Universidad del Desarrollo
Facultad de Ingeniería

CLASIFICACIÓN DE PARENTESCO EN 1er GRADO A PARTIR DE PATRONES DE LLAMADAS.

POR: JOSÉ ALFREDO PÉREZ TORRES

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dra. LORETO BRAVO

Julio 2021

SANTIAGO

Para Pina, Cata, Marita y Laura.

AGRADECIMIENTO

Muchas gracias a mi familia por su cariño, respaldo y tiempo brindado, en esta nueva aventura académica.

Muchas gracias a la UDD y a los profesores que compartieron su tiempo, conocimiento y paciencia conmigo. En particular, muchas gracias a la Dra. Loreto Bravo, que en un momento de dificultad para mí, me brindó su apoyo y me motivó a continuar avanzando.

Muchas gracias a Sebastián Diaz, alumno de doctorado en la UDD, quien me facilitó el set de datos, y además me orientó y enseñó en su exploración y manipulación.

Muchas gracias a todas las compañeras y compañeros del Magíster; sus comentarios y bromas hicieron más interesantes y entretenidas las asignaturas que cursamos juntos.

Muchas gracias a Roy Barrera por su amistad, su compañía, su buena disposición y las enseñanzas de vida compartidas durante este Magíster.

TABLA DE CONTENIDO

RESUMEN **1**

1. INTRODUCCIÓN **3**

2. TRABAJO RELACIONADO **8**

3. HIPÓTESIS Y OBJETIVOS **11**

4. DATOS Y METODOLOGÍA **14**

4.1. DATOS 14

4.2. METODOLOGÍA 20

5. RESULTADOS **31**

6. CONCLUSIONES **43**

6.1. DISCUSIÓN 43

6.2. LIMITACIONES 44

6.3. TRABAJO FUTURO 45

7. BIBLIOGRAFÍA **47**

ANEXO **50**

Resumen

Las empresas de telecomunicaciones utilizan la información de uso de su infraestructura para poder mejorar sus servicios y captar nuevos clientes. Uno de los desafíos que tienen es identificar las relaciones familiares entre los distintos dispositivos de su red. En la literatura existen algoritmos que nos permiten, a partir del patrón de llamadas y sus metadatos (apellidos anonimizados, género y edad), determinar si es que existen relaciones de padre, madre, hermanos. Una de las limitaciones de esta técnica, es que requiere tener los metadatos para ambos dispositivos. Este no es siempre el caso. Por ejemplo, no se cuentan con los metadatos en los siguientes casos: (1) Teléfono de otra compañía, (2) Cuando un cliente tiene múltiples teléfonos, no se sabe qué familiar/amigo utiliza el teléfono.

Es por esto, que la pregunta de investigación de este proyecto es ¿Es posible implementar un modelo de machine learning capaz de identificar la relación familiar entre dos dispositivos cuando se tienen los metadatos de solo uno de los clientes?

Para poder contestar nuestra pregunta, se dispone de un dataset de llamadas de telefonía móvil anonimizadas del año 2015, de la operadora Movistar junto con metadatos asociados a los dispositivos (apellidos anonimizados, género y edad). Estas llamadas de telefonía móvil corresponden a comunicaciones realizadas exclusivamente entre clientes de Movistar para los que se tienen metadatos (de hecho excluye números que están en un plan pero que no se sabe quién es el usuario del teléfono).

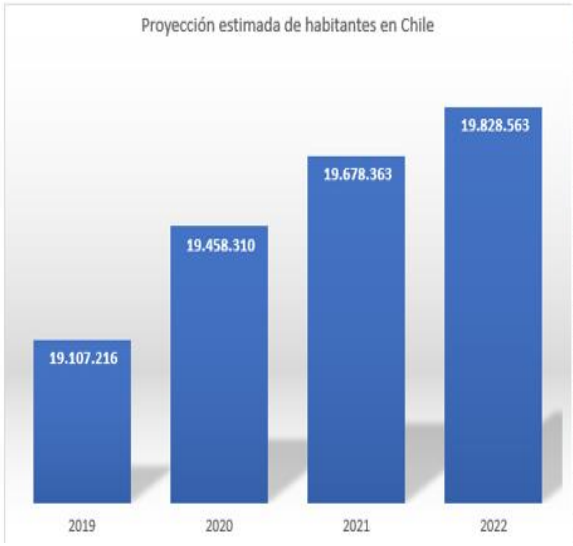
Para poder contestar la pregunta de investigación, se utilizó la técnica de David-Barret y otros al para obtener las relaciones padre-hija, padre-hijo, madre-hija, madre-hijo utilizando los patrones de comunicación con los metadatos para ambos móviles. Luego, se entrenaron diversos modelos considerando distintos niveles de disponibilidad de metadatos para los dos dispositivos.

Si bien es cierto el rendimiento de los modelos de clasificación mejora a medida que se incorporan más metadatos, no se puede considerar que se tiene buenos resultados. En consecuencia, dado el set de datos usados, no es suficiente usar como base los patrones de llamadas para la construcción y entrenamiento de modelos de clasificación para conocer la relación de parentesco en 1er grado Madre o Padre.

1. Introducción

La proyección estimada de la población total en Chile para el año 2020, es de 19.945.216 personas (Proyecciones de Población, 2020). Por otra parte, para el mes de julio 2020 la cantidad total de teléfonos celulares en estado activo registrados en las distintas operadoras de telefonía móvil del país se contabilizan en 29.875.613 (Transporte y comunicaciones-producto, 2020.).

De tal forma que es posible inferir que, a julio de 2020 en promedio cada habitante posee entre 1 y 2 teléfonos celulares activos.



TELEFONÍA MÓVIL, NÚMERO DE MÓVILES POR TIPO DE CLIENTE
JULIO 2020

Mes y año [p]	Total	Móviles Personales	Móviles Empresas
ene-19	29.756.940	26.694.422	3.062.518
feb-19	29.727.952	26.635.208	3.092.744
mar-19	29.873.064	26.716.839	3.156.225
abr-19	29.812.897	26.611.089	3.201.808
may-19	29.904.887	26.676.143	3.228.744
jun-19	29.985.029	26.770.237	3.214.792
jul-19	30.007.831	26.760.765	3.247.066
ago-19	30.026.156	26.749.829	3.276.327
sept-19	30.012.172	26.703.490	3.308.682
oct-19	29.617.384	26.276.820	3.340.564
nov-19	29.534.564	26.177.242	3.357.322
dic-19	30.015.912	26.636.642	3.379.270
ene-20	30.055.567	26.652.143	3.403.424
feb-20	30.057.303	26.626.284	3.431.019
mar-20	29.987.304	26.480.364	3.506.940
abr-20	29.860.261	26.313.241	3.547.020
may-20	29.779.603	26.222.941	3.556.662
jun-20	29.646.591	26.060.984	3.585.607
jul-20	29.875.613	26.237.298	3.638.315

Figura 1: Proyección de Total de habitantes en Chile por año. Contabilización de móviles registrados. Fuente INE.

Desde la óptica del tráfico de telefonía móvil, la cantidad de llamadas realizadas mensualmente desde enero 2020 a julio 2020 totalizan 8.542.994.000 llamadas. Donde el mes de abril 2020 tuvo menos llamadas con un total de 1.077.339.000 llamadas, y el mes de marzo 2020 registró 1.357.175.000 llamadas (Transporte y comunicaciones-producto, 2020)

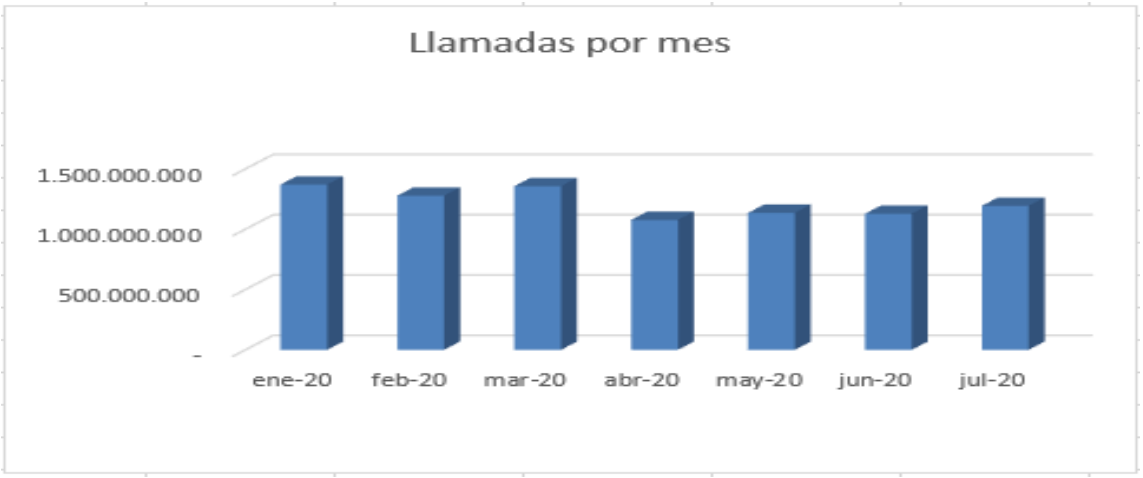


Figura 2: Total de llamadas móviles por mes, desde enero 2020 a julio 2020. Fuente INE.

Estas cifras permiten evidenciar el protagonismo de las comunicaciones usando como canal las redes de telefonía móvil.

En esta era digital, las comunicaciones móviles juegan un rol protagónico. Las personas están hiperconectadas en cualquier momento y en cualquier lugar. Las comunicaciones, entiéndase conversaciones, en la era digital ocurren de manera independiente del horario y del lugar en el mundo en que se encuentren sus interlocutores.

Es relevante para las empresas operadoras de telefonía móvil, tener una visión ampliada de sus clientes, en términos de conocer la relación familiar y/o de amistad que sus clientes tienen con quienes llaman y/o de quienes reciben llamadas. En el entendido que esta visión ampliada de sus clientes se hace extensiva a quienes se encuentran dentro de la red de llamadas, y que son clientes de otras compañías de telefonía móvil.

Las relaciones familiares se agrupan en diversas clasificaciones de acuerdo con sus características. Se entiende como relación familiar con parentesco en 1er grado, a padres, hijos y hermanos (Pariente de primer grado | NHGRI, 2020).

Dentro de las posibles ventajas competitivas que las empresas operadoras de telefonía móvil pueden lograr mediante el conocimiento de las relaciones familiares de parentesco en 1er grado de sus clientes, están las mejoras sobre la oferta comercial y la atención al cliente que se puede brindar. Los planes familiares hacen uso de la necesidad básica del sentido de pertenencia (Sedgwick & Yonge, 2008) de los clientes a un grupo, en este caso familiar. De tal manera que esta modalidad de contrato comercial refuerza la fidelidad de los clientes y disminuye el Churn.

Construir un modelo de clasificación de la relación familiar de parentesco en 1er grado; Madre o Padre, para los clientes de una operadora de telefonía móvil, usando un dataset de los CDR's anonimizados, a partir de los patrones de llamadas (David-Barrett et al., 2016) es el objetivo general propuesto a ser alcanzado. Es decir, lo que se quiere es lograr detectar la relación familiar de parentesco en 1er grado; Madre o Padre, entre clientes de Movistar y clientes de otras operadoras de telefonía Móvil.

Tener presente que las llamadas que realiza y/o recibe un cliente pueden estar tanto dentro como fuera de la red de la empresa operadora de telefonía móvil. Es decir, el modelo de clasificación de parentesco en 1er primer grado; Madre o Padre, para los clientes de la empresa operadora de telefonía móvil se hace extensivo a clientes de otras empresas operadoras de telefonía móvil, y que realizan llamadas entre sí.

En este punto es importante declarar que la información que se tendrá acerca de aquellas personas que no son clientes, en principio corresponde únicamente a la frecuencia y duración de las llamadas cursadas y recibidas. Estas variables forman parte de la base de variables predictoras para los patrones de llamadas en general (David-Barrett et al., 2016).
Para alcanzar el objetivo general anteriormente mencionado, previamente hay objetivos específicos que cumplir. El primer objetivo específico consiste explorar, entender, limpiar, depurar y asegurar la consistencia del dataset que es usado. Un aspecto relevante del dataset original es su gran tamaño, con **8.907.140 vértices y 82.342.782 arcos**.
El segundo objetivo específico dice relación con la definición e implementación de las reglas para determinar las relaciones de parentesco en 1er grado; Madre o Padre. Esto tiene que ver con la creación de una marca – propiedad del arco -, que identifique cual es la relación de parentesco en 1er grado; Madre o Padre.
El tercer objetivo es el entrenamiento de 4 modelos de clasificación, y la comparación de sus resultados y métricas. Tener presente que, como requisito de este tercer objetivo específico, desde el dataset de entrenamiento son eliminadas aquellas variables que únicamente están presentes para los clientes de Movistar.

Son consideradas entonces para el entrenamiento de los modelos clasificatorios aquellas variables que se conocen, tanto de clientes como de no clientes.

En el apartado “Trabajo Relacionado”, es mencionada la bibliografía investigada relevante y que es usada como sustento de este desarrollo.

Si bien es cierto se ha realizado una introducción a los objetivos, en el capítulo “Hipótesis y objetivos”, se entrega un mayor detalle y profundidad acerca de estos tópicos.

En el ítem “Datos y metodología” es detallado el dataset usado, su formato, su volumen, sus variables, a que fecha corresponden sus muestras. Además, se detalla la metodología usada para alcanzar los objetivos ya señalados.

En capítulo “Resultados”, son presentados los resultados obtenidos originalmente, junto con los logrados en las diferentes iteraciones realizadas, en la media en que se van incorporando variables tales como edad, género e indicador de coincidencia de apellidos entre quien hace la llamada y quien la recibe.

Finalmente, en el punto “Conclusión y Trabajo Futuro”, se informa acerca de las inferencias de los resultados obtenidos, de las limitaciones y sugerencias de trabajos futuros que se pueden realizar a partir de este.

2. Trabajo Relacionado

Las comunicaciones móviles están inmersas dentro del quehacer cotidiano, y forman parte de las tecnologías de comunicación (Ling & Horst, 2011), que se arraigan y van tomando espacio dentro del ámbito cultural y geográfico en el mundo.

Resulta interesante conocer y entender el impacto que tienen las comunicaciones móviles en las actividades humanas de la vida cotidiana (Monsivais et al., 2017).

Es posible entonces a partir de las comunicaciones de telefonía móvil, identificar patrones de relaciones sociales entre familiares y/o amistades (David-Barrett et al., 2016), considerando para ello atributos como la edad, el género, y la frecuencia y duración de las llamadas de telefonía móvil.

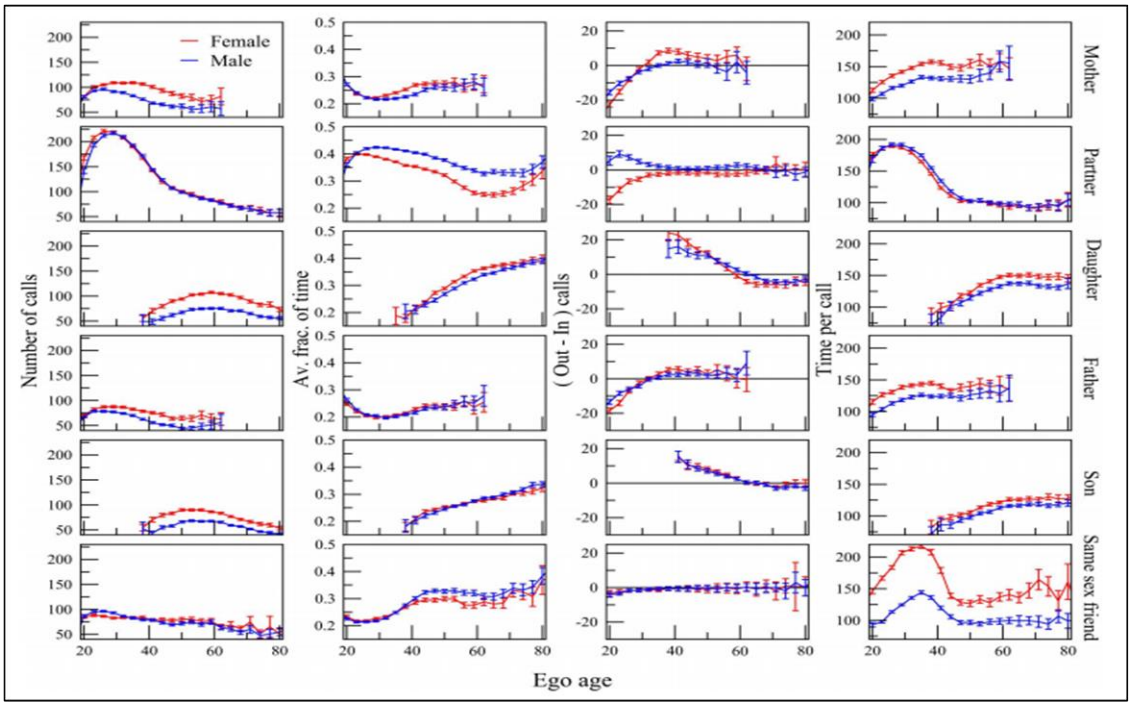


Figura 3: Patrones de comunicación a lo largo de la vida. Fuente https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5156499.

Más aún, el estudio de las comunicaciones móviles posibilita conocer las diferencias entre las relaciones de amistad muy cercanas. Dicho de otra forma, conocer las diferencias entre las relaciones de "mejores amigos", considerando para ello la edad y el género (Palchykov et al., 2013).

Continuando con la exploración de los atributos edad y género, dentro del contexto de las llamadas de telefonía móvil, se observa que ocurre una mayor frecuencia y duración de llamadas móviles para las mujeres de entre 30 y 40 años (David-Barrett et al., 2017).

Si bien es cierto, las comunicaciones móviles tales como voz, mensaje de texto, y mensajes de data (haciendo uso de redes sociales), se pueden realizar desde un mismo teléfono celular, los patrones de comunicación que se establecen son diferentes entre las llamadas móviles y el envío de mensajes de texto (Aledavood et al., 2016).

Al considerar personas de edades similares, en cuanto a su comunicación mediante llamadas de telefonía móvil, se observa que la frecuencia y duración de las llamadas es menor para las personas mayores en comparación con las personas más jóvenes (Fudolig et al., 2020).

En el caso de las personas de edad diferente, dentro del ámbito de las llamadas de telefonía móvil, se observa que personas mayores llaman con mayor frecuencia a las personas más jóvenes. También se ve que existe una tendencia en donde las madres llaman con mayor frecuencia a sus hijas que a sus hijos, y los padres llaman más a sus hijos que sus hijas (Ghosh et al., 2019).

Desde el punto de vista de los géneros, mujeres y hombres tienen distintas formas de comunicarse tanto verbalmente como de manera escrita. Estas diferencias de género también ocurren en lo que a comunicaciones móviles se refiere (Baron & Campbell, 2012).

En lo individual, se establece que los patrones de comunicaciones de telefonía móvil si bien es cierto están inmersos dentro de los ritmos circadianos de cada persona, también consideran una componente social. Esto se explica porque las llamadas son cursadas a una persona en particular y en un horario definido (Aledavood et al., 2015).

En cuanto al uso de las comunicaciones de telefonía móvil, estas pueden enmarcarse en el ámbito familiar, de amistad y laboral. En este sentido, lo que se observa es que hay una tendencia a que el objetivo más relevante corresponda con el de mantener una comunicación continua con los familiares y amistades (Wajcman, et al., 2008).

3. Hipótesis y Objetivos

En el contexto de las comunicaciones telefónicas móviles, es posible inferir patrones de relaciones sociales entre familiares y amistades a lo largo del curso de la vida (David-Barrett et al., 2016). Es posible también entonces modelar y clasificar ciertas relaciones familiares usando para ello como insumo básico las llamadas de telefonía móvil que se registran entre las distintas personas que se comunican usando una red de telefonía móvil particular, sean estas personas clientes o no de una misma empresa operadora de telefonía móvil.

Dicho esto, el interés del proyecto de grado está focalizado en conocer las relaciones familiares de parentesco en 1er grado que existen entre las personas que se comunican usando la red de telefonía móvil de la empresa Movistar, independientemente de que estas personas sean o no clientes de la empresa Movistar.

La compañía telefónica al conocer las relaciones familiares de sus clientes, entonces contará con el conocimiento que le permitirá generar mejores ofertas comerciales, que le permitirán además brindar una mejor atención al cliente. Logrando en consecuencia avanzar hacia un conducta de cliente más fiel, reduciendo el Churn e incluso pudiendo aumentar el "Engagement" con el cliente y además aumentar la captación de nuevos clientes que forman parte de la red familiar de un cliente antiguo de la compañía telefónica.

Entonces a continuación es enunciado formalmente el objetivo general del proyecto de grado:

"*Construir modelos de clasificación de la relación familiar parentesco en 1er grado, usando para ello un dataset de CDR's anonimizado, a partir de los patrones de llamadas*."

Tener presente que el dataset a ser usado, corresponde a llamadas de telefonía móvil registradas por la empresa Movistar, durante el año 2015. Este data-set se encuentra enriquecido con información adicional sensible, que está anonimizada para cumplir con la regulación vigente. Mayores detalles del dataset son señalados en el apartado "4.1. Datos".

Para alcanzar el objetivo general ya enunciado, se requiere cumplir con objetivos específicos, que permitan mediante su sinergia elaborar y conformar el núcleo y sustento del objetivo general.

Entonces a continuación son enumerados los objetivos específicos a ser logrados:

1. En cuanto al dataset usado con los CDR´s anonimizados, explorar, analizar, hacer limpieza y depuración de este, asegurando su consistencia.
2. Puesto que no sé conoce explícitamente quienes tienen una relación de parentesco en 1er grado; Madre o Padre dentro del dataset, entonces es que se debe construir y determinar esa relación. Para ello entonces, se requiere definir e implementar las lógicas asociadas a la determinación de las relaciones de parentesco en 1er grado; Madre o Padre, basándose en la metodología Barret et al.

3. Entrenar y probar 4 modelos de clasificación. Se persigue con esto, aumentar las posibilidades de encontrar un modelo de clasificación que se ajuste de mejor manera al data set empleado en términos lograr un buen resultado.
4. Enriquecer las variables predictoras a ser usadas para la construcción y entrenamiento de los modelos de clasificación, usando para ello técnicas de "Feature Engineering"
5. Determinar el modelo de clasificación con mejor rendimiento, mediante la comparativa de sus métricas.

4. Datos y Metodología

4.1. Datos

Tal como el título del proyecto de grado adelanta, el dataset usado corresponde en su origen a CDR's anonimizados. Ahora es detallado el dataset en términos de la fecha a la cual corresponden las muestras disponibles, el formato del dataset, el tamaño del dataset y las variables que lo conforman.

Se dijo que el dataset corresponde a un grafo, y en ese sentido se debe tener presente que los elementos que lo constituyen corresponden a los vértices y a las aristas (Palchykov et al., 2013). En donde cada vértice equivale a un cliente, que tiene atributos a modo de ejemplo tales como edad, género, etc. Por otro lado, cada arista representa una relación de un cliente con otro cliente. En este caso al ser un grado dirigido (Bollobas, 2013), cada arista representa las llamadas desde un nodo A, hacia un nodo B.

Las aristas también tienen propiedades, como por ejemplo cantidad de llamadas, cantidad de segundos de duración de las llamadas, etc.

La muestra corresponde a las llamadas de telefonía móvil entre clientes de la empresa Movistar, registradas el año 2015. Es decir, en el dataset no hay llamadas telefónicas que estén asociadas a clientes de otra operadora de telefonía móvil. No obstante lo ya antes mencionado, en la figura 4 es mostrado un ejemplo general de red de llamadas que considera clientes de Movistar y también clientes de otras compañías.

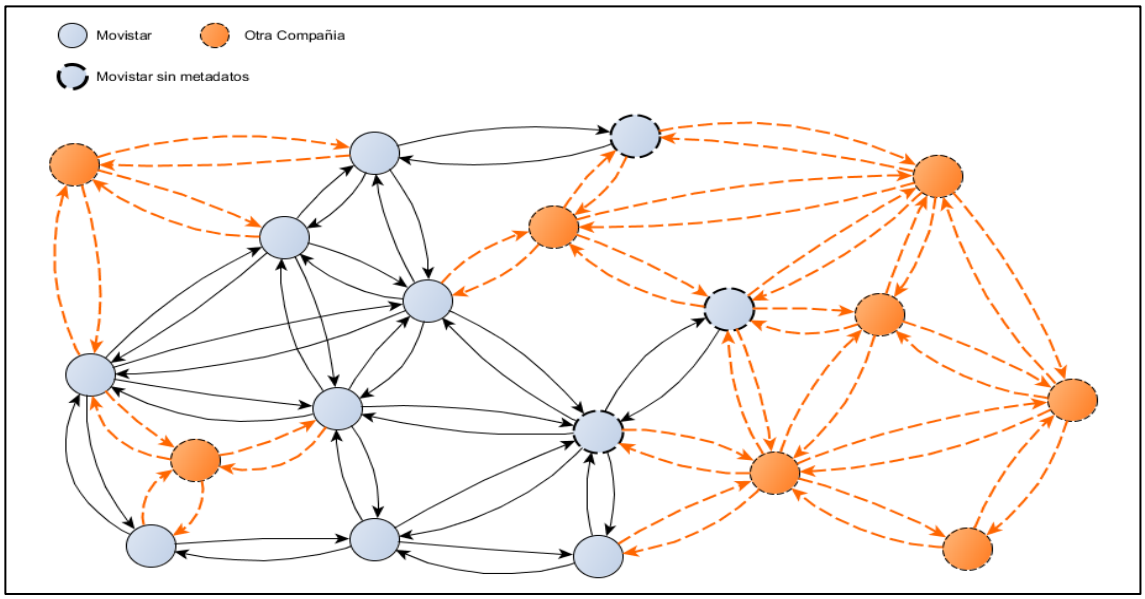


Figura 4: Grafo general de ejemplo de la red de llamadas móviles registradas entre clientes de Movistar y también con clientes de otras compañías. Fuente de elaboración propia.

En la red de llamadas presenta en la figura 4, se ven 3 tipos de clientes:

a) Clientes de Movistar, para los cuales son conocidos metadatos muy relevantes para la identificación de la relación parentesco en 1er grado; Madre o Padre, tales como la edad, su género, y también los apellidos. Tener presente que estos metadatos per se permiten implementar algoritmos para la identificación la relación parentesco en 1er grado; Madre o Padre.

b) Clientes de Movistar, para los cuales no son conocidos algunos o bien la totalidad de los metadatos edad, género y apellidos.

c) Clientes de otras compañías de telefonía móvil, para los cuales no se conocen los metadatos edad, género ni apellidos.

Mencionado lo anterior, es decir que no tenemos metadatos para algunos clientes, no nos es posible deducir el parentesco entre ellos, es por esto que inicialmente nos restringiremos a los clientes para los que cuales tenemos sus metadatos. La figura 5, presenta el grafo ejemplo con la red de llamadas correspondiente al dataset usado.

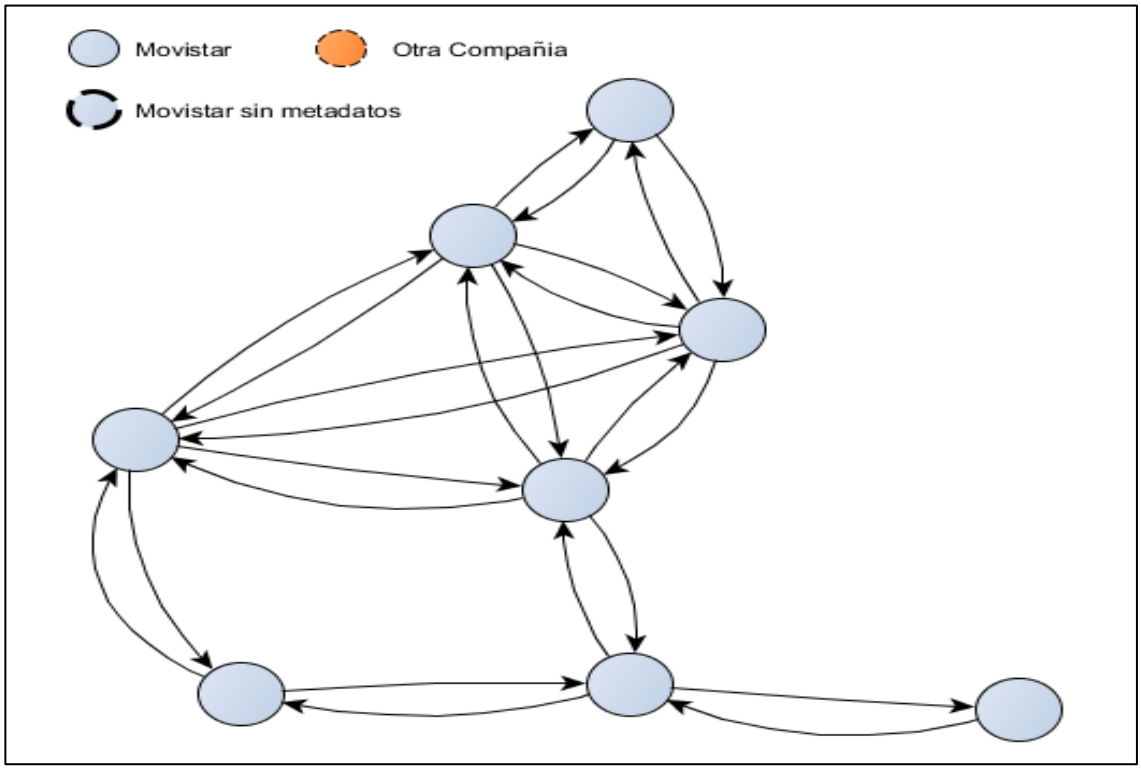


Figura 5: Grafo de la red de llamadas móviles, del dataset usado. Fuente de elaboración propia.

Estas llamadas se encuentran agrupadas por año y en donde los datos sensibles como por ejemplo los números telefónicos, están anonimizados.

Las variables usadas para la agrupación anual ya comentada corresponden a la cantidad de llamadas cursadas hacia algún número telefónico en particular, la cantidad de llamadas recibidas desde algún número telefónico puntual, y junto con ello, también agrupa la duración en segundos de las llamadas cursadas y recibidas, respectivamente.

Resulta relevante explicar que el dataset se encuentra en el formato de archivo ".gt". El formato de archivo ".gt", es un formato simple binario diseñado para almacenar grafos usando la herramienta Graph-Tool (The gt file format — graph-tool 2.35 documentation, 2020).

Esto implica que las actividades de exploración y manipulación del dataset requieren ser realizadas exclusivamente usando la herramienta Graph-Tool. Por cierto, que esto tiene ventajas y limitaciones. Las ventajas dicen relación con la variedad de funciones disponibles y la buena performance asociada al uso de la herramienta Graph-Tool (Quick start using graph-tool — graph-tool 2.35 documentation, 2020). En términos de las limitaciones que tiene el uso de archivos en formato ".gt", la herramienta Graph-Tool, no incorpora algoritmos de Machine Learning. Por lo que se debe realizar una transformación de la data para la construcción de los modelos de clasificación, de acuerdo con el objetivo perseguido.

Cabe mencionar que la herramienta Graph-Tool, no se encuentra dentro del pool de herramientas de uso masificado, en el ámbito de la Ciencias de los Datos. En ese sentido, requiere tiempo para su estudio y familiarización. Y aun cuando existe documentación, los ejemplos son escasos y además hay un incipiente background de foros en el internet que orienten y ayuden a solucionar las dificultades que vayan surgiendo en la medida que se avance en la exploración y manipulación de la data.

A continuación, en las figuras 6 y 7, se muestra el Layout del dataset usado, desagregado por vértice y arista, junto con sus propiedades y sus descripciones asociadas.

	PROPIEDAD	DESCRIPCIÓN
VÉRTICE	Age	Edad
	IdOwner	Id del dueño del plan "hasheado"
	Ln1	1° apellido "hasheado"
	Ln2	2° apellido "hasheado"
	Sex	Género 0: Mujer, 1: Hombre
ARCO	CallsIn	# de llamadas entrantes
	CallsOut	# de llamadas salientes
	Rank	Ranking
	secIn	# segundos de las llamadas entrantes
	secOut	# segundos de las llamadas salientes

Figura 6: Atributos del dataset. Fuente de elaboración propia.

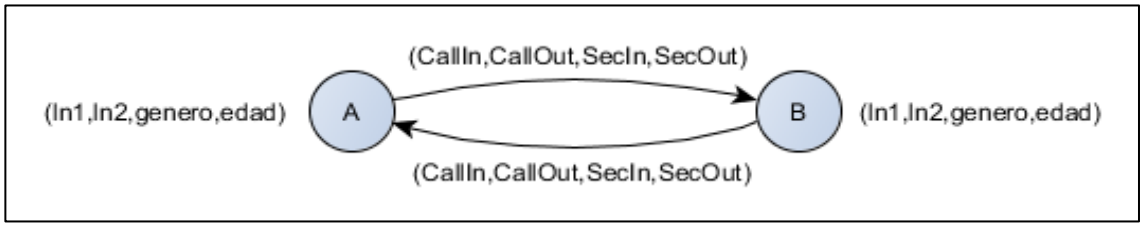


Figura 7: Grafo explicativo del dataset, en términos de las variables predictoras a ser usadas para los modelos de clasificación. Fuente de elaboración propia.

Algunas consideraciones para tener presente acerca de las variables del dataset.

En lo que se refiere a las propiedades asociadas al vértice se tiene:

- Age: corresponde a la edad del cliente. Cuando Age < 18, entonces se refiere a clientes extranjeros, o bien retornados. En el caso de que Age = 0, corresponde a una inconsistencia en la información.

En lo que se refiere a las propiedades asociadas a arista se tiene:

- CallsIn: cantidad total de llamadas recibidas el año 2015, desde un fono en particular.
- CallsOut: cantidad total de llamadas cursadas el año 2015, hacia un fono en particular.
- Rank: es el ranking de frecuencia de llamadas, en donde 1 es la mayor frecuencia de llamadas, 2 es el segundo, y así sucesivamente.
- secIn: cantidad total de segundos de duración de las llamadas recibidas el año 2015, desde un fono en particular.
- secOut: cantidad total de segundos de duración de las llamadas cursadas el año 2015, hacia un fono en particular.

En cuanto al volumen o tamaño del dataset, se tiene la siguiente información:

- Cantidad de vértices: **8.907.140**
- Cantidad de aristas: **82.342.782**
- Tamaño en GB: **3,4GB**

Por último, se aclara que puesto que no se cuenta con los datos base del dataset, entonces el margen de error es desconocido para este caso.

4.2. Metodología

La metodología de trabajo desde el punto de vista de su arquitectura es un híbrido entre tareas secuenciales e iteración de actividades (A hybrid model for IT project with Scrum, 2020). Con ello lo que se busca es asegurar el avance según la planificación, y velar por la mejora continua de los resultados logrados.

Resulta clave profundizar en el objetivo perseguido dentro del contexto de la data con la que cuenta. Esto es, se quieren construir modelos de clasificación para identificar la relación de parentesco en 1er grado; Madre o Padre, de los clientes de Movistar, usando los CDR's registrados y disponibilizados por la empresa. Entonces, hay que tener en cuenta que un cliente de Movistar cursa y recibe llamadas de clientes de otras empresas de telefonía. Por lo tanto, de cara a la construcción de los modelos de clasificación, se deben considerar también las llamadas móviles entre clientes y no clientes de Movistar.

A priori se tiene que el dataset usado considera exclusivamente las muestras de llamadas realizadas entre los clientes de Movistar. Este dataset de llamadas entre clientes de Movistar contiene información tal como por ejemplo primer apellido, segundo apellido, edad, etc., que se conoce dada la característica de ser clientes de Movistar. En el caso de no ser cliente de Movistar, la información que se tiene registrada es cantidad de llamadas recibidas desde un cliente de Movistar, cantidad de llamadas cursadas hacia un cliente de Movistar, y la duración total en segundos asociada.

Por lo tanto, al momento de realizar el entrenamiento de los modelos de clasificación, se deben considerar únicamente como variables predictoras aquellas variables que correspondan a data asociada tanto a clientes como a no clientes de Movistar. Es decir, se deben eliminar aquellas variables obtenidas desde la característica de ser cliente de Movistar. Por ejemplo, se deben eliminar variables tales como edad, apellidos y género. Así como se deben eliminar variables por lo antes explicado, también existe la posibilidad de incorporar nuevas variables usando para ello técnicas de "Feature Engineering" (Zheng & Casari, 2018).

A continuación, se presentan y explican cada una de las etapas que conforman la metodología de trabajo:

1. *Exploración, análisis, limpieza del dataset, selección de vértices y edges*. En esta fase de la metodología de trabajo, se realiza la exploración del dataset de tal manera de conocer y familiarizarse con este. Junto con ello se realiza un primer análisis en términos de sus muestras, los valores asociados, y la consistencia de las mismas. De cara a la consistencia por ejemplo se valida y aplica un filtro para que la edad sea mayor a 0. Es decir que la propiedad Age de los vértices sea mayor a 18. La aplicación del filtro sobre vértices que no cumplan con la condición antes mencionada también debe acompañarse con la eliminación de las aristas que se encuentren relacionadas. Esto tiene como resultado un nuevo dataset, con el mismo Layout que el dataset original y con menos muestras.

2. *Construcción del set de datos de training y test*. Esta fase está compuesta a su vez de sub-fases. En términos generales esta fase genera como resultado el set de datos de entrenamiento y prueba para ser usado por los modelos de clasificación a ser instanciados en una etapa posterior. Son indicadas las sub-fases:

   2.1. *Determinación reglas para la relación de parentesco en 1er grado*. La variable objetivo para el entrenamiento de los modelos de clasificación no existe en el dataset. Es por ello que se realiza la construcción de la variable objetivo, que en este caso corresponde a la relación de parentesco en 1er grado; Madre o Padre.

   Es así como es creada una nueva propiedad PG1 como parte de los atributos que tienen las aristas, en donde su dominio de valores representa lo siguiente:

   - PG1 = 0, sin relación de parentesco en 1er grado. Esto quiere decir que el vértice(cliente) A(origen) no tiene relación de parentesco en 1er grado Madre ni Padre con el vértice(cliente) B(destino).
   - PG1 = 7, es Madre. Esto quiere decir que el vértice(cliente) A(origen) es Madre del vértice(cliente) B(destino).
   - PG1 = 11, es Padre. Esto quiere decir que el vértice(cliente) A(origen) es Padre del vértice(cliente) B(destino).

   Para completar el valor de la propiedad PG1, se debe entonces definir todas las reglas que determinarán como se construyen las estas relaciones Madre o Padre.

Resulta relevante entender que las reglas para determinar la relación de parentesco en 1er grado; Madre o Padre, están basadas en la metodología Barrett et al, pero además incorporan la relación que se establece usando los apellidos.

- Regla para determinar la relación Madre. El vértice A, es Madre del vértice B, si se cumple con las siguientes condiciones:
    - El primer apellido (Ln1) del vértice A es igual al segundo apellido (Ln2) del vértice B
    - La edad (Age) del vértice A es mayor que la edad (Age) del vértice B. Esta diferencia de edades tiene que ser mayor a 15 años.
    - El género del vértice A es mujer (Sex=0).

- Regla para determinar la relación Padre. El vértice A, es Padre del vértice B, si se cumple con las siguientes condiciones:
    - El primer apellido (Ln1) del vértice A es igual al primer apellido (Ln1) del vértice B
    - La edad (Age) del vértice A es mayor que la edad (Age) del vértice B. Esta diferencia de edades tiene que ser mayor a 17 años.
    - El género del vértice A es varón (Sex=1).

Como resultado de estas reglas tenemos entonces un nuevo atributo PG1 a nivel de arista, correspondiente a la variable objetivo Madre, Padre, o Sin Relación, según el cálculo realizado a partir de los metadatos edad (Age), género (Sex) y apellidos (Ln1 y Ln2). Con esta variable objetivo calculada es que se realiza el entrenamiento de los modelos de clasificación.

2.2. *Generación del dataset para entrenamiento y prueba.* En esta sub-fase se tiene como resultado un nuevo dataset a ser usado como base para el posterior entrenamiento y prueba de los modelos de clasificación. El dataset que hasta ahora ha sido usado, deja de ser un grafo de extensión ".gt" y entonces se transforma en un nuevo dataset de extensión ".csv". Es muy importante destacar que por motivos de simplificación, para al entrenamiento y prueba de los modelos de clasificación realmente se usan 2 dataset independientes; en el primero de ellos su variable objetivo contiene las 2 clases "Sin Relación" y "Madre", en tanto que el segundo dataset se consideran las 2 clases "Sin Relación" y "Padre. Tal como se indicó antes, el dataset base es el mismo, por lo tanto si bien es cierto hay un dataset para la clases "Sin Relación" y "Madre" y otro dataset para las clases "Sin Relación" y "Padre", la cardinalidad(cantidad de muestras), de ambos dataset es la misma. Por su puesto que la distribución de las clases "Sin Relación", "Madre" y "Padre" difieren entre ambos dataset. Esto se puede apreciar en el apartado "5.Resultados".

Puesto que el dataset contiene metadata que es propia de los clientes de Movistar, y dado que lo que se quiere hacer entrenar y testear modelos de clasificación tanto para clientes de Movistar como clientes de otras compañías, entonces simulamos una red de llamadas que cumpla con ello. Para eso lo que se hace es eliminar las variables predictoras edad(Age), género(Sex) y apellidos(Ln1 y Ln2) para los clientes destinos.

A continuación se presentan las figuras 8 y 9, en donde son presentados los atributos de los 2 dataset ya mencionados. Se ve que ambos dataset tienen los mismos atributos solo diferenciándose por las clases de la variable objetivo PG1.

VARIABLES	DESCRIPCIÓN
PG1	Valor 0 es clase "Sin Relación", valor 7 es clase "Madre".
Origen	Cliente origen de la llamada, es un índice dentro del dataset.
Ori_Age	Edad del cliente origen
Ori_Sex	Género del cliente origen; Mujer = 0, Varón = 1.
Destino	Cliente destino de la llamada, es un índice dentro del dataset.
CallsIn	Cantidad de llamadas que el cliente origen recibe del cliente destino en 2015.
callsout	Cantidad de llamadas que el cliente origen hace al cliente destino en 2015.
secIn	Total de segundos asociados a CallsIn.
secOut	Total de segundos asociados a secOut.

Figura 8: Dataset en su primera instancia, considerando las clases “Sin Relación” y “Madre”, para su variable objetivo.

VARIABLES	DESCRIPCIÓN
PG1	Valor 0 es clase "Sin Relación", valor 11 es clase "Padre".
Origen	Cliente origen de la llamada, es un índice dentro del dataset.
Ori_Age	Edad del cliente origen
Ori_Sex	Género del cliente origen; Mujer = 0, Varón = 1.
Destino	Cliente destino de la llamada, es un índice dentro del dataset.
CallsIn	Cantidad de llamadas que el cliente origen recibe del cliente destino en 2015.
callsout	Cantidad de llamadas que el cliente origen hace al cliente destino en 2015.
secIn	Total de segundos asociados a CallsIn.
secOut	Total de segundos asociados a secOut.

Figura 9: Dataset en su primera instancia, considerando las clases "Sin Relación" y "Padre", para su variable objetivo.

En una segunda instancia son agregadas nuevas variables predictoras usando para ello la técnica de "Feature Engineering". Estas nuevas variables predictoras son:

- %_CallsIn, corresponde al porcentaje de llamadas recibidas por el fono A desde el fono B, sobre el total de llamadas recibidas por el fono A.
- %_CallsOut, corresponde al porcentaje de llamadas realizadas por el fono A hacia el fono B, sobre el total de llamadas realizadas por el fono A.
- %_SecIn, corresponde al porcentaje de segundos asociados a las llamadas recibidas por el fono A desde el fono B, sobre el total de segundos asociados a todas las llamadas recibidas por el fono A.
- %_SecOut, corresponde al porcentaje de segundos asociados a las llamadas realizadas por el fono A hacia el fono B, sobre el total de segundos asociados a todas las llamadas realizadas por el fono A.

Las siguientes figuras 10 y 11, los nuevos dataset incorporando las variables predictoras %_CallsIn, %_CallsOut, %_SecIn y %_SecOut.

VARIABLES	DESCRIPCIÓN
PG1	Valor 0 es clase "Sin Relación", valor 7 es clase "Madre".
Origen	Cliente origen de la llamada, es un índice dentro del dataset.
Ori_Age	Edad del cliente origen
Ori_Sex	Género del cliente origen; Mujer = 0, Varón = 1.
Destino	Cliente destino de la llamada, es un índice dentro del dataset.
CallsIn	Cantidad de llamadas que el cliente origen recibe del cliente destino en 2015.
callsout	Cantidad de llamadas que el cliente origen hace al cliente destino en 2015.
secIn	Total de segundos asociados a CallsIn.
secOut	Total de segundos asociados a secOut.
%_CallsIn	Porcentaje de llamadas que cliente origen recibe del cliente destino, respecto del total de llamadas que el cliente origen recibe de todos sus destinos en 2015.
%_CallsOut	Porcentaje de llamadas que cliente origen hace al cliente destino, respecto del total de llamadas que el cliente origen hace a todos sus destinos en 2015.
%_SecIn	Porcentaje de segundos asociados a CallsIn respecto del total de segundos asociados a todas las llamadas que el cliente origen recibe en 2015.
%_SecOut	Porcentaje de segundos asociados a secOut respecto del total de segundos asociados a todas las llamadas que el cliente origen realiza en 2015.

Figura 10: Dataset en su primera instancia, considerando las clases "Sin Relación" y "Madre", para su variable objetivo.

VARIABLES	DESCRIPCIÓN
PG1	Valor 0 es clase "Sin Relación", valor 11 es clase "Padre".
Origen	Cliente origen de la llamada, es un índice dentro del dataset.
Ori_Age	Edad del cliente origen
Ori_Sex	Género del cliente origen; Mujer = 0, Varón = 1.
Destino	Cliente destino de la llamada, es un índice dentro del dataset.
CallsIn	Cantidad de llamadas que el cliente origen recibe del cliente destino en 2015.
callsout	Cantidad de llamadas que el cliente origen hace al cliente destino en 2015.
secIn	Total de segundos asociados a CallsIn.
secOut	Total de segundos asociados a secOut.
%_CallsIn	Porcentaje de llamadas que cliente origen recibe del cliente destino, respecto del total de llamadas que el cliente origen recibe de todos sus destinos en 2015.
%_CallsOut	Porcentaje de llamadas que cliente origen hace al cliente destino, respecto del total de llamadas que el cliente origen hace a todos sus destinos en 2015.
%_SecIn	Porcentaje de segundos asociados a CallsIn respecto del total de segundos asociados a todas las llamadas que el cliente origen recibe en 2015.
%_SecOut	Porcentaje de segundos asociados a secOut respecto del total de segundos asociados a todas las llamadas que el cliente origen realiza en 2015.

Figura 11: Dataset en su primera instancia, considerando las clases "Sin Relación" y "Padre", para su variable objetivo.

3. *Entrenamiento de modelos de clasificación.* En esta fase se realiza la instanciación y entrenamiento de los 4 modelos de clasificación seleccionados, Regresión Logística (Li, 2019), LightGBMClassifier (Welcome to LightGBM's documentation! — LightGBM 3.0.0.99 documentation, 2020), XGBClassifier (XGBoost Documentation — xgboost 1.3.0-SNAPSHOT documentation, 2020) y BalancedBaggingClassifier (Imbalanced Datasets with Imbalanced-Learn, 2018). En cuanto a la forma en que se realiza la selección de los datos para el entrenamiento y prueba, en primera instancia se aplica a nivel de fonos, para luego ser traspasada esta selección de entrenamiento y prueba a nivel de las muestras.

Por otra parte, fase de entrenamiento y prueba se realiza con la totalidad de las muestras del dataset. Esto último plantea un desafío técnico.

4. *Incorporación de variables e iteración en el entrenamiento de los modelos de clasificación.* Para mejorar el rendimiento de los modelos de clasificación, entonces se opta por la estrategia de incorporar de manera gradual las variables originalmente presentes en el set de datos.

   Con esto lo que se persigue es ir aumentando el rendimiento de los modelos de clasificación, hasta completar el espectro de las variables que fueron usadas para definir e implementar las reglas con las cuales se logró determinar las relaciones de parentesco en 1er grado, Madre o Padre.

   La incorporación de variables es en forma sucesiva e iterativa es:

   4.1. Iteración 1, se agregan indicadores de Ranking y de Grado

   - Rank, es el ranking de frecuencia de llamadas, en donde 1 es la mayor frecuencia de llamadas, 2 es el segundo, y así sucesivamente.
   - OutDegree, si A es el origen de las llamadas, entonces corresponde a la cantidad total de destinos a los cuales llama A.
   - InDegree, si A es el origen de las llamadas, entonces corresponde a la cantidad de teléfonos que llaman al origen A.

   Tener presente que en el "Anexo" se incorporan nuevas iteraciones, en donde, son agregadas las variables edad, género y apellido, de manera paulatina.

Esto lo que persigue por una parte es ir monitoreando la mejora del rendimiento de los modelos de clasificación en la medida que se van incorporando las variables predictoras ya mencionadas, y por otra parte se quiere comprobar la lógica implementada mediantes algoritmos para construir las variables objetivos Madre y Padre, pero esta vez usando los modelos de clasificación.

5. Resultados

Para presentar los resultados obtenidos, es usada la metodología a modo de guía e hilo conductor de las actividades realizadas. De esta manera se tiene entonces en primera instancia la presentación de los resultados de la actividad "*Exploración, análisis, limpieza del dataset, selección de vértices y edges*"

Las figuras 12 y 13 muestran el resultado alcanzado en términos de la depuración del dataset original. Se ve que al realizar el filtro de los vértices mediante condición la propiedad Age > 18, y a continuación realizando la eliminación de los arcos asociados hay una disminución significativa de vértices desde **8.907.140** a **3.154.126**, es decir se disminuye a un **35%** de la cantidad original de vértices. En tanto que, en el caso de los arcos, estos disminuyen desde **82.342.782** hasta **27.637.782**, es decir, se disminuye a un **34%** de la cantidad original de arcos. La gran diferencia entre las cantidades originales tanto de vértices como de arcos, versus sus cantidades depuradas, se explica porque son eliminados los vértices que cumplen con ser empresa, ser extranjero, ser retornados, o bien en última instancia presentar alguna inconsistencia en su metadata.

	DATASET		
	ORIGINAL	DEPURADO	% REDUCCIÓN
# VÉRTICES	8.907.140	3.154.126	35%
# ARCOS	82.342.782	27.637.782	34%

Figura 12: Tabla del tamaño del dataset original versus el dataset depurado. Fuente de elaboración propia.

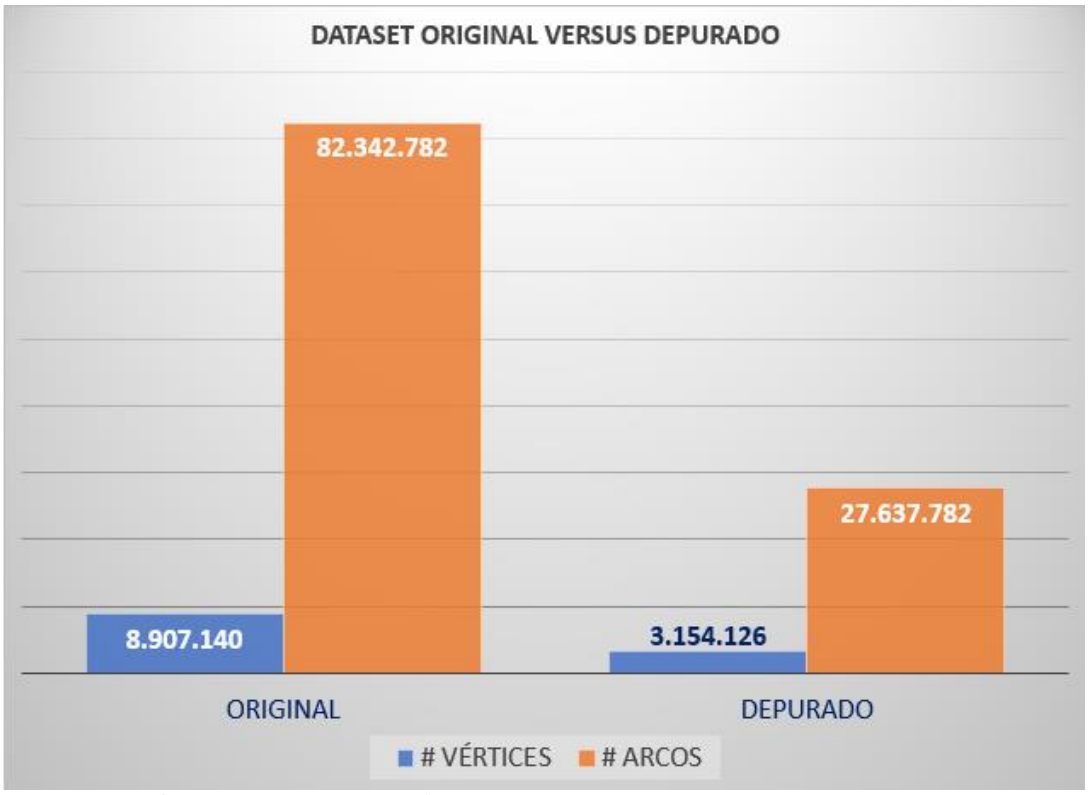


Figura 13: Grafico comparativo del tamaño del dataset original versus el dataset depurado. Fuente de elaboración propia

Los resultados de la segunda actividad de la metodología, “Construcción del set de datos de training y test” consideran la definición e implementación de las reglas para determinación la relación de parentesco en 1er grado; Madre o Padre. Que en definitiva corresponde a la variable ”target” del dataset de cara al entrenamiento de los de los modelos de clasificación.

Desde el punto de vista de los arcos, en las figuras 14 y 15 se observan los valores alcanzados en cuanto a la determinación de las relaciones de parentesco en 1er grado; Madre o Padre.

	PARENTESCO EN 1° GRADO		
	SIN RELACIÓN	MADRE	PADRE
# ARCOS	26.791.268	512.852	333.662
% ARCOS	96,94%	1,86%	1,21%

Figura 14: Tabla de relaciones de parentesco en 1er grado, para arcos. Fuente de elaboración propia.

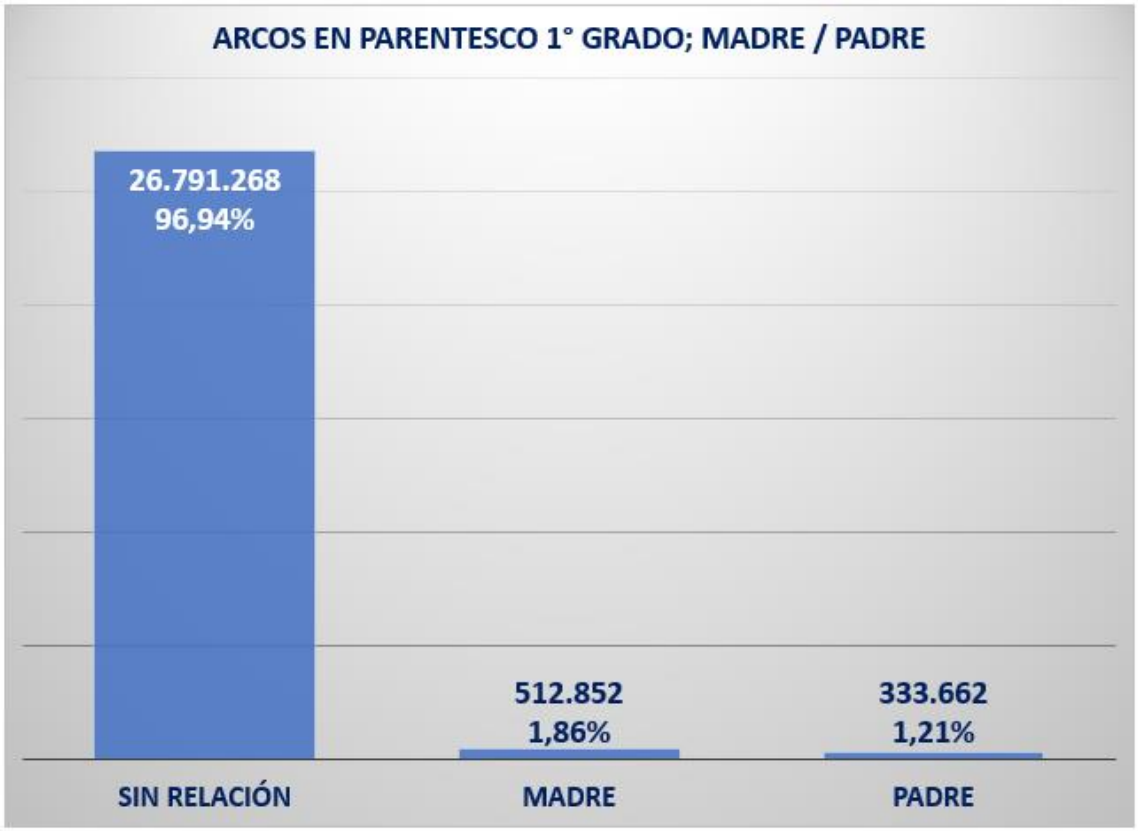


Figura 15: Gráfico de relaciones de parentesco en 1er grado para arcos. Fuente de elaboración propia.

Desde la óptica de los vértices, en las figuras 16 y 17 son ilustrados los valores obtenidos a partir de los algoritmos implementados, usando las reglas ya explicadas para la determinación de las relaciones de parentesco en 1er grado.

	PARENTESCO EN 1° GRADO		
	SIN RELACIÓN	MADRE	PADRE
# VÉRTICES	2.553.230	355.982	244.914
% VÉRTICES	80,95%	11,29%	7,76%

Figura 16: Tabla de relaciones de parentesco en 1er grado, para vértices. Fuente de elaboración propia.

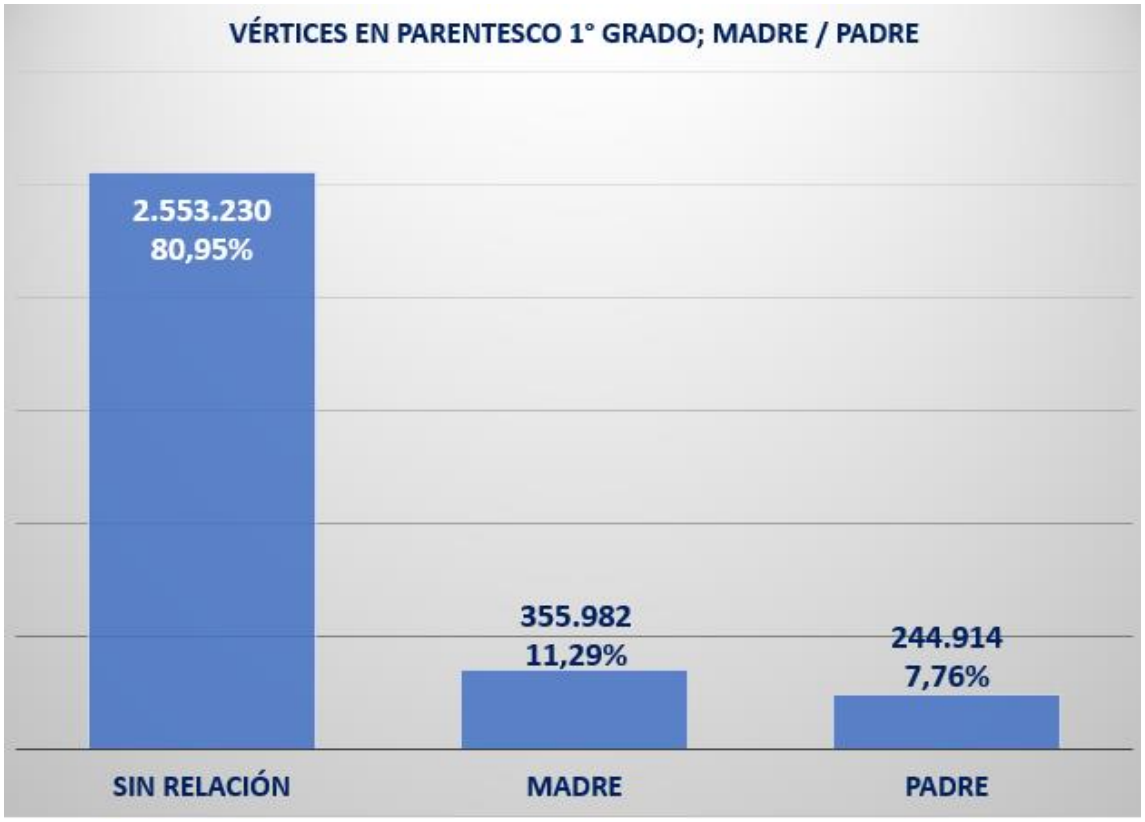


Figura 17: Gráfico de relaciones de parentesco en 1er grado, para vértices. Fuente de elaboración propia.

En lo que se refiere a los resultados obtenidos en la fase "*Entrenamiento de modelos de clasificación*", el entrenamiento y prueba de los modelos de clasificación se realiza con el set de datos completo, es decir, con las **27.637.782** muestras.

Es relevante entender la forma en que se realiza el Split de la data. En una primera instancia esta tarea se efectúa a nivel de fonos que existen en el dataset y que en total corresponden a **3.154.126** fonos. En una segunda instancia, el Split ya realizado es "traspasado" a nivel de muestras, y que en total corresponden a **27.637.782** muestras.

En las figuras 18 y 19, se muestra el primer Split realizado, en donde se agrupa por fono, y además se efectúa una separación, según la relación de parentesco en 1er grado sea Madre o Padre.

	AGRUPADO POR VÉRTICE(FONO-ORIGEN)			
	SIN RELACION	MADRE	%	TOTAL
ENTRENAMIENTO	2.098.622	266.972	13%	2.365.594
PRUEBA	699.522	89.010	13%	788.532
TOTAL	2.798.144	355.982		3.154.126

Figura 18: Tabla con Split de data agrupado por fono origen, cuando la relación es "MADRE". Fuente de elaboración propia.

	AGRUPADO POR VÉRTICE(FONO-ORIGEN)			
	SIN RELACION	PADRE	%	TOTAL
ENTRENAMIENTO	2.182.234	183.360	8%	2.365.594
PRUEBA	726.978	61.554	8%	788.532
TOTAL	2.909.212	244.914		3.154.126

Figura 19: Tabla con Split de data agrupado por fono origen, cuando la relación es "PADRE". Fuente de elaboración propia.

Por otro lado, en las figuras 20 y 21, se muestra el Split desde la óptica de las llamadas telefónicas (las muestras), y también de manera separada, según la relación de parentesco en 1er grado sea Madre o Padre.

	AGRUPADO POR ARCOS(ÓPTICA LLAMADAS)			
	SIN RELACION	MADRE	%	TOTAL
ENTRENAMIENTO	20.349.822	384.642	1,89%	20.734.464
PRUEBA	6.775.108	128.210	1,89%	6.903.318
TOTAL	27.124.930	512.852		27.637.782

Figura 20: Tabla con Split de data, cuando la relación es "MADRE". Fuente de elaboración propia.

	AGRUPADO POR ARCOS(ÓPTICA LLAMADAS)			
	SIN RELACION	PADRE	%	TOTAL
ENTRENAMIENTO	20.484.588	249.876	1,22%	20.734.464
PRUEBA	6.819.532	83.786	1,23%	6.903.318
TOTAL	27.304.120	333.662		27.637.782

Figura 21: Tabla con Split de data, cuando la relación es "PADRE". Fuente de elaboración propia.

Corresponde ahora presentar los resultados alcanzados por los modelos de clasificación ya entrenados. En las figuras 22, 23, 24 y 25 se muestran las métricas para medir el rendimiento de los 4 modelos de clasificación, para el primer experimento realizado. En tanto que las figuras 22 y 23 dice relación con la clasificación de la variable objetivo Madre, las figuras 24 y25 se refieren a la clasificación de la variable objetivo Padre. En ambos experimentos, las variables predictoras usadas son la mismas y también se encuentran ilustradas en cada una de las figuras.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.05	0.00	0.01	128210
	Accuracy			0.98	6903318
	Macro AVG	0.52	0.50	0.50	6903318
	Weighted AVG	0.96	0.98	0.97	6903318
	Accuracy Score		0.9805		
LGBMClassifier	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.31	0.07	0.11	128210
	Accuracy			0.98	6903318
	Macro AVG	0.65	0.53	0.55	6903318
	Weighted AVG	0.97	0.98	0.97	6903318
	Accuracy Score		0.9798		
BalancedBagging Classifier	Sin Relación	1.00	0.79	0.89	6775108
	Madre	0.08	0.95	0.15	128210
	Accuracy			0.80	6903318
	Macro AVG	0.54	0.87	0.52	6903318
	Weighted AVG	0.98	0.80	0.87	6903318
	Accuracy Score		0.7975		
XGBClassifier	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.50	0.01	0.02	128210
	Accuracy			0.98	6903318
	Macro AVG	0.74	0.51	0.51	6903318
	Weighted AVG	0.97	0.98	0.97	6903318
	Accuracy Score		0.9814		

Confusion Matrix		
	Sin Relación	Madre
Sin Relación	6768516	6592
Madre	127834	376

	Sin Relación	Madre
Sin Relación	6755611	19497
Madre	119290	8920

	Sin Relación	Madre
Sin Relación	5384366	1390742
Madre	6557	121653

	Sin Relación	Madre
Sin Relación	6773675	1433
Madre	126766	1444

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Rank	NO
OutDegree	NO
InDegree	NO

Figura 22: Tabla con métricas de los modelos de clasificación para la clase "Madre", y tabla de las variables predictoras usadas. Fuente de elaboración propia.

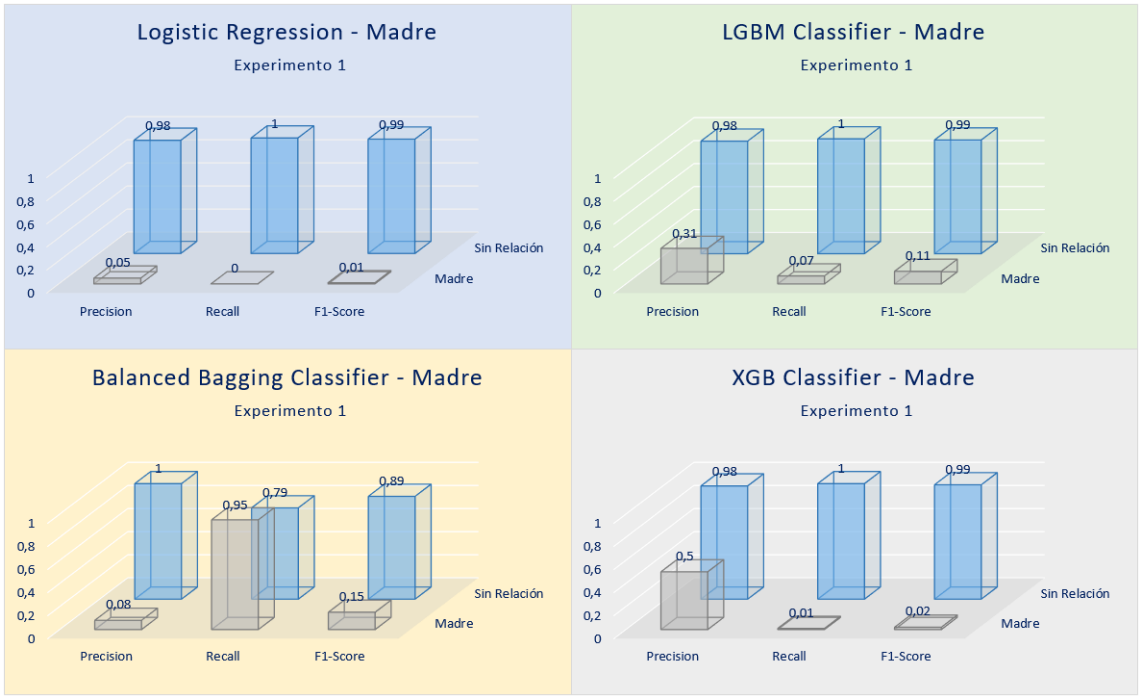


Figura 23: Gráfica con métricas de los modelos de clasificación para la clase "Madre", del experimento 1. Fuente de elaboración propia.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.00	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.49	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.97	6903318
	Accuracy Score		0.9878		
LGBMClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.19	0.07	0.11	83786
	Accuracy			0.98	6903318
	Macro AVG	0.59	0.54	0.55	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9849		
BalancedBagging Classifier	Sin Relación	1.00	0.80	0.89	6819532
	Padre	0.05	0.93	0.10	83786
	Accuracy			0.80	6903318
	Macro AVG	0.53	0.86	0.50	6903318
	Weighted AVG	0.99	0.80	0.88	6903318
	Accuracy Score		0.8044		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.43	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.71	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9878		

Confusion Matrix		
	Sin Relación	Padre
Sin Relación	6819532	0
Padre	83706	0

Confusion Matrix		
	Sin Relación	Padre
Sin Relación	6793179	26353
Padre	77556	6230

Confusion Matrix		
	Sin Relación	Padre
Sin Relación	5475563	1343969
Padre	6203	77583

Confusion Matrix		
	Sin Relación	Padre
Sin Relación	6819363	169
Padre	83658	128

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Rank	NO
OutDegree	NO
InDegree	NO

Figura 24: Tabla con métricas de los modelos de clasificación para la clase "Padre", y tabla de las variables predictoras usadas. Fuente de elaboración propia.

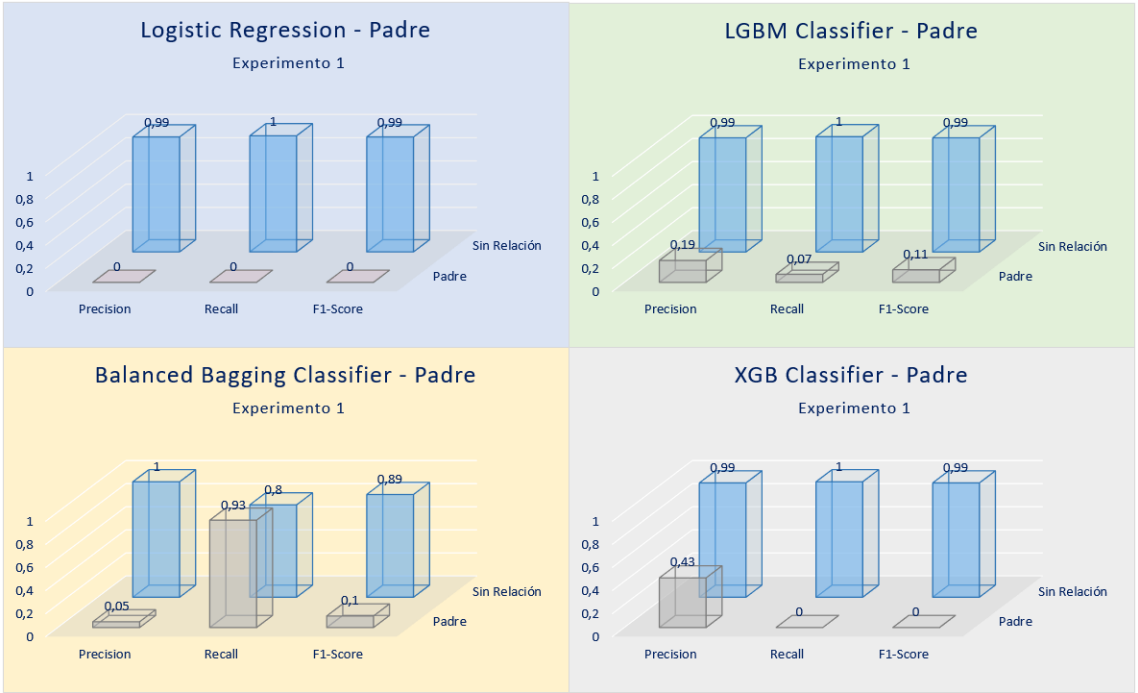


Figura 25: Gráfica con métricas de los modelos de clasificación para la clase "Padre", del experimento 1. Fuente de elaboración propia.

Los resultados obtenidos – figuras 22, 23, 24 y 25 -, muestran métricas de bajo rendimiento. Esto se explica porque para la clase Madre se tiene que *Precision* va desde 0.05 a 0.50, *Recall* va desde 0.00 a 0.95 y *F1-Score* va desde 0.01 a 0.15, en tanto que para la clase Padre se tiene que *Precision* va desde 0.02 a 0.50, *Recall* va desde 0.00 a 0.93 y *F1-Score* va desde 0.00 a 0.10. Se debe tener en cuenta que dada la desigualdad en la distribución de las clases, entonces la métrica *F1-Score* es la que nos ayuda a valorar de mejor manera el rendimiento de los 4 modelos de clasificación. En este primer experimento se tiene que el mayor valor alcanzado para la clase Madre de *F1-Score* es 0.15 y para la clase Padre se tiene que el mayor valor de *F1-Score* es 0.10.

Se realiza una primera iteración en donde se agregan nuevas variables; Rank, OutDegree e InDegree. Mediante las figuras 26, 27, 28 y 29 se ilustran las nuevas métricas logradas agregando las nuevas variables predictoras ya mencionadas. Las figuras 26 y 27 considera la clasificación de la clase “Madre”, y las figuras 28 y 29 corresponde a la clasificación de la clase “Padre”.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.01	0.00	0.00	128210
	Accuracy			0.98	6903318
	Macro AVG	0.50	0.50	0.50	6903318
	Weighted AVG	0.96	0.98	0.97	6903318
	Accuracy Score		0.9812		
LGBMClassifier	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.28	0.08	0.13	128210
	Accuracy			0.98	6903318
	Macro AVG	0.63	0.54	0.56	6903318
	Weighted AVG	0.97	0.98	0.97	6903318
	Accuracy Score		0.9790		
BalancedBagging Classifier	Sin Relación	1.00	0.79	0.88	6775108
	Madre	0.08	0.95	0.15	128210
	Accuracy			0.80	6903318
	Macro AVG	0.54	0.87	0.52	6903318
	Weighted AVG	0.98	0.80	0.87	6903318
	Accuracy Score		0.7964		
XGBClassifier	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.50	0.01	0.02	128210
	Accuracy			0.98	6903318
	Macro AVG	0.74	0.51	0.51	6903318
	Weighted AVG	0.97	0.98	0.97	6903318
	Accuracy Score		0.9814		

Confusion Matrix		
	Sin Relación	Madre
Sin Relación	6774110	998
Madre	128199	11

	Sin Relación	Madre
Sin Relación	6747938	27170
Madre	117576	10634

	Sin Relación	Madre
Sin Relación	5375779	1399329
Madre	5988	122222

	Sin Relación	Madre
Sin Relación	6773637	1471
Madre	126760	1450

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Rank	SI
OutDegree	SI
InDegree	SI

Figura 26: Iteración 1. Tabla con métricas de los modelos de clasificación para la clase "Madre", y tabla de las variables predictoras usadas. Fuente de elaboración propia.

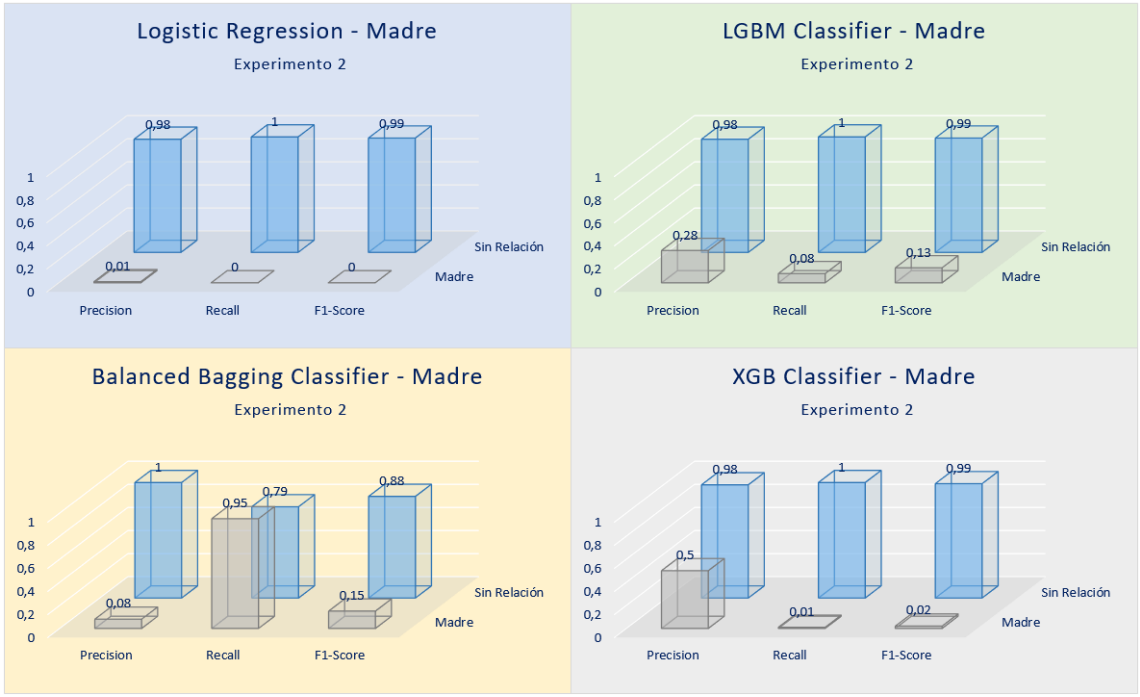


Figura 27: Gráfica con métricas de los modelos de clasificación para la clase "Madre", del experimento 2. Fuente de elaboración propia.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.02	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.50	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.99	6903318
	Accuracy Score		0.9876		
LGBMClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.20	0.06	0.09	83786
	Accuracy			0.99	6903318
	Macro AVG	0.59	0.53	0.54	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9858		
BalancedBagging Classifier	Sin Relación	1.00	0.80	0.89	6819532
	Padre	0.05	0.93	0.10	83786
	Accuracy			0.80	6903318
	Macro AVG	0.53	0.87	0.50	6903318
	Weighted AVG	0.99	0.80	0.88	6903318
	Accuracy Score		0.8029		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.50	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.74	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9878		

Confusion Matrix	
Sin Relación	Padre
6818281	1251
83758	28

Confusion Matrix	
Sin Relación	Padre
6800971	18561
79142	4644

Confusion Matrix	
Sin Relación	Padre
5465359	1354173
5819	77967

Confusion Matrix	
Sin Relación	Padre
6819376	156
83633	153

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Rank	SI
OutDegree	SI
InDegree	SI

Figura 28: Iteración 1. Tabla con métricas de los modelos de clasificación para la clase "Padre", y tabla de variables predictoras usadas. Fuente de elaboración propia.

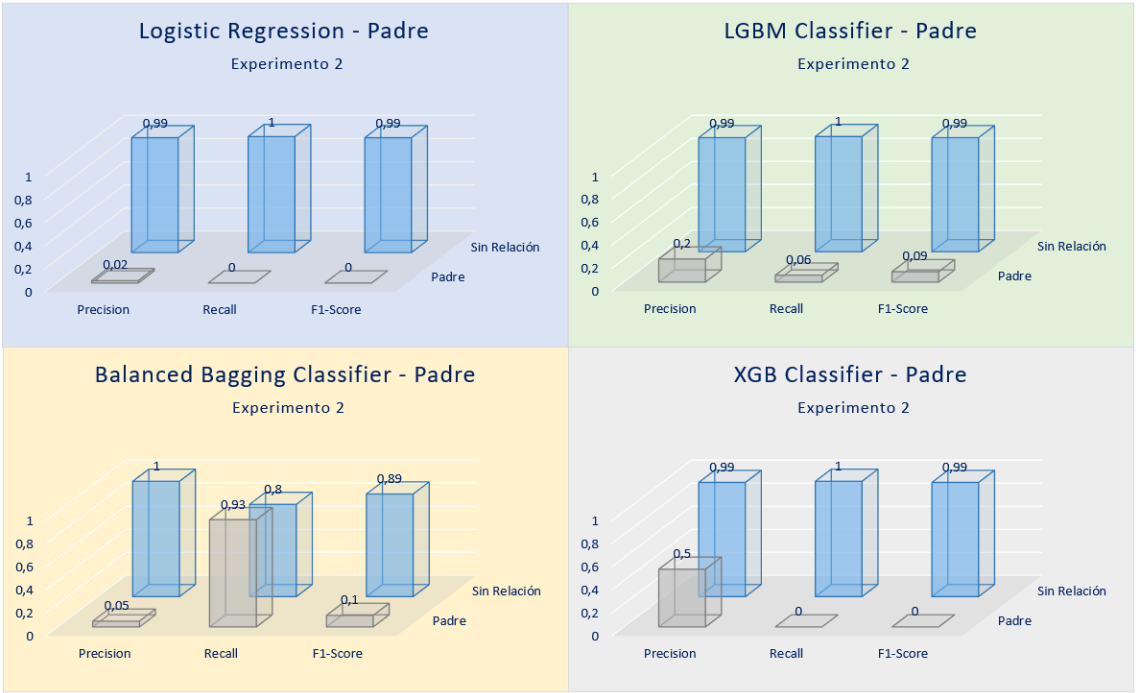


Figura 29: Gráfica con métricas de los modelos de clasificación para la clase "Padre", del experimento 2. Fuente de elaboración propia.

No se ve una mejora significativa en las métricas de rendimiento para los 4 modelos de clasificación. Al mirar la métrica *F1-Score* para la clase Madre se ve que logra un valor máximo de 0.15, en tanto que para la clase Padre el valor máximo logrado es 0.10.

Como una forma de comprobar la lógica implementada para determinar las variables objetivos que se desean clasificar; Madre o Padre, es que se realizan nuevas iteraciones, incorporando las variables edad, género y apellido de manera progresiva. Estos nuevos experimentos son registrados en el apartado “Anexo”.

6. Conclusiones

6.1. Discusión

Se logró proponer e implementar una estrategia metodológica, para definir reglas y construir algoritmos que permiten determinar las relaciones de parentesco en 1er grado; Madre o Padre basándose para ello en la metodología Barret y otros, junto con el uso de atributos tales como apellidos. Todo lo anterior en pos de construir las clases "Sin Relación", "Madre" y "Padre" que a priori no están disponibles en el dataset original. Esto permitió en definitiva contar con dataset para el entrenamiento y prueba de los modelos de clasificación tanto para clientes como no clientes de Movistar.

Se logró establecer que dado el set de datos usado en términos de las llamadas móviles que este contiene, y las variables objetivo construidas; relación de parentesco Madre o Padre, utilizando como base general los patrones de llamadas (David-Barrett et al., 2016), resulta insuficiente el conjunto de variables predictoras seleccionadas para lograr un buen rendimiento en las métricas de los modelos de clasificación entrenados y probados, en el contexto de relaciones de parentesco de 1er grado Madre o Padre, entre clientes de Movistar y clientes de otras compañías.

6.2. Limitaciones

En cuanto al dataset usado, este presenta las siguientes limitaciones:

- El dataset usado considera un conjunto limitado de variables predictoras a ser usadas que permitan realizar el entrenamiento y prueba de los modelos de clasificación. Por ejemplo, el dataset no contiene variables topográficas.
- Las llamadas móviles que contiene son cursadas exclusivamente entre clientes de Movistar. Es por eso es por lo que en principio se realiza una simulación de llamadas entre clientes de Movistar con no clientes de Movistar, eliminando variables tales como edad, género y apellidos.
- Las llamadas móviles que contiene son mutuas. Esto quiere decir que en el dataset únicamente están las llamadas en donde si el fono A cursa llamadas al fono B, es porque también el fono B cursa llamadas al fono A. No están en este dataset aquellas llamadas que se hacen en un único sentido, ósea no están aquellas llamadas que el fono A hace al fono C, y en donde el fono C no cursa llamadas al fono A.
- En el set de datos hay fonos que tienen un único destino de llamadas, en tanto que hay otros fonos que tienen más de 200 destinos diferentes de llamadas. Esto impacta en el balanceo de la data de cara al entrenamiento y prueba de los modelos de clasificación.

Desde el punto de vista técnico se presentó la dificultad de entrenar y probar los modelos de clasificación con un set de datos de **27.637.782** de muestras. Las actuales herramientas de Big Data están limitadas en términos de los modelos de clasificación que disponiblizan para su uso. Es decir, por lo pronto hay una limitante asociada al uso de los modelos de clasificación, que no permite ocupar la potencialidad y escalabilidad que ofrecen las herramientas de Big Data.

6.3. Trabajo Futuro

Sería de interés el poder contar con un set de datos que tuviese tanto llamadas de clientes de movistar como clientes de otras compañías, y que también que no solo contenga las llamadas mutuas. Esto por una parte evitaría simular la existencia de clientes que no son de movistar, tal cual como se hizo en el presente trabajo, y por otro lado eliminaría el actual sesgo que se tiene al solo contar con llamadas móviles mutuas.

Sería de interés profundizar en la dificultad que surge por la naturaleza dispar de las llamadas móviles, asociada a la cantidad de destinos diferentes que pueden tener los distintos fonos que cursan llamadas. Ya se mencionó que hay fonos que tienen un único destino, en tanto que hay otros fonos que tienen más de 200 destinos diferentes.

Sería de interés poder contar con un set de datos que incorpore nuevas variables predictoras tales como variables topológicas.

Sería de interés aplicar Deep Learning como modelo de clasificación, y ver si el uso de redes neuronales para clasificar la relación de 1er grado Madre o Padre, tiene mejores de desempeño que las obtenidas.

Sería interesante contar con un dataset que contenga las edades de quienes reciben las llamadas. Y basandosé en las relaciones de homofilia entres las llamadas telefónicas y junto con ello en la existencia de relaciones disortativas(Brea et al., 2018).

Sería interesante que en un trabajo futuro, además de considerar las actuales relaciones de parentesco en 1er grado Madre y Padre, también se pudiesen agregar las relaciones de parentezco Hermano y Conyuge. Esto con el fin de completar el nucleo familiar de las relaciones de parentesco en 1er grado y considerando la importancia en terminos de la cercanía familiar, que estas relaciones representan.

7. Bibliografía

A hybrid model for IT project with Scrum. Recuperado 27 de septiembre de 2020, de https://www.infona.pl/resource/bwmeta1.element.ieee-art-000005986572

Aledavood, T., López, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., & Saramäki, J. (2015). Daily Rhythms in Mobile Telephone Communication. *PLoS ONE*, *10*(9). https://doi.org/10.1371/journal.pone.0138098

Aledavood, T., López, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., & Saramäki, J. (2016). Channel-Specific Daily Patterns in Mobile Phone Communication. En S. Battiston, F. De Pellegrini, G. Caldarelli, & E. Merelli (Eds.), *Proceedings of ECCS 2014* (pp. 209-218). Springer International Publishing. https://doi.org/10.1007/978-3-319-29228-1_18

Baron, N. S., & Campbell, E. M. (2012). Gender and mobile phones in cross-national context. *Language Sciences*, *34*(1), 13-27. https://doi.org/10.1016/j.langsci.2011.06.018

Bollobas, B. (2013). *Modern Graph Theory*. Springer Science & Business Media.

Brea, J., Burroni J., & Sarraute C. (2018). Inference of Users Demographic Attributes based on Homophily in Communication Networks.

David-Barrett, T., Kertesz, J., Rotkirch, A., Ghosh, A., Bhattacharya, K., Monsivais, D., & Kaski, K. (2016). Communication with Family and Friends across the Life Course. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.0165687

David-Barrett, T., Rotkirch, A., Ghosh, A., Bhattacharya, K., Monsivais, D., Behncke, I., Kertesz, J., & Kaski, K. (2017). Peer relations with mobile phone data: Best friends and family formation. *arXiv:1708.07759 [physics]*. http://arxiv.org/abs/1708.07759

Evaluation Metrics Machine Learning. Recuperado 27 de septiembre de 2020, de https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/?fbclid=IwAR20Eb2b9AjJl9wdjaVedhj3s-Vw2OA6qIrpWrLjxBjJrSe9qjk76hDv81w

Families without Borders: Mobile Phones, Connectedness and Work-Home Divisions—Judy Wajcman, Michael Bittman, Judith E. Brown, 2008. Recuperado 27 de septiembre de 2020, de https://journals.sagepub.com/doi/abs/10.1177/0038038508091620

Fudolig, M. I. D., Monsivais, D., Bhattacharya, K., Jo, H.-H., & Kaski, K. (2020). Different patterns of social closeness observed in mobile phone communication. *Journal of Computational Social Science*, *3*(1), 1-17. https://doi.org/10.1007/s42001-019-00054-8

Ghosh, A., Monsivais, D., Bhattacharya, K., Dunbar, R. I. M., & Kaski, K. (2019). Quantifying gender preferences in human social interactions using a large cellphone dataset. *EPJ Data Science*, *8*(1), 9. https://doi.org/10.1140/epjds/s13688-019-0185-9

Historia – Servicio Electoral de Chile. Recuperado 27 de septiembre de 2020, de https://www.servel.cl/historia/

Imbalanced datasets with imbalanced-learn. (2018, Agosto 6). David Ten. https://xang1234.github.io/louvain/

Li, S. (2019, febrero 27). *Building A Logistic Regression in Python, Step by Step*. Medium. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Ling, R., & Horst, H. A. (2011). Mobile communication in the global south. *New Media & Society*, *13*(3), 363-374. https://doi.org/10.1177/1461444810393899

Monsivais, D., Ghosh, A., Bhattacharya, K., Dunbar, R. I. M., & Kaski, K. (2017). Tracking urban human activity from mobile phone calling patterns. *PLoS Computational Biology*, *13*(11). https://doi.org/10.1371/journal.pcbi.1005824

Palchykov, V., Kertész, J., Dunbar, R., & Kaski, K. (2013). Close Relationships: A Study of Mobile Communication Records. *Journal of Statistical Physics*, *151*(3), 735-744. https://doi.org/10.1007/s10955-013-0705-0

Pariente de primer grado | NHGRI. Genome.gov. Recuperado 27 de septiembre de 2020, de https://www.genome.gov/es/genetics-glossary/Pariente-de-primer-grado

Proyecciones de Población. Default. Recuperado 27 de septiembre de 2020, de http://www.ine.cl/estadisticas/sociales/demografia-y-vitales/proyecciones-de-poblacion

Quick start using graph-tool—Graph-tool 2.35 documentation. Recuperado 27 de septiembre de 2020, de https://graph-tool.skewed.de/static/doc/quickstart.html

Sedgwick, M. G., & Yonge, O. (2008). «We're it», «we're a team», «we're family» means a sense of belonging. *Rural and Remote Health*, *8*(3), 1021.

The gt file format—Graph-tool 2.35 documentation. Recuperado 27 de septiembre de 2020, de https://graph-tool.skewed.de/static/doc/gt_format.html

Transporte y comunicaciones-producto. Default. Recuperado 27 de septiembre de 2020, de http://www.ine.cl/estadisticas/economia/transporte-y-comunicaciones/transporte-y-comunicaciones/transporte-y-comunicaciones-producto

Welcome to LightGBM's documentation! —LightGBM 3.0.0.99 documentation. Recuperado 27 de septiembre de 2020, de https://lightgbm.readthedocs.io/en/latest/index.html

XGBoost Documentation—Xgboost 1.3.0-SNAPSHOT documentation. Recuperado 27 de septiembre de 2020, de https://xgboost.readthedocs.io/en/latest/index.html

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (1st ed.). O'Reilly Media, Inc.

Anexo

A continuación son presentados los experimentos realizados en donde son incorporadas las variables edad destino, género destino y apellidos(origen y destino) de manera paulatina. Son construidas las variables “Gen_=”(de la misma generación), “Gen_-”(de una generación anterior), y “Gen_+”(de una generación posterior) usando la técnica de “feature engineering” basándose para ello en la variable edad.

De tal manera que se realiza una iteración de entrenamiento y prueba, en donde en esta oportunidad se agregan las variables predictoras edad del destino de las llamadas “Des_Age” y género del destino de las llamadas “Des_Sex”, tal como se ilustra en las figuras 30, 31, 32 y 33, y además en donde son ilustradas las nuevas métricas logradas. Las figuras 30 y 31 considera la clasificación de la clase “Madre”, y las figuras 32 y 33 corresponden a la clasificación de la clase “Padre”.

1 Iteración 2

Se agrega la edad y género de los destinos de las llamadas

- Des_Age, si A es el origen de la llamada, entonces es la edad de quien es el destino de la llamada cursada por A.
- Des_Sex, si A es el origen de la llamada, entonces es el género de quien es el destino de la llamada cursada por A. En caso de ser mujer su valor es 0, en caso de ser varón su valor es 1.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.07	0.00	0.01	128210
	Accuracy			0.98	6903318
	Macro AVG	0.53	0.50	0.50	6903318
	Weighted AVG	0.96	0.98	0.97	6903318
	Accuracy Score		0.9808		
LGBMClassifier	Sin Relación	0.99	0.99	0.99	6775108
	Madre	0.42	0.28	0.34	128210
	Accuracy			0.98	6903318
	Macro AVG	0.71	0.64	0.66	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9795		
BalancedBagging Classifier	Sin Relación	1.00	0.90	0.95	6775108
	Madre	0.16	1.00	0.27	128210
	Accuracy			0.90	6903318
	Macro AVG	0.58	0.95	0.61	6903318
	Weighted AVG	0.98	0.90	0.93	6903318
	Accuracy Score		0.9007		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6775108
	Madre	0.64	0.35	0.45	128210
	Accuracy			0.98	6903318
	Macro AVG	0.82	0.67	0.72	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9842		

Confusion Matrix	
Sin Relación	Madre
6770621	4487
127856	354

Sin Relación	Madre
6726621	48487
92415	35795

Sin Relación	Madre
6090000	685108
251	127959

Sin Relación	Madre
6750587	24521
83872	44338

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	SI
Des_Sex	SI
Rank	SI
OutDegree	SI
InDegree	SI
Gen_=	NO
Gen_-	NO
Gen_+	NO
OA1_=_DA2	NO
OA1_=_DA1	NO

Figura 30: Iteración 2. Tabla con métricas de los modelos de clasificación para la clase "Madre", y tabla de las variables predictoras usadas. Fuente de elaboración propia.

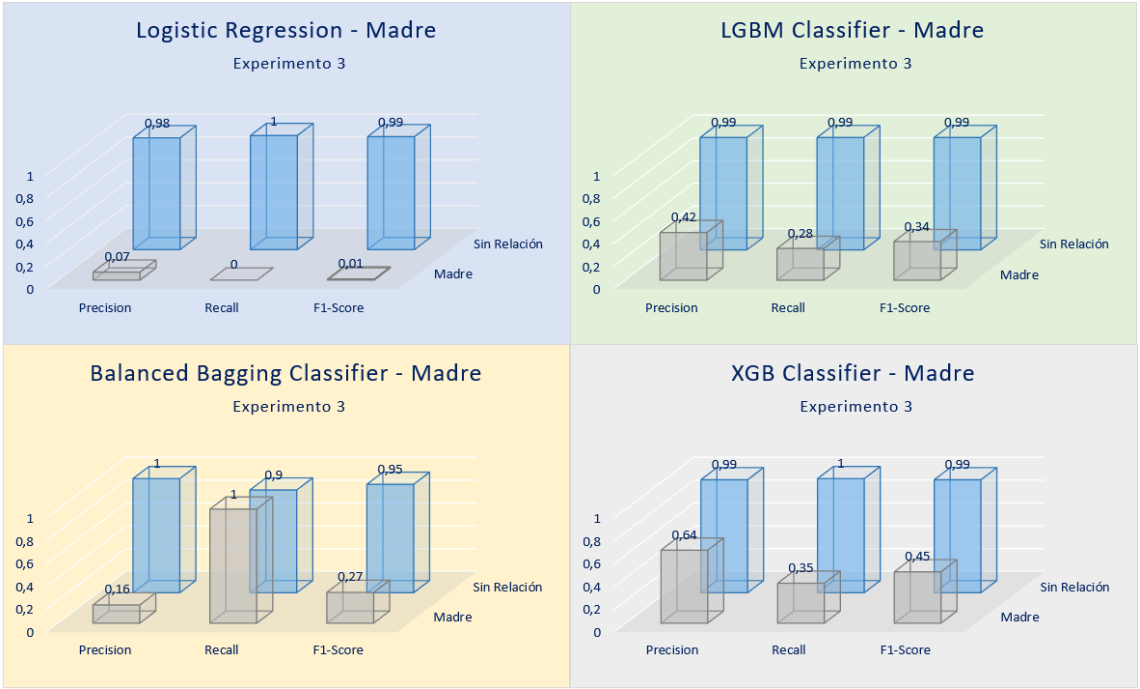


Figura 31: Gráfica con métricas de los modelos de clasificación para la clase "Madre", del experimento 3. Fuente de elaboración propia.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.05	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.52	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9873		
LGBMClassifier	Sin Relación	0.99	0.99	0.99	6819532
	Padre	0.30	0.18	0.23	83786
	Accuracy			0.98	6903318
	Macro AVG	0.65	0.59	0.61	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9849		
BalancedBagging Classifier	Sin Relación	1.00	0.90	0.95	6819532
	Padre	0.11	1.00	0.20	83786
	Accuracy			0.90	6903318
	Macro AVG	0.55	0.95	0.57	6903318
	Weighted AVG	0.99	0.90	0.94	6903318
	Accuracy Score		0.9010		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.58	0.19	0.29	83786
	Accuracy			0.99	6903318
	Macro AVG	0.74	0.60	0.64	6903318
	Weighted AVG	0.99	0.99	0.99	6903318
	Accuracy Score		0.9885		

Confusion Matrix	
Sin Relación	Padre
6815947	3585
83613	173

Sin Relación	Padre
6784106	35426
68502	15284

Sin Relación	Padre
6136931	682601
190	83596

Sin Relación	Padre
6807971	11561
67667	16119

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	SI
Des_Sex	SI
Rank	SI
OutDegree	SI
InDegree	SI
Gen_=	NO
Gen_-	NO
Gen_+	NO
OA1_=_DA2	NO
OA1_=_DA1	NO

Figura 32: Iteración 2. Tabla con métricas de los modelos de clasificación para la clase "Padre", y tabla de variables predictoras usadas. Fuente de elaboración propia.

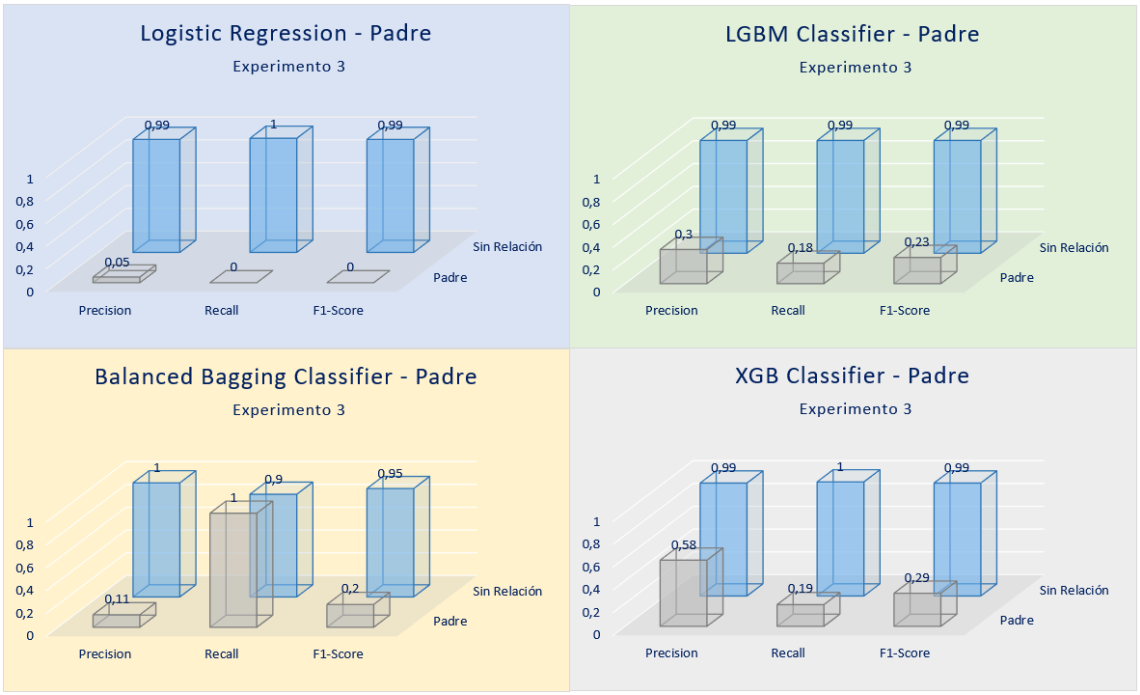


Figura 33: Gráfica con métricas de los modelos de clasificación para la clase "Padre", del experimento 3. Fuente de elaboración propia.

Se ve en las figuras 30 y 31 que la mejor métrica *F1-Score* alcanzada es 0.45 para la clase Madre. En las figuras 32 y 33 se observa que la mejor métrica *F1-Score* lograda es 0.29 para la clase Padre. Si bien es cierto que las métricas mejoran para ambas clases, sus valores aún están por debajo de lo aceptable. Considerando como un valor aceptable a la métrica *F1-Score* mayor o igual a 0.8.

2 Iteración 3

Se incorporan los indicadores de la generación y se eliminan edad y género de los destinos de las llamadas

- Gen_=, quien hace la llamada es de la misma generación de quien recibe la llamada.
- Gen_-, quien hace la llamada es de 1 generación anterior de quien recibe la llamada.
- Gen_+, quien hace la llamada es de 1 generación posterior de quien recibe la llamada.

Las figuras 34, 35, 36 y 37 muestran las nuevas métricas obtenidas. Las figuras 34 y 35 consideran la clasificación de la clase “Madre”, y las figuras 36 y 37 corresponde a la clasificación de la clase “Padre”.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.98	1.00	0.99	6775108
	Madre	0.06	0.00	0.00	128210
	Accuracy			0.98	6903318
	Macro AVG	0.52	0.50	0.50	6903318
	Weighted AVG	0.96	0.98	0.97	6903318
	Accuracy Score		0.9806		
LGBMClassifier	Sin Relación	0.99	1.00	0.99	6775108
	Madre	0.62	0.32	0.42	128210
	Accuracy			0.98	6903318
	Macro AVG	0.80	0.66	0.71	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9837		
BalancedBagging Classifier	Sin Relación	1.00	0.92	0.96	6775108
	Madre	0.20	1.00	0.33	128210
	Accuracy			0.92	6903318
	Macro AVG	0.60	0.96	0.64	6903318
	Weighted AVG	0.99	0.92	0.95	6903318
	Accuracy Score		0.9241		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6775108
	Madre	0.62	0.32	0.42	128210
	Accuracy			0.98	6903318
	Macro AVG	0.80	0.66	0.71	6903318
	Weighted AVG	0.98	0.98	0.98	6903318
	Accuracy Score		0.9837		

Confusion Matrix

	Sin Relación	Madre
Sin Relación	6769748	5360
Madre	127897	313

	Sin Relación	Madre
Sin Relación	6749994	25114
Madre	87381	40829

	Sin Relación	Madre
Sin Relación	6251483	523625
Madre	122	18088

	Sin Relación	Madre
Sin Relación	6750174	24934
Madre	87279	40931

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	NO
Des_Sex	NO
Rank	SI
OutDegree	SI
InDegree	NO
Gen_=	SI
Gen_-	SI
Gen_+	SI
OA1_=_DA2	NO
OA1_=_DA1	NO

Figura 34: Iteración 3. Tabla con métricas de los modelos de clasificación para la clase "Madre", y tabla de las variables predictoras usadas. Fuente de elaboración propia.

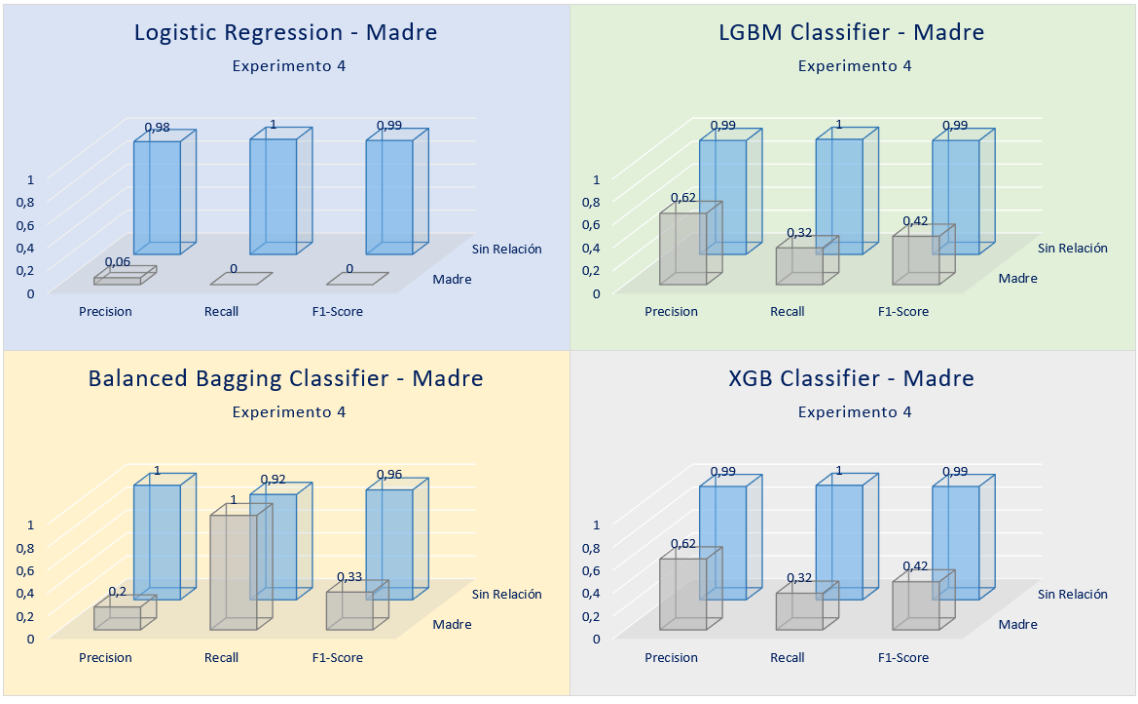


Figura 35: Gráfica con métricas de los modelos de clasificación para la clase "Madre", del experimento 4. Fuente de elaboración propia.

		Precision	Recall	F1-Score	Support
LogisticRegression	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.00	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.49	0.50	0.50	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9878		
LGBMClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.55	0.09	0.16	83786
	Accuracy			0.99	6903318
	Macro AVG	0.77	0.55	0.58	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9880		
BalancedBagging Classifier	Sin Relación	1.00	0.92	0.96	6819532
	Padre	0.13	1.00	0.23	83786
	Accuracy			0.92	6903318
	Macro AVG	0.56	0.96	0.59	6903318
	Weighted AVG	0.99	0.92	0.95	6903318
	Accuracy Score		0.9169		
XGBClassifier	Sin Relación	0.99	1.00	0.99	6819532
	Padre	0.54	0.10	0.16	83786
	Accuracy			0.99	6903318
	Macro AVG	0.77	0.55	0.58	6903318
	Weighted AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9880		

Confusion Matrix

	Sin Relación	Padre
Sin Relación	6819532	0
Padre	83786	0

	Sin Relación	Padre
Sin Relación	6813178	6354
Padre	76155	7631

	Sin Relación	Padre
Sin Relación	6246003	573529
Padre	22	83764

	Sin Relación	Padre
Sin Relación	6812674	6858
Padre	75668	8118

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	NO
Des_Sex	NO
Rank	SI
OutDegree	SI
InDegree	NO
Gen_=	SI
Gen_-	SI
Gen_+	SI
OA1_=_DA2	NO
OA1_=_DA1	NO

Figura 36: Iteración 3. Tabla con métricas de los modelos de clasificación para la clase "Padre", y tabla de variables predictoras usadas. Fuente de elaboración propia.

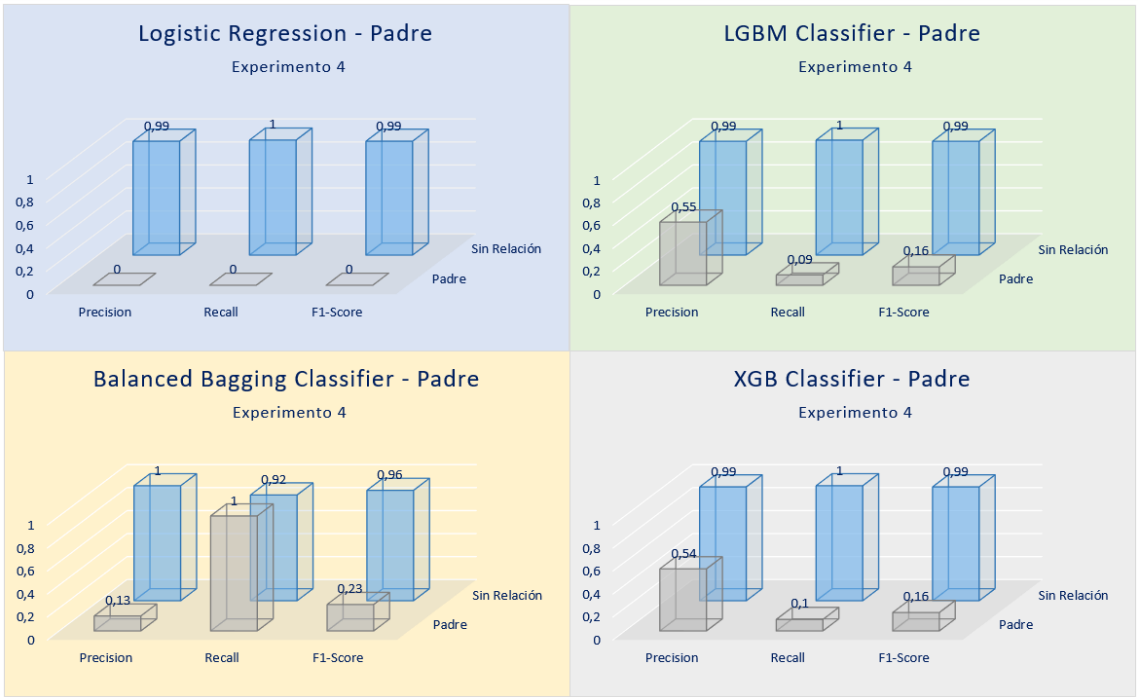


Figura 37: Gráfica con métricas de los modelos de clasificación para la clase "Padre", del experimento 4. Fuente de elaboración propia.

De acuerdo a las figuras 34 y 35, se tiene que la mejor métrica *F1-Score* es 0.42, para la clase Madre, y esta se alcanza para 2 modelos de clasificación. En el caso de las figuras 36 y 37, se ve que la métrica *F1-Score* alcanza un valor máximo de 0.23. Estas métricas son comparativamente inferiores a las obtenidas en el experimento anterior.

3 Iteración 4

Se realiza una última iteración. En esta iteración, se adicionan los indicadores de coincidencia de apellidos, entre origen y destino de llamadas

- OA1_=_DA2, el primer apellido de quien hace la llamada es igual al segundo apellido de quien recibe la llamada. Cuando esto se cumple, el valor del indicador es 1, caso contrario su valor es 0.
- OA1_=_DA1, el primer apellido de quien hace la llamada es igual al primer apellido de quien recibe la llamada. Cuando esto se cumple, el valor del indicador es 1, caso contrario su valor es 0.

Con estos indicadores se incorpora todo el espectro de variables usados para construir la lógica que permitió determinar las relaciones de parentesco en 1er grado, Madre o Padre. En las figuras 38, 39, 40 y 41 muestran las métricas alcanzadas. Las figuras 38 y 39 están asociadas a la variable objetivo Madre y las figuras 40 y 41 están vinculadas a la variable objetivo Padre.

		Precision	Recall	F1-Score	Support
LogisticRegression	Relación	0.98	1.00	0.99	6775108
	Madre	0.05	0.00	0.01	128210
	Accuracy			0.98	6903318
	Macro AVG	0.52	0.50	0.50	6903318
	AVG	0.96	0.98	0.97	6903318
	Accuracy Score		0.9800		
LGBMClassifier	Relación	1.00	1.00	1.00	6775108
	Madre	1.00	1.00	1.00	128210
	Accuracy			1.00	6903318
	Macro AVG	1.00	1.00	1.00	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.9999		
BalancedBagging Classifier	Relación	1.00	1.00	1.00	6775108
	Madre	1.00	1.00	1.00	128210
	Accuracy			1.00	6903318
	Macro AVG	1.00	1.00	1.00	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.9999		
XGBClassifier	Relación	1.00	1.00	1.00	6775108
	Madre	1.00	1.00	1.00	128210
	Accuracy			1.00	6903318
	Macro AVG	1.00	1.00	1.00	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.9999		

Confusion Matrix

	Sin Relación	Madre
Sin Relación	6764680	10428
Madre	127607	603

	Sin Relación	Madre
Sin Relación	6774861	247
Madre	13	128197

	Sin Relación	Madre
Sin Relación	6774850	258
Madre	0	128210

	Sin Relación	Madre
Sin Relación	6774858	250
Madre	25	128185

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	NO
Des_Sex	NO
Rank	SI
OutDegree	SI
InDegree	NO
Gen_=	SI
Gen_-	SI
Gen_+	SI
OA1_=_DA2	SI
OA1_=_DA1	SI

Figura 38: Iteración 4. Tabla con métricas de los modelos de clasificación para la clase "Madre", y tabla de variables predictoras usadas. Fuente de elaboración propia.

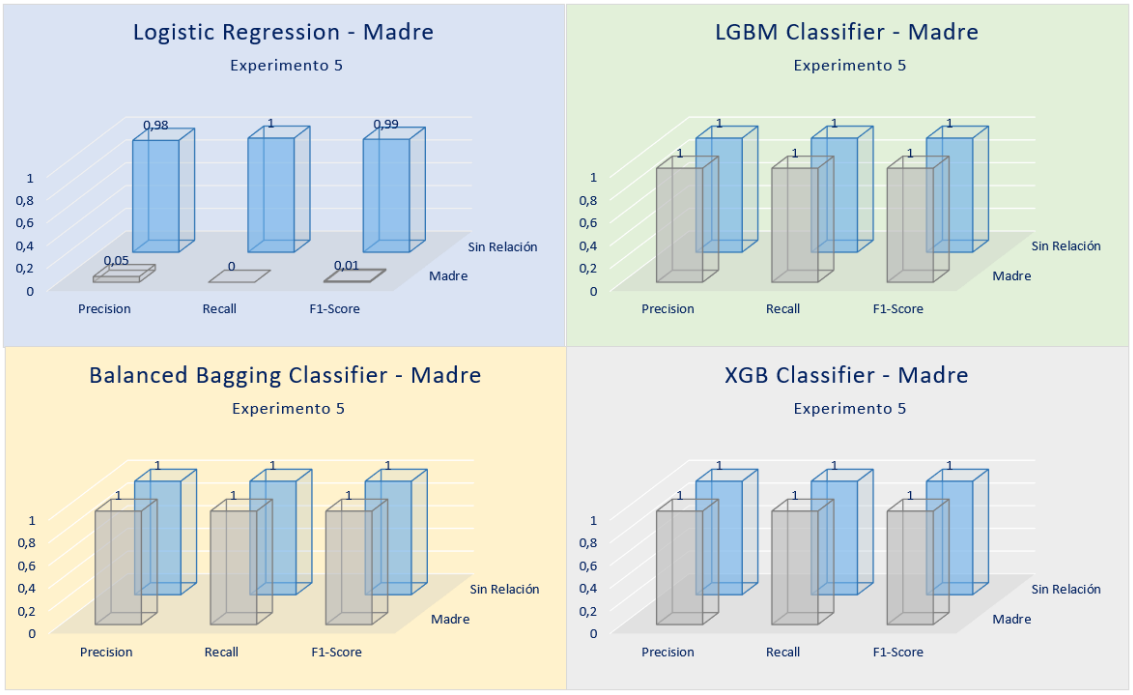


Figura 39: Gráfica con métricas de los modelos de clasificación para la clase "Madre", del experimento 5. Fuente de elaboración propia.

		Precision	Recall	F1-Score	Support
LogisticRegression	Relación	0.99	1.00	0.99	6819532
	Padre	0.00	0.00	0.00	83786
	Accuracy			0.99	6903318
	Macro AVG	0.49	0.50	0.50	6903318
	AVG	0.98	0.99	0.98	6903318
	Accuracy Score		0.9878		
LGBMClassifier	Relación	1.00	1.00	1.00	6819532
	Padre	0.95	1.00	0.97	83786
	Accuracy			1.00	6903318
	Macro AVG	0.97	1.00	0.99	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.9993		
BalancedBagging Classifier	Relación	1.00	1.00	1.00	6819532
	Padre	0.95	1.00	0.97	83786
	Accuracy			1.00	6903318
	Macro AVG	0.97	1.00	0.99	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.9999		
XGBClassifier	Relación	1.00	1.00	0.99	6819532
	Padre	0.95	1.00	0.97	83786
	Accuracy			1.00	6903318
	Macro AVG	0.97	1.00	0.99	6903318
	AVG	1.00	1.00	1.00	6903318
	Accuracy Score		0.999		

Confusion Matrix

	Sin Relación	Padre
Sin Relación	6819532	0
Padre	83786	0

	Sin Relación	Padre
Sin Relación	6815072	4460
Padre	72	83707

	Sin Relación	Padre
Sin Relación	6814996	4536
Padre	0	83786

	Sin Relación	Padre
Sin Relación	6815072	4460
Padre	66	83720

VARIABLES	USADAS
Origen	SI
Ori_Age	SI
Ori_Sex	SI
Destino	SI
CallsIn	SI
callsout	SI
secIn	SI
secOut	SI
%_CallsIn	SI
%_CallsOut	SI
%_SecIn	SI
%_SecOut	SI
Des_Age	NO
Des_Sex	NO
Rank	SI
OutDegree	SI
InDegree	NO
Gen_=	SI
Gen_-	SI
Gen_+	SI
OA1_=_DA2	SI
OA1_=_DA1	SI

Figura 40: Iteración 4. Tabla con métricas de los modelos de clasificación para la clase "Padre", y tabla de variables predictoras usadas. Fuente de elaboración propia.

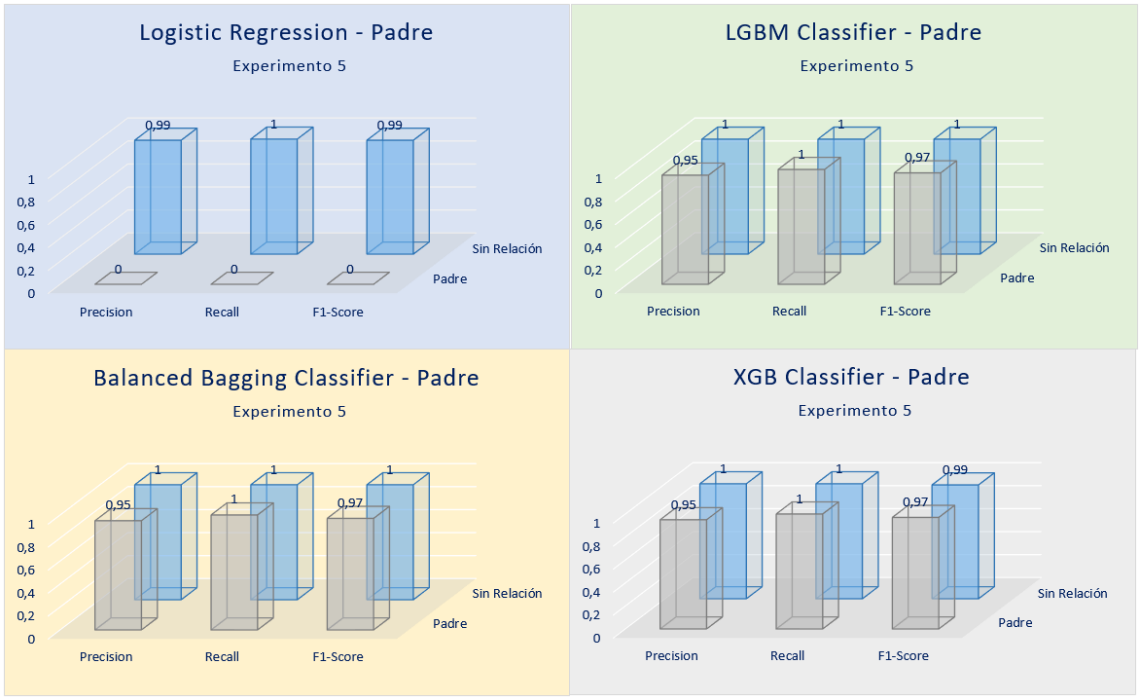


Figura 41: Gráfica con métricas de los modelos de clasificación para la clase "Padre", del experimento 5. Fuente de elaboración propia.

Se puede apreciar en las figuras 38 y 39, que para la clase Madre la métrica *F1-Score* va desde el 0.01 hasta el valor 1.00. Tres de los cuatro modelos de clasificación obtienen un *F1-Score* = 1.00.

Al mirar las figuras 40 y 41, se observa que para la clase Padre la métrica F1-Score va desde el valor 0.00 hasta el valor 0.97. Tres de los cuatro modelos de clasificación alcanzan un F1-Score = 0.97.