Redes neuronales y el pronóstico de ventas de una agencia de seguros de la ciudad de Bogotá D. C.

Neuronal networks and the sales forecast of an insurance agency in the city of Bogotá D. C.

Maria Paula Garzón D'Aleman

Ingeniera Industrial

Estudiante Especialización en Estadística Aplicada

Cristian Camilo Díaz Herrera

Tutor Machine Learning

Fundación Universitaria Los Libertadores

RESUMEN

El presente estudio pretendió establecer un modelo estadístico para el análisis de tendencias de compra con la caracterización de los diferentes tipos de compradores con los que cuenta la empresa analizada, esto con el fin de plantear estrategias que permitan incursionar en el mercado dinámico optimizando los recursos de la empresa.

El análisis de tendencias que se realizó tuvo en cuenta los factores que pueden contribuir a cómo impulsar las ventas y cuáles son aquellos que las afectan; teniendo en cuenta la actividad económica de la empresa analizada, los principales factores de riesgo son servicio al cliente, calidad en el servicio, políticas de renovación y negocios nuevos, tiempos de respuesta, precios de cada una de las soluciones y los beneficios, y por último, el constante cambio en la cultura de asegurabilidad.

Palabras clave: Redes neuronales artificiales, pronóstico, cumplimiento, vivienda, pyme, movilidad

ABSTRACT

The next study purposes to establish a statistical model for the analysis of acquiring trends with the characterization of the different types of buyers that the analyzed company has, the main focus was to propose strategies that allow entering the dynamic market optimizing the resources of the company.

The trend analysis that was carried out would take into account the factors that could have contributed to how increase sales and which ones affected them; taking into account the economic activity of the analyzed company the main risk factors was: customers, service, quality of service, renewal policies and new business, response time, prices of each of the solutions and the benefits, finally, the constant change in the culture of insurability.

Keywords: Artificial neural networks, forecasting, compliance, housing, pyme and mobility.

INTRODUCCIÓN

La empresa analizada es una agencia de seguros que se encuentra constituida y en funcionamiento desde el año 2007 con el fin de brindar tranquilidad y gestionar los riesgos de

cada uno de sus clientes, corporativos como personas naturales por medio de la asesoría en la colocación de seguros generales en las diferentes compañías aliadas.

En el año 2019 inició con la estructuración de su proyecto “MEGA” el cual consistió en establecer la Meta Estratégica Grande y Ambiciosa, y aquellos proyectos que servirían de base para lograr la meta propuesta. La MEGA asignada para el año 2030 es la emisión de primas de seguros generales por COP 100.000.000.00, para lo cual dentro de su estructuración se contempló la creación de una herramienta o modelo para pronosticar las ventas de acuerdo al comportamiento de los clientes en las verticales de cumplimiento, vivienda, pyme y movilidad.

Teniendo en cuenta lo anterior, se tomó la muestra de 780 clientes. Como se observa en la Tabla 1, se contó para cada cliente (los cuales se representaron con su respectivo identificador), la tipificación de su naturaleza jurídica (es decir, si es persona natural o jurídica), si contó con los datos básicos para el contacto como lo son la dirección, el celular y el e-mail, su respectiva clasificación de acuerdo a la edad, sexo, sector económico y región.

Tabla 1

Información básica de la muestra de clientes

IDENTIF	TIPO	DIREC	CEL	EMAIL	CL_EDAD	SEXO	SECTOR ECONÓMICO	REGIÓN
851828	PJ	1	1	0	PRI_DEC	EMPRESA	SERVICIOS	BOGOTA D. C.

En la Tabla 2, se observa la prima por cada una de las verticales, es decir, cumplimiento, vivienda, pyme y movilidad, junto con el número de documentos emitidos por cada una de ellas.

Tabla 2

Información de la adquisición de seguros por cada una de las verticales

CUMPL	CUMPL_CANT	VIVIENDA	VIVIENDA_CANT	PYME	PYME_CANT	MOVILIDAD	MOVILIDAD_CANT
	0		0	2837522	1		0

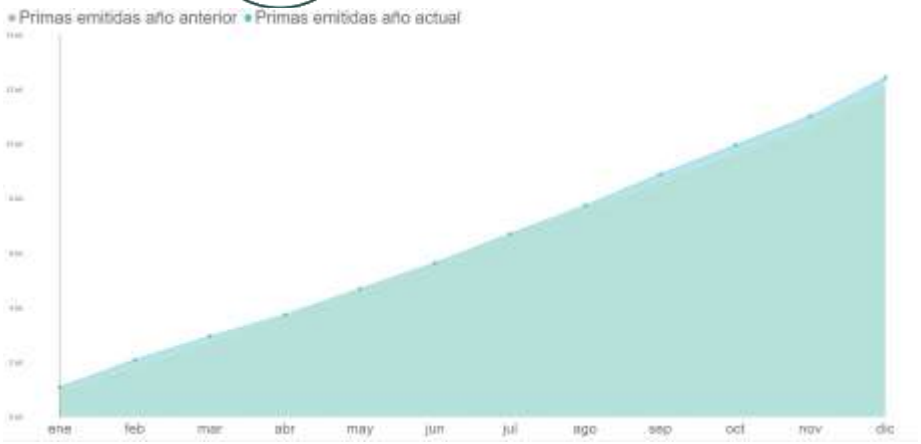
La información de adquisición de seguros comprendió los movimientos desde el 01 de enero del año 2020 hasta el 31 de diciembre del año 2020, con el fin de analizar el ejercicio anual de la empresa. En este ejercicio anual se observó la venta de COP 1.596.695.346 en primas y a su vez, 2.956 documentos; adicionalmente se evidenció que el 45.3% de los datos se encontraron en la variable de cumplimiento, es decir, en el año 2020 está vertical fue la de mayor contribución a la emisión de primas; por otra parte, la vertical que menos aportó a la emisión de primas con el 13% fue movilidad teniendo en cuenta el confinamiento presentado a raíz del COVID-19.

REFERENTES TEÓRICOS

La Federación de Aseguradores Colombianos, FASECOLDA, es aquella entidad sin ánimo de lucro que se encarga de promover las condiciones para el desarrollo sostenible del sector asegurador en Colombia, por lo cual es un referente importante para conocer la situación pasada versus la situación actual del comportamiento del sector junto con sus diferentes tipos de seguros. A continuación, en la imagen 1 se observa el crecimiento de la emisión de primas a través de cada uno de los meses del año 2020.

Figura 1.

Crecimiento de la emisión de primas en el año 2020. Fuente: FASECOLDA



La base teórica de este proyecto se encuentra en la agrupación de los términos de: Redes neuronales, aprendizaje no supervisado, k-means y seguros generales.

Redes neuronales

Tomando el concepto de "Una red neuronal es un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona" (Matich, 2001, p. 4), se determina que se puede asociar la red neuronal a un proceso de emulación de características propias de los seres humanos como recordar, asociar situaciones y analizar su comportamiento; por lo cual, se consideró el modelo ideal para analizar el comportamiento actual de los clientes e ir modificándose de acuerdo a las diferentes variables y entornos que se puedan presentar.

Aprendizaje no supervisado

En el aprendizaje no supervisado a diferencia del aprendizaje supervisado, no se observan etiquetas, se agrupan de acuerdo a los clústeres que posteriormente serán analizados para validar si son homogéneos o heterogéneos, para este caso particular, se utilizó el agrupamiento por k-means.

"El aprendizaje supervisado son un conjunto de técnicas que permiten inferir modelos para extraer conocimiento de conjuntos de datos donde a priori se desconoce. Las técnicas de aprendizaje no supervisado se pueden aplicar sin necesidad de tener los datos etiquetados para el entrenamiento." (Calvo, 2019, p.1)

Agrupamiento clúster por k-means

López y Fernández (2018) afirman que k-means "Corresponde a la teoría de Máquinas de Aprendizaje, su funcionamiento también es través de una función de costo que está determinada por la expresión:

$$Min\, J(C^{(1)}, \dots, C^{(m)}, \mu_1, .., \mu_k)$$

Esto contribuye a la determinación de que es un método que permite la agrupación de modo que se minimice la variación total dentro del conjunto de datos. La asignación inicial de los clústeres es importante para el objetivo del análisis. El algoritmo estándar Hartigan-Wong, define la variación total dentro del conglomerado como la suma de distancias al cuadrado, las distancias euclidianas entre los elementos y el centroide, tal como se observa en la siguiente fórmula:

Fórmula K-Means

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

x_i: Diseñar un punto de datos perteneciente al clúster.
μk: Es el valor medio de los puntos asignados al clúster

Para obtener la agrupación de datos por k-means se realizan los siguientes pasos:

1. Definir número de agrupaciones (k)
2. Selección aleatoria por parte del algoritmo de los centroides

3. Los elementos restantes se asignan a cada centroide
4. Se calcula el nuevo valor medio de cada grupo
5. Validación de la ubicación y reubicación, hasta que se logra la convergencia

Seguros generales

Seguro: "10. m. Contrato por el que alguien se obliga mediante el cobro de una prima a indemnizar el daño producido a otra persona, o a satisfacerle un capital, una renta u otras prestaciones convenidas. El seguro del coche, de accidentes, del hogar." – RAE.

METODOLOGÍA

Se realizó el análisis del comportamiento de compra de los clientes de la agencia de seguros en la ciudad de Bogotá, donde el objetivo es identificar si existen patrones de consumo y el acierto que se puede presentar en las predicciones; para lo cual, se ejecutaron cinco fases como se detalla a continuación:

Fase 1: Análisis descriptivo multivariado

Se inicia con el análisis de la estructura de la base de datos, en este caso se encontró que se tienen 17 variables junto con 780 observaciones. Se visualizó el tipo de dato al que pertenece cada variable, la base de datos que se manipuló contenía variables de tipo carácter y numéricas, tal como lo refleja la figura 2.

Figura 2

Estructura de la base de datos inicial.

```
'data.frame':   780 obs. of  16 variables:
 $ TIPO            : chr  "PN" "PN" "PN" "PN" ...
 $ DIREC           : num  1 1 1 1 1 1 1 0 1 1 ...
 $ CEL             : num  1 1 1 1 1 1 1 0 0 0 ...
 $ EMAIL           : num  1 1 1 1 1 1 1 1 0 1 ...
 $ CL_EDAD         : chr  "GEN_Y" "BABY_BOOM" "GEN_Y" "BABY_BOOM" ...
 $ SEXO            : chr  "FEMENINO" "MASCULINO" "FEMENINO" "MASCULINO" ...
 $ SECTOR.ECONÓMICO: chr  "SERVICIOS" "SERVICIOS" "SERVICIOS" "SERVICIOS" ...
 $ REGIÓN          : chr  "BOGOTA D. C." "BOGOTA D. C." "BOGOTA D. C." "BOGOTA D. C."
 ...
 $ CUMPL           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CUMPL_CANT      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ VIVIENDA        : num  0 0 730000 766831 1691782 ...
 $ VIVIENDA_CANT   : num  0 0 1 1 2 0 0 0 0 0 ...
 $ PYME            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PYME_CANT       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ MOVILIDAD       : num  114900 200200 212700 212700 212700 ...
 $ MOVILIDAD_CANT  : num  1 1 1 1 1 1 1 1 1 1 ...
```

En el análisis descriptivo multivariado se identificó el promedio de las diferentes variables, para el caso de cumplimiento el promedio de las primas emitidas fue de COP 928.129, para vivienda COP 406.885, para pyme COP 445.975, y por último, para movilidad COP 266.056. Luego, se puede observar en la matriz de varianzas y covarianzas que el siguiente par de variables poseía una relación inversa, es decir, que mientras una variable crecía, la otra decrecía: Cumplimiento-Vivienda.

Figura 3

Matriz de varianzas y covarianzas

	CUMPL	CUMPL_CANT	VIVIENDA	VIVIENDA_CANT	PYME	PYME_CANT	MOVILIDAD	MOVILIDAD_CANT
CUMPL	8717422419410	10690179.2540	-291899462914	-223373.39641	1187358772296	80250.357737	1187708450829	-380443.070070
CUMPL_CANT	10690179	38.5022	-766769	-0.62903	2487293	0.027305	-214082	-1.162498
VIVIENDA	-291899462914	-766768.6987	1971635400404	468652.34589	1084072574626	199121.783751	134111820752	330611.036699
VIVIENDA_CANT	-223373	-0.6290	468652	0.30088	15212	0.027364	24857	0.161468
PYME	1187358772296	2487292.9201	1084072574626	15211.67088	1139782632597O	855392.550201	610403246881	-81830.621821
PYME_CANT	80250	0.0273	199122	0.02736	855393	0.161961	284488	0.009748
MOVILIDAD	1187708450829	-214081.9428	134111820752	24857.23881	610403246881	284487.716497	7151036433777	252603.007330
MOVILIDAD_CANT	-380443	-1.1625	330611	0.16147	-81831	0.009748	252603	0.494222

Posteriormente, en la revisión de la matriz de correlaciones se observa que no hay variables que posean asociación de acuerdo a la correlación superior al 0.5, la imagen que se encuentra a continuación confirma lo enunciado anteriormente:

Figura 4

Matriz de correlación

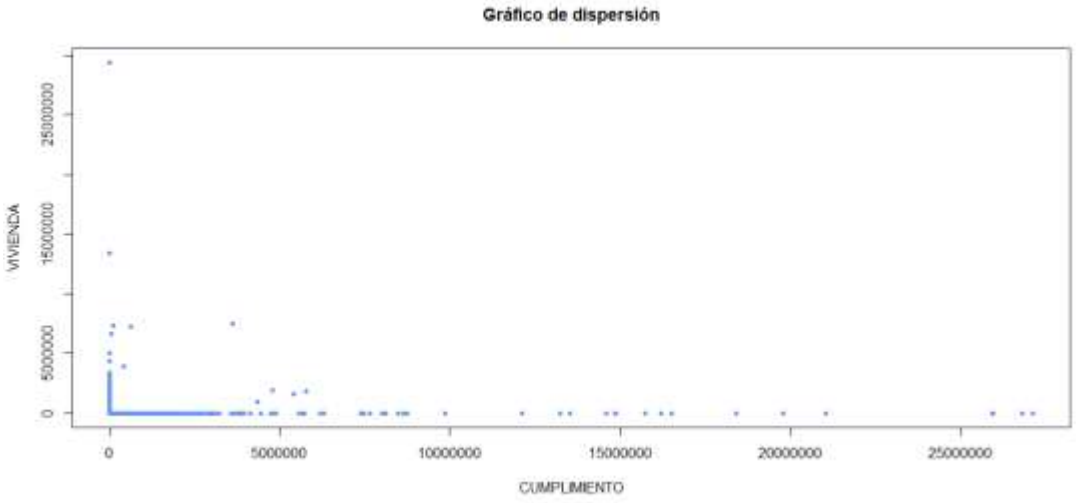
	CUMPL	CUMPL_CANT	VIVIENDA	VIVIENDA_CANT	PYME	PYME_CANT	MOVILIDAD	MOVILIDAD_CANT
CUMPL	1.00000	0.59443	-0.07041	-0.137925	0.119118	0.06754	0.15043	-0.18329
CUMPL_CANT	0.59443	1.00000	-0.08801	-0.184814	0.118733	0.01093	-0.01290	-0.26649
VIVIENDA	-0.07041	-0.08801	1.00000	0.608475	0.228683	0.35237	0.03572	0.33492
VIVIENDA_CANT	-0.13792	-0.18481	0.60848	1.000000	0.008214	0.12396	0.01695	0.41873
PYME	0.11912	0.11873	0.22868	0.008214	1.000000	0.62958	0.06761	-0.03448
PYME_CANT	0.06754	0.01093	0.35237	0.123960	0.629577	1.00000	0.26435	0.03445
MOVILIDAD	0.15043	-0.01290	0.03572	0.016946	0.067612	0.26435	1.00000	0.13437
MOVILIDAD_CANT	-0.18329	-0.26649	0.33492	0.418726	-0.034478	0.03445	0.13437	1.00000

Los gráficos de dispersión que se relacionarán junto con su justificación, apoyaron al entendimiento del comportamiento de cada una de las variables y combinación de las mismas, es importante tener en cuenta que se tomaron aquellas variables que contenían el valor en COP de la prima emitida en el período de tiempo que se estableció en el estudio:

En la figura 5, se observa el comportamiento de las primas para la variable de cumplimiento combinada con la variable de vivienda; para este caso sólo se visualiza un punto que cumple ambas variables.

Figura 5

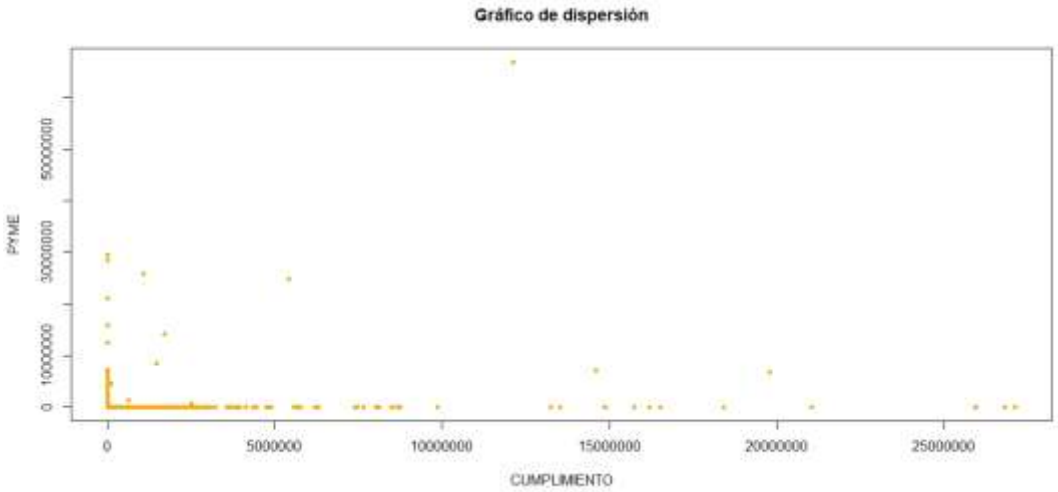
Gráfico de dispersión de las variables Cumplimiento y Vivienda



Para el caso de la figura 6, se visualiza que existen más puntos donde se comparte la emisión de ambas variables.

Figura 6

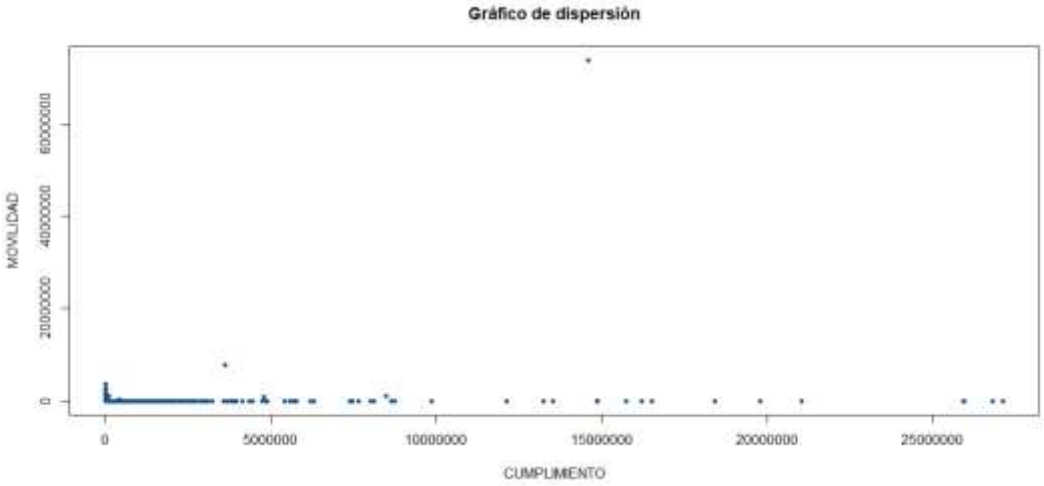
Gráfico de dispersión de las variables Cumplimiento y Pyme



En la combinación de las variables de cumplimiento y movilidad como lo representa la figura 7, sólo hay dos puntos que poseen ambas clases de seguros, esto quiere decir, que se debe tener en cuenta que es una combinación por explotar y aumentar en la emisión para el siguiente período a evaluar.

Figura 7

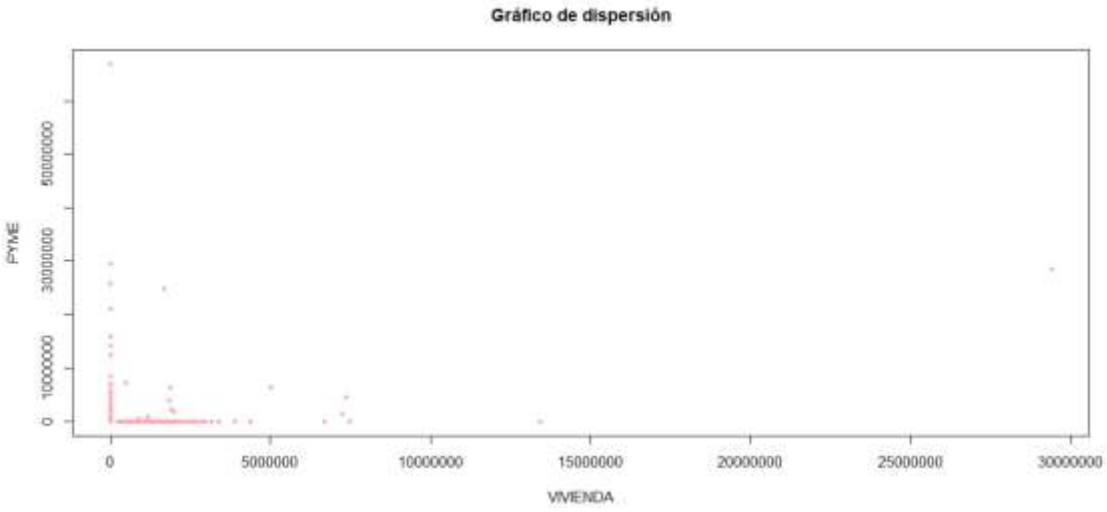
Gráfico de dispersión de las variables Cumplimiento y Movilidad



Para el caso de las variables de vivienda y pyme aquellos individuos que coinciden con ambas variables continúan siendo muy pocos, es decir, es otra combinación que puede ser incluida dentro de las estrategias de aumento de emisión de primas.

Figura 8

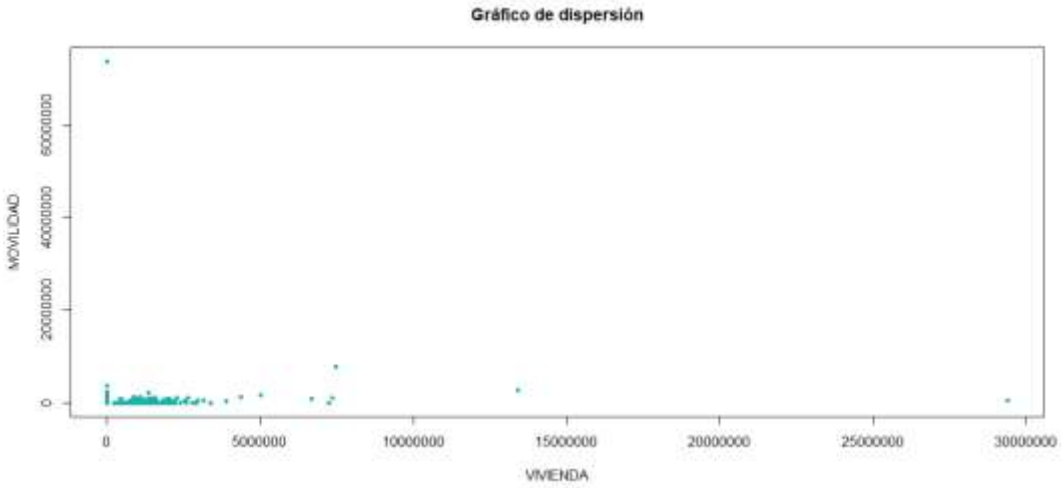
Gráfico de dispersión de las variables Vivienda y Pyme



En el ejercicio de la agencia se evidencia en la figura 9, la tendencia de que las emisiones de primas correspondientes a la variable de vivienda tienen cierta relación con la emisión de primas de la variable de movilidad en pequeños valores.

Figura 9

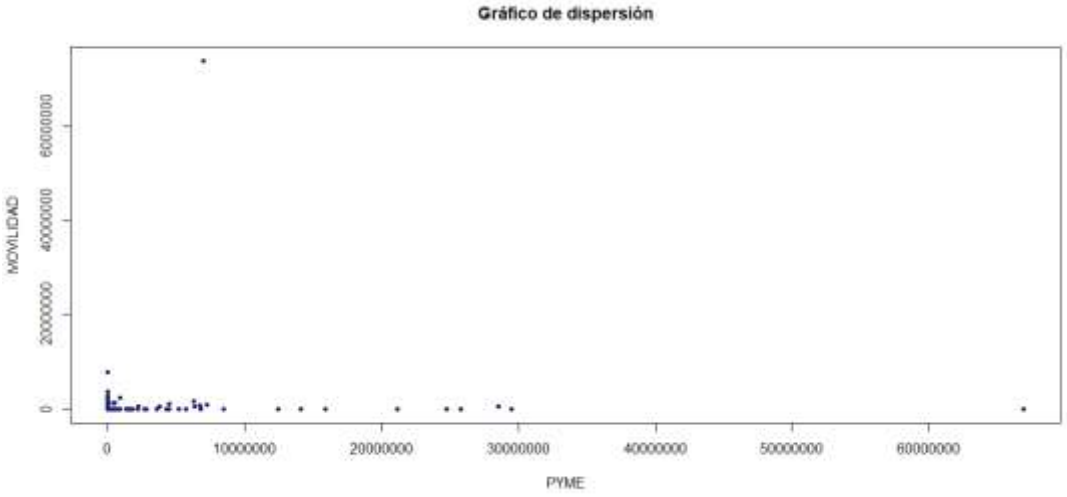
Gráfico de dispersión de las variables Vivienda y Movilidad



En el análisis de la combinación de las variables de pyme y movilidad, solo se observa un caso donde se comparte la emisión en ambas variables, para este caso, se puede considerar la integración de pólizas colectivas de movilidad en los clientes que emitieron primas en la variable de pyme, esto potencializa la fidelización con la empresa.

Figura 10

Gráfico de dispersión de las variables Pyme y Movilidad



Fase 2: Análisis de componentes principales

Para realizar el análisis de componentes principales se tomarán las variables de cumplimiento, cumplimiento cantidad, vivienda, vivienda cantidad, pyme, pyme cantidad, movilidad y movilidad cantidad. De acuerdo a la gráfica presentada y los resultados del código eig.val, se determinó las dimensiones que poseen mayor participación, son desde la dimensión 1 hasta la dimensión 5, dado que alcanza en su porcentaje acumulativo el 86.94%. Dicha información se puede visualizar en la gráfica Scree Plot.

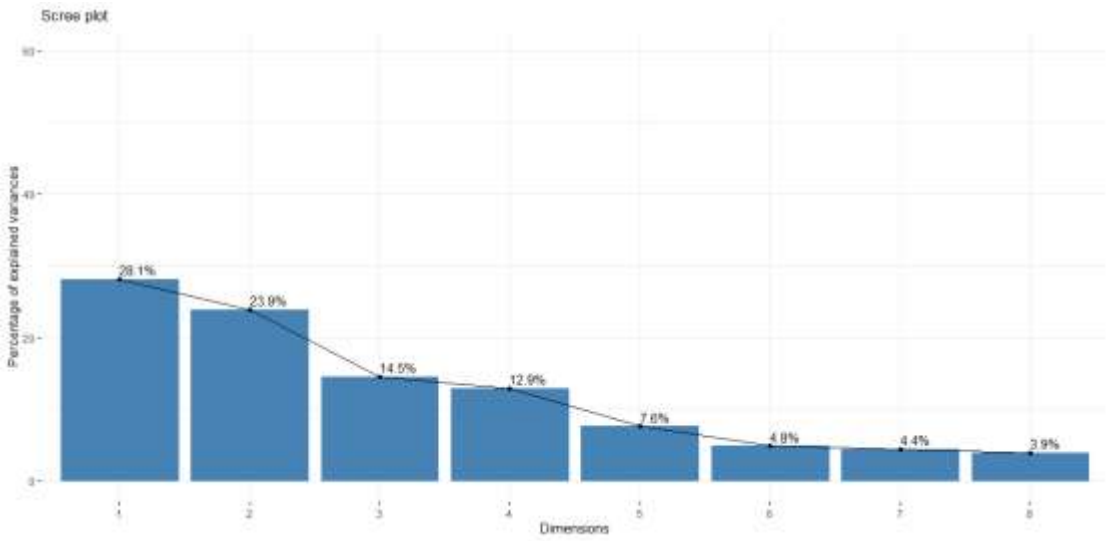
Figura 11

Resultados de los valores propios

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.2484	28.105	28.10
Dim.2	1.9081	23.851	51.96
Dim.3	1.1620	14.525	66.48
Dim.4	1.0324	12.905	79.39
Dim.5	0.6046	7.557	86.94
Dim.6	0.3851	4.814	91.76
Dim.7	0.3497	4.371	96.13
Dim.8	0.3098	3.872	100.00

Figura 12

Scree Plot



A continuación, se encuentra el detalle del análisis de componentes principales para las variables, es importante resaltar la calidad de la representación de los datos en cada una de las dimensiones, en el caso de la dimensión 1, la variable que es más representativa es vivienda, para la dimensión 2 es la variable de pyme, en la dimensión 3 se encuentra la variable de cumplimiento, en la dimensión 4 la variable pyme, y por último, en la dimensión 5 la variable de pyme se repite, en este caso no se tiene en cuenta movilidad cantidad dado que es una variable que sólo indica la cantidad de documentos emitidos y no como tal la prima emitida que es el objeto de estudio.

De acuerdo al resultado de la contribución, se concluye que en la dimensión 1 se encuentra la variable de vivienda, en el caso de la dimensión 2 la variable de pyme, se encuentra en la dimensión 3 la variable de cumplimiento, la variable de movilidad se encuentra representada en la dimensión 4, y para finalizar, en la dimensión 5 la variable pyme (para este caso se tiene el concepto anterior presente, es decir, el objeto de estudio es la emisión de primas).

Figura 13

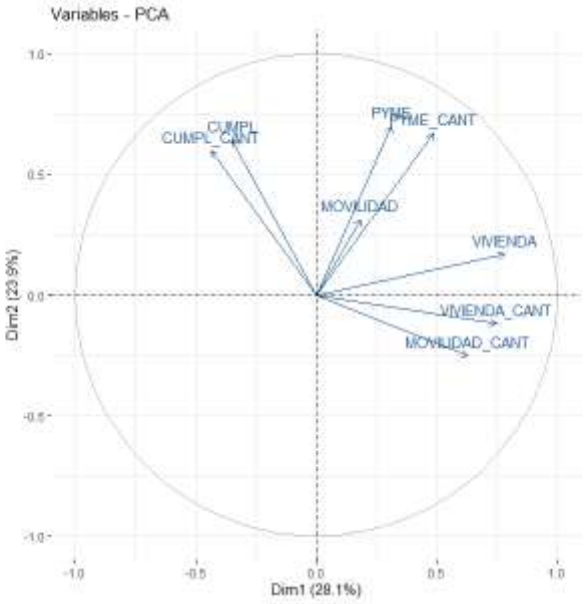
Resultados del análisis de componentes principales

```
> #Coordinates
> var$coord
                Dim.1   Dim.2    Dim.3     Dim.4    Dim.5
CUMPL         -0.3481  0.6403  0.51985  0.099747  0.06836
CUMPL_CANT    -0.4343  0.5969  0.48963 -0.151395  0.07498
VIVIENDA       0.7838  0.1670  0.27546 -0.242378 -0.22520
VIVIENDA_CANT  0.7488 -0.1201  0.42981 -0.142850 -0.23112
PYME           0.3117  0.7000 -0.40885 -0.232426  0.24564
PYME_CANT      0.4862  0.6696 -0.38434  0.007078 -0.03929
MOVILIDAD      0.1828  0.3115 -0.01953  0.889605 -0.21863
MOVILIDAD_CANT 0.6297 -0.2533  0.27591  0.273662  0.61680
> #Cos2: quality on the factore map
> var$cos2
                Dim.1   Dim.2     Dim.3     Dim.4    Dim.5
CUMPL         0.12119 0.41002 0.2702402 0.0099495 0.004673
CUMPL_CANT    0.18861 0.35627 0.2397379 0.0229205 0.005621
VIVIENDA      0.61428 0.02789 0.0758775 0.0587473 0.050717
VIVIENDA_CANT 0.56073 0.01442 0.1847335 0.0204060 0.053418
PYME          0.09715 0.49002 0.1671596 0.0540219 0.060337
PYME_CANT     0.23641 0.44834 0.1477193 0.0000501 0.001543
MOVILIDAD     0.03340 0.09702 0.0003816 0.7913969 0.047798
MOVILIDAD_CANT 0.39658 0.06414 0.0761288 0.0748907 0.380444
> #Contributions to the principal components
> var$contrib
               Dim.1   Dim.2    Dim.3     Dim.4   Dim.5
CUMPL          5.390 21.4883 23.25691  0.963743  0.7730
CUMPL_CANT     8.389 18.6714 20.63187  2.220153  0.9298
VIVIENDA      27.321  1.4614  6.53003  5.690456  8.3892
VIVIENDA_CANT 24.940  0.7556 15.89819  1.976596  8.8359
PYME           4.321 25.6808 14.38578  5.232734  9.9805
PYME_CANT     10.515 23.4964 12.71274  0.004853  0.2553
MOVILIDAD      1.485  5.0844  0.03284 76.657304  7.9064
MOVILIDAD_CANT 17.639  3.3617  6.55166  7.254162 62.9298
```

Se observa en la gráfica "Variables - PCA" la similitud del par de variables de Movilidad y Pyme en la dimensión 2; también se concluyó que no existe similitud entre las variables de Cumplimiento y Vivienda.

Figura 14

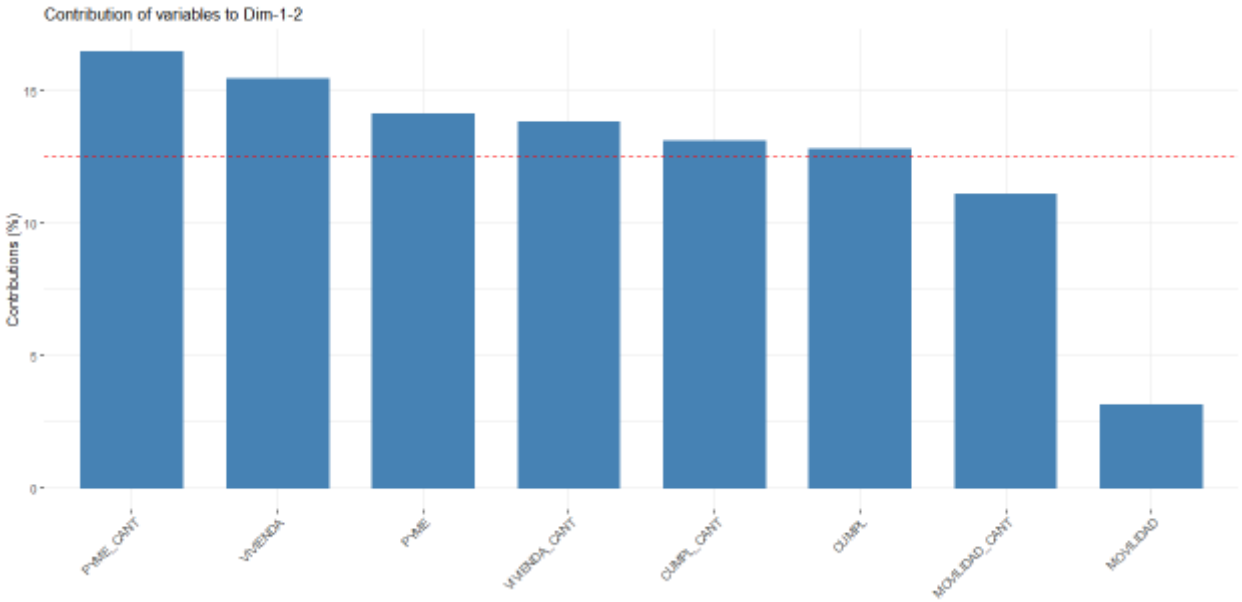
Variables - PCA



Se observa en la gráfica de contribución a las dimensiones 1 y 2 que las variables de vivienda, pyme y cumplimiento son aquellas que más aportan, lo cual confirma, el análisis anteriormente descrito.

Figura 15

Gráfico de contribución de las variables a las dimensiones 1 y 2

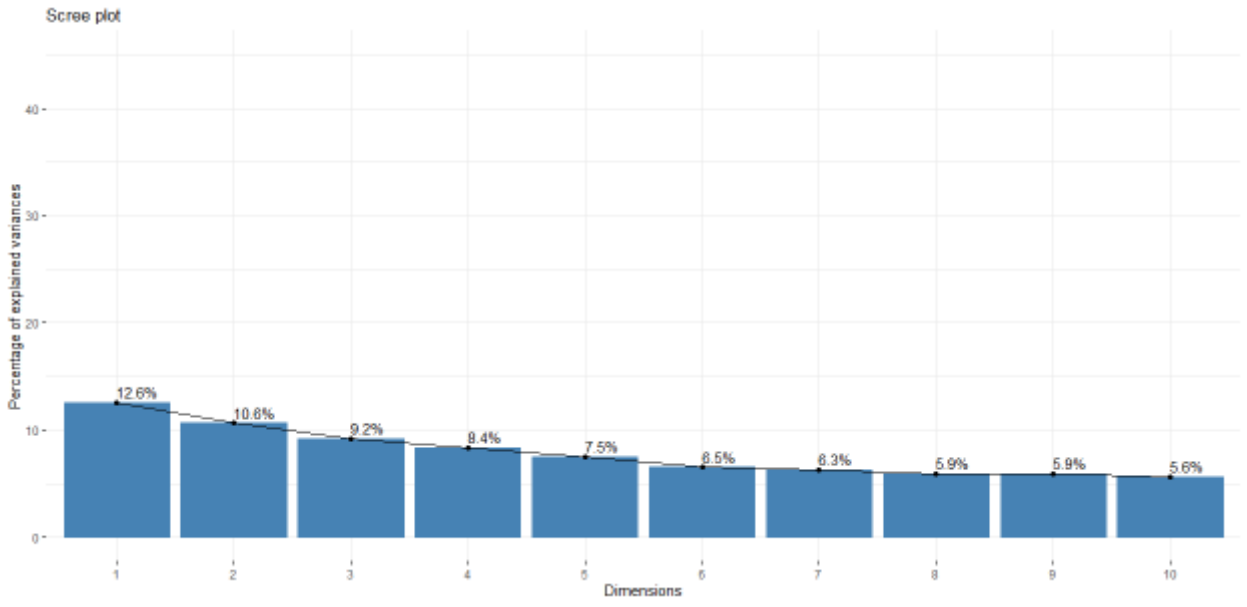


Fase 3: Análisis de correspondencias múltiples

Para iniciar con el análisis de correspondencias múltiples se seleccionar las variables a tener en cuenta, para este caso, se utilizaron las variables de clasificación de edades, sexo, sector económico y región; adicionalmente, los 150 primeros registros de la base de datos se seleccionaron como datos activos. Se realizó el gráfico Scree Plot con el fin de determinar cuáles eran las dimensiones que se iban a tener en cuenta y se definió tomar las cinco primeras dimensiones.

Figura 16

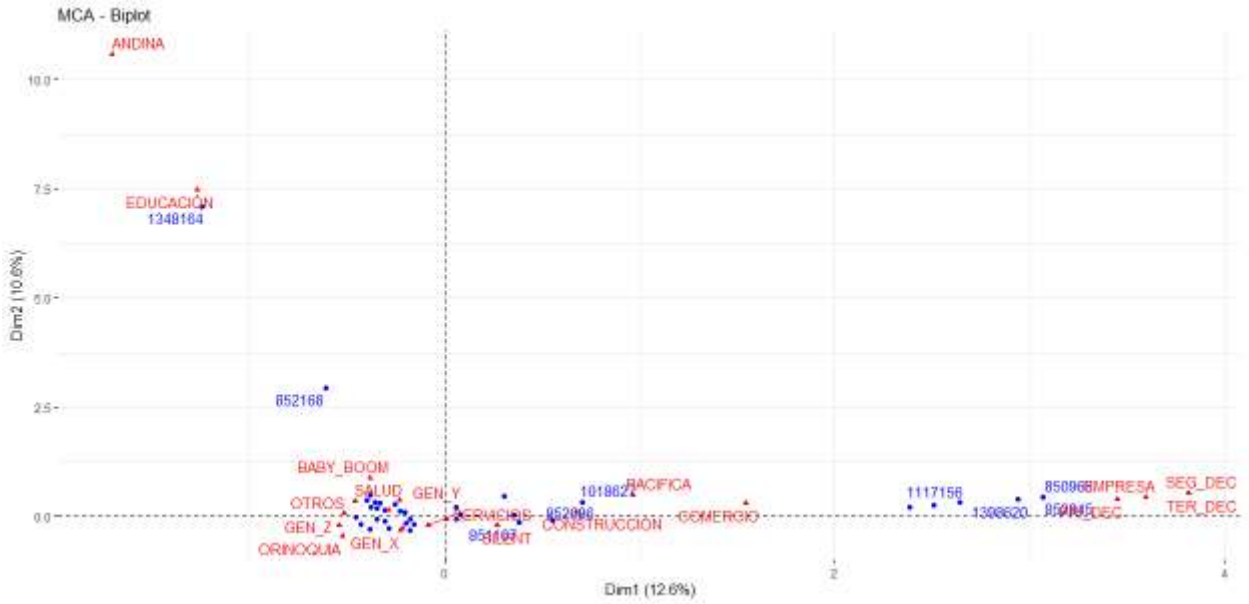
Scree Plot



En el Biplot que se encuentra a continuación se observa la cercanía de las empresas con la clasificación de primera, segunda y tercera década desde la constitución; de igual manera, la cercanía de los sectores económicos de comercio, construcción y servicios con la región pacífica.

Figura 17

Gráfico Biplot del análisis de correspondencias múltiples



En el análisis de correspondencias múltiples para las variables, se puede concluir lo siguiente con base en las figuras 18, 19 y 20, referente a los datos por dimensiones:

1. Se relacionan las coordenadas de cada una de las variables categóricas por dimensión.

   Figura 18

   Coordenadas de la variable de clasificación de edad

```
> # Coordinates
> var$coord
                Dim 1    Dim 2      Dim 3    Dim 4    Dim 5
BABY_BOOM   -0.381540  0.97645  0.4802032 -0.20835 -0.07168
GEN_X       -0.224910 -0.32087 -0.4118776 -0.09024 -0.31133
GEN_Y       -0.230345  0.36990  0.1648911  1.55001  1.34243
GEN_Z       -0.541139 -0.19244  2.8742218 -2.00908  0.99494
PRI_DEC      3.448222  0.38764 -0.4742004 -1.04561  1.56951
SEG_DEC      3.814352  0.53171  0.4383136  0.03471 -1.56334
SILENT       0.269832 -0.20445  3.8150406  0.76727 -2.42574
TER_DEC      3.814352  0.53171  0.4383136  0.03471 -1.56334
EMPRESA      3.594674  0.44527 -0.1091948 -0.61348  0.32637
FEMENINO    -0.217006 -0.27763  0.3622814  0.26849 -0.24434
MASCULINO   -0.286580  0.15256 -0.2580617 -0.12468  0.14221
COMERCIO     1.542568  0.30387  1.0588454  1.44533 -1.42887
CONSTRUCCIÓN 0.674379  0.10546 -0.6353689  0.27906  2.51528
EDUCACIÓN   -1.268457  7.48300 -0.8212220 -0.16849 -0.59939
OTROS       -0.515088  0.06654  2.5334877 -1.25896  0.51983
SALUD       -0.460484  0.35878  0.3999466  1.72164  2.07647
SERVICIOS   -0.084369 -0.19374 -0.3450812 -0.11025 -0.14931
ANDINA      -1.705017 10.57133 -1.0567047 -1.19432 -1.91521
BOGOTA D. C.  0.008608 -0.07220  0.0003376 -0.03948  0.02902
ORINOQUIA   -0.524610 -0.44964 -1.7937607 -0.77438 -1.15317
PACIFICA     0.964260  0.49198  2.8028364  7.77265 -1.19788
```

2. En el caso de la calidad de la representación de los datos en las diferentes dimensiones, se concluye que la variable que mejor se representa en la dimensión 1 es la categoría de empresa, para el caso de la dimensión 2 es el sector económico educación, en la dimensión 3 son aquellas actividades económicas que se clasificaron en Otros, en la dimensión 4 se encuentra la región pacífica, y para concluir el análisis de la calidad de representación, en la dimensión 5 se observa la presencia de la Generación Y.

   Figura 19

   Calidad en la representación de los datos de la variable de clasificación de edad en cada una de las dimensiones

```
> # Cos2: quality on the factore map
> var$cos2
                Dim 1     Dim 2       Dim 3      Dim 4     Dim 5
BABY_BOOM   0.0250203 0.1320291 0.039633536 0.00746087 0.0008831
GEN_X       0.0780199 0.1587995 0.261653067 0.01256006 0.1494970
GEN_Y       0.0076956 0.0198454 0.003943454 0.34846060 0.2613763
GEN_Z       0.0122040 0.0015431 0.344214630 0.16818339 0.0412458
PRI_DEC     0.4954264 0.0062610 0.009369418 0.04555435 0.1052731
SEG_DEC     0.1966119 0.0038204 0.002596201 0.00001628 0.0330273
SILENT      0.0009766 0.0005649 0.196682907 0.00795554 0.0795164
TER_DEC     0.1966119 0.0038204 0.002596201 0.00001628 0.0330273
EMPRESA     0.9229771 0.0141615 0.000851679 0.02688283 0.0077021
FEMENINO    0.0313943 0.0513848 0.087498559 0.04805640 0.0398006
MASCULINO   0.0938607 0.0266004 0.076109549 0.01776558 0.0231117
COMERCIO    0.1883069 0.0073072 0.088724386 0.16531554 0.1615716
CONSTRUCCIÓN 0.0189494 0.0004634 0.016820567 0.00324479 0.2636103
EDUCACIÓN   0.0217430 0.7566941 0.009113589 0.00038361 0.0048549
OTROS       0.0230709 0.0003850 0.558135668 0.13782497 0.0234976
SALUD       0.0088352 0.0053635 0.006664886 0.12350186 0.1796546
SERVICIOS   0.0217389 0.1146334 0.363679901 0.03712278 0.0680858
ANDINA      0.0195106 0.7500203 0.007522522 0.00957322 0.0246176
BOGOTA D. C. 0.0036307 0.2554425 0.000005585 0.07638502 0.0412721
ORINOQUIA   0.0018471 0.0013569 0.021594478 0.00402455 0.0089249
PACIFICA    0.0062402 0.0016245 0.052724107 0.40546330 0.0096303
```

3. En el análisis de la contribución de las variables a las diferentes dimensiones, se relacionó la mejor contribución de la siguiente manera:

   a. Dimensión 1: La categoría de empresa
   b. Dimensión 2: El sector económico de educación
   c. Dimensión 3: Aquellas actividades económicas que fueron clasificadas en el sector económico de Otros

d. Dimensión 4: La región Pacífica

e. Dimensión 5: El sector económico de construcción

Figura 20

Contribución de los datos de la variable de clasificación de edad en cada una de las dimensiones

```
> varScontrib
                 Dim 1    Dim 2        Dim 3     Dim 4    Dim 5
BABY_BOOM      0.99965  6.24113  2.159872458  0.448381  0.05905
GEN_X          1.43682  3.46007  6.572538864  0.347929  4.60782
GEN_Y          0.31467  0.96009  0.219939598 21.432463 17.88747
GEN_Z          0.54854  0.08206 21.103142844 11.370844  3.10280
PRI_DEC       22.26833  0.33296  0.574421152  3.079920  7.91938
SEG_DEC        9.08275  0.20881  0.163589455  0.001132  2.55356
SILENT         0.04512  0.03087 12.393205606  0.552813  6.14794
TER_DEC        9.08275  0.20881  0.163589455  0.001132  2.55356
EMPRESA       40.33341  0.73219  0.050764386  1.767055  0.56331
FEMENINO       0.88194  1.70790  3.352731156  2.030677  1.87130
MASCULINO      2.05082  0.68766  2.268258305  0.583881  0.84516
COMERCIO       8.17010  0.37510  5.250650662 10.788860 11.73251
CONSTRUCCIÓN   0.85174  0.02465  1.031236902  0.219380 19.83062
EDUCACIÓN      1.00445 41.35860  0.574257235  0.026656  0.37536
OTROS          0.99378  0.01962 32.792474153  8.930054  1.69400
SALUD          0.39712  0.28523  0.408611481  8.349935 13.51488
SERVICIOS      0.25106  1.56638  5.728967335  0.644896  1.31604
ANDINA         0.90741 41.27082  0.477205057  0.669717  1.91621
BOGOTA D. C.   0.00340  0.28301  0.000007134  0.107591  0.06468
ORINOQUIA      0.08591  0.07466  1.369885628  0.281547  0.69471
PACIFICA       0.29022  0.08939  3.344651135 28.365139  0.74962
```

Fase 4: Red Neuronal

Con la información previa, se tomó la variable de CANAL.VENTA, la cual fue la transformación de la variable inicial CL_EDAD; teniendo en cuenta que existe variabilidad entre las escalas de los valores máximos y mínimos, se procedió a utilizar la función *scale* para cumplir con el requerimiento de la red neuronal.

Figura 21

Resumen de la base de datos con los valores escalados

```
> summary(bd1)
     DIREC                CEL                EMAIL           CANAL.VENTA          SEXO          
 Min.   :-4.6811   Min.   :-1.7610   Min.   :-1.6411   Min.   :-1.2607   Min.   :-0.8956  
 1st Qu.: 0.2133   1st Qu.: 0.5671   1st Qu.:-1.6411   1st Qu.:-1.2607   1st Qu.:-0.8956  
 Median : 0.2133   Median : 0.5671   Median : 0.6086   Median : 0.7922   Median :-0.8956  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.2133   3rd Qu.: 0.5671   3rd Qu.: 0.6086   3rd Qu.: 0.7922   3rd Qu.: 1.4724  
 Max.   : 0.2133   Max.   : 0.5671   Max.   : 0.6086   Max.   : 0.7922   Max.   : 1.4724  
 SECTOR.ECONÓMICO       REGIÓN              CUMPL            CUMPL_CANT          VIVIENDA       
 Min.   :-1.5315   Min.   :-4.2207   Min.   :-0.3144   Min.   :-0.46964   Min.   :-0.2898  
 1st Qu.:-1.0399   1st Qu.: 0.1426   1st Qu.:-0.3144   1st Qu.:-0.46964   1st Qu.:-0.2898  
 Median : 0.4349   Median : 0.1426   Median :-0.3144   Median :-0.46964   Median :-0.2898  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
 3rd Qu.: 0.9264   3rd Qu.: 0.1426   3rd Qu.:-0.1502   3rd Qu.: 0.01384   3rd Qu.:-0.1110  
 Max.   : 1.4180   Max.   : 6.6877   Max.   : 8.8769   Max.   :13.06780   Max.   :20.6620  
 VIVIENDA_CANT          PYME             PYME_CANT          MOVILIDAD          MOVILIDAD_CANT    
 Min.   :-0.5165   Min.   :-0.1321   Min.   :-0.2326   Min.   :-0.099492   Min.   :-0.7094  
 1st Qu.:-0.5165   1st Qu.:-0.1321   1st Qu.:-0.2326   1st Qu.:-0.099492   1st Qu.:-0.7094  
 Median :-0.5165   Median :-0.1321   Median :-0.2326   Median :-0.099492   Median :-0.7094  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.000000   Mean   : 0.0000  
 3rd Qu.: 1.3065   3rd Qu.:-0.1321   3rd Qu.:-0.2326   3rd Qu.:-0.002676   3rd Qu.: 0.7131  
 Max.   : 8.5989   Max.   :19.6942   Max.   : 9.7067   Max.   :27.489680   Max.   : 6.4029  
```

Luego, se creó el conjunto de entrenamiento y de test con el fin de validar la calidad de clasificación que se desarrolló en la red neuronal. Como se observa en la figura 22, se tomó el 80% de los datos para el conjunto de entrenamiento y el restante, es decir, el 20% para el conjunto de test; se debe tener en cuenta que los datos asociados al número 1 corresponden a canal de venta tradicional y el número 2 al canal de venta digital.

Figura 22

Distribución de los canales de venta

```
> table(training_set$CANAL.VENTA)

  1   2 
237 387 
> table(test_set$CANAL.VENTA)

 1  2 
64 92 
```

Se utilizó el paquete “H2O” para la implementación dado que posee funciones para crear redes neuronales y modelos de Deep Learning. Luego de realizado el proceso de cargue y

predicción, se analizó la matriz de confusión que arrojó el modelo; para este caso, se observaron 156 predicciones correctas.

Figura 23

Matriz de confusión

```
> cm
          y_pred
y_test_set   1
         1 156
```

RESULTADOS

A continuación se relacionan los puntos clave que se resaltaron como resultados del desarrollo y análisis de este trabajo de grado:

- En el período 01-01-2020 a 31-12-2020 no se encontraron relaciones directas entre las primas emitidas por cada una de las variables.
- La variable de cumplimiento en la emisión de primas tiene participación del 45.3%.
- El canal de venta tradicional aplica tanto para las personas que se encuentran clasificadas dentro de los grupos de generaciones BABY_BOOM, SILENT, GENERACIÓN X.
- El canal de venta digital fue dirigido más para aquellos clientes que se encuentran clasificados en las generaciones restantes dado que poseen mayor conectividad.
- Se obtuvieron 156 predicciones correctas, lo cual debe aportar en la generación de estrategias para fortalecer las ventas en los diferentes canales de ventas.

CONCLUSIONES

Se realizó la recomendación de la actualización periódica de datos de los diferentes clientes con el fin de tener mejores pronósticos e indicar cuál es el mejor canal de venta para cada uno.

Adicionalmente, se sugirieron a la empresa analizada las siguientes opciones de estrategia para aumentar las ventas por cada uno de los canales:

Opción 1 – Canal de venta tradicional: Reforzar con aumento de personal dedicado a la atención y desarrollo de las capacidades del personal actual, esto mejorará el indicador de satisfacción del cliente y a su vez, aumento en las ventas.

Opción 2 – Canal de venta digital: Desarrollo de las capacidades y medios tecnológicos para el aumento de atención de casos por medio del canal digital; esta alternativa lo que ocasionará es potencializar el conocimiento del trámite en tiempo real y estandarización de los tiempos de respuesta.

REFERENCIAS BIBLIOGRÁFICAS

Matich, D. (2001). Redes neuronales: Conceptos básicos y aplicaciones. Rosario.

Carvajal, P. (2003). Estudio del pronóstico de la demanda de energía eléctrica, utilizando modelos de series de tiempo. Scientia et Technica Año IX, No 23, Diciembre 2003. UTP. ISSN 0122-1701.

ATRIA Innovation (2019). Qué son las redes neuronales y sus funciones. Recuperado de https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

Pérez, F y Fernández, H. (2007). Las redes neuronales y la evaluación del riesgo de crédito. Rev. ing. univ. Medellin vol.6 no.10 Medellín Jan./June 2007. On-line version ISSN 2248-4094.

Calvo, D. (2019). Aprendizaje no supervisado. Recuperado de https://www.diegocalvo.es/aprendizaje-no-supervisado/

Toro, E., Mejía, D., y Salazar, H. (2004). Pronóstico de venta usando redes neuronales. Scientia Et Technica, vol. X, núm. 26, diciembre, 2004, pp. 25-30. ISSN: 0122-1701

Roman, V. (2019). Aprendizaje no supervisado en Machine Learning: Agrupaciones. Recuperado de https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc

Picazo, A. (1995). Eficiencia en los Seguros. Revista de economía aplicada Número 8 (vol. III), 1995, págs. 197 a 215.

Data Science con R. Recuperado de https://bookdown.org/dparedesi/data-science-con-r/aprendizaje-no-supervisado.html

Piedrahita, G. (2020). Gestión del conocimiento en la gerencia Proyecto Core Seguros Generales de Seguros Sura Colombia. Recuperado de https://repository.eafit.edu.co/handle/10784/24153

López D. y Fernández A. (2018). Aplicación en los medios de prensa de un agrupamiento K-Means (Clustering K-Means). Recuperado de https://rches.utem.cl/articulos/aplicacion-en-los-medios-de-prensa-de-un-agrupamiento-k-means-clustering-k-means/