultura:** diferentes actividades como música, pintura, cerámica, escultura, fotografía, grabado; convenios con descuentos para compras de libros y descuentos por la nómina.

**Hogar:** descuentos y tarifas especiales para los colaboradores en diferentes productos para el hogar, electrodomésticos, tecnología, carpintería, mercado y aseo.

En la evaluación de la usabilidad de los convenios por tipo, se observó que Ahorro y Finanzas tiene el mayor porcentaje de uso frente a las demás tipologías en cada uno de los clústers; este resultado es coherente con el acompañamiento que le brinda la compañía a sus empleados mediante el programa de salud financiera, el cual entrega herramientas que permiten una correcta administración de los recursos económicos y lograr una mejor toma de decisiones financieras para cumplir los sueños sin comprometer el equilibrio entre la salud física y mental.

En la Figura 16, se comparó la usabilidad de los convenios por tipología en los diferentes clústers; se observó un comportamiento similar, ya que al graficarse las frecuencias y evaluar los tipos de convenios más usados y menos usados, cada tipología se encuentra con frecuencias similares para los tres clústers; sin embargo, el Clúster Millennials tiene un comportamiento diferente para los convenios del tipo Salud, ya que estos son los segundos más usados. Para los clusters Familias Nucleares y Cambios y Permanencias se encontró que los productos y servicios de la tipología Acompañamiento del Duelo son los segundos más usados.

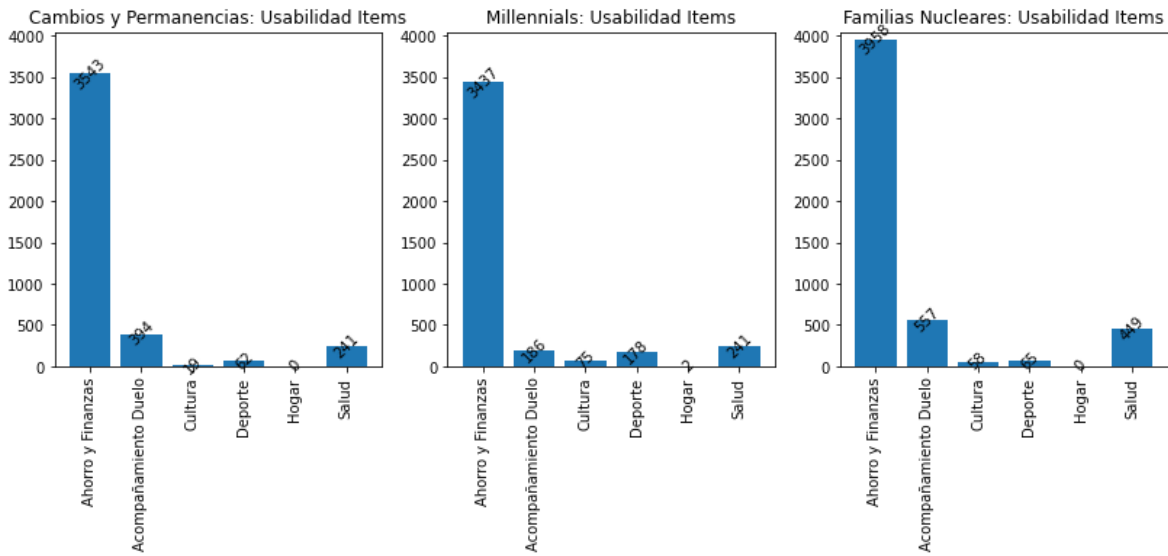


Figura 16. comparación usabilidad de los convenios por tipología y clúster.

## 5.2 Definición de las Metas de Uso

La meta se definió basándose en el comportamiento y el tamaño de cada clúster. Esta meta se estableció como duplicar los usuarios actuales por cada tipo de convenio. Si el número de usuarios establecidos como meta supera la cantidad de personas que pertenecen al clúster, entonces la meta se reduce al total de personas del clúster. Por ejemplo, los convenios de Ahorros y Finanzas son utilizados por más del 50% de las personas que pertenecen a cada clúster, y en consecuencia la meta se fija en el 100% de las personas del clúster. En el caso de que no haya usabilidad de un tipo de convenio en el clúster, se fija la meta en 2 personas.

En las Figuras 17, 18 y 19, se muestra el porcentaje de uso de los convenios por tipo en cada uno de los clústers, donde las barras azules indican el % total de uso y las barras naranjas indican el % de la meta propuesta.

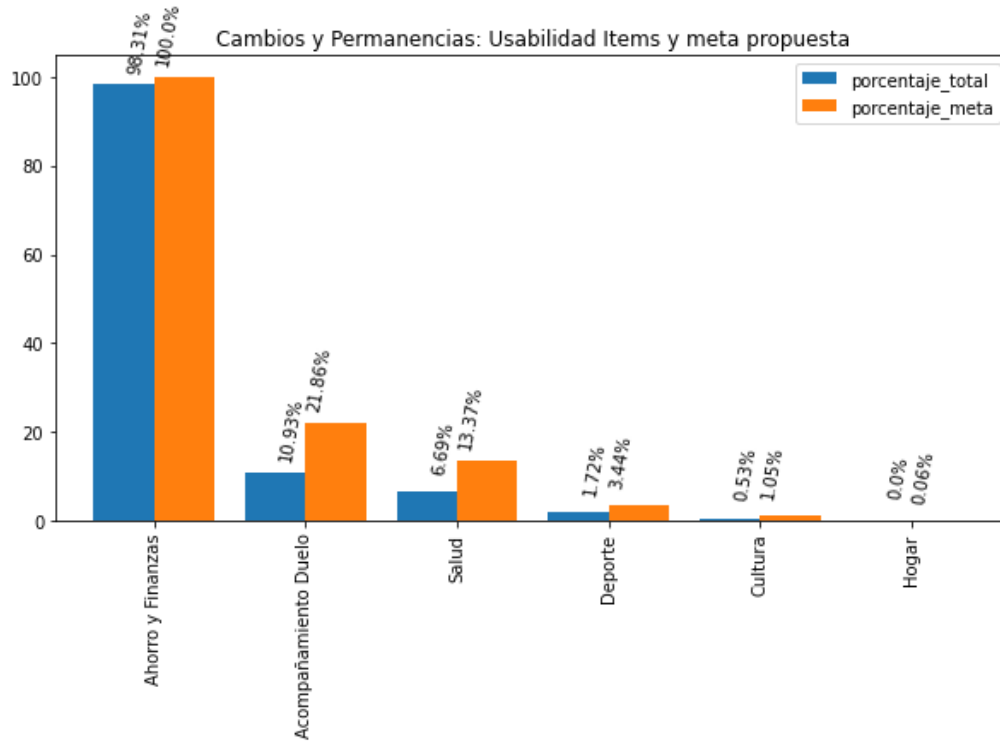


Figura 17. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Cambios y Permanencias.

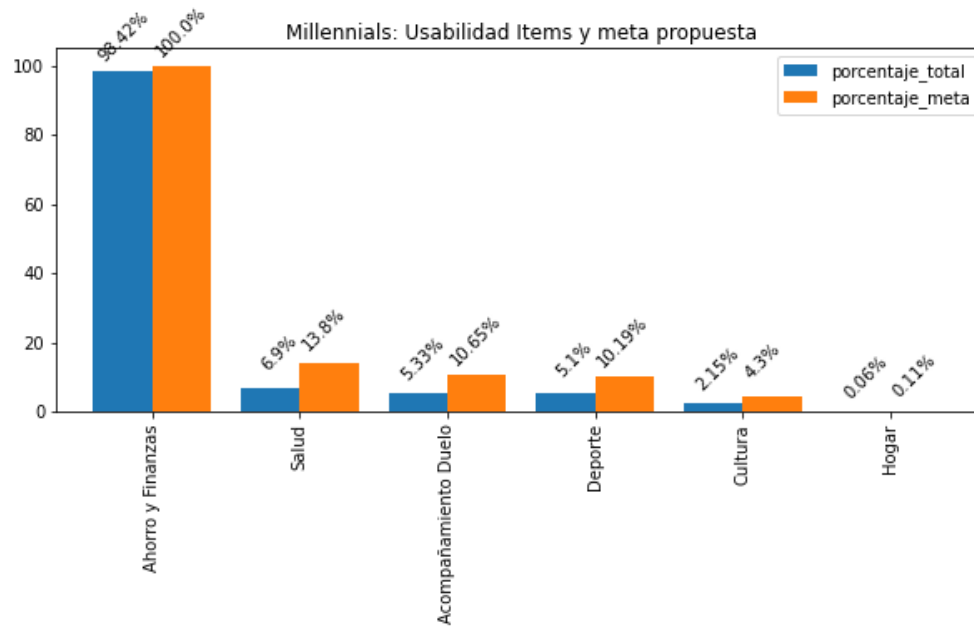


Figura 18. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Millennials.

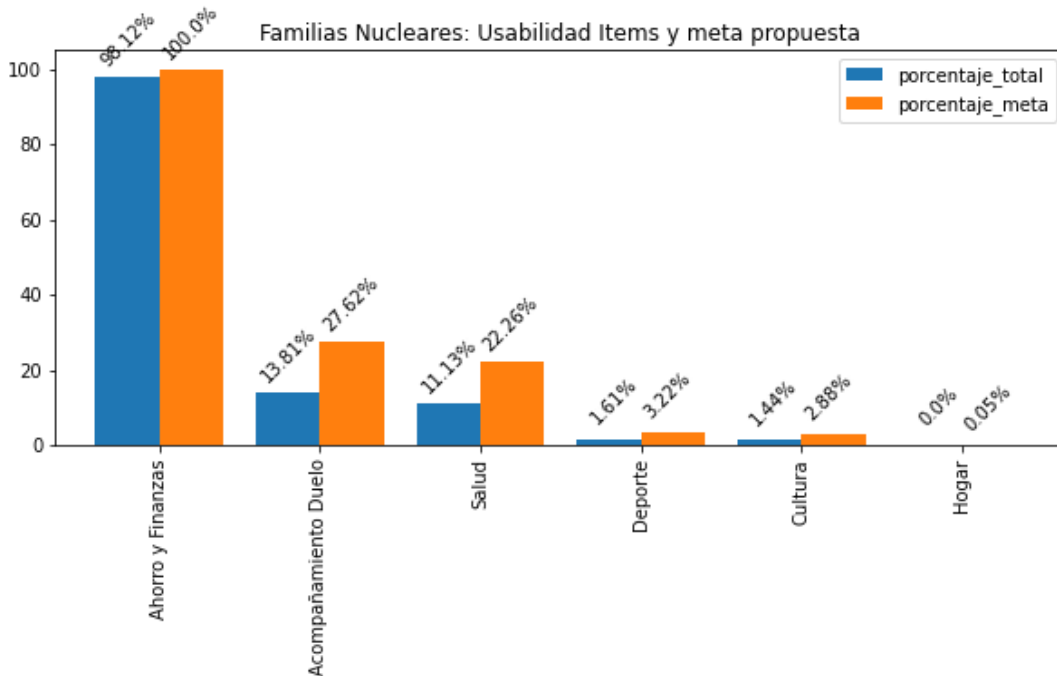


Figura 19. Porcentaje de usabilidad meta propuesta por tipo de convenio para el clúster Familias Nucleares.

### 5.3 Pronóstico de Uso de un Tipo de Convenio

Teniendo definida la meta a alcanzar para cada uno de los clústers, se evaluó la probabilidad de que un empleado use o no un tipo de convenio, con el fin de apoyar las recomendaciones que se realicen en trabajos futuros, para esto se usó el algoritmo Árbol de Decisión.

Inicialmente se dividió en nuevos conjuntos de datos el dataframe utilizado para la caracterización de los clústers de la siguiente forma:

Las características (variables demográficas y socio-demográficas) se tomaron como variables de entrada y el tipo de convenio como la variable de salida, resultando así un total de 6 dataframes, uno por cada tipo de convenio. De acuerdo con la usabilidad graficada en la Figura 16, donde se observa que algunos tipos de convenios no son muy usados y en otros su usabilidad se acerca al 100% de los empleados, se obtiene que los conjuntos de datos contienen clases desbalanceadas, lo que perjudicaría las clases minoritarias; teniendo este escenario, para la construcción del modelo se empleó en el árbol, el parámetro tamaño de las clases, balanced con el fin de equilibrarlas.

Al correrse el modelo para cada tipo de convenio se obtuvo los siguientes resultados para los que se midieron el Score del modelo y se usó la métrica Balanced Accuracy Score, ideal para problemas de clasificación binaria y conjuntos de datos desequilibrados.

Tabla 5. Resultados de los modelos

<b>Tipo de Convenio</b>	<b>Score</b>	<b>Balanced Accuracy Score</b>
Acompañamiento del Duelo	0.8206	0.6902
Ahorro y Finanzas	0.8277	0.7063
Cultura	0.7151	0.7911
Deporte	0.7049	0.7097
Hogar	0.9997	0.9996
Salud	0.8152	0.5657

Teniendo estos resultados se contará con el soporte para la construcción de las diferentes recomendaciones que sean planteadas para los empleados de acuerdo con sus intereses actuales, lo que haría parte de un trabajo futuro.

## 6 Dashboard de Resultados Prototipo

Se creó un dashboard utilizando la herramienta de Microsoft Power BI, para la presentación de los resultados del prototipo del sistema de recomendación. Para la construcción de cada una de las visuales, se extrajeron los conjuntos de datos que se obtuvieron en la construcción del modelo.

La Figura 20, corresponde a la primera hoja del dashboard, la cual contiene una exploración inicial de los datos que están compuestos por las principales características demográficas y socio-demográficas que fueron recopiladas para el desarrollo de este trabajo.

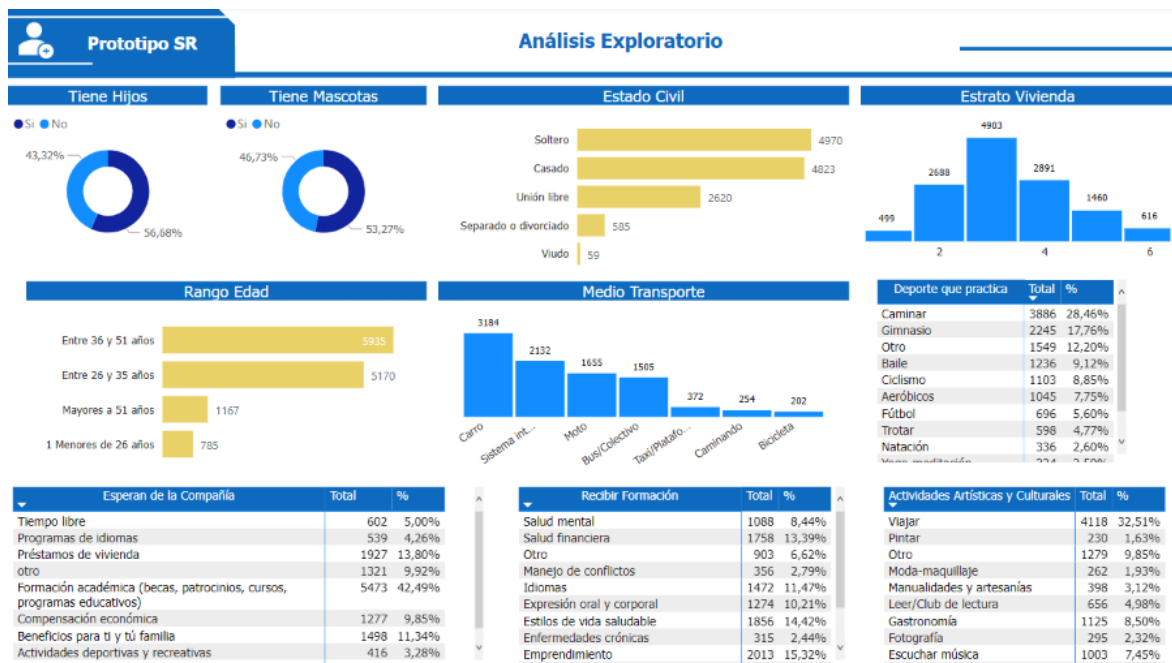


Figura 20. Dashboard, análisis exploratorio de las variables.

Las figuras 21 y 22 corresponden a la presentación de la caracterización de los empleados, construida a partir del proceso de clusterización que se corrió con el algoritmo de K-MODES, allí se comparan las principales características para los diferentes clústers obtenidos, las cuales contienen información que permite describir 3 grupos diferentes de empleados, estas características son:

- Rango de edad
- Estado civil
- Estrato socioeconómico
- Tiene hijos
- Tiene Mascotas
- Medio de transporte
- Intereses en formación

- Actividades artísticas y culturales
- Oferta de la compañía

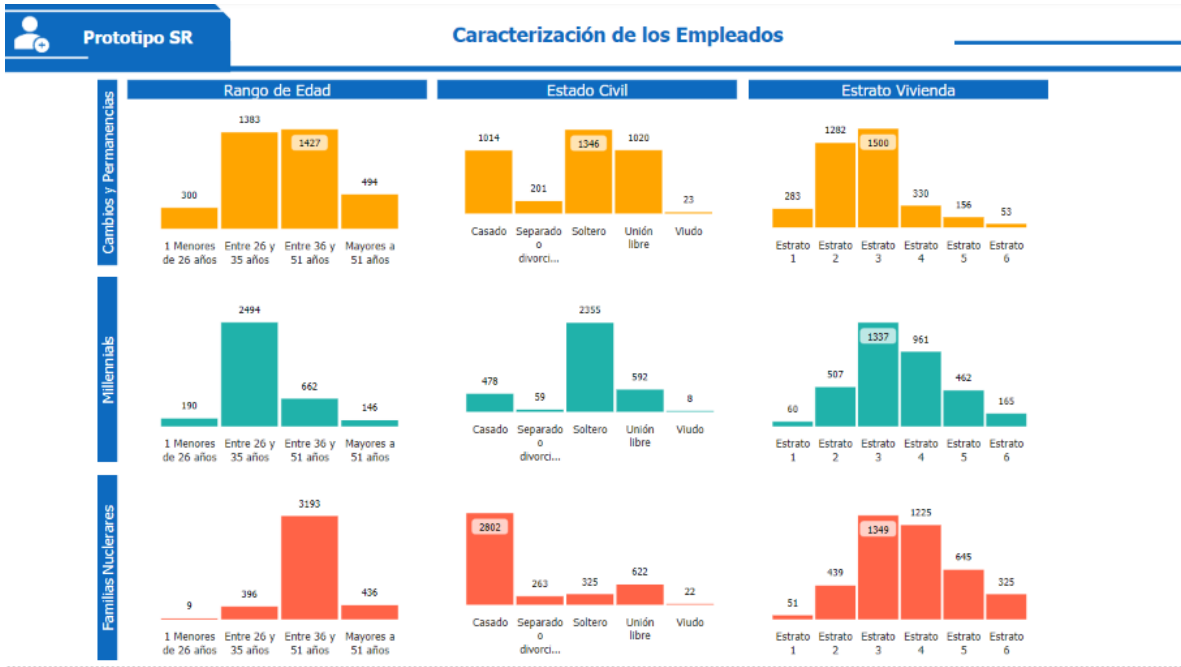


Figura 21. Dashboard, caracterización variables demográficas.

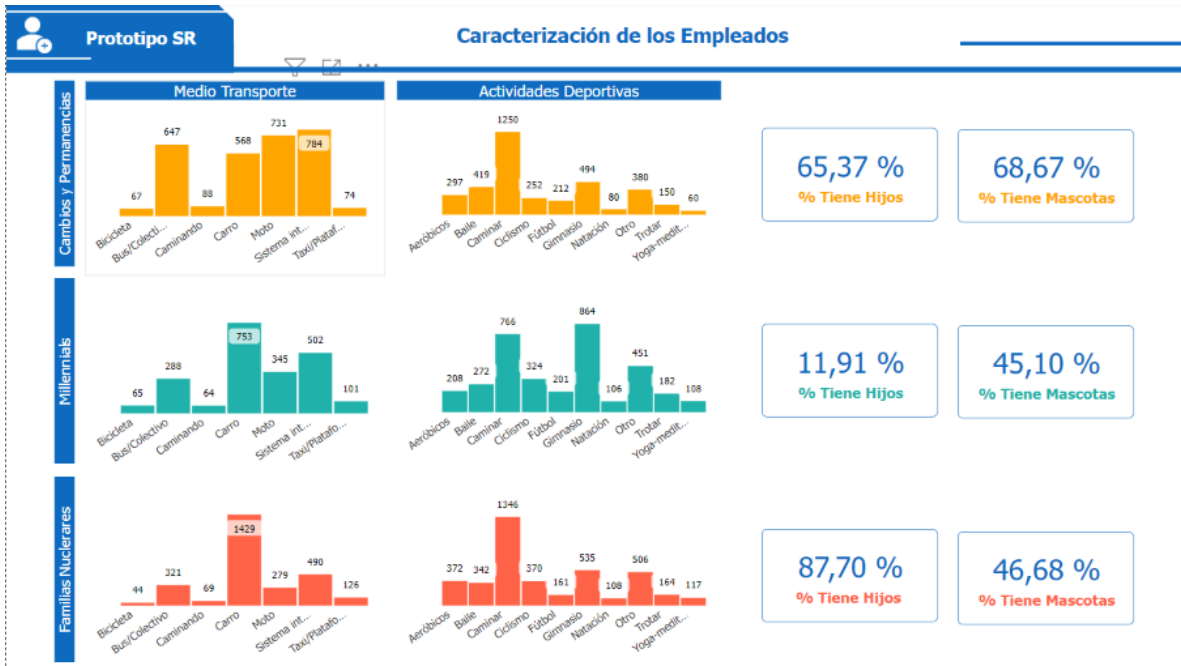


Figura 22. Dashboard, caracterización de las variables socio-demográficas.

Finalmente, en la Figura 23, se presentó la usabilidad actual de los convenios por tipología y por clúster, los porcentajes de uso de los convenios y la meta a alcanzar, la cual fue definida basándose en el comportamiento y el tamaño de cada clúster, estableciendo esta como duplicar los usuarios actuales por cada tipo de convenio.

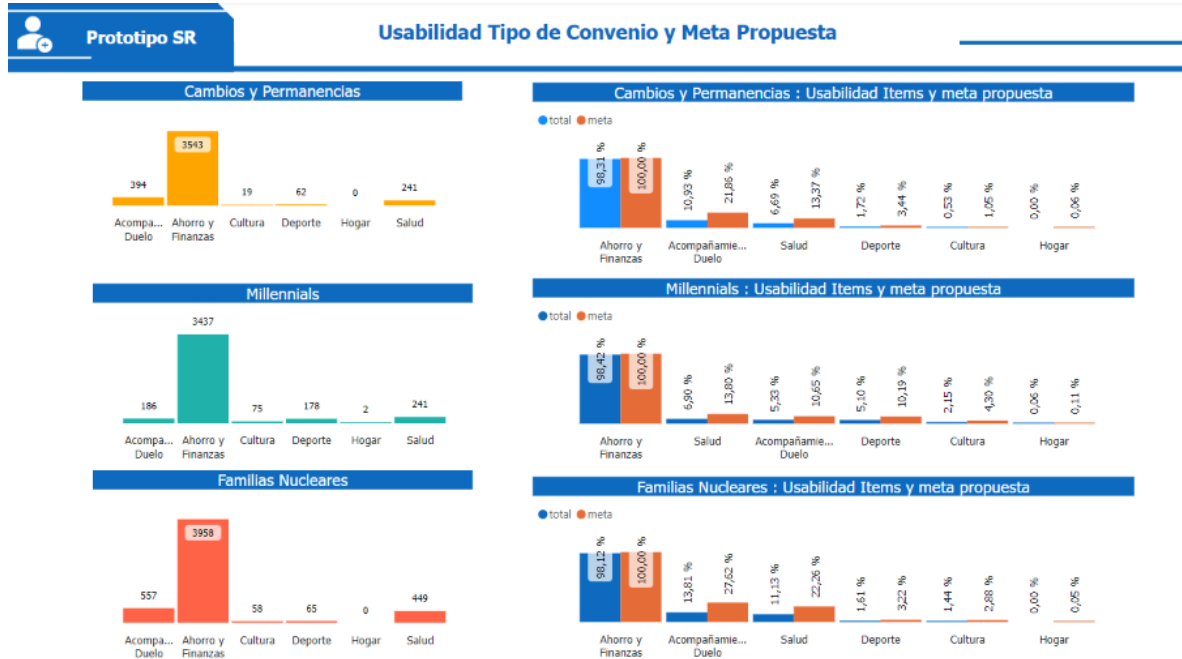


Figura 23. Dashboard, usabilidad por tipo de convenio y meta propuesta.

## 7 Conclusiones

### 7.1 Cumplimiento de Objetivos

Se recopiló la información de los empleados a través de las bases de datos de Talento Humano, las cuales contienen la información demográfica y socio-demográfica de los empleados, una vez recopilada la información, se depuró y preparó los conjuntos de datos a utilizar en este proyecto.

Se realizó un análisis exploratorio de las características de los empleados, logrando un entendimiento de su comportamiento, la composición de la población de Suramericana; se logró explorar la composición de los hogares de los empleados, la distribución por rangos de edad, género, la selección de preferencias e intereses, entre otras.

Se caracterizó mediante técnicas de clustering a los empleados de la compañía. Utilizando el algoritmo K-MODES se obtuvo un resultado de 3 clústers, los cuáles se nombraron de la siguiente manera: Cambios y Permanencias, Millennials y Familias Nucleares. Para cada clúster se realizó una descripción donde se explicó mediante un conjunto de variables la caracterización de cada uno de estos grupos. El modelo de clusterización se apoyó del modelo Árbol de Decisión con el fin de encontrar y evaluar aquellas variables que explicarían el modelo.

Se analizó para cada clúster la utilización de los convenios actuales, estos convenios fueron evaluados agrupándose por la tipología que los clasifica, donde se observó que la mayoría de los tipos de convenios tienen poca usabilidad lo que soporta el planteamiento del problema de este proyecto.

Para cada clúster obtenido se proponen metas de utilización por tipo de convenio basándose en el comportamiento y el tamaño de cada clúster. Esta meta se estableció como duplicar los usuarios actuales por cada tipo de convenio. Se construyó un árbol de decisión que permitiera saber si una persona usaría o no un tipo de convenio al momento de la entrega de una recomendación.

Se desarrolló un prototipo de un sistema de recomendación, del cual se presentan los resultados mediante un dashboard, en el que se observa el análisis exploratorio de la población, los resultados de la caracterización de los empleados a través de los clústers obtenidos, el uso de los convenios por tipología y la presentación de las metas de uso a alcanzar en la oferta de convenios de los empleados del Grupo Suramericana, teniendo en cuenta sus intereses actuales.

### 7.2 Principales Conclusiones y Trabajo Futuro

Cuando se consulta en la bibliografía sobre la construcción de sistemas de recomendación, estos en su mayoría hablan de recomendaciones basadas en ítems de los que ya se ha capturado la percepción de los usuarios y en otros casos recomendaciones basadas en las similitudes entre los mismos ítems, para lo que utilizan diferentes tipos de filtrado en la construcción de una recomendación. De acuerdo con las características que componen el

problema planteado en este trabajo, el cual es un caso que está enfocado al plan de beneficios de los empleados de una organización, caso de sistema de recomendación que no es común encontrar en la bibliografía, fue necesario plantear el prototipo del sistema de recomendación, basándose en diferentes algoritmos de Machine Learning, partiendo de una clusterización de la población para conocer sus gustos y apoyarse de un modelo como lo fue el árbol de decisión que permitiera explicar el modelo y además evaluar si esa recomendación a entregar sería usada o no por el empleado; por lo que este problema permitió plantear un sistema de recomendación diferente a los que habitualmente son desarrollados.

En el presente trabajo se logró realizar una evaluación de los empleados, partiendo de una exploración inicial y mediante el uso de diferentes técnicas de análisis como lo fue el proceso de clusterización, se logró una evaluación más profunda de la población, lo que permitió identificar su comportamiento a través de las diferentes características que diferenciaron cada uno de los clústers.

Con la caracterización realizada se cuenta ahora con una mirada global de los diferentes públicos que componen la compañía, lo que permitirá llegar a cada uno de ellos de manera más asertiva ya que se tiene conocimiento de sus intereses.

A partir de lo desarrollado en este trabajo la compañía cuenta actualmente con nuevas herramientas que servirán de soporte para la toma de decisiones y la construcción de recomendaciones que serán la base para trabajos futuros. El paso que sigue después de la entrega de los resultados de este prototipo del sistema de recomendación es automatizar y llevar a producción el sistema de recomendación de la oferta de convenios que tienen los empleados y de esta forma puedan hacer mayor uso de ellos a partir de las sugerencias que se entreguen.

Para recomendaciones futuras es ideal recopilar más información de la usabilidad de los convenios, y así apoyar también la construcción de las recomendaciones en lo que los empleados usan, a demás es importante para este tipo de sistemas de recomendación recopilar las percepciones que tienen los usuarios de los convenios y beneficios que se están entregando.

## 8 Bibliografía

- [1] Comfama. People Analytics: Equipos Humanos que hacen empresas sostenibles. Obtenido de Memorias Eventos: Comfama octubre digital. 9 de octubre de 2019
- [2] Ruta N, #4Revolución: principales tecnologías. Marzo de 2020. [En línea]. Disponible en: <https://www.rutanmedes/cuarta-revolucion-industrial>
- [3] Performance Magazine, The KPI Institute. How to Measure the Performance of your Human Resources Department. Mayo de 2017. [En línea]. Disponible en: <https://www.performancemagazine.org/performance-human-resources/>
- [4] IBM Knowledge center. CRISP-DM Help. Septiembre de 2020. [En línea]. Disponible en: [https://www.ibm.com/support/knowledgecenter/SS3RA7\\_15.0.0/com.ibm.spss.crispdm.help/crisp\\_overview.htm](https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm)
- [5] J. Herrera, E. Castro, "Sistema de Recomendación de Matrículas en Asignaturas Basados en Perfiles de Docentes, Alumnos y Asignaturas en la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín". Julio de 2018. [En línea]. Disponible en: <https://www-scopus-com.ezproxy.unal.edu.co/record/display.uri?eid=2-s2.0-85057477385&origin=resultslist&sort=plf-f&src=s&st1=sistema+recomendacion&st2=&sid=2d43abf3ff65ae53db41dc7b053d57ce&sot=b&sdt=b&sl=36&s=TITLE-ABS-KEY%28sistema+recomendacion%29&relpos=8&citeCnt=0&searchTerm=>
- [6] M. Caro, J. Hernandez, "Diseño de un Sistema de Recomendación en Repositorios de Objetos de Aprendizaje Basado en la Percepción del Usuario: Caso Rodas". Enero de 2011. [En línea]. Disponible en: [http://www.scielo.org.co/scielo.php?script=sci\\_arttext&pid=S0124-81702011000100004&lang=es](http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-81702011000100004&lang=es)
- [7] H. Yate, "Elección de población susceptible a campañas de comunicación mediante sistemas de recomendación en las droguerías Colsubsidio". 2020. [En línea]. Disponible en: <http://eds.b.ebscohost.com/eds/detail/detail?vid=1&sid=4d281688-dc54-45f4-94c9-b28d40141b06%40pdc-v-sessmgr04&bdata=Jmxhbmc9ZXMmc2l0ZT1lZHMtbGl2ZQ%3d%3d#AN=edsbas.87F717C&db=edsbas>
- [8] S. Oliva, "Implementación y desarrollo de sistemas de recomendación para profundización de clientes y cruce de productos en un banco de retail". 2017. [En línea]. Disponible en: <http://eds.b.ebscohost.com/eds/detail/detail?vid=5&sid=4ffe05a3-7c95-4429-a6a0-395c45a08377%40pdc-v-sessmgr05&bdata=Jmxhbmc9ZXMmc2l0ZT1lZHMtbGl2ZQ%3d%3d#AN=edsbas.7786E2EA&db=edsbas>

[9] D. Arocha, "Sistemas de recomendación de productos para banca empresarial y corporativa: Un análisis comparativo". Noviembre de 2018. [En línea]. Disponible en: <http://eds.b.ebscohost.com/eds/detail/detail?vid=6&sid=4ffe05a3-7c95-4429-a6a0-395c45a08377%40pdc-v-sessmgr05&bdata=Jmxhbm9ZXMmc2l0ZT1lZHMtbGl2ZQ%3d%3d#AN=edsbas.1835B43B&db=edsbas>

[10] D. Morales, D. Martín, "Integración de modelos de agrupamiento y reglas de asociación obtenidos de múltiples fuentes de datos". Octubre de 2011. [En línea]. Disponible en: <http://www.scielo.org.mx/pdf/cys/v16n2/v16n2a5.pdf>

[11] D. Galli, Soledad, "Python Feature Engineering Cookbook". Enero de 2020.

[12] D. Gonzalez, Ligdi, "Reducción de la dimensionalidad". Abril de 2020. [En línea]. Disponible en: <https://aprendeia.com/reduccion-de-la-dimensionalidad-machine-learning/>

[13] D. Ashish, Ashydv, "Bank Customer Clustering (K-Modes Clustering)". SF. [En línea]. Disponible en: <https://www.kaggle.com/ashydv/bank-customer-clustering-k-modes-clustering>

[14] D. Bonthu, Harika, "KModes Clustering Algorithm for Categorical data". Junio de 2021. [En línea]. Disponible en: <https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/>

[15] D. Amat, Joaquin, "KModes Clustering Algorithm for Categorical data". Octubre de 2020. [En línea]. Disponible en: [https://www.cienciadedatos.net/documentos/py07\\_arboles\\_decision\\_python.html](https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html)

[16] D. Aguila, Juan, "Webinar: Crea un motor de recomendación usando Machine Learning". Julio de 2020. [En línea]. Disponible en: <https://www.youtube.com/watch?v=NCQZuprocXk>

[17] D. Abella, Blanca, 2021. "Mejora de las predicciones en muestras desbalanceadas". Tesis de Ingeniería Informática. Universidad Autónoma de Madrid. [En línea]. Disponible en: [https://repositorio.uam.es/bitstream/handle/10486/697900/abella\\_miravet\\_blanca\\_tfg.pdf?sequence=1](https://repositorio.uam.es/bitstream/handle/10486/697900/abella_miravet_blanca_tfg.pdf?sequence=1)

[18] Velasquez, Juan, "Métricas básicas de clasificación". Septiembre de 2021. [En línea]. Disponible en: [https://jdvelasq.github.io/courses/notebooks/sklearn\\_model\\_selection\\_and\\_evaluation/2-12\\_mtricas\\_de\\_clasificacion.html](https://jdvelasq.github.io/courses/notebooks/sklearn_model_selection_and_evaluation/2-12_mtricas_de_clasificacion.html)

[19] Gerón, Aurélien, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems”. 2019.

[20] Velasquez, Juan, “Inducción de reglas con el algoritmo Apriori”. 2019. [En línea]. Disponible en: <https://jdvelasq.github.io/mineria-de-datos/notebooks/sklearn/apriori/01-induccion-de-items-apriori.html>