

PROYECTO INTEGRADOR 2.

**PREVENIR NIVELES ALTOS DE ROTACIÓN DE KONECTA DESDE LA
ANALÍTICA**

ESTUDIANTES:

DIANA C. VELAZQUES GAVIRIA

ID: 43.872.224

SEVASTIAN MORENO ZAPATA

CRISTIAN D. MUÑOZ MORA

ID: 1.020.417.894

ID: 1.152.458.017

JULIAN CASTELBLANCO B.

JUAN C. CEBALLOS ARIAS

ID: 1.152.189.889

ID: 1.037.614.958

MAESTRIA EN CIENCIA DE LOS DATOS Y ANALÍTICA

UNIVERSIDAD EAFIT

17/11/2019

Medellín

La especialización de BPO es el desarrollo de proyectos que incrementen la productividad y rendimiento de los negocios de nuestros clientes, para así obtener una ventaja competitiva en su mercado.

Resumen

El presente trabajo se enmarca en una de las empresas de servicios integrales de BPO y contact center a nivel mundial. Aquí se detallaran actividades realizadas por la organización en los últimos años, y especialmente su gestión humana, donde se puede apreciar un volumen importante de personal, que tiene condiciones no siempre favorables para el trabajador, lo que ocasiona índices de rotación bastante altos. Posteriormente, se estudiará la problemática de la rotación de personal con todas sus implicaciones, analizando desde diferentes variables como el salario, adherencia a los horarios y turnos de trabajo, edades de los asesores, encuestas de retiro dentro de la compañía. Siendo el centro de análisis del documento, con la intención de diagnosticar si dicho modelo es o no el más conveniente para la organización, analizando las bases de datos de las variables anteriormente mencionadas; generando un punto de partida a la insatisfacción del personal que abandona la compañía.

Palabras Claves: Rotación de personal, adherencia, variable, análisis, desempeño.

Abstract

The given document is part of a global BPO and contact center services company. Here are activities carried out by the organization in recent years and especially its human management, where you can see a significant volume of personnel, which has conditions not always favorable for the worker, which causes high attrition rates. Subsequently, the problem of personnel attrition will be studied with all its implications, the analysis of variables such as salary, adherence to schedules and work, the ages of the representatives, the quit surveys within the company. Being the center of analysis of the document, with the intention of diagnosing and saying whether the model is or is not the most convenient for the organization, analyzing the data bases of the variables previously mentioned; generating a point of departure for the dissatisfaction of the personnel leaving the company

Keywords: Staff rotation, adherence, variable, analysis, performance.

1. Introducción

Los altos niveles de rotación generan problemas de servicio (velocidad y calidad) no deseados por la organización, los clientes corporativos y los usuarios finales del servicio. La rotación de los empleados operativos requiere la asignación de altos costos de reemplazo, los cuales afectan los resultados financieros de cada uno de los clientes y el de la compañía; adicional a la dedicación que cada una de las áreas debe invertir para mitigar o controlar este indicador y conociendo la importancia del activo del conocimiento en propiedad de estos empleados e incluso del impacto que este proceso provoca en la calidad del servicio y ambiente laboral en su interior. Para suplir el gran número de solicitudes de asesores, contratan personal en grandes cantidades, teniendo en cuenta que buena parte de ese personal abandonará la operación más pronto que tarde. Específicamente, en este sector, ha sido la constante desde hace varios años como lo manifiesta (Morgan), es uno de los "...factores que afectan la capacidad de cumplir las métricas establecidas de producción, teniendo un efecto económico y comercial intangible e inmediato en el desempeño financiero de la industria y causa inestabilidad en el ambiente de trabajo".

Por lo tanto el objetivo de este documento es analizar la rotación de asesores en un servicio de Konecta Colombia, con el fin de identificar los causales de retiro e identificar las variables que más influencia tienen en este fenómeno, segmentar los asesores en diferentes grupos de riesgo y finalmente predecir las probabilidades de que un asesor permanezca un al menos 12 meses y que características presentan aquellos que se retiran antes de dicho tiempo.

Para llevar a cabo el análisis se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).

Fase I. Business Understanding. Definición de necesidades del cliente (comprensión del negocio)

La necesidad que se tiene es la de tener un diagnóstico del la rotación de asesores para un servicio particular de Konecta, con el fin de entender los drivers que apalancan el indicador, poder predecir que agentes tienen la mayor probabilidad de retirarse y así poder tomar decisiones concretas que mitiguen la rotación no deseada fomentando la permanencia en la compañía.

Fase II. Data Understanding. Estudio y comprensión de los datos

Para llevar a cabo el análisis, se obtuvo una base de datos de Konecta que contiene el histórico de personas retiradas de los últimos 3 años. (2017 a 2019). Esta base de datos tiene un total de 2669 registros y 35 variables, entre las que se encuentra información personal, sociodemográfica, desempeño laboral e ingresos del empleado. Adicionalmente a partir de resultados históricos de salario y desempeño se crearon nuevas variables como: Devengado promedio, devengado máximo, devengado mín, y de esta misma manera con las variables de salario variable y deducciones. También se calcularon el total de meses en los que se le midió el desempeño y las probabilidades de quedar en los diferentes niveles (Sobresaliente, en objetivo, fuera de objetivo y deficiente), entre otras.

Primero que todo se realizó un entendimiento general de los datos. Para esto se realizó una visualización inicial y un resumen general de las variables de la base como se observa a continuación en la tabla 1.

	cedula	mes_baja	dia_baja	xt_days	xt_month	edad_	devengado_mean	devengado_sd	devengado_median	devengado_min
count	2.669000e+03	2669.000000	2669.000000	2669.000000	2669.000000	2669.000000	2.665000e+03	2659.000000	2.665000e+03	2.598000e+03
mean	1.407878e+12	6.260397	14.536905	492.596103	17.201574	25.152866	8.116172e+05	313362.383894	7.852886e+05	3.357984e+05
std	3.633440e+13	3.325091	8.489424	524.975418	17.278840	4.741110	1.453937e+05	104533.644048	1.162076e+05	2.196920e+05
min	4.145180e+05	1.000000	1.000000	14.000000	1.000000	18.000000	1.420830e+05	11486.949660	1.420830e+05	1.000000e+00
25%	1.020432e+09	3.000000	7.000000	122.000000	5.000000	22.000000	7.267270e+05	246500.449700	7.455295e+05	1.423235e+05
50%	1.037650e+09	6.000000	15.000000	283.000000	10.000000	24.000000	8.237004e+05	305820.068400	7.974040e+05	3.253150e-05
75%	1.097995e+09	9.000000	21.000000	721.000000	25.000000	27.000000	9.107278e+05	370963.619900	8.468660e+05	5.169488e+05
max	9.720000e+14	12.000000	31.000000	4399.000000	145.000000	59.000000	1.530132e+06	772567.801700	1.457971e+06	1.029325e+06

Tabla 1. Descriptiva de los datos.

deducción_sd	deducción_média	deducción_min	deducción_max	CONTEO MESES DE GARANTIZADO	CONTEO MESES DE DESEMPEÑO	PROP SOB	PROP OBJ	PROB FO	PROB DEF
2658.000000	2665.000000	2665.000000	2.665000e+03	2665.000000	2665.000000	2561.000000	2561.000000	2561.000000	2561.000000
59978.847875	124768.079174	38801.807129	2.529874e+05	0.278424	7.748593	0.192821	0.148619	0.206328	0.452232
49959.535869	92963.269554	51765.732997	2.321621e+05	1.188567	7.476127	0.302277	0.211866	0.267925	0.358475
313.955411	5006.000000	-55156.000000	5.006000e+03	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
23484.349385	62476.000000	9624.000000	8.990800e+04	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000
42843.555520	89140.000000	24090.000000	1.426480e+05	0.000000	5.000000	0.000000	0.050000	0.105263	0.500000
83516.458668	157706.000000	47296.000000	3.680330e+05	0.000000	11.000000	0.250000	0.250000	0.333333	0.764706
441190.799600	869011.000000	407492.000000	2.214723e+06	18.000000	31.000000	1.000000	1.000000	1.000000	1.000000

Tabla 2. Continuación descriptiva de los datos.

Adicionalmente se analiza si las variables son numéricas o categóricas para hacer las transformaciones iniciales pertinentes como imputación de datos, estandarización y conversión a indicadoras de las variables categóricas.

Dentro de las variables categóricas se tienen las siguientes en el grafico 1:

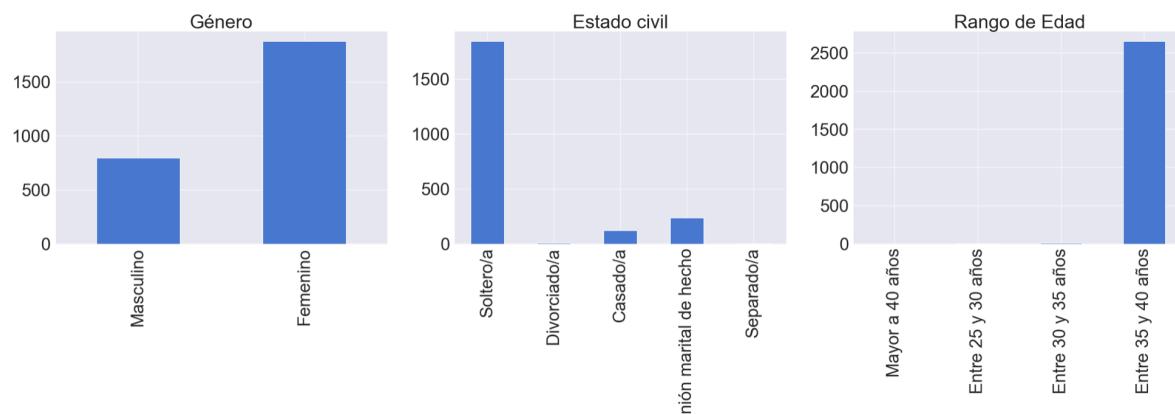


Grafico 1. Variables categoricas

Adicionalmente una vez el empleado se retira diligencia una encuesta donde tipifica el causal del retiro. A continuación se analiza esta variable en el grafico 2:

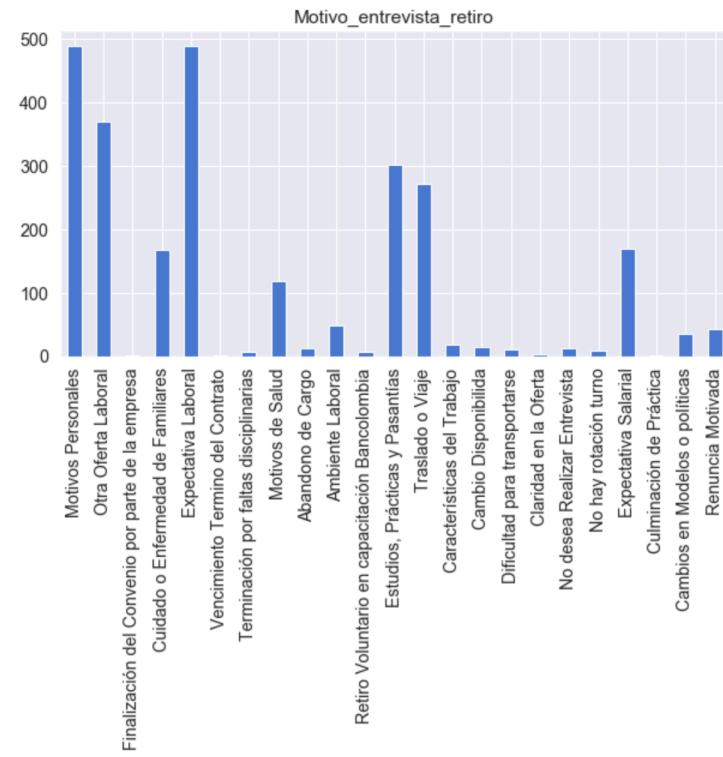


Grafico 2. Motivo retiro.

Se observa que los motivos más frecuentes de retiro voluntario hacen referencia a motivos personales, otra oferta laboral, expectativas laborales, estudios, traslados o viajes y expectativa salarial.

Mediante una prueba de Tukey (SCHEFFÉ, 1953) se evalúa en cada variable cuáles de las clases son significativas para la conversión a variables dummy y se obtiene que sólo Estado Civil y Género se deben categorizar.

Posteriormente y con el fin de poder predecir la probabilidad de permanencia de los agentes a diferentes períodos de tiempo se crearon 4 variables respuesta a partir de la variable “x_month” que representa la cantidad de meses que el empleado estuvo en la compañía. Con las probabilidades de permanencia baja (inferior a 3 meses) se pueden generar alertas para que el área de recursos humanos pueda preparse y desde un inicio contratar personas de holgura para dar cubrimiento a los requerimientos, adicionalmente con las probabilidades de permanencia superiores a 12 meses, se pueden identificar características de los empleados para tener en cuenta en el ciclo de vida del agente en la compañía y poder identificar con tiempo quienes si no cumplen ciertas condiciones tendrían alta probabilidad de retirarse. El comportamiento de dichas variables es el siguiente en el grafico 3:

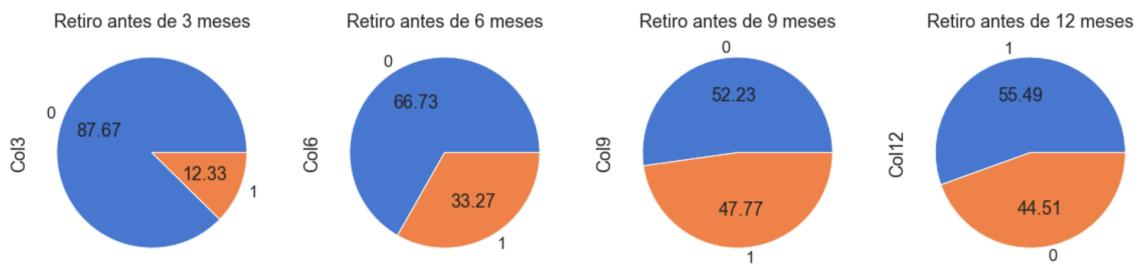


Grafico 3. Cantidad de meses empleado dentro de la compañía.

Se observa como disminuye la probabilidad de permanencia a medida que aumentan los meses, siendo muy similar a los 9 y a los 12 meses.

Con base en esto, se analiza la media de cada una de las variables numéricas de las personas que se retiran y que permanecen en estos períodos de tiempo:

	cedula	mes_baja	dia_baja	xt_days	xt_month	edad_	devengado_mean	devengado_sd	devengado_median	devengado_min	...
Col3											
0	1.190305e+12	6.218803	14.116667	553.470085	19.224359	25.194017	837330.911758	314973.605067	802854.089003	333183.786374	...
1	2.955359e+12	6.556231	17.525836	59.632219	2.814590	24.860182	628406.737294	301668.520912	660134.504573	354213.941176	...
2 rows x 36 columns											
<code>datadummy.groupby('Col6').mean()</code>											
	cedula	mes_baja	dia_baja	xt_days	xt_month	edad_	devengado_mean	devengado_sd	devengado_median	devengado_min	...
Col6											
0	1.563587e+12	6.173498	14.322291	690.750702	23.733296	25.480629	872556.735515	323442.424293	815559.734814	331428.163783	...
1	1.095583e+12	6.434685	14.967342	95.171171	4.101351	24.495495	689463.218564	233019.237664	724609.825254	344569.190972	...
2 rows x 36 columns											
<code>datadummy.groupby('Col9').mean()</code>											
	cedula	mes_baja	dia_baja	xt_days	xt_month	edad_	devengado_mean	devengado_sd	devengado_median	devengado_min	...
Col9											
0	6.709949e+11	6.176471	14.429699	824.659254	28.132712	25.704448	893116.785963	329693.694586	823350.279612	332602.548126	...
1	2.213537e+12	6.352157	14.654118	129.540392	5.250196	24.549804	722364.833149	295392.782161	743606.276730	339314.590946	...
2 rows x 36 columns											
<code>datadummy.groupby('Col12').mean()</code>											
	cedula	mes_baja	dia_baja	xt_days	xt_month	edad_	devengado_mean	devengado_sd	devengado_median	devengado_min	...
Col12											
0	7.871713e+11	6.335017	14.330808	915.479798	31.122054	25.964646	902376.439175	330700.081784	828604.195451	331439.509450	...
1	1.905784e+12	6.200540	14.702228	153.375422	6.035111	24.501688	738727.278288	299381.509305	750501.314953	339336.560669	...

Tabla 3. Media de los devengados asesores.

En la tabla 3, puede observarse las diferencias en los diferentes rangos de permanencia, por ejemplo lo sucedido con el salario devengado, en todos los rangos el salario es inferior para las personas retiradas que para las que permanecen más tiempo, estando entre 630 mil y 680 mil los que permanecieron

hasta los 3 y 6 meses y con una mayor variabilidad en esos mismos rangos de permanencia (ver devengado sd)

Y con respecto a las otras variables:

PROB DEF	cat_estado_civil_Casado/a	cat_estado_civil_Otros	cat_estado_civil_Soltero/a	cat_estado_civil_Unión marital de hecho	cat_genero_Femenino	cat_genero_Masculino
0.479176	0.048291	0.002564	0.717094	0.083333	0.701709	0.298291
0.217803	0.030395	0.006079	0.510638	0.127660	0.708207	0.291793
<code>datadummy.groupby('Col6').mean()</code>						
PROB DEF	cat_estado_civil_Casado/a	cat_estado_civil_Otros	cat_estado_civil_Soltero/a	cat_estado_civil_Unión marital de hecho	cat_genero_Femenino	cat_genero_Masculino
0.553629	0.051656	0.002807	0.724312	0.083661	0.715890	0.284110
0.233436	0.034910	0.003378	0.626126	0.099099	0.675676	0.324324
<code>datadummy.groupby('Col9').mean()</code>						
PROB DEF	cat_estado_civil_Casado/a	cat_estado_civil_Otros	cat_estado_civil_Soltero/a	cat_estado_civil_Unión marital de hecho	cat_genero_Femenino	cat_genero_Masculino
0.612758	0.050933	0.002869	0.727403	0.081062	0.725968	0.274032
0.268160	0.040784	0.003137	0.652549	0.097255	0.676863	0.323137
<code>datadummy.groupby('Col12').mean()</code>						
PROB DEF	cat_estado_civil_Casado/a	cat_estado_civil_Otros	cat_estado_civil_Soltero/a	cat_estado_civil_Unión marital de hecho	cat_genero_Femenino	cat_genero_Masculino
0.638749	0.047138	0.003367	0.728956	0.081650	0.728956	0.271044
0.296824	0.045240	0.002701	0.661715	0.094531	0.681296	0.318704

Tabla 4. Probabilidad de desempeño.

Los que llevan 3 meses, tienen una probabilidad de quedar en su desempeño en deficiente de 0.47, con respecto a los que llevan al menos 12 meses donde dicha probabilidad aumenta 0.63. Adicionalmente se observa que las personas que duraron menos de 12 meses tienen una menor probabilidad (de 0.296) de estar en deficiente que las que permanecieron más de ese tiempo. (de 0.638), lo que pareciera ser que en esa antigüedad las personas no se van por esta causa.

Tambien se realizan boxplot para detectar outliers y conocer la forma de distribución de las variables:



Grafica 4. Distribución y detección de datos atípicos.

Y de la misma manera los histogramas para conocer mejor las variables analizadas:



Grafico 5. Comportamiento de variables en análisis.

Fase III. Descriptive statistics. Estadística de los datos.

Inicialmente se quiere observar el comportamiento de los datos a travez de los diferentes año desde el 2017 hasta el 2019 en gestión y cómo ha venido cambiado sus indicadores. Con esto, en el grafico 6 se observa como ha venido cambiando el indicador de bajas, en este se quiere resaltar de igual forma el servicio que hace parte de las bajas generales de la compañía.

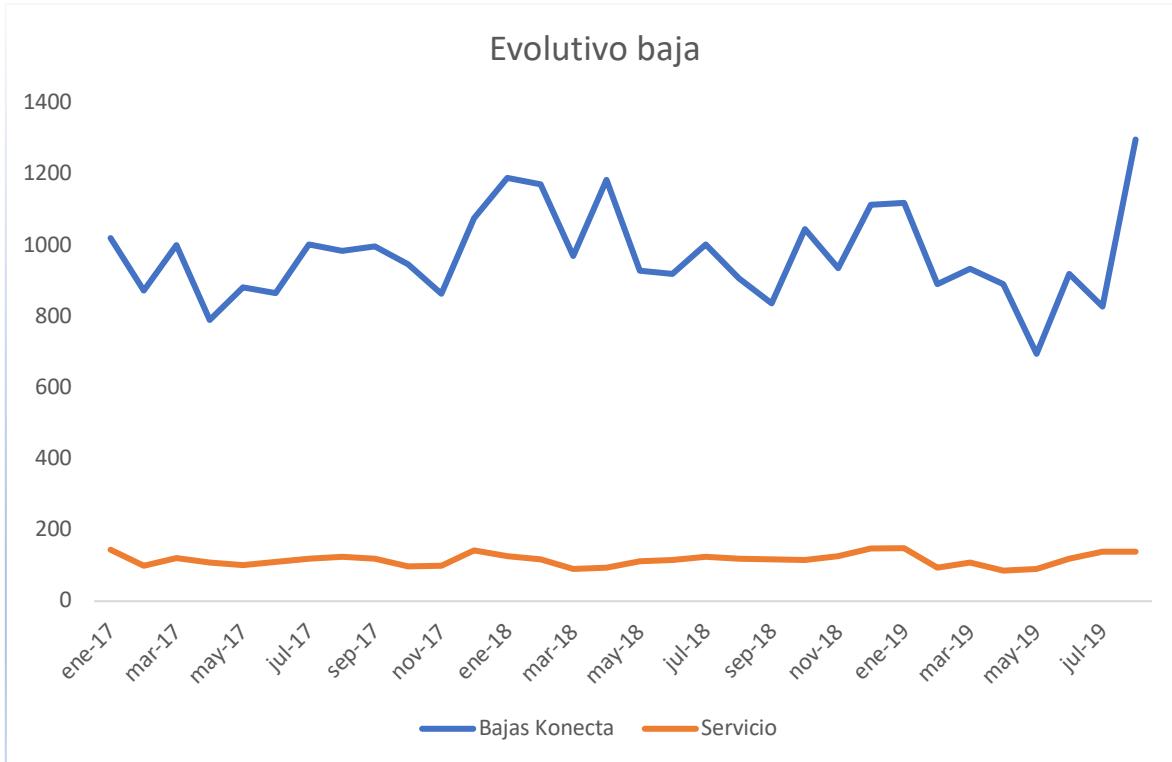


Grafico 6. Evolutivo de bajas en Konecta y bajas en el servicio.

En el grafico anterior, se puede observar como el servicio tiene una gran participación en el indicador general de bajas en la compañía significando el 12% en promedio de este. Donde, el promedio de bajas en Konecta se refleja 971 promedio de empleados en baja mensual y el servicio con un 117 promedio mensual. Si esto es llevado a un indicador de rotación de la compañía, se tiene que konecta tiene un procedimiento del 4,74% de nivel de rotación promedio desde enero de 2017 hasta julio de 2019. Como se puede observar en el grafico 7.

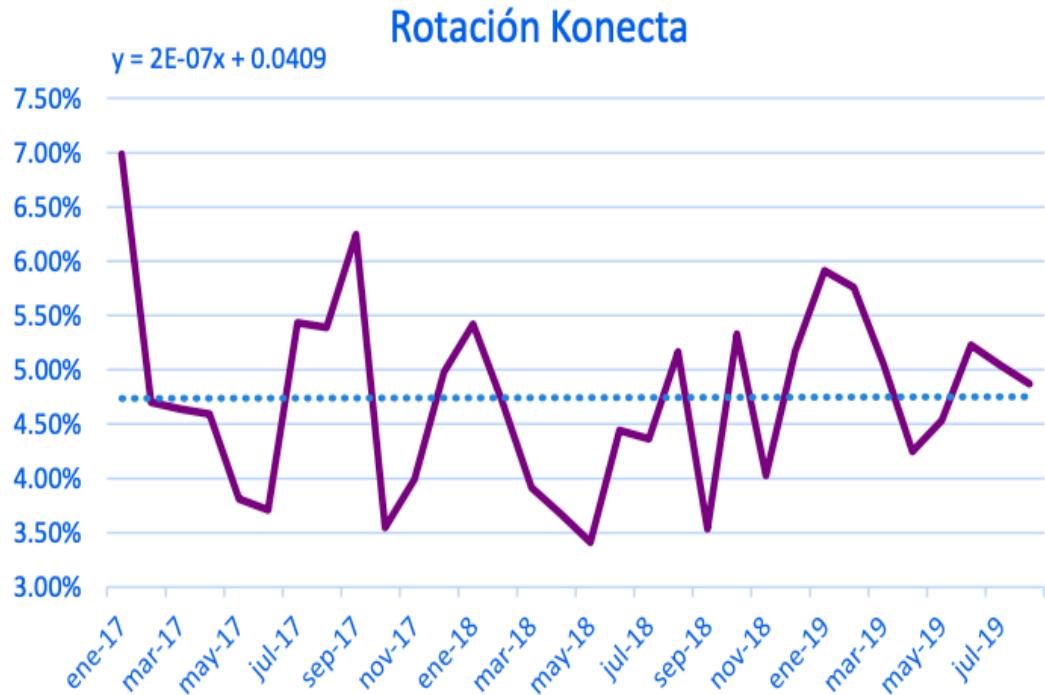


Grafico 7. Nivel de rotación konecta

Según (Ponce, 1995) la rotación de personal se puede presentar tanto interna como externamente, es decir, de forma interna es cuando la misma organización promueve o impulsa a que sus empleados de la entidad cubran las ausencias que estos dejan o cuando es necesario realizar modificaciones en el puesto actual y de manera externa, cuando el personal por alguna razón toma la decisión de romper relaciones con la compañía.

Debido a esto, existen varias formas de que el empleador corte relaciones laborales con la compañía, debido a esto, es que se tiene este indicador.

A pensar de esto, por ser una compañía BPO, se tiene estacionalidad en el año de acuerdo al cliente y sus necesidades. Por ende, se debe de manejar un control de gestión capaz de mantener operativas todas las operaciones de la compañía. Donde cada mes del año tiene su requerimiento. Con esto, se quiso analizar la estacionalidad de las bajas dentro de los meses del año. Para de esta forma evaluar su comportamiento.

Resumen Mensual

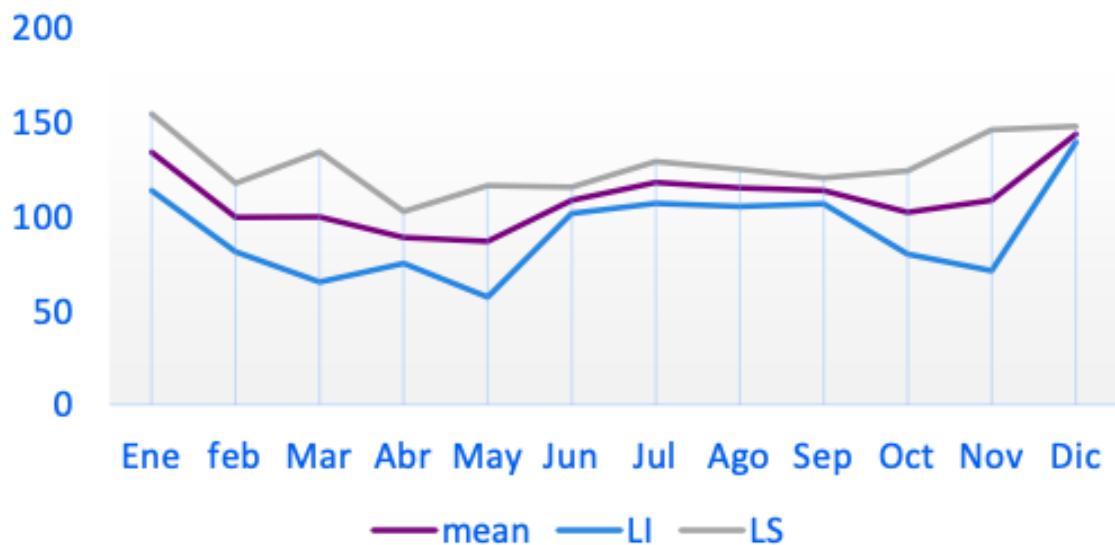


Grafico 8. Estacionalidad en los meses del año.

Particularmente los meses de diciembre y enero, tienen promedios de Bajas mucho mas grandes que el resto del año, durante los meses de junio a septiembre las bajas en volúmenes durante el periodo de tiempo analizado parece ser constante. En la grafica, tambien se puede observar sus limites inferiores y superiores, bajados en intervalos de confianza donde se puede observar que Mayo tiene una alta desviación de su media en el limite inferior.

Mes	mean	sd	IQR	Min	25%	Median	75%	Max	Sum
Ene	134	10	9	122	131	140	140	140	420
feb	99	9	9	93	94	95	103	110	312
Mar	100	18	18	81	92	102	109	116	320
Abr	89	7	7	82	85	88	92	96	289
May	87	30	28	53	76	98	104	109	270
Jun	109	4	3	106	107	109	110	111	226
Jul	118	6	4	114	116	118	120	122	245
Ago	115	5	0	115	115	115	115	115	245
Sep	114	4	3	111	112	114	115	116	238
Oct	102	11	8	94	98	102	106	110	212
Nov	109	19	14	95	102	109	115	122	226
Dic	144	2	2	142	143	144	144	145	291

Tabla 5. Estacionalidad mensuales.

En la tabla se observa el comportamiento de los meses con su descriptiva, donde como se dijo anteriormente, parece ser un indicador constante.

Donde al final, los motivos que apalancan el indicador de bajas son los siguientes:

Motivos de baja	Media	Desviación	IQR	Min	25%	Mediana	75%	Max	n
Renuncia Voluntaria	523	548	620	9	133	313	753	4399	2983
Despedido Justa Causa	550	477	532	63	207	392	739	2879	289
Terminación de contrato	420	436	183	120	182	243	365	2680	60
Abandono de Cargo	276	365	199	47	82	134	280	1515	27
Despido sin justa Causa	424	325	317	64	196	356	512	1060	16
Fallecimiento	984	1126	796	188	586	984	1382	1780	2

Tabla 6. Motivos bajas.

De estos motivos, el que se quiere entrar analizar, son las renuncias voluntarias ya que estas, son las que la compañía no tiene como controlarlas, ya que son decisiones propias del trabajador fuera de la compañía. Estas, representan un alto costo para compañía y por ende alteran el indicador de bajas de la compañía y al final empezar otra curva de aprendizaje con un nuevo asesor.

Debido a esto, se analizaron los periodos en los que los asesores renuncia a la compañía por voluntad propia. Donde se quiso enfatizar en los meses que la relación entre la compañía y el asesor se vuelve robusta.

- 0 a 3 meses: El asesor puede estar en periodo de prueba, las condiciones pactadas inicialmente no fueron cumplidas por el empleador, el trabajo no es el esperado, ambiente laboral. Estos meses son muy importantes para la compañía, ya que el asesor tiene un tiempo de entrenamiento.
- 3 a 6 meses: El asesor ya conoce el comportamiento de la compañía, el asesor ya tiene buenas prácticas, el ambiente laboral no es de su gusto, vive lejos del lugar de trabajo
- 6 a 9 meses: El asesor ya esta devengando aparte de su salario base, un salario variable, Puede que halla tenido problemas con su jefe inmediato, no le gusta el salario.
- 9 a 12 meses: La relación con la compañía es bastante madura, ambiente laboral, cambio de condiciones de trabajo, monotonía, no le gustan las instalaciones.
- 12 a 18 meses: Ya paso la primera navidad con la compañía, horario, estudio, familia.
- 18 a 24 meses: Ya creo un buen ambiente con la compañía, problemas con el salario variable, enfermedad de familiares.
- 24 a 36 meses: Ya puede postularse a otros cargos dentro de la compañía, Estudios, prácticas o pasantías, traslado o viaje, otra oferta laboral
- Más de 36 meses: Es una persona bastante madura en la compañía, que podría marcharse por salario, otra oferta laboral, familia, traslado o viaje.

Con esto, siendo nuestro 100% el motivo de motivos voluntarios, este es el comportamiento por los segmentos dados a través del tiempo.

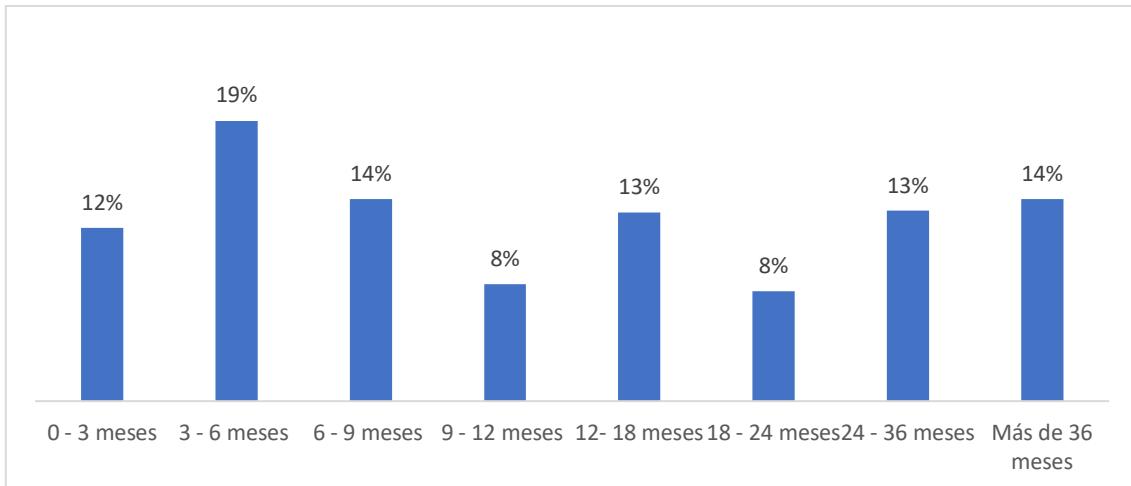


Grafico 9. Comportamiento motivo voluntario

Y siguiente a la grafica, la tabla 7 que muestra los motivos más frecuentes al momento del retiro voluntario, en los segmentos de los meses dentro de la compañía.

En ellos se observa que el 22.4% de las personas que se retiran en menos de 3 meses, son expectativas laborales, esto significaría, que el asesor no estaba agusto con el trabajo o que las condiciones iniciales no fueron constantes.

Tipología de la Medida	0 a 3 meses	3 a 6 meses	6 a 9 meses	9 a 12 meses	12 a 18 meses	18 a 24 meses	24 a 36 meses	Más de 36 meses	Total
Motivos Personales	21.6%	19.6%	22.3%	23.3%	22.4%	22.0%	16.8%	9.6%	19.1%
Expectativa Laboral	22.4%	16.1%	17.6%	16.6%	14.0%	17.4%	19.6%	24.3%	18.5%
Otra Oferta Laboral	10.6%	14.1%	14.1%	13.5%	14.3%	13.3%	17.1%	15.4%	14.3%
Estudios, Prácticas y Pasantías	10.6%	10.8%	10.7%	12.6%	16.2%	11.2%	10.8%	9.9%	11.5%
Traslado o Viaje	8.2%	8.7%	8.2%	12.1%	9.2%	12.4%	8.5%	8.2%	9.1%
Cuidado o Enfermedad de Familiares	4.7%	5.6%	5.5%	4.0%	5.3%	5.4%	6.0%	11.3%	6.2%
Expectativa Salarial	7.5%	9.4%	7.4%	2.2%	4.8%	3.7%	5.5%	5.0%	6.2%

Tabla 7. Motivos retiro voluntario.

Dado que en la tabla presenta expectativa salarial como el menos relevante, es uno de los motivos que la compañía puede contratar y aparte representa valor monetario para la empresa. Con esto, se entra analizar en los mismos segmentos el salario devengado promedio de los asesores. (Para facilidad de programación, se le modificaron algunos nombres a los segmentos de los meses, a continuación serán mostrados)

- R1: 0 a 3 meses.
- R2: 3 a 6 meses.
- R3: 6 a 9 meses
- R4: 9 a 12 meses.
- R5: 12 a 18 meses.
- R6: 18 a 24 meses
- R7: 24 a 36 meses
- R8: Más de 36 meses

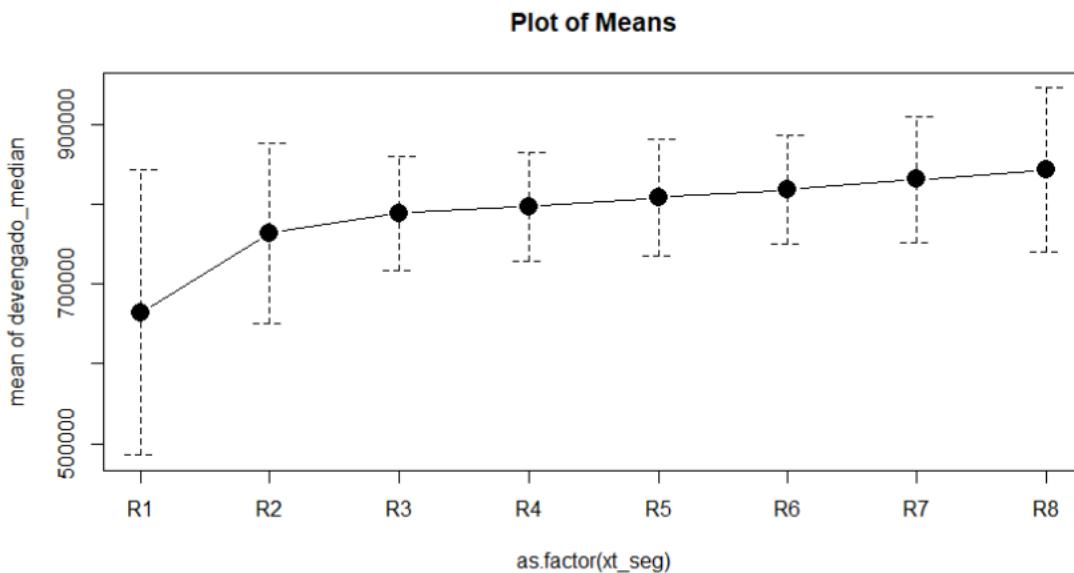


Grafico 10. Salario basico promedio.

Rango	mean	sd	IQR	min	25%	median	75%	max	n	Var
0 a 3 meses	\$ 663 956	\$ 178 505	\$ 265 766	\$ 142 083	\$ 551 578	\$ 698 775	\$ 817 345	\$ 1 004 462	371	
3 a 6 meses	\$ 763 363	\$ 112 949	\$ 107 533	\$ 30 086	\$ 721 072	\$ 783 466	\$ 828 605	\$ 1 426 624	626	14.0%
6 a 9 meses	\$ 788 283	\$ 71 705	\$ 86 881	\$ 489 821	\$ 744 270	\$ 785 586	\$ 831 151	\$ 1 265 738	476	3.2%
9 a 12 meses	\$ 796 509	\$ 68 449	\$ 78 984	\$ 572 585	\$ 757 062	\$ 789 475	\$ 836 046	\$ 1 216 612	278	1.0%
12 a 18 meses	\$ 808 065	\$ 73 659	\$ 93 573	\$ 460 953	\$ 761 009	\$ 804 610	\$ 854 582	\$ 1 042 376	447	1.4%
18 a 24 meses	\$ 817 710	\$ 67 648	\$ 78 177	\$ 566 854	\$ 773 682	\$ 812 628	\$ 851 859	\$ 1 092 613	247	1.2%
24 a 36 meses	\$ 830 557	\$ 78 150	\$ 95 819	\$ 457 595	\$ 778 531	\$ 823 944	\$ 874 350	\$ 1 261 915	434	1.6%
Más de 36 meses	\$ 842 591	\$ 102 536	\$ 110 580	\$ 100 879	\$ 782 817	\$ 830 413	\$ 893 397	\$ 1 457 971	469	1.4%

Tabla 8. Salario basico.

Análizando el salario promedio basico de los asesores que tiene según su tiempo en la compañía. Se logra identificar que el salario basico promedio de los asesores en los primeros tres meses es mucho menor que en el resto del tiempo, esto es debido a que aún no se han medido algunos indicadores por estar en entrenamiento. Pero, después de que pasan los 3 meses la variación en su salario es del 14% más y de ahí en adelante, la variación es muy constante en promedio 1,5% trimestral. Esto, debido a que tiene algunas desviaciones frente a la media, puede ganar mucho más de lo presentado en la tabla.

Y frente al salario variable, siendo este un salario extra al salario basico el cual es devengado por los asesores que presenten buen desempeño en la compañía y la operación en la que se encuentre.

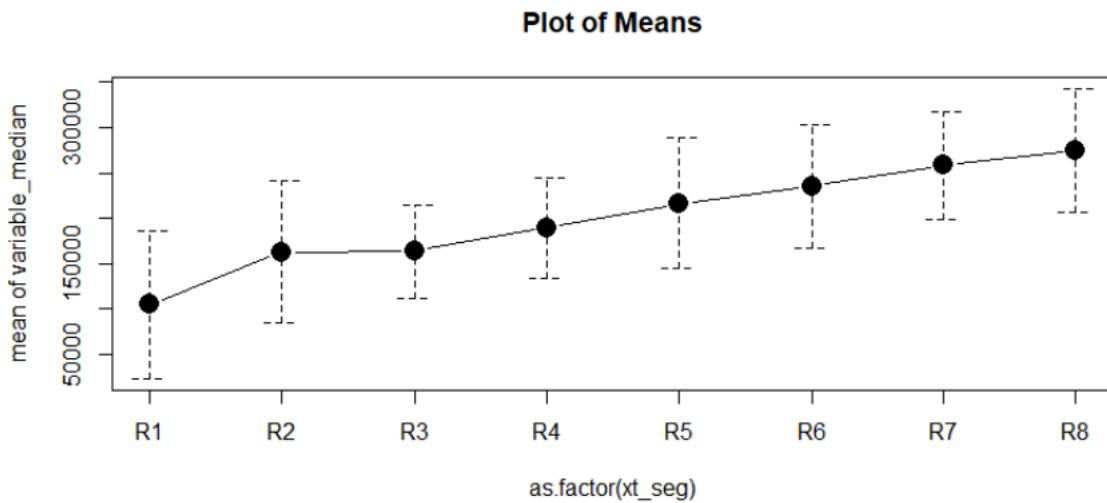


Grafico 11. Promedio salario variable.

Rango	mean	sd	IQR	min	25%	median	75%	max	n	Var
0 a 3 meses	\$ 104 993	\$ 81 126	\$ 130 213	\$ 2 920	\$ 29 761	\$ 96 805	\$ 159 974	\$ 294 491	371	
3 a 6 meses	\$ 162 598	\$ 78 337	\$ 105 656	\$ 11 311	\$ 102 925	\$ 150 600	\$ 208 581	\$ 513 209	626	43.7%
6 a 9 meses	\$ 163 630	\$ 51 483	\$ 77 686	\$ 25 606	\$ 123 310	\$ 168 256	\$ 200 995	\$ 497 510	476	0.6%
9 a 12 meses	\$ 189 575	\$ 55 446	\$ 63 426	\$ 43 054	\$ 161 475	\$ 193 197	\$ 224 901	\$ 436 771	278	14.7%
12 a 18 meses	\$ 216 444	\$ 72 027	\$ 96 991	\$ 10 633	\$ 170 815	\$ 225 130	\$ 267 806	\$ 389 362	447	13.3%
18 a 24 meses	\$ 235 289	\$ 68 246	\$ 103 857	\$ 62 688	\$ 183 912	\$ 252 897	\$ 287 769	\$ 377 404	247	8.3%
24 a 36 meses	\$ 258 585	\$ 59 218	\$ 73 138	\$ 19 442	\$ 228 910	\$ 268 759	\$ 302 048	\$ 384 716	434	9.4%
Más de 36 meses	\$ 274 482	\$ 67 716	\$ 62 652	\$ 26 177	\$ 251 084	\$ 288 970	\$ 313 736	\$ 757 077	469	6.0%

Tabla 9. Distribución salario variable.

A pesar de que este salario es variable y en algunas ocasiones se puede devengar más salario que en otra ocasiones. Este, presenta inicialmente el mismo comportamiento del salario basico, donde en los primeros 3 meses el salario a devengar es mejor en promedio que en el resto del tiempo. En el cual, pasando de los 3 meses iniciales se incrementa en un 43% en salario variable.

Fase IV. Data clusterization. Creación de grupos dentro del conjunto de datos.

Para identificar inicialmente el comportamiento de la rotación de la población de empleados de la organización y de esa forma, encontrar homogenización o patrones dentro de las bajas en la compañía, se corrieron diferentes técnicas que clusterización trabajando con todas las variables mediante el método de programación Rstudio. Las cuales, los métodos de agrupación trabajados fueron: Kmeans, Kmedoids, Clara, Fuzzy C-means y Jerárquico.

Para ejecutar el procedimiento se cargó la base de datos y se analizó las variables categóricas con la técnica mencionada anteriormente Tukey, con el fin de identificar

cuáles eran representativas para los datos en función de la rotación, al tener identificadas estas variables se procedió a transformarlas en dummies, posterior, se analizaron las variables numéricas para encontrar registros faltantes (NA), al observar que la cantidad de faltantes era representativa, se decidió a utilizar el método de Mice, el cual consiste en la imputación de los registros faltantes de cada variable al promedio, es decir, los datos faltantes se llenaron con el promedio de cada columna. Al tener la base de datos sin faltantes y variables categóricas, se continuo con la reducción de dimensión de los datos con la técnica de embebimiento de TSNE (los puntos cercanos en el espacio se atraen y las distancias se repelen). Luego de esto, se aplico la matriz de distancias con el método de Kendall (**Heijden, 2018**) para observar si entre los datos existe correlación y agrupamientos

Matriz de Distancias con método Kendall: Como se observa en la gráfica, los datos poseen correlación y tiende a reunirse en grupos, por ende, si se puede ejecutar los Clusters



Grafico 12. Metodo Kendall.

Después de ejecutar la matriz de distancias, se utilizó el test de Hubert, el cual ayuda a determinar el número óptimo de grupos. En el grafico de Hubert se busca el codo más alto el cual corresponde a un aumento significativo del valor de las medidas, es decir, el pico más significativo.

En el grafico 13, se muestra el comportamiento de la agrupación de los Clusters de dos al número diez

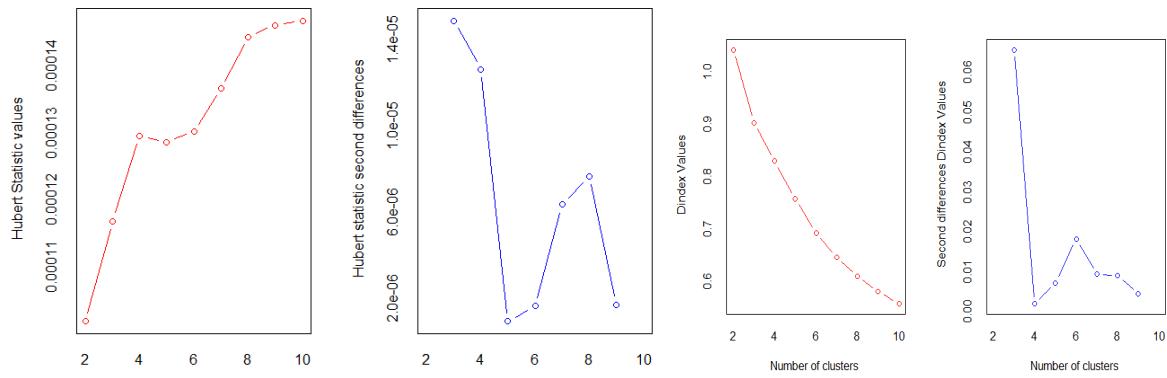


Grafico 13. Comportamiento de 2 a 10 cluster.

Se observa las diferencias significativas que se tienen en cada uno de los cluster frente a otros

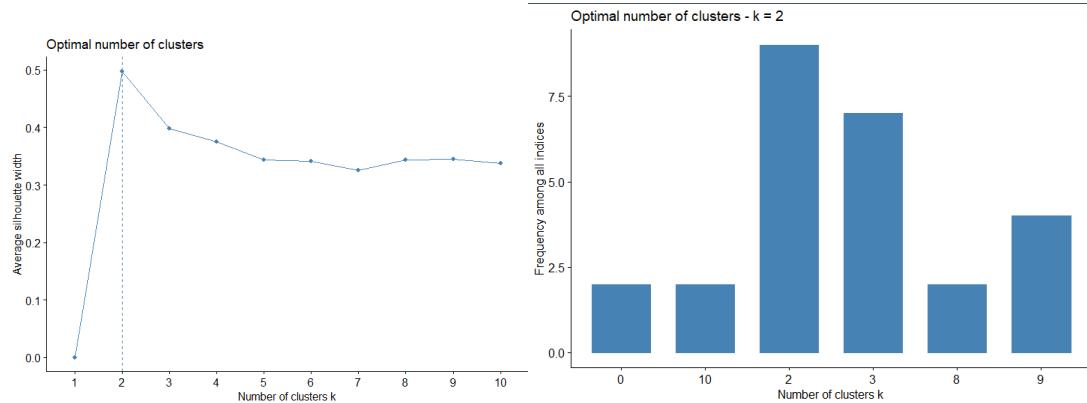


Grafico 14. Test de codo y diferencia intracluster.

El test de Hubert indica que el número óptimo de clusters son dos (2). El cual dentro de ellos dos, habra una diferencia significativa que pueda diferenciarlos.

Al observar las gráficas se evidencia que entre dos y tres Clusters no se tiene una diferencia considerable, es decir, es viable trabajar con tres grupos, para tomar la decisión del número de Clusters se procede con la representación del conjunto de datos con dos y tres Cluster.

Agrupación con dos Clusters: En la grafica 15, se puede observar el agrupamiento que se tiene al crear dos grupos bajo el metodo K-Means.

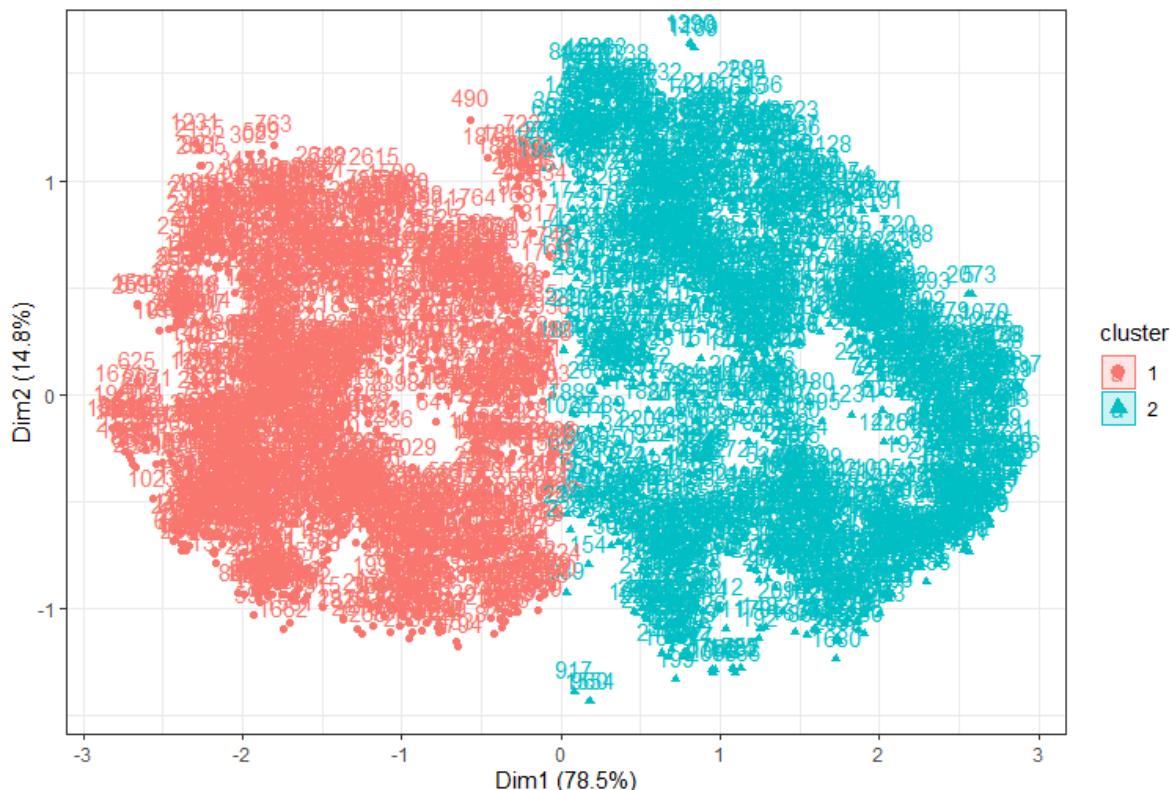


Grafico 15. Agrupación con 2 grupos.

Al crear los 2 grupos, se percibe en el segundo Cluster (verde) un tercer grupo, debido a que los datos al lado derecho parte inferior se encuentran alejados y dispersos de los demás. Por ende, como acto de prueba, se realiza la visualización de los datos con tres Clusters.

Agrupación con tres Clusters: La representación de los tres cluster, se validara por medio del mismo metodo utiizado previamente con los 2 clusters (K-means).

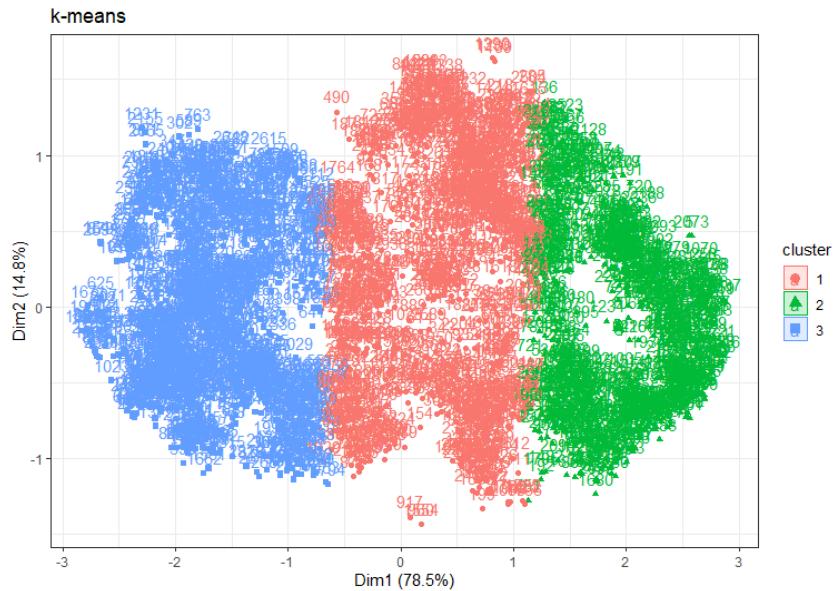


Grafico 16. K-means 3 clusters.

Con la representación del conjunto de datos en los tres Clusters, los datos se observan compactos en cada uno de sus grupos, por ende, se decide realizar el ejercicio con tres grupos.

K-means es una técnica de agrupación que tiene como objetivo partir un conjunto de datos de N registros en K grupos donde cada observación pertenece a un grupo cuyo valor medio es más cercano a su grupo y más diferente a los demás grupos, y se define como:

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

donde $\boldsymbol{\mu}_i$ es la media de puntos en S_i .

Formula 1. Definición K-means

En la tabla demuestra como estan distribuidos los asesores en los 3 grupos

Grupo	Asesores	Mediana_Mes
1	1,035	4
2	909	15
3	725	25

Tabla 10. Distribución de los asesores k-means

Agrupación por el Método Fuzzy C-Means: Esta técnica es una variación difusa de K-means, el cual surge de la necesidad de resolver la deficiencia del agrupamiento; este método está basado en una partición difusa que representa la división de cada muestra en todos los grupos utilizados, la pertenencia de un grupo toma valores entre cero y uno. Para identificar si los grupos son representativos se aplicó la técnica de silhouette, la cual varía entre menos uno y uno, entre más cercano a uno los clusters son representativos. Y se representa:

$$J_m(U, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m d_{ik}^2$$

Formula 2. Definición Fuzzy

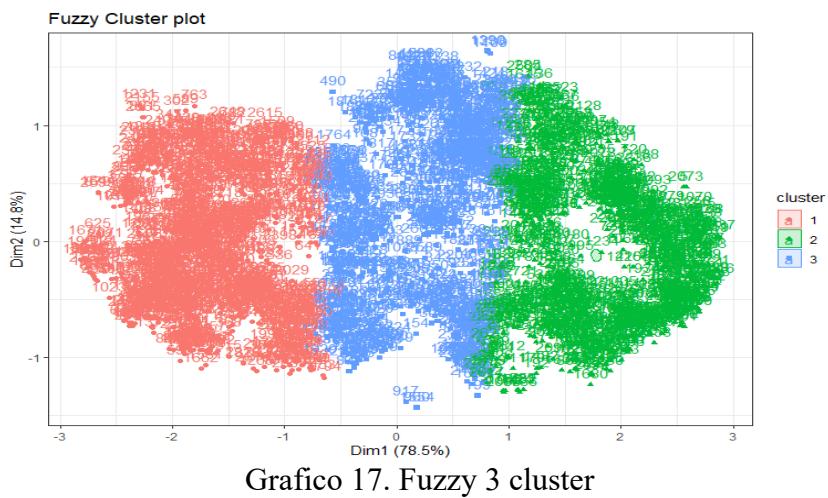


Grafico 17. Fuzzy 3 cluster

Adicionalmente, se analizó la técnica Silhouette. La cual, mide la diferencia de los datos intracluster creada por el algoritmo.

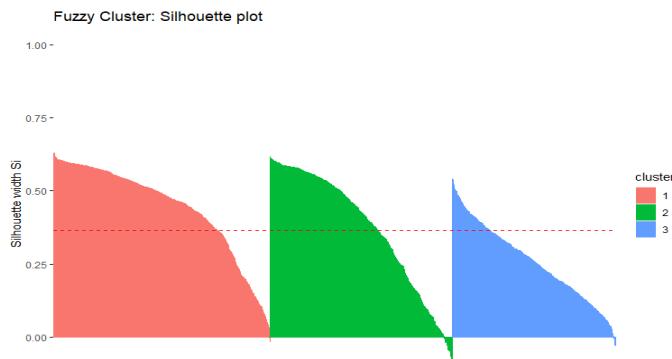


Grafico 18. Silhouette Fuzzy

Y al final, se observa en la tabla 6, la distribución de los asesores que tuvieron en la agrupación mediante este metodo.

Fuzzy		
Cluster	Size	Score
1	1,027	0.44
2	867	0.37
3	775	0.26

Tabla 11. Distribución de los asesores Fuzzy.

Agrupación por el Método K-medoids: Este método divide la base de datos en grupos, buscando minimizar las distancias entre los puntos de cada grupo frente a su punto centro y maximizar la distancia de los puntos de un grupo con respecto a los puntos de los demás grupos. La cual, su función de costo se representa de la siguiente forma:

$$\text{costo}(x, c) = \sum_{k=1}^d |x_i - c_i|$$

Formula 3. Definición costo K-medoids

Donde x es cualquier objeto, c es el medoid, y d es la dimensión del objeto

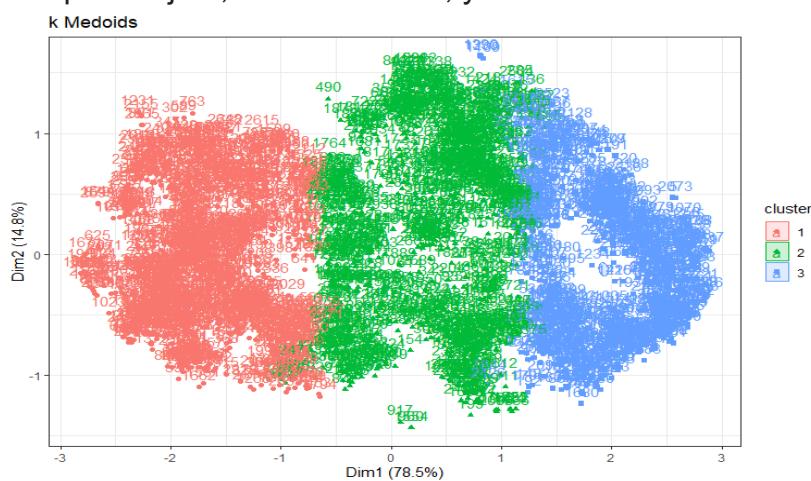


Grafico 19. K-medoids 3 cluster

Igualmente, se analizó la técnica Silhouette. La cual, mide la diferencia de los datos intracluster creada por el algoritmo.

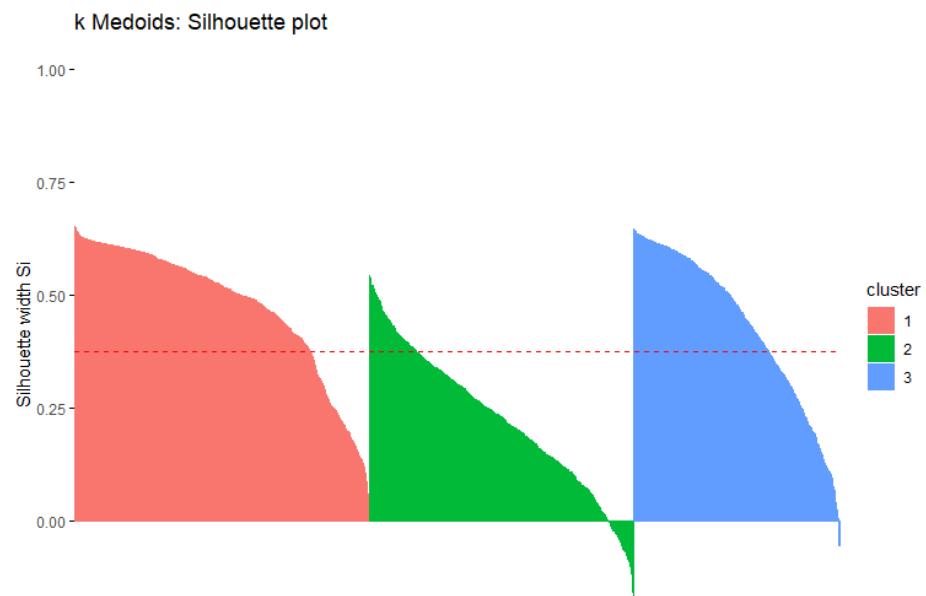


Grafico 20. Silhouette K-medoids

Y al final, se observa en la tabla 7, la distribución de los asesores que tuvieron en la agrupación mediante este metodo.

k-Medoids		
Cluster	Size	Score
1	1,027	0.48
2	922	0.22
3	720	0.43

Tabla 12. Distribución de los asesores K-Medoids.

Agrupación por el Método Clara: (CLustering LARge Applications)

Clara se basa en el enfoque de muestreo para manejar grandes bases de datos, esta técnica extrae una pequeña muestra del conjunto de datos y aplica el algoritmo de K-medoids para generar el número óptimo para la muestra, la calidad de los medoides resultante se mide por la dispersión promedio entre cada dato en el conjunto de datos en su grupo. Y es representado por la siguiente función de costo:

$$Cost(M, D) = \frac{\sum_{i=1}^n dissimilarity(O_i, rep(M, O_i))}{n}$$

Formula 4. Definición costo CLARA.

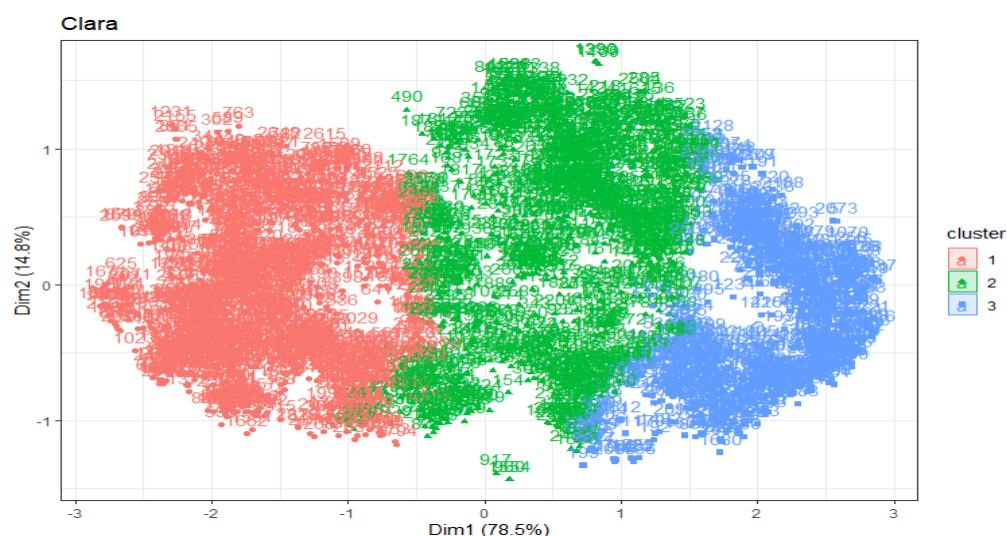


Grafico 21. K-medoids 3 cluster.

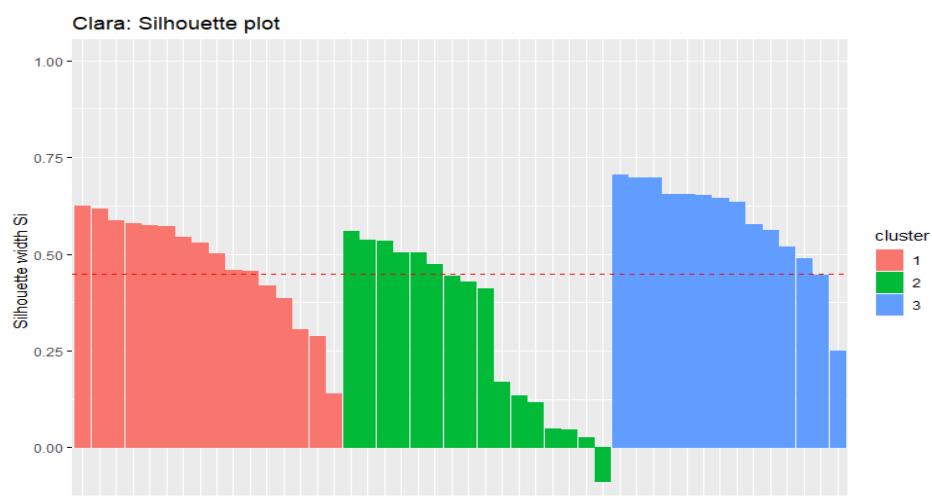


Grafico 22. Silhouette K-medoids

Y al final, se observa en la tabla 8, la distribución de los asesores que tuvieron en la agrupación mediante este metodo.

Clara		
Cluster	Size	Score
1	929	0.47
2	928	0.3
3	812	0.58

Tabla 13. Distribución de los asesores Clara.

Agrupación por el Método Jerárquico: Este método va generando grupos en cada una de las fases del proceso de búsqueda de Cluster, esta técnica fija por si solo el número óptimo de grupos.

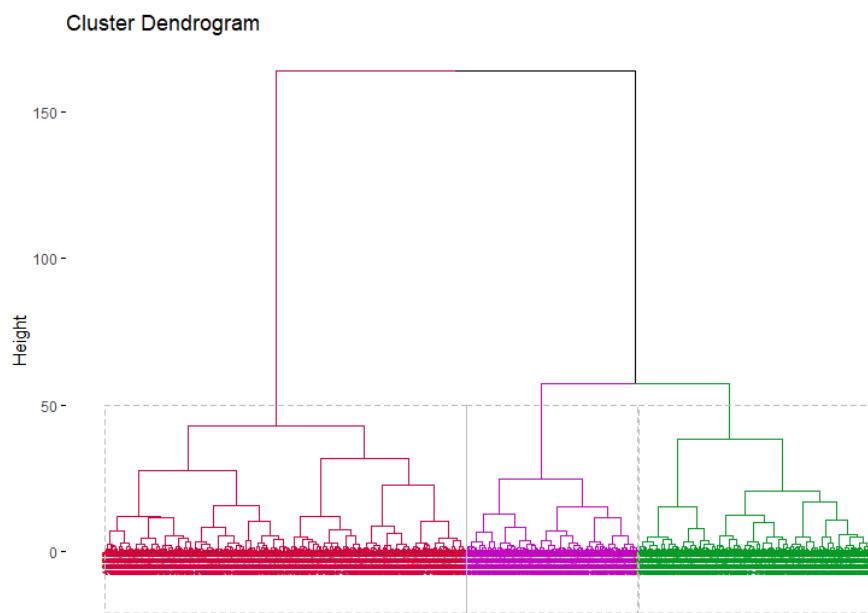


Grafico 23. Dendogram 3 Cluster

Grupo	Asesores	Mediana_Mes
1	1,262	5
2	810	22
3	597	23

Tabla 14. Distribución de los asesores Jerarquico

Para tomar la decisión por cual metodología se iba a realizar el análisis de supervivencia, se evaluaron los Cluster con respecto a las siguientes métricas: Connectivity, Dunn y Silhouette, los resultados que se obtuvieron fueron los siguientes:

	Medida	Score
hierarchical	Connectivity	56.6833
	Dunn	0.0281
	Silhouette	0.3654
kmeans	Connectivity	185.271
	Dunn	0.0129
	Silhouette	0.3953
Fuzzy	Connectivity	189.5313
	Dunn	0.0042
	Silhouette	0.3658
Clara	Connectivity	271.4484
	Dunn	0.0069
	Silhouette	0.3853
Kmedoids	Connectivity	189.7429
	Dunn	0.0065
	Silhouette	0.3756

Técnica Seleccionadas

	Score	Cluster
Connectivity	56.6833	hierarchical
Dunn	0.0281	hierarchical
Silhouette	0.3953	kmeans

Tabla 15. Resultados modelos.

Para realizar el análisis de supervivencia es necesario conocer la distribución de comportamiento de los datos, para esto se evaluaron 5 distribuciones que representan el conjunto de datos, las distribuciones escogidas son: Exponencial, Weibull, Logística, Gompertz y Genf, para escoger la mejor representación de los datos, las distribuciones mencionadas se evaluaron de acuerdo a su AIC

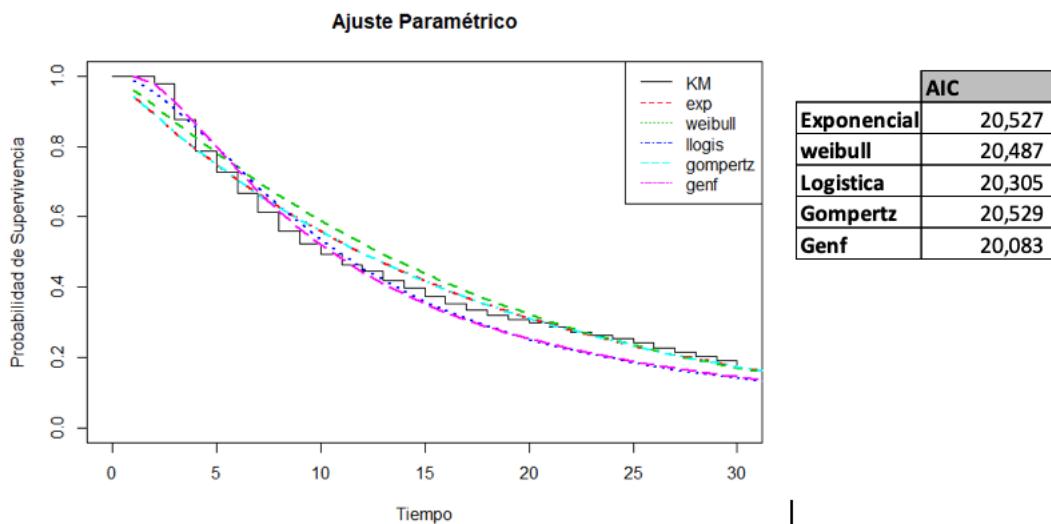


Grafico 24. Distribucion de los datos en los clusters.

De acuerdo con los resultados del AIC se escoge la distribución Genf para realizar el análisis de supervivencia a la organización.

Análisis de supervivencia: Este análisis es una técnica inferencial que consiste en modelizar el tiempo que se tarda en que ocurra un determinado suceso. Los resultados de la selección de cluster mostro que las mejores técnicas son K-means y el jerárquico. Por ende, se decidió correr el análisis de supervivencia con las dos siguientes metodologías.

- 1. Análisis de Supervivencia con K-means:** Como se observa en el análisis de supervivencia con Kmeans, los tres grupos se identifican por los periodos de abandono de los empleados. El grupo objetivo es el Cluster uno, el cual indica que el 70% de los empleados se retiran de la compañía en un tiempo menor o igual a cinco meses.

Grupo	Asesores	Mediana_Mes
1	1,035	4
2	909	15
3	725	25

Tabla 10. Distribución de los asesores k-means

K-means: Comparación de clusters

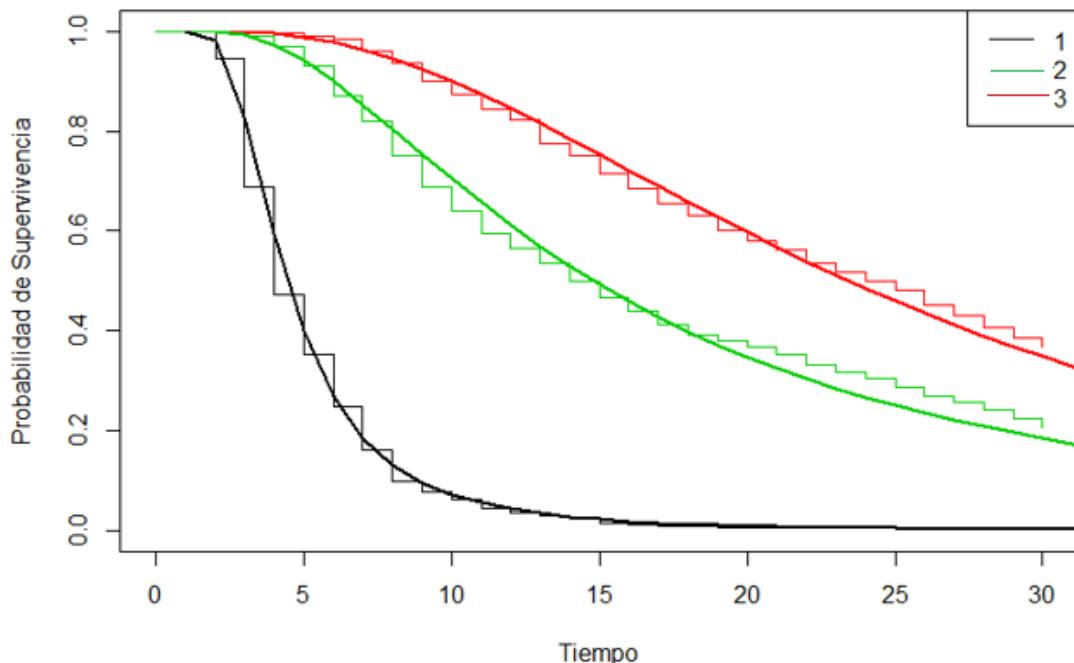


Grafico 25. Comparación intracluster K-Means.

Análisis de Supervivencia con Jerárquico: En la representación visual del análisis de supervivencia con la metodología del Jerárquico se observa que el grupo dos y tres se comportan como si fuera un solo conjunto, el cluster objetivo es el uno, el cual indica que el 70% de los empleados se retiran de la compañía en un tiempo menor o igual a cinco meses.

Grupo	Asesores	Mediana_Mes
1	1,262	5
2	810	22
3	597	23

Tabla 14. Distribución de los asesores Jerarquico

Hierarchical: Comparación de clusters

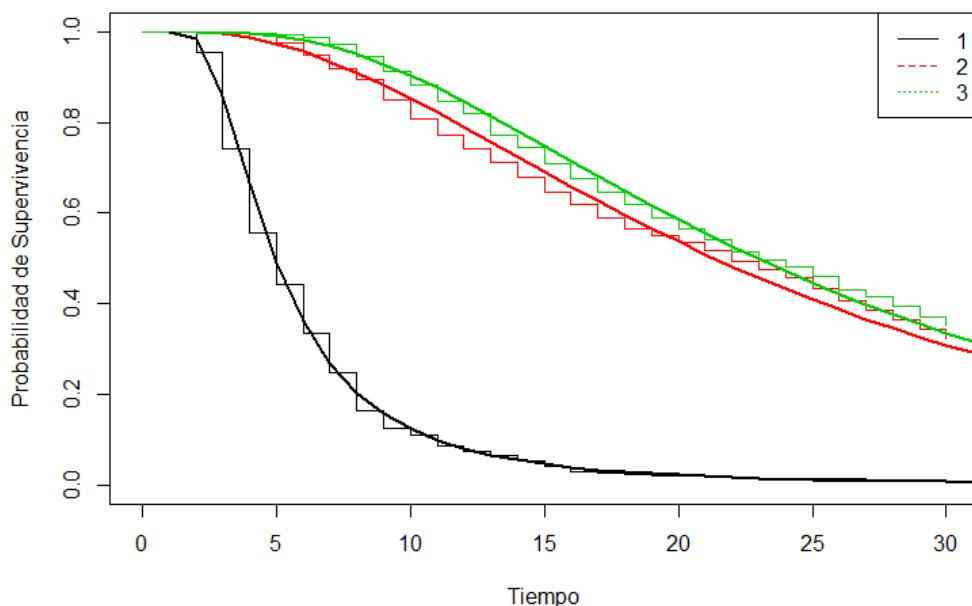


Grafico 26. Comparación intracluster Jerarquico

Con los resultados que se obtuvieron, se tomo la decisión de realizar el análisis de los grupos con la metodología de K-means.

grupo	Meses_Promedio	Meses_Mediana	Devengado_Medio	Devengado_min	Devengado_max	Devengado_sd	Cantidad
1	5	4	682,669	341,471	956,709	265,917	1,035
2	29	25	953,033	394,189	2,122,310	393,930	725
3	21	15	845,799	282,258	1,447,151	302,766	909

Tabla 15. Resultado final de clusterización.

Fase V. Data Preparation. Análisis de los datos y selección de características.

Partición del dataset para el análisis:

Una vez terminado el análisis inicial, se procede con la partición de la base de datos. Se realiza utilizando una estructura 60 – 20 – 20: 60% para entrenar, 20% para testear el entrenamiento y otro 20% final para validar la capacidad de generalización del modelo y se continúa con la selección de variables.

Selección de variables:

Antes de realizar todo el proceso de modelación, se dedicó un buen tiempo a la selección de variables. Para esto se utilizaron diversas técnicas, tales como: filter methods (correlación, mutual information,), y embedding methods (lasso regularisation, elasticnet regularisation, y RFECV (Feature ranking with recursive feature elimination and cross validation), este último va realizando una eliminación recursiva con base en la métrica de accuracy cada que se estima un determinado modelo, en nuestro caso un random forest y el criterio usado fue la entropía.

Con base en lo anterior, como primer paso se analiza un matriz de correlación entre las variables con el fin de concluir acerca de altas correlaciones lineales entre ellas y evitar problemas futuros de multicolinealidad:

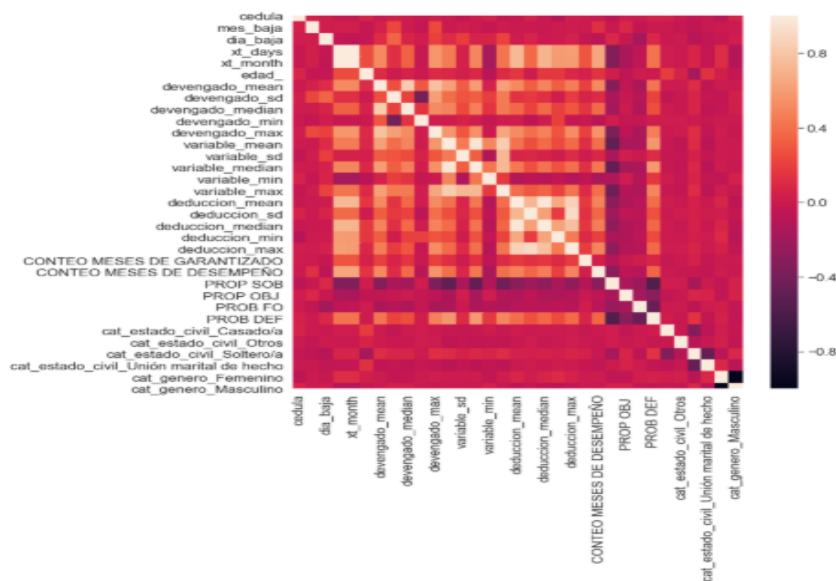


Grafico 27. Correlación de las variables.

Se observa que efectivamente algunas de las variables están altamente correlacionadas entre sí, así que se decide aplicar un primer método de selección que consiste en evaluar grupos de características altamente correlacionadas y mediante la asignación de importancias a esas variables por medio de análisis con bosques aleatorios, se extraen las que menor información aportan a la variable respuesta.

Con base en este primer criterio, se seleccionan 25 variables, posteriormente se utilizaron las técnicas de Lasso y Elasticnet con un modelo logístico y el resultado fue el siguiente:

RFECHV	LASSO	ELASTICNET	MUTUAL INFORMATION
mes_baja	mes_baja	mes_baja	
dia_baja	dia_baja	dia_baja	
edad_	edad_	edad_	edad_
devengado_mean	devengado_mean	devengado_mean	devengado_mean
devengado_sd	devengado_sd	devengado_sd	devengado_sd
devengado_min	devengado_min	devengado_min	devengado_min
devengado_max	devengado_max	devengado_max	devengado_max
variable_sd	variable_sd	variable_sd	variable_sd
variable_median	variable_median	variable_median	'variable_median',
variable_min	variable_min	variable_min	variable_min
variable_max	variable_max	variable_max	'variable_max',
deduccion_mean	deduccion_mean	deduccion_mean	deduccion_mean
deduccion_min	deduccion_min	deduccion_min	'deduccion_min',
deduccion_max	deduccion_max	deduccion_max	'deduccion_max',
CONTEO MESES DE GARANTIZADO			
CONTEO MESES DE DESEMPEÑO	CONTEO MESES DE DESEMPEÑO	CONTEO MESES DE DESEMPEÑO	'CONTEO MESES DE DESEMPEÑO',
PROP SOB	PROP SOB	PROP SOB	PROP SOB
PROP OBJ	PROP OBJ	PROP OBJ	PROP OBJ
PROB FO	PROB FO	PROB FO	PROB FO
PROB DEF			'PROB DEF',
cat_estado_civil_Casado/a	cat_estado_civil_Casado/a	cat_estado_civil_Casado/a	
cat_estado_civil_Soltero/a	cat_estado_civil_Soltero/a	cat_estado_civil_Soltero/a	
cat_estado_civil_Unión marital de hecho			
cat_genero_Femenino	cat_genero_Femenino	cat_estado_civil_Unión marital de hecho	cat_estado_civil_Unión marital de hecho
cat_genero_Masculino.	cat_genero_Masculino.	cat_estado_civil_Unión marital de hecho	cat_estado_civil_Unión marital de hecho
25	24	24	21

Tabla 16. Evaluación modelos estadísticos.

Cómo se observa en la tabla anterior básicamente las variables seleccionadas por los diferentes métodos son casi las mismas, solo se diferencian en que el método RFECHV selecciona adicionalmente la variable 'PROB DEF', que corresponde a la probabilidad de que el empleado obtenga un desempeño de deficiente, lo que lo lleva a no ganar ningún salario variable y el método de mutual information es el que menor número de características selecciona, descartando algunas categorías del estado civil y del mes y día de baja. (retiro del empleado).

Fase VI. Modeling. Modelado

Una vez terminados los procesos de preparación y exploración de los datos y la selección de características se realizó el proceso de modelado. Se utilizaron 7 modelos de aprendizaje sobre los 3 subset mencionados anteriormente: Regresión logística, Gradient boosting, Decission Tree, Random Forest y suport vector machines y k nearest neighbors y gaussian process clasifier. Para evaluar el rendimiento de los modelos se analizó el ROC AUC en el subconjunto de test y también se analizaron otras métricas como el accuracy y la precision, el recall y el f1score que contiene las dos anteriores.

Se utilizó la librería Gridsearch de python para variar los hiperparámetros de dichos modelos y escoger el de mejor AUC.

Fase VII. Evaluation. Evaluación (obtención de resultados)

Los mejores resultados fueron los arrojados por el gradient boosting para las bases de datos con los 4 conjuntos de variables seleccionadas con los siguientes resultados:

BestX1: {'gradient': 0.948}

BestX2: {'gradient': 0.961}

BestX3: {'gradient': 0.951}

BestX4: {'gradient': 0.951}

Sin embargo tanto en AUC como en las otras métricas la regresión logística presentaron resultados muy similares, por lo tanto teniendo en cuenta que el modelo logístico es un modelo más sencillo y de más fácil interpretación se seleccionó este modelo para predecir la probabilidad de permanencia del empleado superior a 12 meses en el conjunto de variables más pequeño. (el obtenido con el método de mutual information feature selection).

Por último para comprender mejor el resultado de cada uno de los modelos en los 4 datasets se realizó un comparativo de las curvas ROC como se observa a continuación.

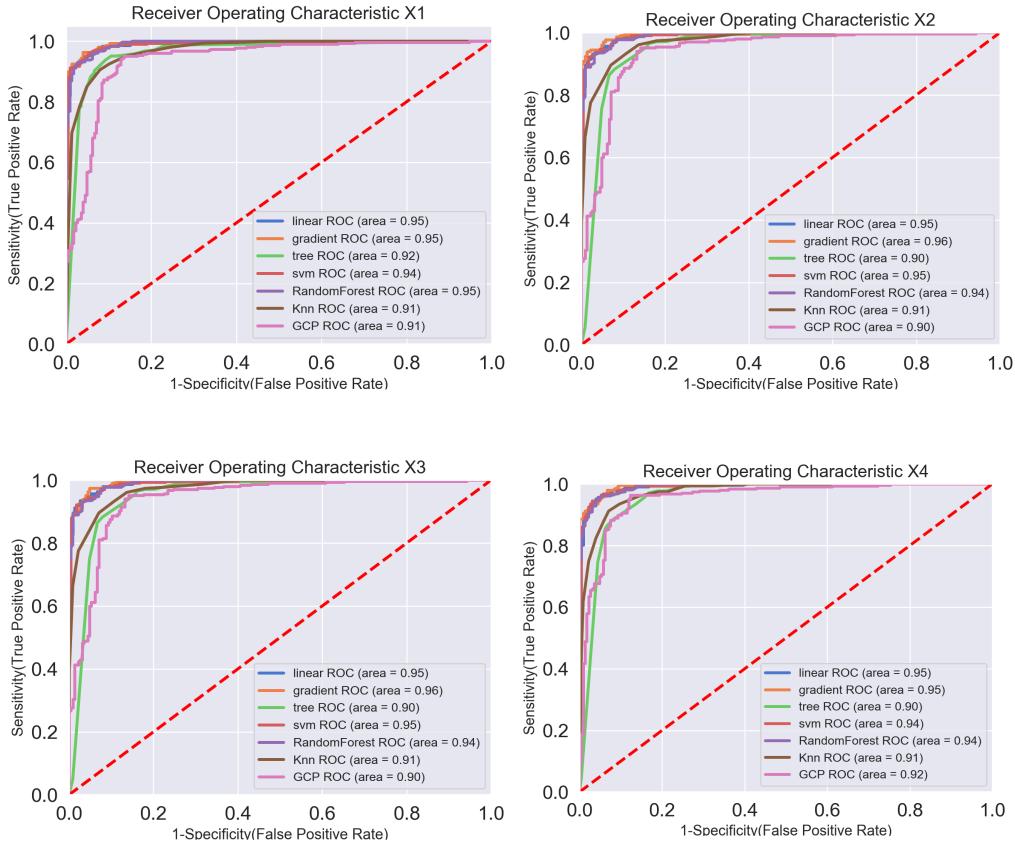


Grafico 28. Curvas ROC

Fase VIII. Deployment. Despliegue (puesta en producción)

Esta sería la última fase del proceso, y es la puesta en producción del modelo. Teniendo en cuenta todo el análisis anterior el modelo recomendado para la puesta en producción sería el obtenido con la regresión logística con la bd1 ya que presenta los mejores resultados de AUC. También puede utilizarse el random forest y tomar el mejor resultado entre ambos, ya que dependiendo de la selección del conjunto de train y test los resultados pueden variar.

La rotación de personal para la empresa Konecta, es un indicador que afecta de gran manera, no solo por lo que representa en costos sino por la experiencia y conocimiento de su personal que sale de ella. Variables tales como las analizadas en este documento (salario, horarios, turnos, relacionamiento con sus líderes, edades, adherencia) deben estar constantemente controladas, para que el personal de la compañía se sienta parte fundamental de la compañía y se encuentren motivados para realizar sus funciones. Una alta probabilidad de retiro se presenta al estar por debajo del salario mínimo, cuando presentan bajos niveles de adherencia y cuando son asesores que tienen su primera experiencia laboral.

Una de los resultados importante es el comportamiento homogéneo que tienen los asesores en los meses que el modelo predictivo predijo que renunciarían, este comportamiento tiene que ver con su desempeño, con sus indicadores los días antes de tomar la decisión. Por ende, se quiere tomar acciones para evitar esta baja realizando mejoras en la operación y dentro de la compañía.

Y (Castillo, 2006) Sostiene que por más evolución que tenga una organización en cuanto a tecnología y demás procesos, dependen de la gestión que se lleve a cabo sobre el activo más importante que puede tener una compañía: el potencial humano. Por ende, respaldar la buena y potencializar el personal activo de Konecta para evitar fugas de la compañía.

Conclusiones

- En síntesis, el conocimiento y la preparación de los datos juegan un papel fundamental en los resultados que se buscan obtener, ya que si no se toman las decisiones oportunas sobre las variables que intervienen en el modelo, los resultados podrán conducir a decisiones erradas.
- Al segmentar los datos en tres clusters con la técnica Kmeans, se permite identificar dos grupos los cuales son objetivos para la organización, dichos grupos son el uno y el dos, donde el primero el abandono de empleados está entre cero y cinco meses y en el grupo dos están los empleados que se retiran entre los cero y quince meses.
- Al identificar las medidas de cada uno de los grupos de retiro de los empleados, se pueden ejecutar acciones sobre estas, con el fin de mejorarlas y así disminuir la rotación temprana en la organización.
- Luego de realizar todo el análisis, correr diferentes modelos, con los diferentes conjuntos de variables seleccionadas, se observa que el modelo recomendado para predecir la variable 'yL' es la regresión logística con la bd1, ya que se obtiene un nivel del AUC promedio el 0.7 y el mejor valor de la métrica Recall con respecto a los demás modelos.
- Durante todo el proceso de modelado se pudo observar la sensibilidad de los resultados al variar la partición de la muestra (train, test), por lo tanto para evitar el sobreajuste y escoger el mejor modelo se realizaron diferentes particiones (10 en total) y se compararon los resultados del AUC. Al final se tomó la decisión con el valor promedio del AUC en todas las corridas realizadas

Bibliografía

- SCHEFFÉ, H. (1953). A METHOD FOR JUDGING ALL CONTRASTS IN THE ANALYSIS OF VARIANCE*. *Biometrika*, 40(1-2), 87–110.
- Heijden, P. G. (2018). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 249-273.
- Ponce, A. R. (1995). *ADMINISTRACIÓN DE PERSONAL*. México: Limusa Noriega.
- Castillo, J. A. (2006). *Administración de personal: un enfoque hacia la calidad*. ECOE.