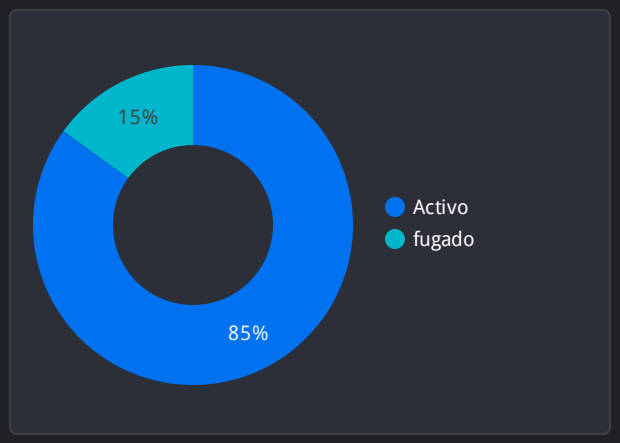
Análisis Previo de Variables

Afiliados

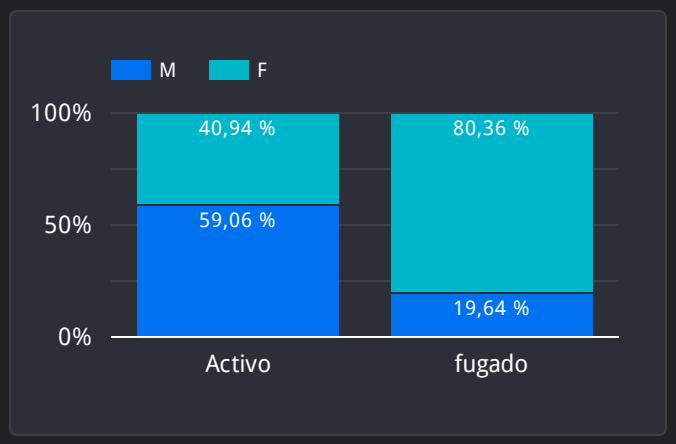
Total
56.224

Hallazgos:
-Se tomó cada fila como una observación. No hubo necesidad de eliminar registros debido a duplicados o valores faltantes.
-Se fugan más mujeres que hombres. Ver gráfico 2.
-El número de fugados ha incrementado en los últimos años. Ver gráfico 3.
-Los fugados se sienten menos satisfechos con el servicio. Ver gráfico 4.
-Los fugados radican más PQR. Ver gráfico 5.
-Los fugados tienen en su mayoría un salario entre 1 y 2 millones COP. Ver gráfico 6.
-Solo se fuga gente entre los 25 y 40 años. Ver gráfico 7.
-La gente que viene de Protección se fuga menos, mientras que los que vienen de Porvenir se fugan más. Ver gráfico 8.

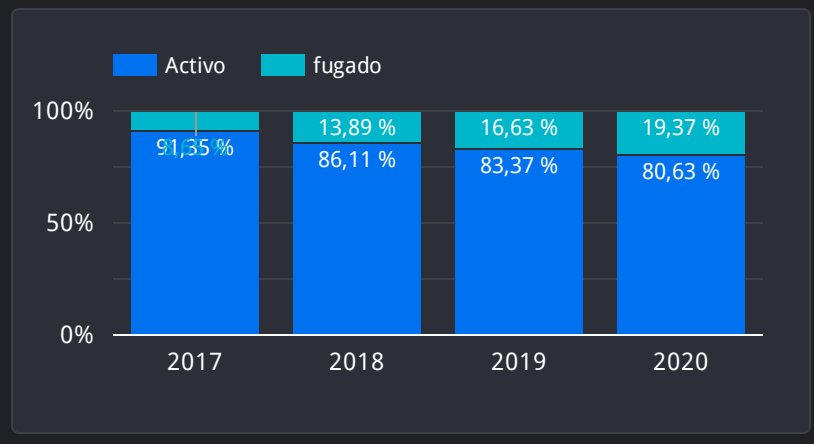
1. Estado Afiliados



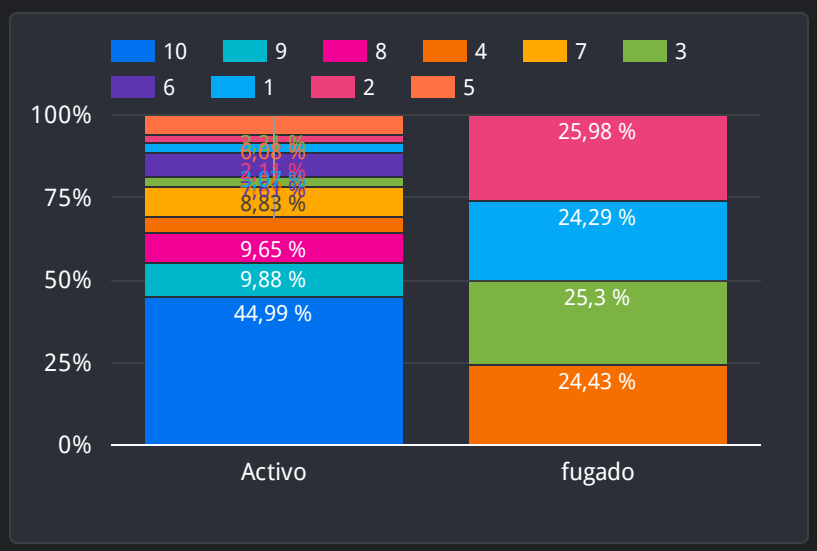
2. Genero Afiliados



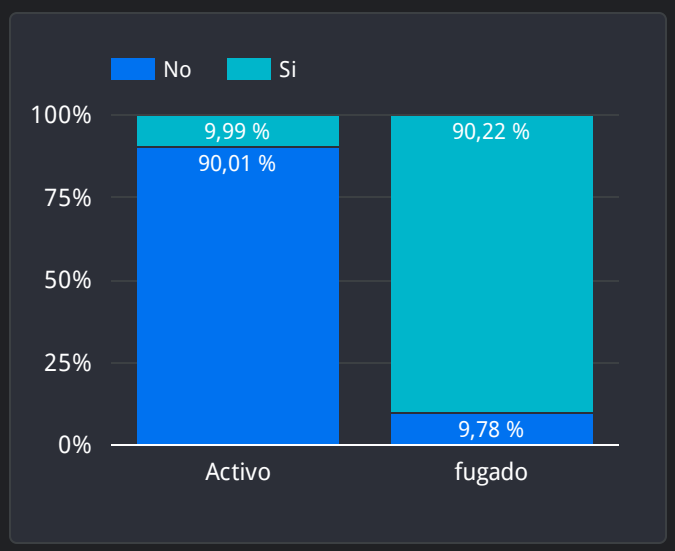
3. Tendencia Fugados



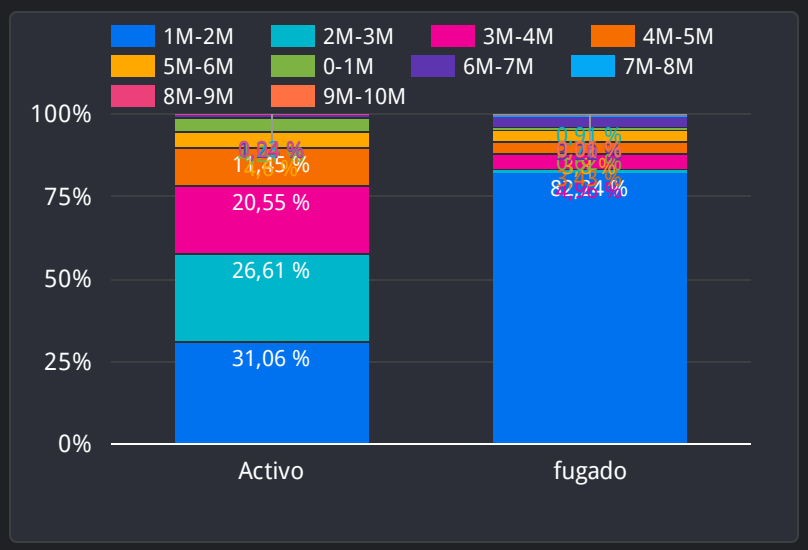
4. Nivel de Satisfacción



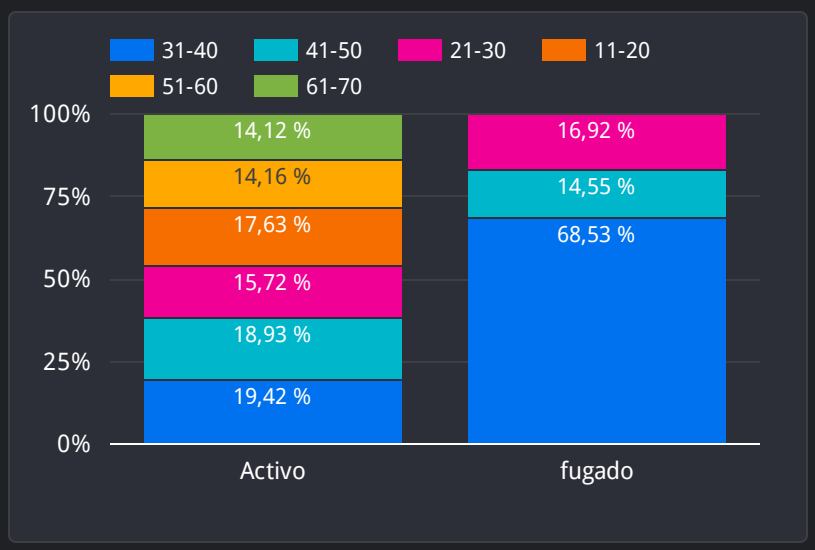
5. Radicación de PQRs



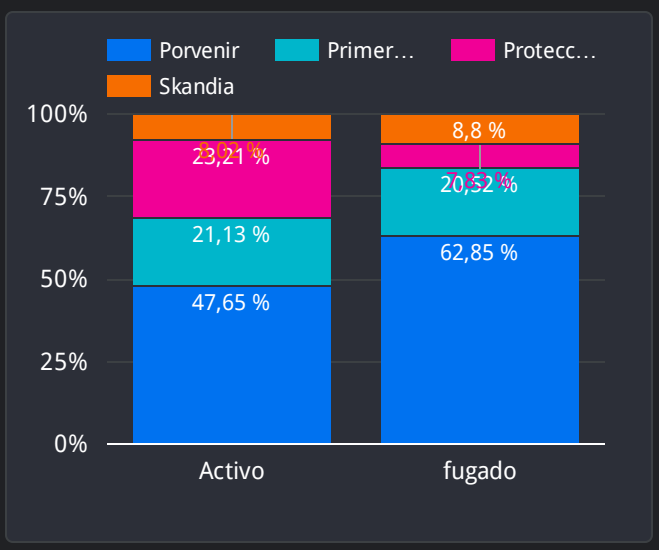
6. IBC



7. Edades



8. Origen Afiliado



Modelo Estadístico

Consideraciones:

-La variable a estimar es si un afiliado se fugo o no, Y=1 si se fugó, Y= 0 de lo contrario
-Con base en el análisis previo de variables se eligieron las siguientes variables explicativos: Genero, Nivel_de_Satisfaccion, PQR, Año, Edad (se crean categorías agrupando cada 10 años), IBC (se crean categorías agrupando cada 1 millon de pesos).
-Para ahorrar recursos en términos computacionales y de tiempo se seleccionaron las mencionadas variables.
-Para las variables explicativas que eran categóricas, se uso el método de one_hot_encoding para representarlas como variables numéricas.
-En este caso se estimará la probabilidad de que un afiliado se fugue. Los modelos sugeridos por la literatura, y por lo tanto se van a usar son: Regresión Logística, Árbol de Decisión, Support Vector Machine y Random Forest.
-Se creó una muestra de train (70% de la población) y una de test (30% de la población), con la primera se entrenarán los modelos, con la segunda se obtendrán las métricas de desempeño con base en las predicciones hechas sobre esta muestra.
-Se elegirá el modelo con mejores métricas de desempeño, siendo la más importante la métrica 'Precision', la siguiente 'Recall'.

Modelo a Estimar:

$$Y = \beta_0 + \beta_1 Genero + \beta_2 NivelSatisfaccion + \beta_3 PQR + \beta_4 Año + \beta_5 Edad + \beta_6 IBC$$

$$Y = \begin{cases} 1, & si\ el\ afiliado\ se\ fugó \\ 0, & de\ lo\ contrario \end{cases}$$

Resultados de las estimaciones

Modelo: Regresión Logística
Precisión: 0.98
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.988720	0.987267	0.987993	14294.000000
1	0.929865	0.937451	0.933643	2574.000000
accuracy	0.979666	0.979666	0.979666	0.979666
macro avg	0.959293	0.962359	0.960818	16868.000000
weighted avg	0.979739	0.979666	0.979699	16868.000000

Modelo: Árbol de Decisión
Precisión: 0.98
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.988614	0.990136	0.989374	14294.000000
1	0.944749	0.936674	0.940694	2574.000000
accuracy	0.981978	0.981978	0.981978	0.981978
macro avg	0.966682	0.963405	0.965034	16868.000000
weighted avg	0.981920	0.981978	0.981946	16868.000000

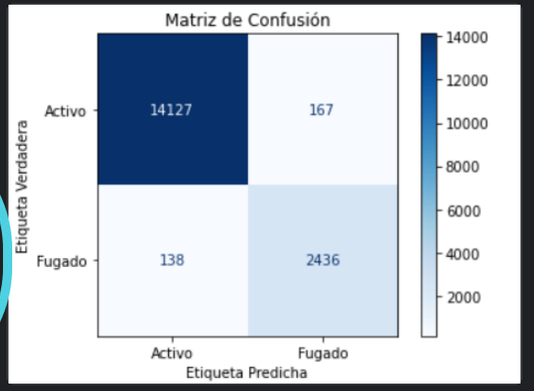
Modelo: SVM
Precisión: 0.85
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.847403	1.000000	0.917399	14294.000000
1	0.000000	0.000000	0.000000	2574.000000
accuracy	0.847403	0.847403	0.847403	0.847403
macro avg	0.423702	0.500000	0.458700	16868.000000
weighted avg	0.718092	0.847403	0.777407	16868.000000

Modelo: Random Forest
Precisión: 0.98
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.990326	0.988317	0.989320	14294.000000
1	0.935843	0.946387	0.941086	2574.000000
accuracy	0.981918	0.981918	0.981918	0.981918
macro avg	0.963085	0.967352	0.965203	16868.000000
weighted avg	0.982012	0.981918	0.981960	16868.000000

Matriz de Confusion Random Forest



Conclusiones

-El modelo que arrojó la mejor 'precision' fue el Random Forest, por lo cual se selecciona este modelo como el mejor, sin embargo los demás modelos, sobre todo el árbol de decisión, también arrojaron buenas estimaciones.
-El modelo de random forest logró identificar correctamente al 91% de fugados.
-La variable que más determinante en la decisión de fuga de los afiliados es el nivel de satisfacción.
-Para futuros análisis se podrían incluir las demás variables.
-Para futuros análisis se podría implementar un modelo de redes neuronales.

Importancia Variables Random Forest