UNIVERSIDAD NACIONAL

Facultad de Ciencias Exactas y Naturales

ESCUELA DE INFORMÁTICA

Modelo de Caracterización de Individuos Morosos Utilizando Algoritmos de Minería de Datos

Para optar al grado de Licenciado en
Informática con énfasis en desarrollo web

Ing. Gerson Vargas Gálvez
Ing. Frander Ramírez Villalobos

Heredia, Costa Rica

Agradecimientos y dedicatorias

Dedicamos este trabajo a nuestros padres que han sido un soporte a lo largo de la vida, destacando siempre que con esfuerzo y disciplina se pueden derribar barreras y alcanzar las metas. A nuestros profesores que han posibilitado con su experiencia y fecunda labor en la enseñanza, poder alcanzar y desarrollar habilidades que posibilitan tanto el crecimiento personal como profesional.

TABLA DE CONTENIDOS

Tablas

Figuras

Gráficos

CAPÍTULO I
INTRODUCCIÓN

CAPÍTULO I: INTRODUCCIÓN

1. Antecedentes

El presente proyecto propone un modelo predictivo de contribuyentes morosos utilizando técnicas de minería de datos en la Municipalidad de San Antonio de Belén, para ello resulta importante analizar los acontecimientos de la morosidad, así se detallan a continuación los siguientes apartados:

A través de los años, la morosidad ha sido un tema en cuestión a nivel de economía en las organizaciones y gobiernos; viendo su impacto a nivel mundial, un ejemplo de esta situación se puede ver reflejada en España, que ha sido un país que durante el tiempo ha presentado índices de morosidad elevados. "En España, se calcula que entre el 80% y el 90% de las empresas sufre problemas de retrasos en los cobros o tiene problemas financieros derivados de la morosidad de sus clientes." (Brachfield, 2000, p.33). Lo anterior refleja una situación alarmante en el área de finanzas organizacionales, además muestra la importancia de realizar una gestión de cobros eficiente.

A pesar de la preocupación que genera la alta morosidad, las organizaciones han buscado un incremento en sus ingresos mediante la prestación de bienes y servicios mediante el crédito a sus clientes, teniendo como principal ventaja el aumento de los capitales.

En Costa Rica a nivel municipal el tema de morosidad ha sido un motivo de discusión que preocupa los gobiernos locales, según la Contraloría General de la República, ente encargado de la evaluación de gestión municipal, gran parte de los municipios presentan debilidades en el tema de gestión de cobros.

"En los últimos cinco años (2010-2014), la Municipalidad de Santa Cruz no ha logrado avanzar en la solución de los problemas vinculados con la gestión de cobro de los tributos municipales. En ese sentido, el pendiente de cobro pasó de ¢1.648,32 millones al 31 de diciembre de 2010 a ¢3.415,2 millones al 31 de diciembre de 2014, situación que coloca a ese Gobierno Local entre los ayuntamientos con mayor morosidad del sector. Al 31 de diciembre de 2014, existen deudas de los contribuyentes con tres o más años de

atraso que ascienden a unos ¢724,6 millones, de los cuales unos ¢412,0 millones están en riesgo de prescripción.” (Contraloría General de la República, 2015).

En la Figura 1, se muestra la situación de morosidad que enfrenta la Municipalidad de Santa Cruz en los periodos comprendidos entre 2010 y 2014, se describe que el pendiente de cobro aumentó de ¢1.648 millones en 2010 (una morosidad del 38%) a ¢3.415 millones en 2014 (una morosidad del 44%).

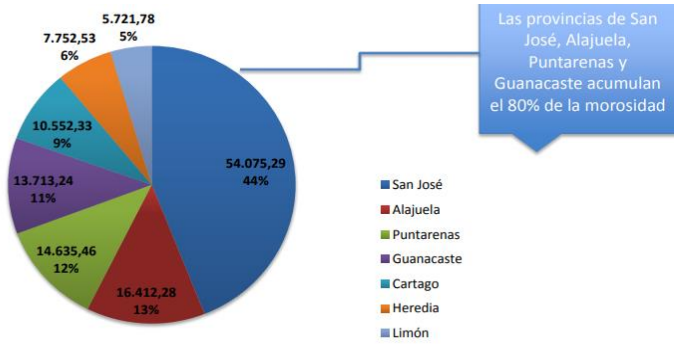
Tabla 1. Representación sobre cobros de tributos al contribuyente en la Municipalidad de Santa Cruz, en el año 2014.

Periodo	Puesto al cobro	Total Recaudado	Pendiente	%Pendiente
2010	4.314,02	2.665,70	1.648,32	38,21%
2011	5.566,21	3.084,25	2.481,96	44,59%
2012	5.591,73	3.529,18	2.062,55	36,89%
2013	6.891,35	3.824,65	3.066,70	44,50%
2014	7.650,45	4.235,24	3.415,20	44,64%

Fuente*:* Obtenido del Informe de Gestión de Cobro de los Tributos Municipales en la Municipalidad de Santa Cruz de la Contraloría General de la República, año 2014.

En el Gráfico 1, se muestran los porcentajes y cifras en millones de colones que representan la morosidad que enfrentan las provincias de Costa Rica, cabe destacar que las provincias de San José, Alajuela, Puntarenas y Guanacaste acumulan el 80% total de la morosidad.

Gráfico 1. **Provincia de Costa Rica Sumas de morosidad y su relación con respecto al total de morosidad del sector municipal 2014.**



Fuente. Recuperado del Informe de Gestión de cobro de tributos y morosidad en el sector municipal costarricense en el 2014.

Ante la morosidad que enfrentan las municipalidades, en 2018 se abrió la opción de realizar la condonación de impuestos a los contribuyentes mediante el proceso llamado Amnistía Tributaria, los municipios son voluntarios para acoger esta disposición, en la actualidad más de 18 municipalidades se han acogido a aplicar la amnistía.

Lo anterior refleja la difícil situación que presentan las municipalidades del país en temas de morosidad, por esta razón se expone a continuación los problemas que se obtienen como consecuencia del aumento en los índices de morosidad.

2. Planteamiento del problema

2.1 Problemas en la gestión Municipal

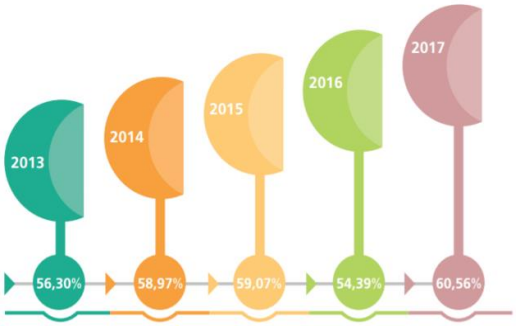
En nuestro país, según informe realizado por la Contraloría General de la República en el año 2017, el promedio general de gestión municipal alcanza un 60,56%, esto representa un índice bajo a pesar de la mejora mostrada en el informe:

“La calificación promedio de las 81 municipalidades evaluadas con el IGM-2017 fue de 60,56 puntos de 100 posibles, exhibiendo una mejoría al relacionar a los resultados obtenidos en periodos anteriores. En comparación con el año 2016, el IGM aumentó en

6,17 puntos, mientras que para el periodo comprendido entre 2015 y 2016 había decrecido en 4,68 puntos.” (CGR, 2017).

En la Figura 1, se representan los porcentajes del índice de gestión municipal obtenidos en la evolución realizada por la Contraloría General de la República a las 81 municipalidades del país en los años comprendidos entre el 2013 y 2017.

Figura 1. **Índice de gestión municipal, calificación promedio de las municipalidades, años 2013-2017**



Fuente. Recuperado del Informe de Índice de Gestión Municipal periodo 2017, Contraloría General de la República.

En la Figura 2, se muestra la ubicación de las municipalidades de Costa Rica según estratificación con respecto a la calificación obtenida en la evaluación de la CGR.

Dicha evaluación contempla 14 áreas de evaluación y 61 indicadores distribuidos en cinco ejes: Desarrollo y gestión institucional; Planificación, participación ciudadana y rendición de cuentas; Gestión de desarrollo ambiental; Gestión de servicios económicos (gestión vial) y; Gestión de servicios sociales.

Figura 2. **Índice de Gestión Municipal, estratificación por nota.**

Fuente*:* Recuperado del Informe de Índice de Gestión Municipal periodo 2017, Contraloría General de la República.

2.2 Percepción de ingresos por cobros de servicios:

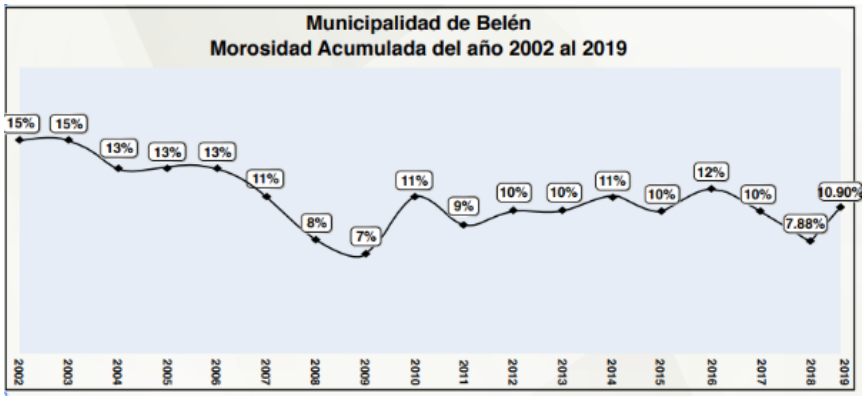
Según los informes de la Controlaría General de la República, se hace un llamado urgente al sector municipal a mejorar sus procesos, siendo uno de ellos la gestión de cobros, el cual presenta índices que afectan el desarrollo cantonal.

La mejora en la gestión municipal abarca muchas áreas, principalmente el sector de servicios que se brindan a la ciudadanía de acuerdo con las necesidades de cada cantón del país. Entre los servicios ofrecidos se destaca el área de cementerio, cruz roja, alcantarillado sanitario, recolección de residuos, limpieza de vías, mantenimiento de parques, patentes, bienes inmuebles, entre otros. Por cada servicio se perciben ingresos que posibilitan actividades a favor del desarrollo de la población; sin embargo, el proceso de recolección de estos fondos no es simple de resolver, el problema se agrava principalmente cuando no se poseen las herramientas necesarias para hacer una gestión de cobros eficiente, provocando que los índices de ingresos que se dejan de percibir en el tiempo adecuado se mantengan en niveles elevados.

En el siguiente gráfico se puede observar el comportamiento de la morosidad en la Municipalidad de San Antonio de Belén, se observa que la morosidad acumulada de los

impuestos y servicios municipales, el año 2019, fue del 10.9%, aumenta en 3.02 puntos porcentuales, con respecto del 2018.

Gráfico 2. **Morosidad acumulada del año 2002 al 2019**



Fuente*:* Informe de labores, Municipalidad de San Antonio de Belén.

2.3 Insatisfacción de contribuyentes:

En la actualidad, los contribuyentes municipales no tienen acceso a la información digital referente a estadísticas de pagos de servicios brindados por el municipio, por lo cual se tiene un grado de insatisfacción de servicio a la ciudadanía.

2.4 Afectación en la toma de decisiones y oportunidades de desarrollo:

Adicionalmente, ante la incapacidad de estudiar tendencias y caracterización de individuos mediante el análisis de datos; la toma decisiones para la implementación de nuevas oportunidades de negocio que permitan fortalecer el marco de desarrollo municipal, genera una ineficiente estimación en la gestión de nuevos proyectos, además imposibilita una acertada identificación de contribuyentes, lo cual afecta en gran medida los ingresos percibidos por los municipios y por consiguiente su gestión, limitando de esta manera el avance, incursión e innovación de nuevas oportunidades en el mercado lo cual impide un mejor posicionamiento de las instituciones municipales, afectando así la imagen del gobierno y el servicio brindado a la población en general.

3. Justificación

Se considera la realización de este proyecto como una posible solución para ayudar a disminuir los índices de morosidad existente en las municipalidades, “Casi la mitad de los municipios del país afrontaron serios problemas en sus finanzas en los últimos años, siendo la morosidad uno de los principales factores que influyeron en esta situación trágica que sufre el país” (Rodríguez, 2014). Este tema ha afectado durante mucho tiempo la estabilidad y el poder de acción de los gobiernos locales, evitando así que se puedan generar nuevos proyectos por la falta de presupuesto disponible.

La recolección de impuestos y cobro de servicios por parte de las municipalidades es de suma importancia y es un tema que se le debe dar un énfasis prioritario. En la actualidad existen muchas soluciones tecnológicas que sin duda ayudarían a mitigar esta situación; sin embargo, se percibe resistencia al cambio por parte de las instituciones públicas y por ello se deja de lado la realización de proyectos innovadores que involucren nuevas tecnologías, con el fin de crecer y así contar con un sistema público más eficiente.

Tomando la idea principal en la que se basa el proyecto y la cual busca facilitar una solución tecnológica basada en técnicas de minería de datos usando algoritmos predictivos, esta implementación vendría a apoyar la toma de decisiones en temas de categorización de los clientes, permitiendo categorizar contribuyentes por medio de determinadas variables predictivas definidas tomando como principal fuente de aprendizaje el comportamiento de pagos de otros individuos, y así lograr identificar si será morosa o bien si en su defecto no lo será.

“Las mejores armas contra el moroso son la perseverancia, la constancia y la insistencia. Para evitar llegar a situaciones críticas y conseguir cobrar a los clientes morosos se deben tener en cuenta varios factores: El tiempo, periodos de riesgo y factor confianza.” (Gala, 2008). Esta sería una funcionalidad automatizada que facilitaría el proceso de cobros, ya que alertaría a las personas encargadas y así puedan tomar decisiones antes de que un individuo incurra en mora.

Por tanto, el beneficio esperado mediante la implementación de esta solución tecnológica consiste en:

- Disminución de los índices de morosidad, dando valor al aumento de los ingresos económicos del municipio
- Aumentar el presupuesto disponible en las municipalidades, esto por consiguiente de la mejora en la recaudación de ingresos por cobros de servicios.
- Mejorar la gestión municipal aumentando los montos disponibles para la ejecución de proyectos tanto actuales como nuevas oportunidades.
- Finalmente, a raíz de una mejor gestión municipal y aumento en el presupuesto, se espera ver reflejada una mejora en los servicios brindados por los gobiernos locales a la población, contribuyendo de esta forma a la mejora en la calidad de vida y desarrollo de los pueblos, obteniendo así un beneficio a la población contribuyente.

4. Objetivos del Proyecto

4.1 Objetivo general

Facilitar la caracterización de contribuyentes morosos utilizando modelos de aprendizaje supervisado en las municipalidades.

4.2 Objetivos específicos

1. Analizar la problemática de morosidad en las organizaciones municipales y la posible aplicación de la minería de datos, mediante métodos de aprendizaje supervisado.
2. Proponer un modelo de minería de datos de aprendizaje supervisado basado en el análisis efectuado para la caracterización de individuos morosos.
3. Evaluar el modelo propuesto y analizar los resultados en función de su utilidad para facilitar la disminución de morosidad en la Municipalidad de San Antonio de Belén.

CAPÍTULO II
MARCO TEÓRICO

CAPÍTULO II: MARCO TEÓRICO

1. Municipalidades en Costa Rica

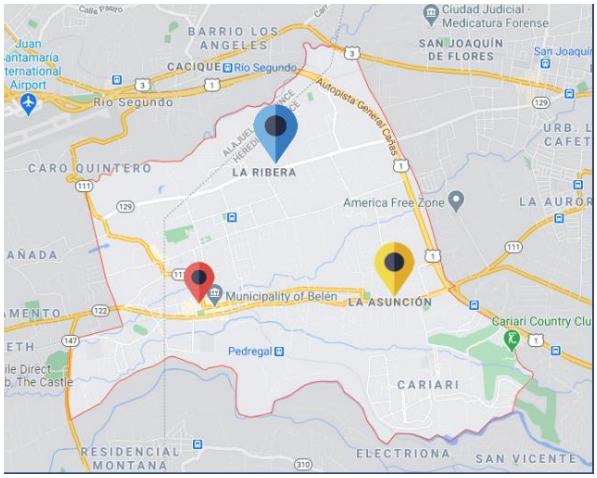
En Costa Rica, el territorio nacional está dividido constitucionalmente de la siguiente manera: “Provincias, cantones y distritos…”. (Constitución Política de Costa Rica, 1949, artículo 168). Consecuentemente, a través de la historia, en Costa Rica se han establecido siete provincias, y ochenta y un cantones, esto implica que el país cuenta con la misma cantidad de gobiernos locales.

En cuanto a la administración, se definen a las municipalidades como las instancias gubernamentales encargadas de la gestión del cantón: “La administración de los intereses y servicios locales en cada cantón, estará a cargo del Gobierno Municipal, formado de un cuerpo deliberante, integrado por regidores municipales de elección popular, y de un funcionario ejecutivo que designará la ley.” (Constitución Política de Costa Rica, 1949, artículo 169). Es de esta forma que los gobiernos municipales se les encargan la función principal de gestionar los servicios básicos que requiere la población para desarrollar un modo de vida óptimo, ofreciendo estrategias para el desarrollo cantonal.

2. Municipalidad de San Antonio de Belén

El cantón de Belén está constituido por tres cantones a saber: San Antonio, La Ribera y La Asunción como se puede ver en la siguiente imagen:

Figura 3. **Distritos del cantón de Belén.**



Fuente: Recuperado de la página web de la Municipalidad de San Antonio de Belén.

Misión: Somos una institución autónoma territorial que promueve el desarrollo integral y equitativo, administra servicios de manera innovadora, eficiente y oportuna, con el propósito de contribuir al bienestar de sus habitantes.

Visión: Ser una institución que, mediante un desarrollo integral, equitativo y equilibrado, garantice el bienestar de sus habitantes.

3. Gestión Municipal

La Real Academia Española define la palabra gestión como "la acción y efecto de gestionar o administrar" RAE (2019). Según Asensio, R. (2012), la gestión a nivel de municipios comprende las acciones, actividades o procedimientos que realizan las entidades u organismos municipales encaminados al logro de objetivos y cumplimiento de metas. (p. 5). Así pues, se obtiene que la gestión municipal comprende la administración de los recursos del cantón con el propósito de alcanzar las metas que propicien el desarrollo de los intereses del pueblo.

Siguiendo este punto, en Costa Rica según la CGR en el Informe de Índices de Gestión Municipal del Año 2017, reporte que toma en cuenta las deferentes municipales establecidas en el país, establece que: "Los gobiernos locales son los responsables de gestionar los recursos cantonales y propiciar condiciones de calidad, innovación,

participación e inclusión que deparen en bienestar general para todos los munícipes.” (CGR, 2017). Es de esta forma que los gobiernos cantonales tienen el reto de llevar a cabo una administración que propicie el bienestar de la población logrando el crecimiento y avance íntegro de la ciudadanía.

4. Morosidad

La mora se puede definir de la siguiente manera: “Se llama constitución en mora del deudor a aquella situación jurídica en la que se encuentra el obligado tras el incumplimiento de la obligación que le incumbía, porque su acreedor al haberle reclamado la prestación le ha colocado en esta situación de "especial responsabilidad" (Picazo Giménez, 1994, pág. 563). En el concepto anterior se especifica que una persona se convierte en morosa ante su acreedor, en el momento en el que incumple las condiciones que se definieron, en este momento el deudor se expone a una penalización por medio de intereses moratorios aplicados sobre el total de la deuda.

Siendo la morosidad un problema que afecta la gestión en las municipalidades, el presente trabajo plantea soluciones basadas en minería de datos, por lo cual es indispensable conocer los siguientes términos:

5. Base de datos

Bob Bryla (2015), sostiene que una base de datos es una colección de datos en disco que se encuentran almacenados en uno o más archivos, dichos archivos se encuentran en servidor de base de datos, (p. 1). Para efectos del presente estudio, se toma el término de base de datos como elemento fundamental de almacenamiento de datos relacionados. Además, se introduce la definición de Sistema de Gestión de Base de Datos, el cual es el elemento principal que permite a las bases de datos gestionar su función. “El SGBD crea y organiza la base de datos, y además atiende todas las solicitudes de acceso hechas a la base de datos tanto por los usuarios como por las aplicaciones”. (Cabello, 2010, p. 23).

6. Minería de datos

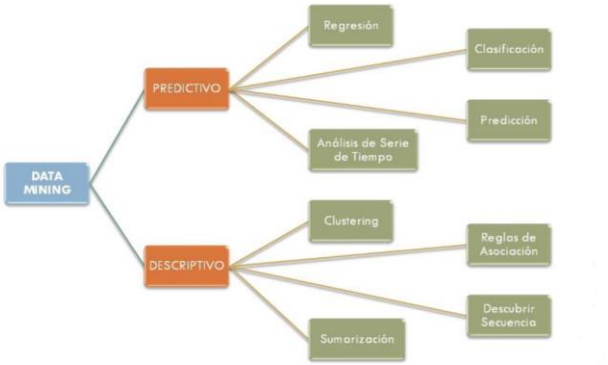
Rodríguez, (2013), menciona que la minería de datos es un proceso no elemental de búsqueda de relaciones, correlaciones, dependencias, asociaciones, modelos, estructuras, tendencias, clases (clústeres), segmentos, los cuales se obtienen de grandes juegos de datos, los cuales pueden estar presentes en bases de datos relacionales, o bien no relacionales. Es así como se puede decir que la minería de datos es un proceso que busca la información útil y utilizable en las organizaciones a partir de grandes colecciones de datos.

Siguiendo la definición de Rodríguez, la minería de datos se confluencia con áreas de conocimiento como: Tecnologías de bases de datos, estadística, visualización, ciencias de la información, matemáticas, entre otras.

Para efectos del presente trabajo, se plantea la aplicación de la minería de datos como un pilar de apoyo y facilitador en la obtención de modelos de datos, en particular en la caracterización de individuos morosos, tomando como principal fuente de información la base de datos de contribuyentes de ayuntamientos municipales.

En la siguiente figura, se muestran las diferentes técnicas utilizadas en la minería de datos, evidenciando la separación en modelos predictivos los cuales tienen un aprendizaje supervisado y los descriptivos o aprendizaje no supervisado.

Figura 4. **Técnicas de minería de datos.**



Fuente. Canal educativo de YouTube de Oldemar Rodriguez Rojas, 2013.

- **Aprendizaje supervisado**: Según Rodríguez, (2013), el aprendizaje supervisado en minería de datos permite predecir acciones futuras que sean de importancia para la empresa, por ejemplo: predecir si un cliente va a ser buen pagador o no.
- **Aprendizaje no supervisado**: Siguiendo el punto de Rodríguez, (2013), el aprendizaje no supervisado permite la segmentación y clusterización de los datos mediante el uso de diferentes métodos.

De esta manera, a continuación, se definen algunos métodos comprendidos en el aprendizaje supervisado:

- **Redes Neuronales**: Salas (2014), define una red neuronal de la siguiente forma:

  "Una red neuronal artificial (ANN) es un esquema de computación distribuida inspirada en la estructura del sistema nervioso de los seres humanos. La arquitectura de una red neuronal es formada conectando múltiples procesadores elementales, siendo éste un sistema adaptivo que posee un algoritmo para ajustar sus pesos...". (p. 1).
- **Bosques Aleatorios**: Según Media y Chacón (2017), definen este modelo: "Básicamente, selecciona aleatoriamente un número de variables con las

que se construye cada árbol individual y se hacen predicciones con estas variables que luego se ponderarán a través del cálculo de la clase más votada de estos árboles que se generaron, para finalmente hacer la predicción por Random Forest.”. (p. 165-189).

- **Árboles de Decisión**: Han y Kamber, (2012), explican este método de la siguiente manera “Imagine que cada uno de los clasificadores en el conjunto es un clasificador de árbol de decisión, de modo que la colección de clasificadores es un "bosque". Los árboles de decisión individuales se generan utilizando una selección aleatoria de atributos en cada nodo para determinar la división”, (p. 383).
- **Máquinas de Soporte Vectorial**: Según Rodríguez, (2013), las máquinas de soporte vectorial tratan de encontrar el hiperplano que separe a las clases con el mayor “margen” posible.

7. Over-fitting y under-fitting del modelo

Al trabajar con modelos predictivos, uno de los temas que se deben tener en cuenta al elegir y por consiguiente proponer un modelo de aprendizaje supervisado es el tema de over-fitting y under fitting o bien perfecto ajuste o desajuste del modelo con respecto a los datos utilizados, Fiels, Miles y Field, lo exponen de la siguiente forma: “Existen un peligro de over-fitting (se tienen muchas variables en el modelo que no generan contribución a la predicción del modelo) y under-fitting (no considerar variables predictivas importantes) del modelo.” (Andy Fiel, Jeremy Miles, Zoe Field, 2012).

8. Validación Cruzada

Según Fiels, Miles y Field definen la validación de la siguiente forma: “Conocer la precisión de un modelo utilizando diferentes grupos de muestras se define como validación cruzada” (Andy Fiel, Jeremy Miles, Zoe Field, 2012).

9. Matriz de confusión

Con el fin de conocer la precisión de un modelo se utiliza la matriz de confusión, el objetivo es obtener los índices de predicción de un determinado modelo para un conjunto de datos de prueba y de aprendizaje:

Veamos el siguiente ejemplo propuesto por Grus: “Dado un un conjunto de datos etiquetados y un modelo predictivo, cada registro puede estar en una de las siguientes categorías:

- Verdaderos positivos: “Un mensaje que es spam, y el modelo lo clasifica como spam de forma correcta.”
- Falso positivo (error tipo 1): “Un mensaje que no es spam, pero el modelo lo clasifica como spam”
- Falso negativo (error tipo 2): “Un mensaje que es spam, pero se predice como no spam.”
- Verdadero negativo: “Es un mensaje que no es spam y el modelo lo predice de forma correcta como no spam.”” (Grus, 2019).

En la siguiente figura, se muestra lo descrito anterior en una matriz de confusión:

Figura 5. **Matriz de confusión**

	Spam	not Spam
predict “Spam”	True Positive	False Positive
predict “Not Spam”	False Negative	True Negative

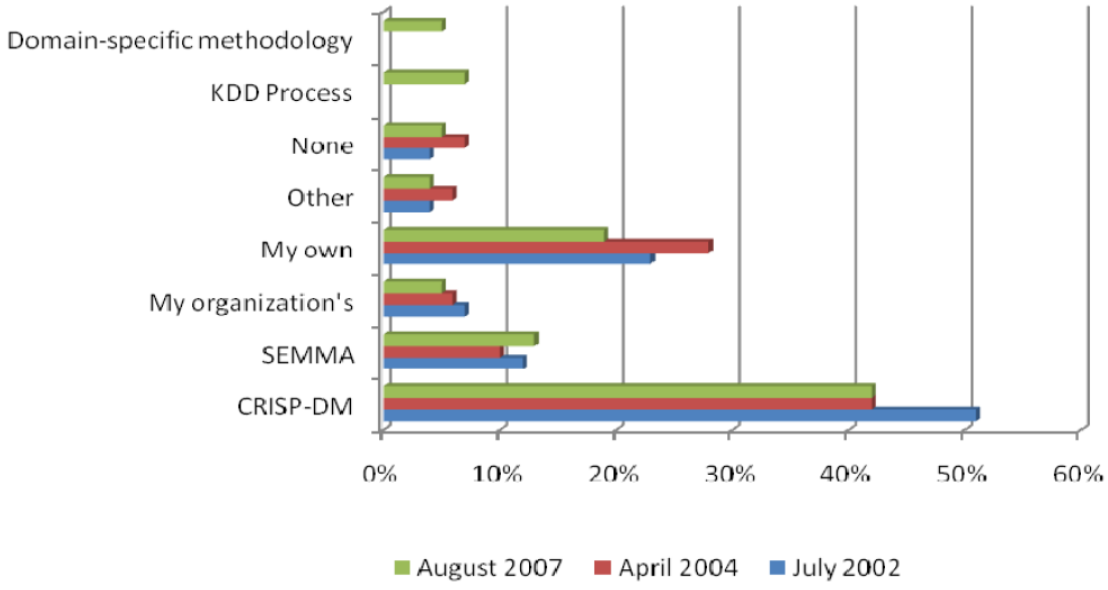
Fuente: Extraído del libro “Data Mining from Scratch”, (Grus, 2019).

10. Metodologías de minería de datos

En minería de datos se utilizan diferentes metodologías, la siguiente gráfica representa el resultado de encuestas realizadas en el año 2007 en donde se evidencia el grado de utilización de las diferentes guías de desarrollo de proyectos aplicados a data

mining. Se obtiene como resultado que la metodología CRISP-DM es la que más se utiliza:

Gráfico 3. **Metodologías utilizadas en data mining**



Fuente: Extraído de la página web kdnuggets, 2007.

Cabe mencionar que el presente proyecto implementa la metodología CRISP-DM como guía de trabajo.

A continuación, se describen algunas metodologías utilizadas en el proceso de minería de datos:

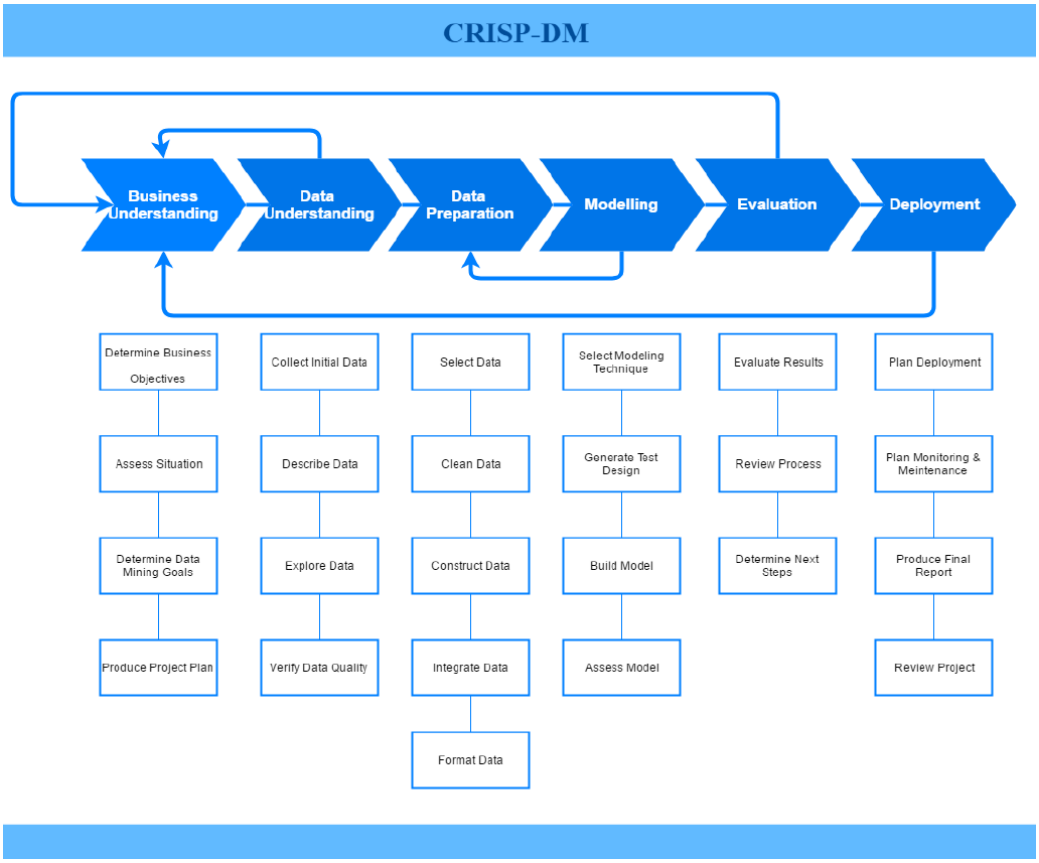
10.1 CRISP-DM

Cross Industry Standard Process for Data Mining es un modelo de proceso de minería de datos, cuyo origen se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases) [Reinartz, 1995], [Adraans, 1996], [Brachman, 1996], [Fayyad, 1996], el desarrollo de una guía de

referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining).

La siguiente figura, muestra las fases de la metodología CRISP-DM.

Figura 6. **Fases de la metodología CRISP-DM.**



Fuente. Óscar Marbán, G. M. (enero, 2009).

A continuación, se describen las fases de la metodología CRISP-DM:

10.1.1 Comprensión del negocio

Según Pete Chapman “Esta es la fase inicial se enfoca en el entendimiento de los objetivos y requerimientos del proyecto con una perspectiva empresarial” (NCR, 2000, p. 10).

Según Rodríguez, en su proyecto realizado con base a la tésis: “Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM)” de José Alberto Gallardo Arancibia” hay preguntas que se deben de responder en esta fase, las cuales son:

¿Cuál es el conocimiento previo disponible acerca del problema?

¿Se cuenta con la cantidad de datos requerida para resolver el problema?

¿Cuál es la relación coste beneficio de la aplicación de DM?

Indicando que en esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de minería de datos.

10.1.2 Comprensión de los datos:

Según Rodríguez, esta fase comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema.

10.1.3 Preparación de los datos:

En esta fase se realizan una serie de actividades concernientes a la construcción de un conjunto de datos. Incluye diferentes tareas como lo son selección de tablas, herramientas de modelado y carga de datos, variables, registros y la limpieza de los datos.

10.1.4 Modelado:

Según Rodríguez, en esta fase se seleccionan las técnicas de modelado que sean atinadas al problema propuesto, para ello se deben considerar los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

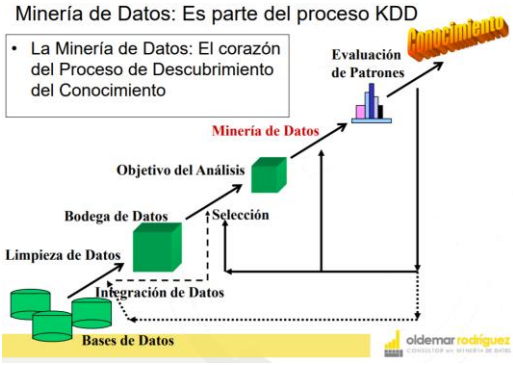
10.1.5 Evaluación:

En esta fase del proyecto se ha desarrollado un modelo (o modelos) que parece tener una buena calidad, desde un punto de vista de análisis de datos.

10.2 Descubrimiento de conocimiento en bases de datos (KDD)

El descubrimiento de conocimiento en bases de datos o KDD (Knowledge Discovery in Data Bases), es el proceso de identificar patrones característicos o especiales a partir de los datos, que incluye varios pasos a saber, tal y como se muestra en la siguiente figura:

Figura 7. **Representación gráfica de KDD.**



Fuente*:* Canal educativo de YouTube de Oldemar Rodríguez Rojas, publicado en el 2013.

En la Figura 5, se pueden observar los pasos que conlleva el KDD, el cual tiene como principal proceso el de minería de datos.

11. Estudios realizados aplicando minería de datos

La minería de datos en contraste con su aplicación en las organizaciones y como un recurso facilitador en la toma de decisiones, se encuentra su utilidad en estudios realizados y evaluados en diferentes organizaciones:

Minería de datos en el área de la municipalidad catastral de Perú

En el estudio de tesis realizado en Perú por Antezana Bustamante, cuyo tema se titula: “Impacto de la implementación de minería de datos en el mantenimiento y análisis

de la información catastral en una municipalidad distrital". (Antezama, 2018, p.1). En este estudio se aplican técnicas de minería de datos en el área de catastro de la municipalidad distrital de Perú. Como resultado principal, en esta investigación se logran evidenciar oportunidades de mejoras gracias al análisis aplicando la minería de datos:

"Optimizando el procedimiento "Inspección y Fiscalización Tributaria" mediante el uso de la minería de datos se contribuye, también, a otros procedimientos como la atención de requerimientos de información y de recomendaciones, reclamaciones tributarias y la proposición de políticas y normas para los procesos tributarios de Registro, Recaudación, Fiscalización y Cobranza.". (Daniel Antezama, 2018, p.230).

Minería de datos para mejorar la seguridad en el tránsito:

Además, en su estudio Antezama menciona el estudio realizado por Scott Salomón (2005) "Using Data Mining to improve traffic safety programs", en donde según se indica, el objetivo principal consistía en usar diferentes técnicas de minería de datos que permitieran mejorar la seguridad del tráfico en función de la reducción de víctimas mortales mediante una evaluación efectiva del monitoreo de las cámaras de seguridad en las intersecciones de semáforos en los Estados Unidos. Como parte de los resultados obtenidos permitieron atender los factores de riesgo mayor en las intersecciones, conductor y vehículos que causaban accidentes, logrando una eficacia e impacto directo en la disminución de víctimas en las carreteras de este país.

En los casos anteriores se logra evidenciar la facilidad que ofrece la minería de datos en diferentes tipos de organizaciones y en específico distintas fuentes de datos, siendo el mayor valor el apoyo en la toma de decisiones lo cual permite lograr una mejora en los procesos.

CAPÍTULO III
METODOLOGÍA

CAPÍTULO III: METODOLOGÍA

1. Tipo de investigación

Para el desarrollo de este proyecto se realizó una investigación de tipo aplicada, en la misma se busca determinar de una manera efectiva la posibilidad de que los contribuyentes incurran en mora, aplicando técnicas de minería de datos en los registros de información personal de los contribuyentes y su historial de pago de servicios en la Municipalidad.

En este estudio se busca aportar de gran manera al desarrollo de la institución, específicamente en el departamento de gestión de cobros, y así proponer una herramienta que facilite el arduo trabajo que implica determinar los posibles clientes que incurrirán en morosidad.

El departamento de gestión de cobros no utiliza modelos predictivos para poder determinar posibles clientes en riesgo de morosidad, de igual forma no se realiza una agrupación de consumidores entre morosos y no morosos, con el fin de generar proyecciones a futuro que les brinden información indispensable para la toma de decisiones de negocio acertadas y así mitigar el problema que viven en la actualidad.

2. Población y muestra

La población de los datos que se utilizaran para el presente estudio tiene en consideración los contribuyentes del cantón de Belén el cual contiene un aproximado de 22,000 habitantes de los cuales un número cercano a los 8,000 son clientes activos de la Municipalidad, además se comprenden los registros disponibles en el sistema municipal SIGMB desde el año 2017 al 2021. Por otro lado, se hace uso del padrón electoral el cual contiene la población votante de Costa Rica, también se accede a los archivos maestros de nacimientos, matrimonios y defunciones del país, facilitados por el Tribunal Supremo de Elecciones.

Para efectos del aprendizaje supervisado propuesto en este estudio, se utilizan funcionalidades del lenguaje de programación R y sus funcionalidades que permiten

segmentar los datos en conjuntos de aprendizaje que comprende un total de 70% y el 30% restante se aplica para efectos de pruebas.

3. Descripción de instrumentos

En el presente proyecto se utilizaron los datos de contribuyentes del municipio de Belén, dichos datos fueron facilitados por la organización, los mismos son generados como parte del proceso de gestión municipal. Así mismo, se facilitan datos obtenidos mediante convenios entre la Municipalidad y el Registro Civil, estas bases de datos fueron facilitadas en archivos de texto separados por comas (.csv), las mismas contienen información de nacimientos, matrimonios y defunciones de Costa Rica. Además, se obtienen los datos del padrón electoral en archivos en formato texto separados por comas (.csv).

4. Procedimientos para analizar la información del diagnóstico

En este proyecto con el fin de analizar y tratar el proceso de minería de datos bajo un estándar bien organizado, se decidió implementar la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) con el que se generó la revisión de la información recolectada por el sistema SIGMB del municipio de Belén.

Este modelo es muy reconocido y cubre todas las fases necesarias para explorar, analizar y manejar los datos que se utilizaron en este proyecto, así como la generación de tareas requeridas, brindando así un orden consecutivo.

La metodología contempla el proceso de análisis de datos de forma estructurada en el desarrollo del proyecto, esto se logra estableciendo un contexto mucho más rico, que influye en la elaboración de los modelos de minería de datos.

CAPÍTULO IV
PROPUESTA DE SOLUCIÓN

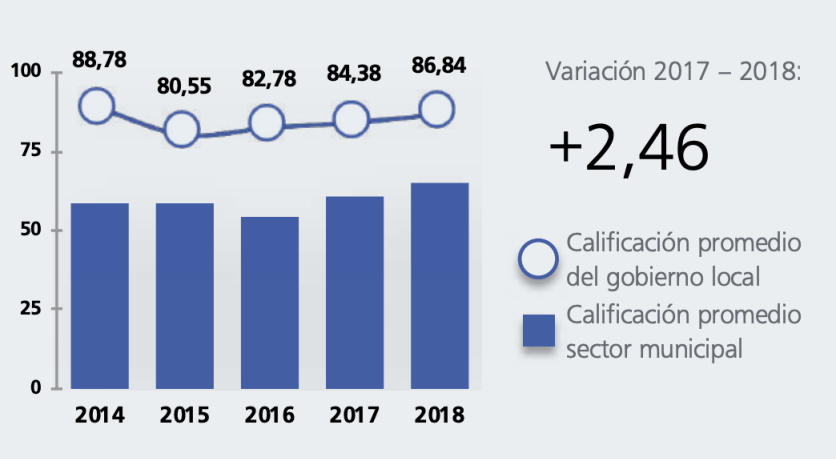
CAPÍTULO IV: PROPUESTA DE SOLUCIÓN

1. Diagnóstico

En la actualidad las organizaciones municipales se ven en la obligación de brindar servicios de alta calidad y de esta forma facilitar una buena prestación de servicios a la población.

Según el informe de la Contraloría General de La República publicado en al año 2018 la Municipalidad es una de las mejores evaluadas en cuanto a índice de gestión municipal, obteniendo un puntaje de 86,84:

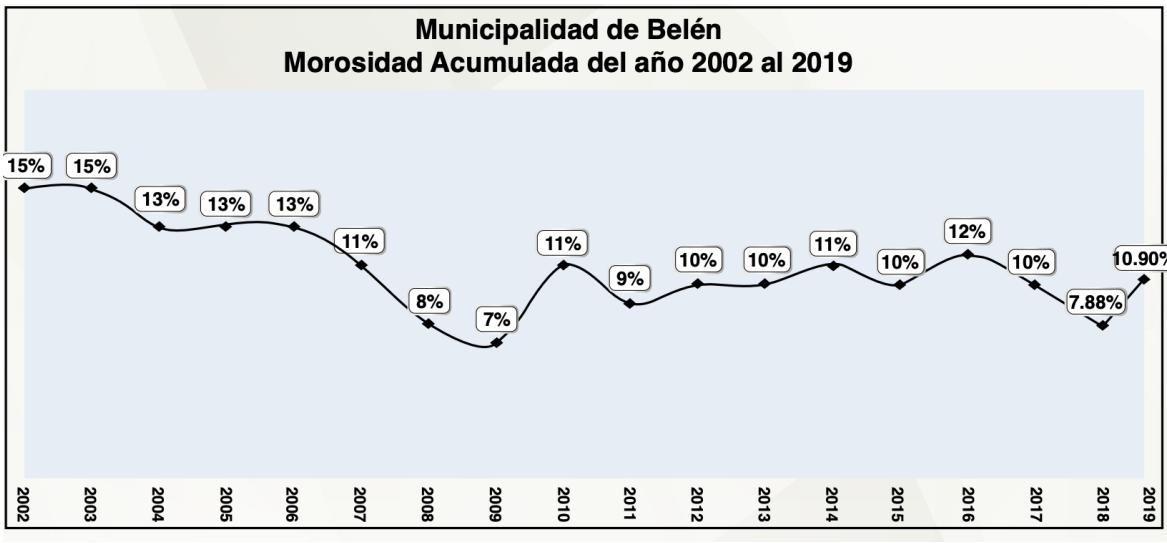
Figura 8. **Índice de gestión municipal de Belén**



Fuente: Informe de Índice de Gestión Municipal 2018, Contraloría General de la República.

En la ilustración anterior, se puede ver que organización municipal tiene índices de gestión municipal que superan el 80% los últimos 5 años, lo cual demuestra los esfuerzos que se realizan por lograr una buena gestión municipal en este municipio. A pesar de ello, los municipios presentan retos importantes en el ámbito de gestión de cobros, es a pesar de los esfuerzos que se realizan día a día. Según el informe de labores expuesto por la Municipalidad en el año 2019, se evidencia que no están exentos de la problemática de morosidad, para este informe en el año 2019 se presenta una morosidad acumulada del 10.9% evidenciando un aumento del 3.02% con respecto al año anterior.

Gráfico 4. **Morosidad acumulada del año 2002 al 2019**



Fuente: Informe de labores 2019, Municipalidad.

Con respecto al proceso de cobros de servicios, la municipalidad identifica los contribuyentes morosos mediante informes generados del Sistema de Gestión Municipal (SIGMB por sus siglas), una vez se tiene el listado el departamento realiza notificaciones a las personas, dando prioridad a las cuentas que están cercanas a prescribir, este proceso es manual y las medidas que se toman es una vez que se sabe que el contribuyente presenta problemas de morosidad, evidenciando que existe una limitante en cuanto a la toma de decisiones preventivas que permitan identificar los contribuyentes morosos con antelación.

Tabla 2. Contribuyentes morosos según SIGMB

-

AUX_CONTAB	NUM_CUENTA	TIP	PERIODO	NUM_PERSON	MON_INTERE	CEDULA	MON_PERIOD	DIA_VENCIM	NOMBRE_COM
CUF	12309	LVP	2018.04	5	4.69	104141038	29.25	642	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.05	5	4.69	104141038	29.25	611	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.06	5	4.69	104141038	29.25	581	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.07	5	3.94	104141038	29.25	550	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.08	5	3.94	104141038	29.25	519	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.09	5	3.94	104141038	29.25	489	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.1	5	3.19	104141038	29.25	458	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.11	5	3.19	104141038	29.25	428	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2018.12	5	3.19	104141038	29.25	397	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2019.01	5	2.46	104141038	29.25	366	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2019.02	5	2.46	104141038	29.25	338	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2019.03	5	2.46	104141038	29.25	307	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2019.04	5	1.72	104141038	29.25	277	CRISTHIAN FERNANDEZ AGUERO
CUF	12309	LVP	2019.05	5	1.72	104141038	29.25	246	CRISTHIAN FERNANDEZ AGUERO

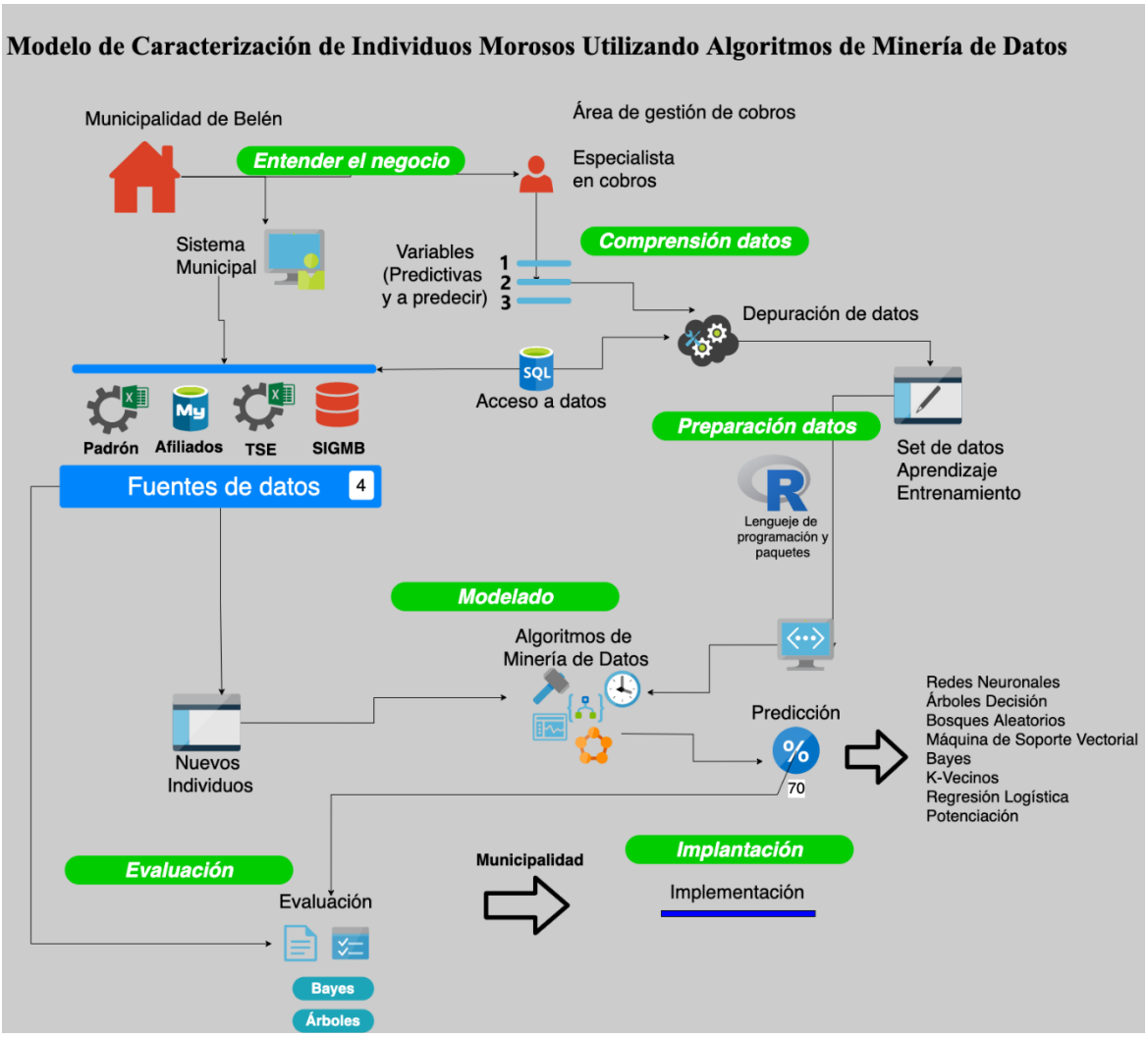
Fuente: Elaboración propia, obtenido del informe de morosidad generado por el SIGMB.

Conociendo de esta forma el diagnóstico de la situación que enfrentan los municipios tanto a nivel de gestión municipal como en el área de gestión de cobros por servicios, se procede con la propuesta de solución del presente proyecto:

2. Propuesta de solución

En este proyecto se ha implementado y desarrollado la metodología CRISP-DM, así pues, como parte de la propuesta de solución resulta importante comprender la implementación de esta metodología en el proyecto, en el siguiente gráfico se muestra un resumen del experimento propuesto y cada una de sus fases iniciando por la etapa de compresión del negocio y finalizando con la evaluación e implementación:

Gráfico 5. **Resumen de la propuesta de solución**



Fuente: Elaboración propia.

De esta manera, teniendo un visualización de la propuesta, se procede con el desarrollo de la solución:

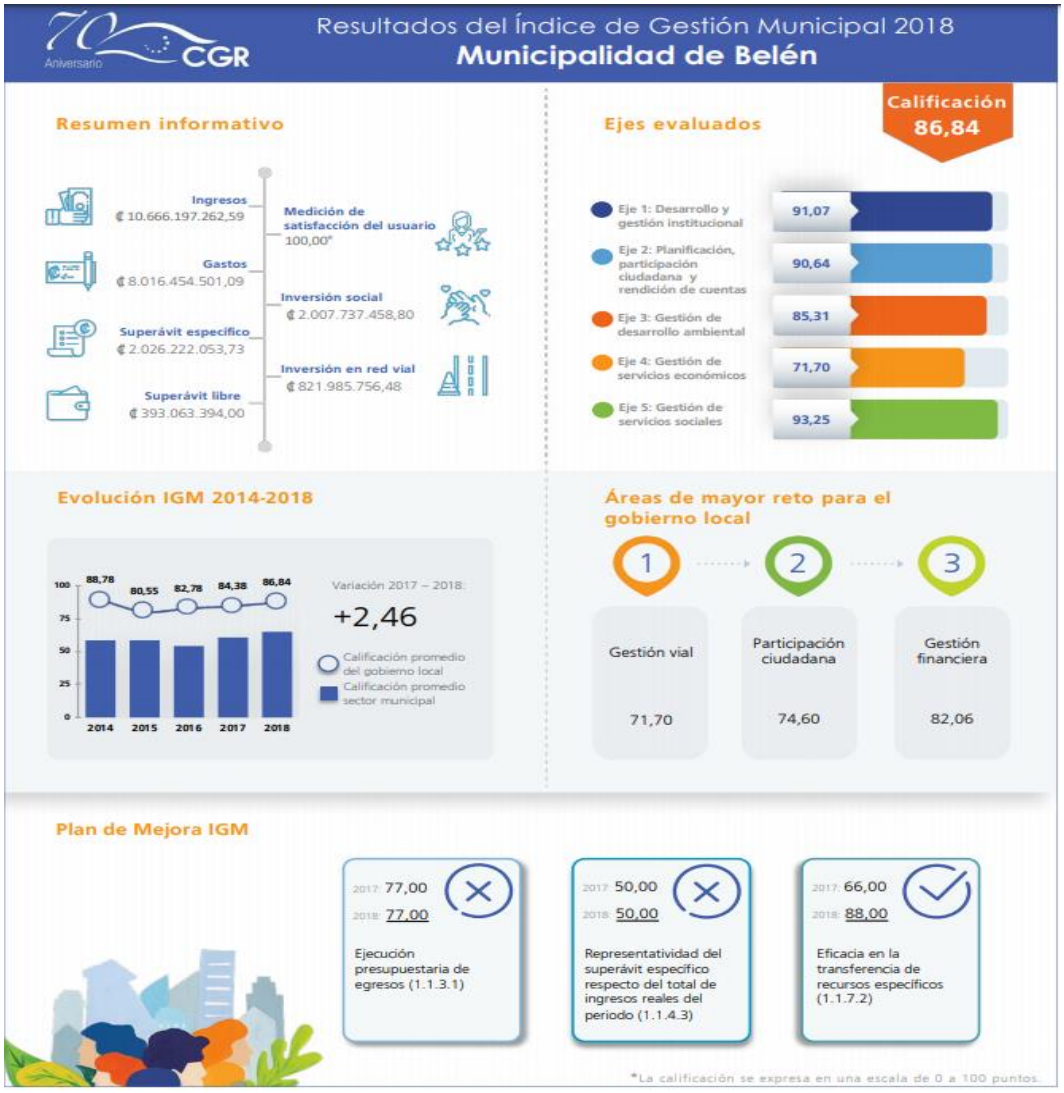
2.1 Comprensión del negocio:

Esta fase ha sido cubierta a lo largo del presente documento, al inicio del presente proyecto se han planteado los antecedentes, problemática y justificación, los cuales permiten conocer y entender el modelo de negocio de la empresa y que permiten la realización del presente proyecto.

Como se conoce, las municipalidades son las principales entidades en brindar bienestar y propiciar el desarrollo de los pueblos mediante la gestión de los recursos de cada cantón. La Municipalidad es uno de los municipios de mejor gestión municipal según la evaluación de la Contraloría General de La República:

En la figura 8, se pueden ver los índices de gestión municipal obtenidos por la Municipalidad en el año 2018:

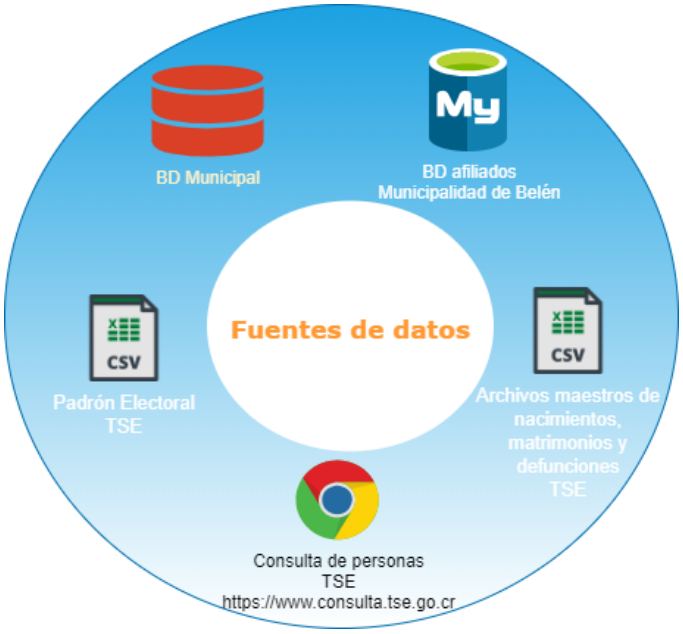
Figura 9. **Índice de gestión municipal Municipalidad, 2018**



Fuente. Informe de Índices de Gestión Municipal, CGR, 2018.

Por otro lado, para efectos del presente proyecto, el departamento de informática facilita el acceso a los datos, esto posible mediante bases de datos locales, o bien accesibles mediante convenios de colaboración con entidades gubernamentales:

Figura 10. **Fuentes de datos utilizadas:**



Fuente. Elaboración propia.

- Base de datos Oracle de la Municipalidad.

  Se facilita un respaldo de la base de datos del sistema integrado de gestión municipal:

  Respaldo de base de datos municipal:

```
 11G -rw-r--r--. 1 oracle oinstall  11G Feb  6 16:02 ORCL_28-01-21.dmp
2.8M -rw-r--r--. 1 oracle oinstall 2.8M Feb  6 17:29 IMP-ORCL-DEC-06-02-21.log
[oracle@VM-LIC-UNA DMP]$
```

Fuente. Elaboración propia.

Esta base de datos tiene la siguiente cantidad de objetos: un total de 1611 tablas de base de datos:

Figura 11. **Respaldo de la base de datos municipal:**

1	PROCEDURE	11261
2	VIEW	2499
3	TABLE	1611
4	PACKAGE	2
5	TYPE	1
6	PACKAGE BODY	1

Fuente. Elaboración propia.

- Base de datos de contribuyentes.

  Archivo .csv obtenido de la base de datos MySQL de la Municipalidad:

Figura 12. **Base de datos de afiliados de la municipalidad.**

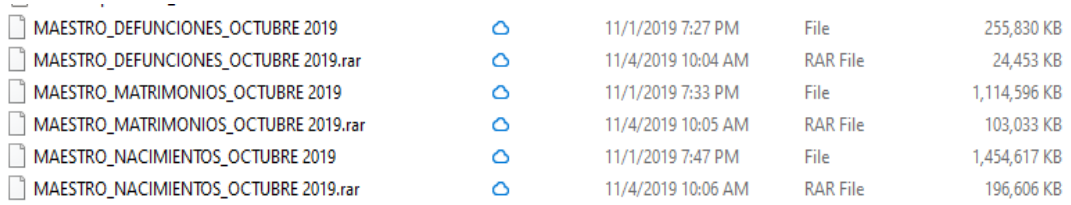
	A	B	C	D
	104800475	ebrown_1@hotm	83842686	A
	3101570085	cuentasporpagar(	22933211	A
	401021263	laboratorio1al@h	88867207	A
	1388248	info@villasdecari	22391341	A
	29	jesus_garcia@hotmail.com		A
	00000000520897A	paolaceccondin@	70128823	A
	17140611	dianahaskour@h	87231193	A
	48446903	info@romany asc	22393874	A
	0000000AO241900	hosterialasvegas@hotmail.(		A
	0000000AS850680	hosterialasvegas@hotmail.(		A
	0000000C1057361	sdebreuning@gm	22907787	A

Fuente. Elaboración propia.

- Base de datos de nacimientos, matrimonios y defunciones del Registro Civil.

  Se facilitan los archivos maestros de defunciones, matrimonios y nacimientos según el Registro Civil:

Figura 13. **Archivos de maestros de nacimientos, matrimonios y defunciones**



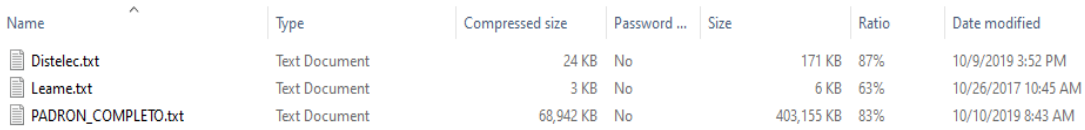
MAESTRO_DEFUNCIONES_OCTUBRE 2019	11/1/2019 7:27 PM	File	255,830 KB
MAESTRO_DEFUNCIONES_OCTUBRE 2019.rar	11/4/2019 10:04 AM	RAR File	24,453 KB
MAESTRO_MATRIMONIOS_OCTUBRE 2019	11/1/2019 7:33 PM	File	1,114,596 KB
MAESTRO_MATRIMONIOS_OCTUBRE 2019.rar	11/4/2019 10:05 AM	RAR File	103,033 KB
MAESTRO_NACIMIENTOS_OCTUBRE 2019	11/1/2019 7:47 PM	File	1,454,617 KB
MAESTRO_NACIMIENTOS_OCTUBRE 2019.rar	11/4/2019 10:06 AM	RAR File	196,606 KB

Fuente. Elaboración propia.

- Padrón electoral del Tribunal Supremo de Elecciones.

  Se obtiene el padrón electoral de Costa Rica:

  Figura 14. **Padrón electoral de Costa Rica**



Name	Type	Compressed size	Password ...	Size	Ratio	Date modified
Distelec.txt	Text Document	24 KB	No	171 KB	87%	10/9/2019 3:52 PM
Leame.txt	Text Document	3 KB	No	6 KB	63%	10/26/2017 10:45 AM
PADRON_COMPLETO.txt	Text Document	68,942 KB	No	403,155 KB	83%	10/10/2019 8:43 AM

Fuente. Elaboración propia.

- Consulta de personas por cédula del Tribunal Supremo de Elecciones.

  Se utiliza el portal de consultas de personas por cédula del Tribunal Supremo de Elecciones con el fin de consultar y validar datos de personas:

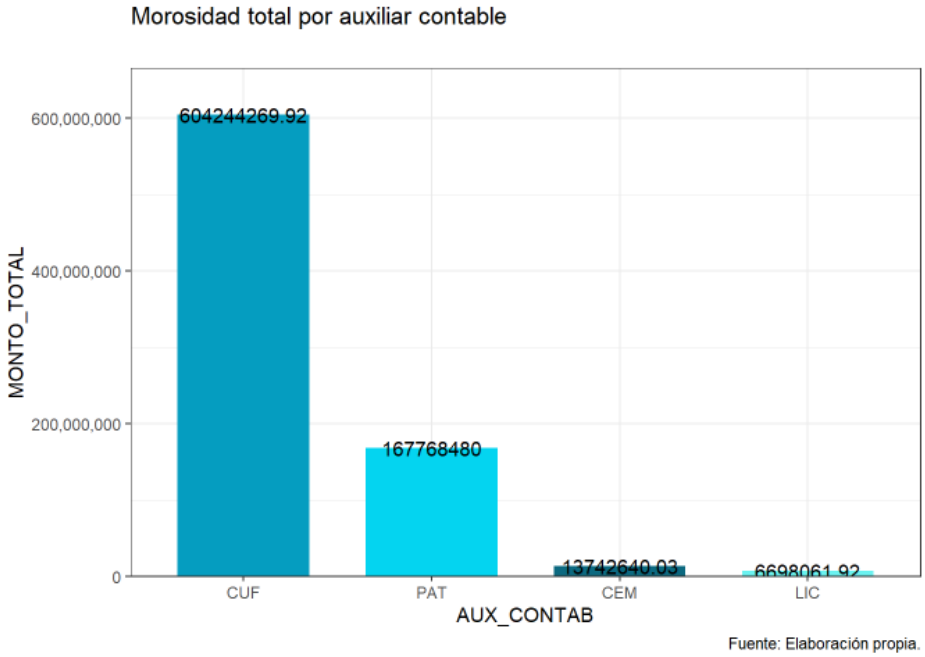
Figura 15. **Portal de consultas a civiles, Tribunal Supremo de Elecciones**

Fuente. Portal de consultas civiles, Tribunal Supremo de Elecciones.

2.1.1 Exploración inicial de morosidad

En temas de morosidad, la Municipalidad presenta una morosidad con montos importantes como se puede ver en el siguiente gráfico, se puede observar que el auxiliar contable que más problemas de morosidad presenta es CUF con un monto pendiente de 604,244,269.92 millones de colones.

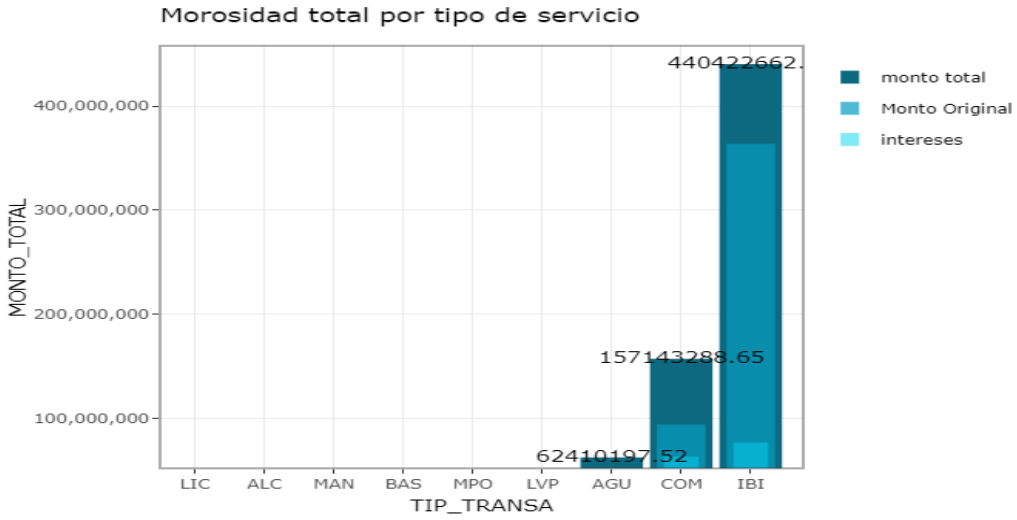
Gráfico 6. **Morosidad total por auxiliar contable**



Fuente: Elaboración propia.

Con respecto a los servicios que presentan mayores problemas de morosidad se tiene que el servicio de bienes inmuebles es el servicio que presenta el monto mayor de deuda con 440,422,662.44 millones de colones, seguidamente el servicio de patentes comerciales con monto pendiente de 157,143,288.65 millones.

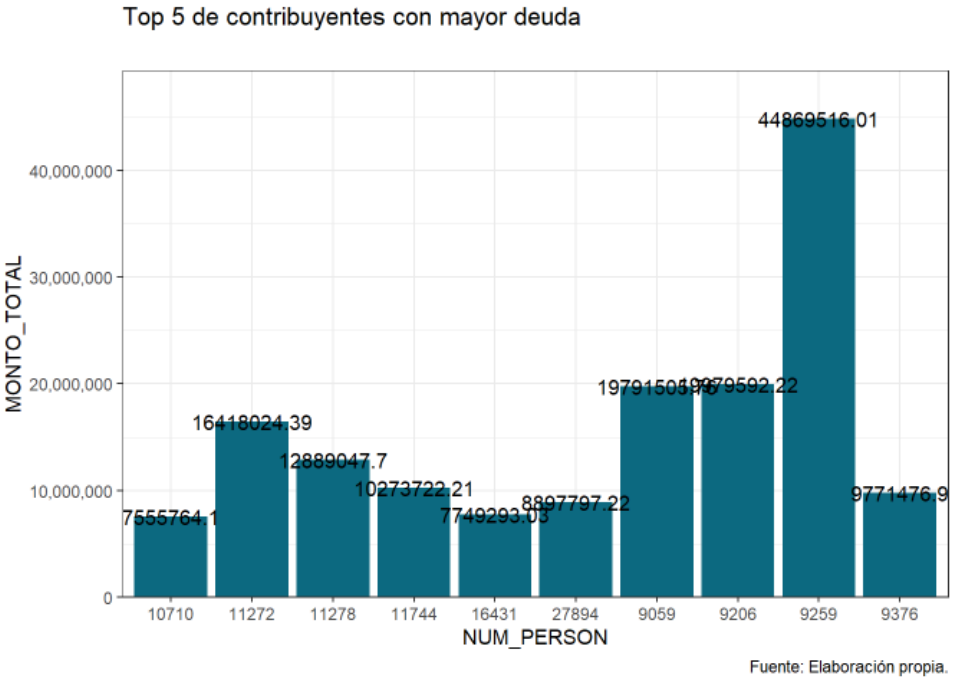
Gráfico 7. **Morosidad total por tipo de servicio**



Fuente: Elaboración propia.

En cuanto a las personas que presentan los montos mayores de deuda, se observa que hay una persona con número 9259 que tiene una deuda de 44, 869,516.01 millones de colones.

Gráfico 8. **Top 5 de contribuyentes con mayor deuda**



2.2 Comprensión de los datos:

Con la finalidad de poder comprender y acceder los datos se pueden mencionar las siguientes herramientas:

- Equipo local: Se hace uso de una laptop con capacidad de 16 GB de memoria RAM y 8 núcleos.
- Servidor Virtual: Se crea una máquina virtual en Virtual Box y se le instala el sistema operativo Oracle Linux Server.
- Base de datos Oracle: Se configura e instala una base de datos Oracle 11G R2.
- Office 365: Se utilizan herramientas de office como: Excel y Word, con el fin de crear, editar, acceder archivos como documentos formales, set de datos, entre otros.
- Servicios en la web: Se hace uso de servicios disponibles como lo es la plataforma de consulta de personas del Tribunal Supremo de Elecciones: https://www.consulta.tse.go.cr/consulta_persona/consulta_cedula.aspx.
- Base de datos MySQL: Se obtienen datos provenientes de base de datos MySQL.
- SQL: Para acceder a los datos, se utilizan sentencias SQL.
- R y RStudio: El presente proyecto hace uso de R como lenguaje de desarrollo y RStudio como IDE de desarrollo.
- Oracle Toad y SQL Developer: Herramientas utilizadas para acceder a la base de datos Oracle.

Figura 16. **Principales herramientas tecnológicas utilizadas en el proyecto.**

Fuente. Elaboración propia.

De igual forma, se utilizan paquetes en R, los principales se muestran en la siguiente figura:

Figura 17. **Paquetes utilizados en R.**

Fuente. Elaboración propia.

A continuación, se mostrará una breve descripción de cada uno de los paquetes utilizados en la herramienta R Studio:

R: Según (Bates, Bengtsson, & Bivand, s.f.) nos indica que R es un lenguaje y un entorno para la computación y los gráficos estadísticos. Es un proyecto GNU que es similar al lenguaje y entorno S. R proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento…) y técnicas gráficas, y es altamente extensible. El lenguaje S es a menudo el vehículo de elección para la investigación en metodología estadística, y R proporciona una ruta de código abierto para participar en esa actividad.

TidyVerse: Según (Wickham, s.f.) El tidyverse es un conjunto de paquetes que funcionan en armonía porque comparten representaciones de datos comunes y diseño de API. El paquete **tidyverse** está diseñado para facilitar la instalación y la carga de paquetes principales desde tidyverse con un solo comando.

Snow: El paquete Snow proporciona una interfaz de alto nivel para el uso de una estación de trabajo de clúster para cálculos paralelos en R . Snow se basa en el modelo de comunicación Maestro / Esclavo en el que un dispositivo o proceso (conocido como maestro) controla uno o más dispositivos o procesos (conocidos como esclavos).

Ggplot2: Según (Bates, Bengtsson, & Bivand, s.f.) describe el paquete como un sistema para crear gráficos de forma declarativa, basado en The Grammar of Graphics . Usted proporciona los datos, le dice a ggplot2 cómo asignar variables a la estética, qué primitivas gráficas usar y se encarga de los detalles.

Plotly: Según (Parmer, Parmer , & Johnson, 2013) la biblioteca de gráficos R de Plotly crea gráficos interactivos con calidad de publicación. Ejemplos de cómo hacer diagramas de líneas, diagramas de dispersión, gráficos de áreas, gráficos de barras, barras de error, diagramas de caja, histogramas, mapas de calor, subtramas, múltiples ejes y gráficos 3D (basados en WebGL).

2.2.1 Etiquetado de individuos

La variable que se desea predecir es “moroso”, así para etiquetar a los individuos se utiliza los siguientes métodos:

1. Reporte de morosidad: este reporte permite obtener los individuos que tienen deuda con la municipalidad, el mismo tiene en cuenta la fecha de vencimiento de cada recibo de pago de servicios.
2. Regla definida por la persona analista de cobros: Según la experiencia de la persona analista de cobros, los individuos adultos mayores que residen en el cantón de Belén y además tienen asignado el servicio de agua, estas personas son catalogadas como no morosas.

Figura 18. **Regla de negocio aplicada a los datos.**

```
datos[datos$IND_AFILIADO=="S" & datos$COD_SERVIC_AGU=="S" &
        (datos$EDAD>59 ) &
        (datos$VOTO_DISTRITO =="SAN_ANTONIO"|
           datos$VOTO_DISTRITO =="LA_RIBERA" |
           datos$VOTO_DISTRITO =="LA_ASUNCION"),]$MOROSO="N"
table(datos$MOROSO)
```

Fuente: Elaboración propia.

La siguiente tabla muestra la cantidad de datos disponibles para efectos de este proyecto:

Tabla 3. Conteo de datos por tablas y vistas

Tabla	Cantidad Registros	Descripción
CUF_TARIFA	109	Tarifas establecidas por cada servicio brindado por la municipalidad.
COM_PERSON	23455	Datos de personas físicas y jurídicas.
PADRON	3440248	Datos de personas votantes en Costa Rica.
CUF_TARMED	184	Tarifas medidas.
CUF_SEROCU	67517	Datos de los servicios por ocupación.
CUF_PROPIE	15917	Datos de propiedades del cantón.

Tabla	Cantidad Registros	Descripción
CUF_PERCOH	1452	Datos históricos de los periodos de cobros.
CUF_PERCOB	209	Datos actuales de los periodos de cobros.
CUF_PATLIC	135	Datos de las patentes de licores.
CUF_PATENT	2364	Datos de las patentes comerciales.
CUF_HIDROM	8147	Hidrómetros municipales.
CUF_DECBIE	5603	Declaraciones de bienes inmuebles.
CUF_CEMENT	1344	Datos del servicio de cementerio.
CUF_AVALUO	5578	Datos de avalúos de las fincas y/o propiedades.
COM_INFPER	11877	Información detallada de personas.
CUF_PERCOS	236	Periodos de cobros.
DIST_ELECT	2133	Códigos de los lugares de votación.
AFILIADOS	4689	Datos de los contribuyentes afiliados.
TSE_NACIMIENTOS	5773362	Nacimientos según el Registro Civil.
TSE_MATRIMONIOS	1915010	Matrimonios según el Registro Civil.
TSE_DEFUNCIONES	1043701	Defunciones según el Registro Civil.

La siguiente tabla muestra las vistas utilizadas, cantidad de datos y la descripción:

Tabla 4. Descripción de las vistas y conteo de datos

Tabla	Cantidad Registros	Descripción
V_CANT_PROPIEDADES	10508	Retorna la cantidad de propiedades de una persona.
PROPIE_SIN_SENAS	632	Retorna las propiedades que no tienen indicada la dirección exacta.

Tabla	Cantidad Registros	Descripción
V_SERVICIO	47938	Retorna la lista de servicios.
V_CANTIDAD_CUENTAS	11009	Retorna la cantidad de cuentas por persona.
CUF_TTRAIN	655058	Lista de cuentas con montos pendientes.
V_AFILIADOS	4689	Lista de personas afiliadas.
COM_PERDET	30884	
CUF_INMALF	23307	
V_MOROSOS	3732	Vista que retorna la lista de contribuyentes que tienen montos pendientes de pago.
V_CUF_CERTIF	149278	Vista que retorna la lista de contribuyentes con montos pendientes.
CUF_PREPTN	826	Lista de cuentas por prescribir.
CUF_PREPTO	1477	Lista de cuentas por prescribir.
CUF_CUENTA	27907	Lista de cuentas municipales.
CUF_CTAAUX	5	Cuentas auxiliares.
CUF_CERTIF	169168	Vista que retorna la lista de contribuyentes con montos pendientes.
V_CUENTA_SERVICIO	67517	Retorna la cuenta y el servicio que tiene asignado.
V_SERVICIOS_X_CUENTA	10484	Retorna los servicios asignados a una cuenta.
V_TARIFA	27681	Retorna la tarifa de una cuenta.

Tabla	Cantidad Registros	Descripción
V_PERSONA_X_TARIFA	10484	Retorna la tarifa asignada a una persona.
V_NACIMIENTO_TSE	13758	Retorna los nacimientos.
V_VALORES_PROPIES	9687	Retorna el valor fiscal de las propiedades.
MATRIMONIO_X_CED	5295	Retorna los matrimonios de una persona.
MATRIMONIOS_HOMBRES	5301	Retorna los matrimonios de los hombres.

MATRIMONIO_X_CED_MUJER	5123	Retorna los matrimonios dada una cédula de una mujer.
MATRIMONIOS_MUJERES	5128	Retorna los matrimonios de las mujeres.
V_HIJOS_MUJERES	4247	Retorna los hijos de las mujeres.
V_HIJOS_HOMBRES	4209	Retorna los hijos de los hombres.
V_HIJOS	8323	Lista de hijos.
V_MATRIMONIOS	10426	Lista de matrimonios.
V_PATENTE_COMER	1084	Lista de patentes comerciales.
V_PATENTE_LIC	73	Lista de patentes de licores.
V_CANT_PAT_COMER	1032	Cantidad de patentes comerciales de una persona.
V_CANT_PAT_LIC	66	Cantidad de patentes de licores de una persona.
CUF_CONSFI	8500	Retorna las construcciones que tiene una finca.

La siguiente tabla muestra la lista de columnas y la descripción de estas:

Tabla 5. Descripción de las columnas o variables

Variable	Tipo	Descripción
TIP_PERSON	Categórica	Indica el tipo de persona, si es física o jurídica.
COD_PROVIN	Categórica	El código de la provincia donde vive el contribuyente
COD_CANTON	Categórica	El código del cantón donde vive el contribuyente
COD_DISTRI	Categórica	El código del distrito donde vive el contribuyente
VOTO_DISTRITO	Carácter	El distrito donde vota la persona contribuyente.
V_PROVINCIA	Categórica	Provincia donde vota la persona.
V_CANTON	Categórica	Cantón donde vota la persona.
COD_TARIFA	Categórica	Código de la tarifa asignada al servicio.
COD_SERVIC	Categórica	Código del servicio que paga la persona.
MONTO_FINCA	Numérica	Monto/valor de la finca.
MONTO_IMPONIBLE	Numérica	Valor imponible de la finca.
EDAD	Numérica	Edad del contribuyente.
SEXO	Categórica	Sexo del contribuyente.
N_HIJOS	Numérica	Números de hijos según el Registro Civil.
ESTADO_CIVIL	Categórica	Estado civil del contribuyente.
TIPO_RELACION	Categórica	Tipo de relación que mantiene la persona.
CANT_CUENTAS	Numérica	Cantidad de cuentas registras en la municipalidad.

Variable	Tipo	Descripción
N_PROPIEDADES	Numérica	Números de propiedades que tiene la persona en el cantón de Belén.
N_PAT_COMER	Numérica	Cantidad de patentes comerciales que registra la persona.
N_PAT_LIC	Numérica	Cantidad de patentes de licores que tiene la persona en la municipalidad.
IND_AFILIADO	Categórica	Indica si la persona está afiliada o no para recibir estados de cuentas e información de contacto.
PROP_SENAS	Categórica	Indica si la propiedad tiene señas o dirección exacta.
CONSTRUC_FINCA	Categórica	Indica si tiene construcciones en las propiedades.
MOROSO	Categórica	Variable a predecir, indica si la contribuyente está o no moroso ante la municipalidad.

A continuación, se describen los auxiliares contables utilizados en la Municipalidad:

Tabla 6. Auxiliares contables

Auxiliar	Descripción
CUF	Auxiliar contable utilizado para servicios generales como recolección de residuos sólidos, cruz roja, servicio de agua, impuestos de bienes inmuebles.
CEM	Auxiliar contable utilizado para el servicio de cementerio.
PAT	Auxiliar contable utilizado para patentes comerciales.
LIC	Auxiliar contable utilizado para patentes de licores.

A continuación, se describen los servicios principales que brinda la Municipalidad, la lista complete se puede encontrar en el anexo 1:

Tabla 7. Servicios brindados por la Municipalidad

AUX_CONTAB	COD_SERVIC	DES_SERVIC
PAT	COM	PATENTE COMERCIAL
PAT	LIC	PATENTES LICORES
CUF	MIC	MULTA IMPUESTO CONSTRUCCION
CUF	AGF	SERV. AGUA FIJO
CUF	GES	SERVICIOS AMBIENTALES
CUF	IBI	IMPUESTO DE BIENES INMUEBLES
CEM	DER	DERECHO CEMENTERIO
CUF	LVP	LIMPIEZA DE VIAS Y SITIOS PUBLICOS
CUF	MPO	MANT. PARQ. Y OBRAS DE ORNATO
CUF	CCR	CONTRIBUCION CRUZ ROJA
CUF	RBR	RECOLECCION DE BASURA RECICLABLE
CUF	AGU	SERV. AGUA POTABLE
CEM	CEM	DERECHO CEMENTERIO
CEM	MAN	MANTENIMIENTO CEMENTERIO
CUF	BAS	RECOLECCION RESIDUOS SOL. Y VALORIZABLES
CUF	ALC	ALCANTARILLADO SANITARIO Y PTAR
CUF	IVA	IMPUESTO VALOR AGREGADO

Categoría de servicios: La municipalidad brinda los servicios con las siguientes categorías:

Tabla 8. Categorías de los servicios brindados por la Municipalidad

Código	Descripción
REP	Servicio brindado en negocios.
ORD	El servicio se brinda en locales comerciales o condominios.
SOC	Servicios brindados en lugares con fines sociales.
DOM	Servicios residenciales.
IND	El servicio se brinda en zonas industriales.
PRE	El servicio es brindado como preferencial: Escuelas, Ministerios Públicos.

Fuente: Elaboración propia.

2.3 Preparación de los datos

Se ha configurado y preparado la base de datos de datos Oracle, los usuarios principales que serán utilizados en este proyecto son los siguientes:

Tabla 9. Usuarios y esquemas de base de datos

Usuario	Descripción
system	Usuario de administración de base de datos
sys	Usuario de administración de base de datos
mineria	Usuario utilizado para la carga de datos del Tribunal Supremo de Elecciones, afiliados, y Registro Civil.
dec	Usuario y esquema principal de la base de datos municipal.

Fuente: Elaboración propia.

Se realiza la restauración de la base de datos:

2.3.1 Restauración respaldo de base de datos municipal

Inicialmente los datos fueron brindados en un archivo .dmp archivo de respaldo de base de datos.En la siguiente figura se puede observar que el esquema de base de datos que se debe restaurar es el esquema DEC.

Figura 19. **Importar los datos con data pump**

```
Connected to: Oracle Database 11g Release 11.2.0.4.0 - 64bit Production
Master table "SYS"."SYS_IMPORT_SCHEMA_02" successfully loaded/unloaded
Starting "SYS"."SYS_IMPORT_SCHEMA_02":  "/******** AS SYSDBA" directory=BACKUPS DUMPFILE=ORCL_28-01-21.dmp LOGFILE=IMP-ORCL-DEC-06-02-21.log
 schemas=DEC remap_tablespace=TAB_01:SUB,TAB_02:SUB,TAB_03:SUB,TAB_04:SUB,TAB_05:SUB
Processing object type DATABASE_EXPORT/SCHEMA/USER
Processing object type DATABASE_EXPORT/SCHEMA/GRANT/SYSTEM_GRANT
```

Fuente. Elaboración propia.

En la siguiente figura, se muestran los objetos de base de datos que fueron recuperados:

Figura 20. **Objetos de base de datos**

1	PROCEDURE	11261
2	VIEW	2499
3	TABLE	1611
4	PACKAGE	2
5	TYPE	1
6	PACKAGE BODY	1

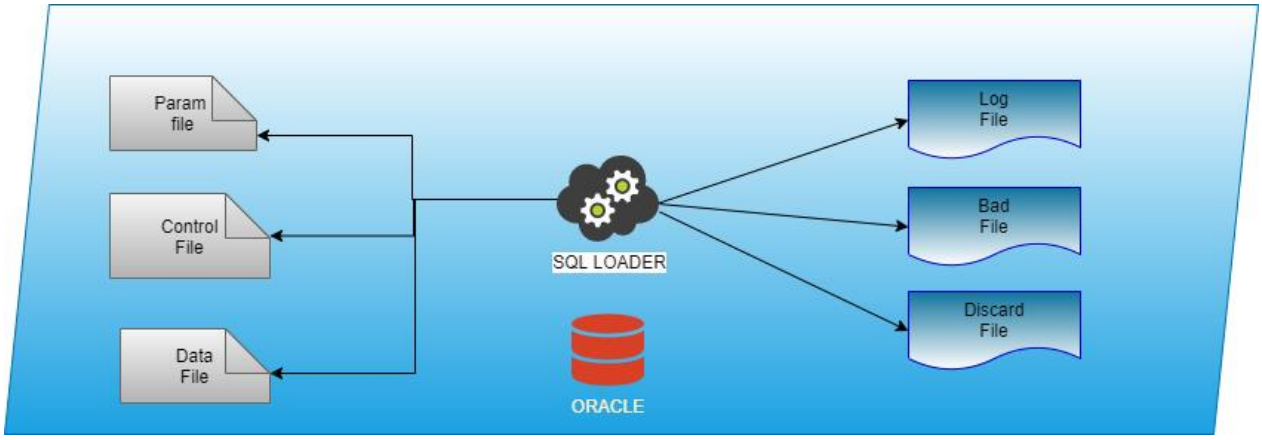
Fuente. Elaboración propia.

2.3.2 Restauración de datos del Registro Civil

Se realiza la restauración de datos de matrimonios, defunciones y nacimientos.

En este caso se utiliza la herramienta SQL Loader:

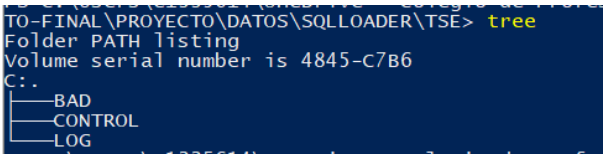
Figura 21. **Funcionamiento general de SQL*Loader**



Fuente. Elaboración propia.

En la siguiente figura, se muestra la estructura general de archivos para la carga de datos con SQL*Loader, se pueden identificar 3 carpetas:

Figura 22. **Estructura de archivos para carga de datos SQL*Loader**



Fuente. Elaboración propia.

- BAD: Este directorio contiene los reportes de los datos que no son cargados correctamente.
- Control: Contiene los archivos de control con las sentencias para realizar la carga de datos.

Figura 23. **Archivo de control para la carga de la tabla de datos de nacimientos**

```
# control
OPTIONS ( DIRECT=TRUE)
load data
infile '/datos/SQLLOADER/TSE/DATA/MAESTRO_NACIMIENTOS_OCTUBRE2019.dat'
into table MINERIA.TSE_NACIMIENTOS_TMP
TRUNCATE
trailing nullcols
(
VALUE position(1:257)
)
```

Fuente. Elaboración propia.

- LOG: Contiene los archivos de log general de la carga.

Figura 24. **Log de la carga de datos**

```
# LOGs

Tabla "MINERIA"."TSE_NACIMIENTOS":
  1043701 Filas cargadas correctamente.
  0 Filas no cargadas debido a errores de datos.
  0 Filas no cargadas porque todas las cláusulas WHEN han fallado.
  0 Filas no cargadas porque todos los campos eran nulos.


  Table MINERIA.TSE_DEFUNCIONES_TMP:
  1043701 Rows successfully loaded.
  0 Rows not loaded due to data errors.
  0 Rows not loaded because all WHEN clauses were failed.
  0 Rows not loaded because all fields were null.



Table MINERIA.TSE_MATRIMONIOS_TMP:
  1915010 Rows successfully loaded.
  0 Rows not loaded due to data errors.
  0 Rows not loaded because all WHEN clauses were failed.
  0 Rows not loaded because all fields were null.
```

Fuente. Elaboración propia.

Como se puede observar para estas tablas se han recuperado la siguiente cantidad de registros:

Nacimientos: 5,773,362

Matrimonios: 1,915,010

Defunciones: 1,043,701

- Script: Se crea una rutina en Linux para ejecutar la carga de todas las tablas:

Figura 25. **Script de sistema operativo para realizar la carga de datos**

```
#sh script
export tnsname=unalic</br>
export username=mineria</br>
export password=gbsystem01</br>
export NLS_LANGUAGE=SPANISH</br>

sqlldr $username/$password@$tnsname control='CONTROL/CARGA_NAC.ctl' log='LOG/CARGA_NAC.log'</br> bad='BAD/CARG
A_NAC.bad'</br>
sqlldr $username/$password@$tnsname control='CONTROL/CARGA_DEF.ctl' log='LOG/CARGA_DEF.log'</br> bad='BAD/CARG
A_DEF.bad'</br>
sqlldr $username/$password@$tnsname control='CONTROL/CARGA_MAT.ctl' log='LOG/CARGA_MAT.log'</br> bad='BAD/CARG
A_MAT.bad'</br>
```

Fuente. Elaboración propia.

2.3.3 Restauración de datos del padrón electoral

De igual forma para este set de datos la carga se realiza con la herramienta SQL*Loader: Del set de datos del padrón electoral se recupera una totalidad de: 3,440,248 registros.

Figura 26. **Carga de datos del padrón electoral**

```
OPTIONS ( DIRECT=TRUE)
load data
infile '/datos/SQLLOADER/PADRON/DATA/PADRON_COMPLETO.txt'
into table MINERIA.PADRON
TRUNCATE
fields terminated by ',' optionally enclosed by '"'
trailing nullcols
(
CEDULA,
CODELEC,
SEXO,
FECHACADUC,
JUNTA,
NOMBRE,
APELLIDO_1,
APELLIDO_2
)

# LOG

Table MINERIA.PADRON:
  3440248 Rows successfully loaded.
  0 Rows not loaded due to data errors.
  0 Rows not loaded because all WHEN clauses were failed.
  0 Rows not loaded because all fields were null.
```

Fuente. Elaboración propia.

Restauración de datos de afiliados: Estos datos inicialmente están en un gestor de base de datos MySQL, para poder acceder a estos se utiliza una sentencia SQL la cual retorna un set de datos el cual es exportado como csv para su carga en Oracle:

Figura 27. **Reporte de datos de afiliados municipales**

104800475	ebrown_1@hotmail.(	83842686	A
3101570085	cuentasporpagar@clu	22933211	A
401021263	laboratorio1al@hotm	88867207	A
1388248	info@villasdecariair.(	22391341	A
29	jesus_garcia@hotmail.com		A
00000000520897A	paolaceccondin@out	70128823	A
17140611	dianahaskour@hotm	87231193	A
48446903	info@romany asoc.cc	22393874	A
0000000AO241900	hosterialasvegas@hotmail.com		A
0000000AS850680	hosterialasvegas@hotmail.com		A
0000000C1057361	sdebreuning@gmail	22907787	A

Fuente. Elaboración propia.

Una vez obtenido el reporte, se procede con la carga en Oracle la cual se realiza con SQL*Loader, en la Figura 26, se puede observar que estos datos son recuperados en la tabla “AFILIADOS” un total de 4,689 registros:

Figura 28. **Carga de datos de afiliados**

```
OPTIONS ( DIRECT=TRUE)
load data
infile '/datos/SQLLOADER/PADRON/DATA/MB_AFILIADOS_2019-11-05T13_31_55.csv'
into table MINERIA.AFILIADOS
TRUNCATE
fields terminated by ',' optionally enclosed by '"'
trailing nullcols
(
IDENTIFICACION,
CORREO,
TELEFONO,
ESTADO
)
```

```
Table MINERIA.AFILIADOS:
  4689 Rows successfully loaded.
  0 Rows not loaded due to data errors.
  0 Rows not loaded because all WHEN clauses were failed.
  0 Rows not loaded because all fields were null.
```

Fuente. Elaboración propia.

2.3.4 Limpieza de datos

Los datos de matrimonios, nacimientos y defunciones fueron cargados en una tabla de una sola columna, según las siguientes indicaciones:

Datos de defunciones:

Tabla 10. Descripción de columnas

Descripción	Tamaño	Tipo	Observaciones
Cita de Defunción	12	Alfanumérico	Formato PTTTTAAAA P = Provincia T = Tomo A = Asiento
Cita de Nacimiento (# de cédula)	20	Alfanumérico	Formato PTTTTAAAA P = Provincia T = Tomo A = Asiento ó número de pasaporte
Conocido como Nombre	20	Alfanumérico	
Nombre	50	Alfanumérico	
Conocido como	13	Alfanumérico	

Descripción	Tamaño	Tipo	Observaciones
Primer Apellido			
Primer Apellido	26	Alfanumérico	
Conocido como Segundo Apellido	13	Alfanumérico	
Segundo Apellido	26	Alfanumérico	
Sitio de la defunción	29	Alfanumérico	
Fecha de la Defunción	8	Numérico	formato aaaammdd
Hora de la Defunción	4	Numérico	
Indicador de Extranjero	1	Numérico	0 = Nacional 1 = Extranjero
Indicador de Nacimiento	1	Numérico	0 = Tiene Nacimiento 1 = No tiene Nacimiento
Sexo	1	Numérico	1 = Hombre 2 = Mujer
Relleno	26	Alfabético	

Datos de matrimonios:

Tabla 11. Descripción de los datos de matrimonios

CAMPOS	LARGO	TIPO	OBSERVACIONES
Cita de Matrimonio	13	Numérico	Formato PTTTTFFFAAAAR P = Provincia T = Tomo F = Folio A = Asiento R = Tipo de Relación
Cédula del Hombre	20	Alfanumérico	Formato PTTTTAAAA P = Provincia A = Asiento T = Tomo y número de pasaporte
Nombre del Hombre	50	Alfanumérico	
Primer Apellido del Hombre	26	Alfanumérico	
Segundo Apellido del Hombre	26	Alfanumérico	
Conocido Como (Hombre)	29	Alfanumérico	
Padre del Hombre	50	Alfanumérico	
Nacionalidad del Padre del Hombre	3	Numérico	
Madre del Hombre	50	Alfanumérico	

CAMPOS	LARGO	TIPO	OBSERVACIONES
Nacionalidad de la Madre de la Mujer	3	Numérico	
Indicador de extranjero del Hombre	1	Numérico	0 = Nacional 1 = Extranjero
Nacionalidad del Hombre	3	Numérico	
Estado Civil del Hombre	1	Numérico	1 = Soltero 2 = Casado 3 = Separado 4 = Divorciado 5 = Viudo 6 = Célibe 7 = Reconciliación Judicial 8 = Anulado
Edad del Hombre	2	Numérico	
Indicador de Defunción del Hombre	1	Numérico	0 = No tiene Defunción 1 = Tiene defunción
Cédula de la Mujer	20	Alfanumérico	
Nombre de la Mujer	50	Alfanumérico	
Primer Apellido de la Mujer	26	Alfanumérico	

CAMPOS	LARGO	TIPO	OBSERVACIONES
Segundo Apellido de la Mujer	26	Alfanumérico	
Conocido como (Mujer)	29	Alfanumérico	
Padre de la Mujer	50	Alfanumérico	
Nacionalidad del Padre de la Mujer	3	Numérico	
Madre de la Mujer	50	Alfanumérico	
Nacionalidad de la Madre de la Mujer	3	Numérico	
Indicador de Extranjero de la mujer	1	Numérico	0 = Nacional 1 = Extranjero
Nacionalidad de la Mujer	3	Numérico	
Estado Civil de la Mujer	1	Numérico	1 = Soltero 2 = Casado 3 = Separado 4 = Divorciado 5 = Viudo 6 = Célibe 7 = Reconciliación Judicial 8 = Anulado
Edad de la Mujer	2	Numérico	

CAMPOS	LARGO	TIPO	OBSERVACIONES
Indicador de Defunción de la Mujer	1	Numérico	0 = No tiene Defunción 1 = Tiene defunción
Lugar del Suceso	29	Alfanumérico	
Fecha de Entrada	8	Numérico	FORMATO AAAAMMDD
Fecha de Suceso	8	Numérico	FORMATO AAAAMMDD
Provincia	1	Numérico	
Código de Cantón	2	Numérico	
Distrito Administrativo	2	Numérico	
Tipo de Matrimonio	1	Numérico	1 = Católico 2 = Civil 3 = Pastor (Juez de Paz)
Tipo de Relación	1	Numérico	2 = Matrimonio 3 = Separación Judicial 4 = Divorcio 5 = Viudez 7 = Reconciliación Judicial 8 = Anulación de Matrimonio

Datos de nacimientos:

Tabla 12. Descripción de los datos de nacimientos

DESCRIPCIÓN	LARGO	TIPO	OBSERVACIONES
Número de Cédula	9	Numérico	Formato PTTTTAAAA P = Provincia T = Tomo A = Asiento
Número de Folio	3	Numérico	
Número de Cédula del Padre	9	Numérico	
Número de Cédula de la Madre	9	Numérico	
Código del Hospital	3	Numérico	
Hora del Suceso	4	Numérico	
Fecha del Suceso	8	Numérico	Formato AAAAMMDD
Sexo	1	Numérico	1= Hombre 2 = Mujer 3 = Indefinido
Nacionalidad del Inscrito	1	Numérico	0 = Costarricense 1 = Por Opción 2 = Naturalizado 3 = No indica nacionalidad 4 = Conserva nacionalidad extranjera 5=Conserva

DESCRIPCIÓN	LARGO	TIPO	OBSERVACIONES
			nacionalidad costarricense
Indicador de Defunción	1	Numérico	0 = No tiene Defunción 1 = Tiene defunción
País de Procedencia del Padre	3	Numérico	
País de Procedencia de la Madre	3	Numérico	
Campo de Relleno	2	Numérico	
Provincia y Cantón de Procedencia de la Madre	3	Numérico	
Fecha de Naturalización	8	Numérico	Formato AAAAMMDD
Campo de Relleno	1	Alfanumérico	
Primer Apellido	26	Alfanumérico	
Segundo Apellido	26	Alfanumérico	
Nombre	50	Alfanumérico	
Nombre del Padre	29	Alfanumérico	
Nombre de la Madre	29	Alfanumérico	
Lugar de Nacimiento	29	Alfanumérico	

De esta manera, se definen tres sentencias para realizar limpiar estos datos y darle el formato esperado.

En la figura siguiente se puede observar la sentencia para darle el formato a la tabla de matrimonios:

Figura 29. **Sentencia SQL para brindar el formato a los datos**

```
INSERT INTO mineria.TSE_MATRIMONIOS
select
TRIM(TO_CHAR(substr(value,1,13))) cita_matrimonio,
TRIM(TO_CHAR(substr(value,14,20))) cedula_hombre,
TRIM(TO_CHAR(substr(value,34,50))) nombre_hombre,
TRIM(TO_CHAR(substr(value,84,26))) app1_hombre,
TRIM(TO_CHAR(substr(value,110,26))) app2_hombre,
TRIM(TO_CHAR(substr(value,136,29))) conocido_cm_hombre,
TRIM(TO_CHAR(substr(value,165,50))) padre_hombre,
TRIM(TO_CHAR(substr(value,215,3))) nac_padre_hombre,
TRIM(TO_CHAR(substr(value,218,50))) madre_hombre,
TRIM(TO_CHAR(substr(value,268,3))) nac_madre_mujer,
TRIM(TO_CHAR(substr(value,271,1))) ind_extran_hombre,
TRIM(TO_CHAR(substr(value,272,3))) nacionalidad_hombre,
TRIM(TO_CHAR(substr(value,275,1))) estado_civil_hombre,
TRIM(TO_CHAR(substr(value,276,2))) edad_hombre,
TRIM(TO_CHAR(substr(value,278,1))) ind_def_hombre,
TRIM(TO_CHAR(substr(value,279,20))) cedula_mujer,
TRIM(TO_CHAR(substr(value,299,50))) nombre_mujer,
TRIM(TO_CHAR(substr(value,349,26))) app1_mujer,
TRIM(TO_CHAR(substr(value,375,26))) app2_mujer,
TRIM(TO_CHAR(substr(value,401,29))) conoc_cm_mujer,
TRIM(TO_CHAR(substr(value,430,50))) padre_mujer,
TRIM(TO_CHAR(substr(value,480,3))) nacion_padre_mujer,
TRIM(TO_CHAR(substr(value,483,50))) madre_mujer,
TRIM(TO_CHAR(substr(value,533,3))) nacion_madre_mujer,
TRIM(TO_CHAR(substr(value,536,1))) ind_extran_mujer,
TRIM(TO_CHAR(substr(value,537,3))) nacionalidad_mujer,
TRIM(TO_CHAR(substr(value,540,1))) estado_civil_mujer,
TRIM(TO_CHAR(substr(value,541,2))) edad_mujer,
TRIM(TO_CHAR(substr(value,543,1))) ind_def_mujer,
TRIM(TO_CHAR(substr(value,544,29))) lugar_suceso,
TRIM(TO_CHAR(substr(value,573,8))) fecha_entrada,
TRIM(TO_CHAR(substr(value,581,8))) fecha_suceso,
TRIM(TO_CHAR(substr(value,589,1))) provincia,
TRIM(TO_CHAR(substr(value,590,2))) canton,
TRIM(TO_CHAR(substr(value,592,2))) distrito,
TRIM(TO_CHAR(substr(value,594,1))) tipo_matrimonio,
TRIM(TO_CHAR(substr(value,595,1))) tipo_relacion
from MINERIA.TSE_MATRIMONIOS_TMP;
```

Fuente: Elaboración propia.

Además, se realiza la corrección de los nombres de los distritos:

Figura 30. **Expresión regular para limpiar datos con fin de línea**

```
# Se limpian end-lines

UPDATE MINERIA.DIST_ELECT
   SET distrito = REGEXP_REPLACE (distrito, '(^[[:space:]]*|[[:space:]]*$)');
```

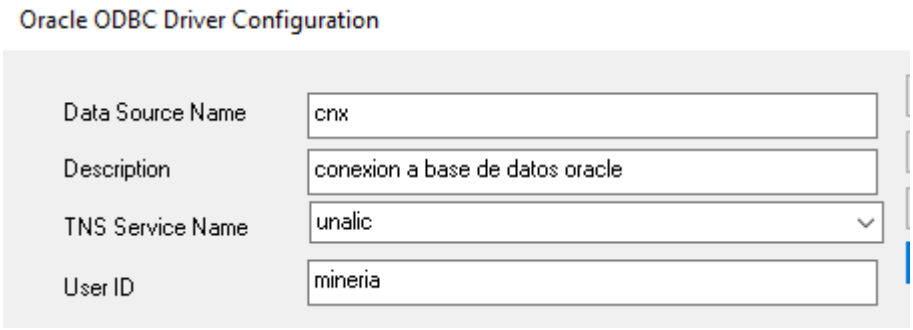
2.3.5 Configuración y conexión a la base de datos

Una vez se han recuperado los datos, se procede a realizar la conexión con la base de datos desde R, para ello se configura una conexión ODBC:

TNS names: unalic = (DESCRIPTION = (ADDRESS_LIST = (ADDRESS = (PROTOCOL = TCP)(HOST = 192.168.100.127)(PORT = 1521))) (CONNECT_DATA = (SERVICE_NAME = unalic)))

Configuración del ODBC:

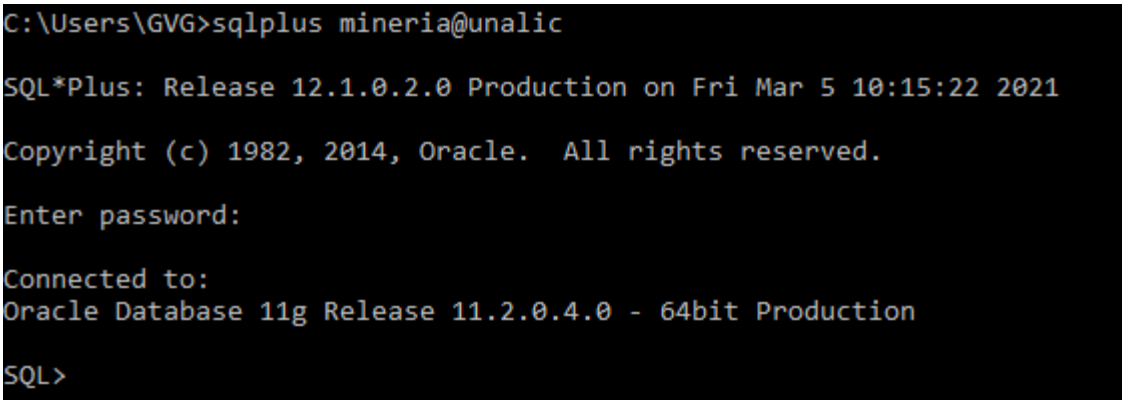
Figura 31. **Configuración del ODBC**



Fuente: Elaboración propia.

Prueba de conexión con SQL*Plus: Para verificar que la base de datos está disponible se realiza una prueba de conexión desde SQL*Plus:

Figura 32. **Conexión a la base de datos desde SQLPlus**



Fuente: Elaboración propia.

2.3.6 Conexión desde R

Para efectos de conectarse a la base de datos de RStudio, se ha creado un ODBC, además del usuario: minería

Figura 33. **Detalles de conexión a la base de datos desde R**

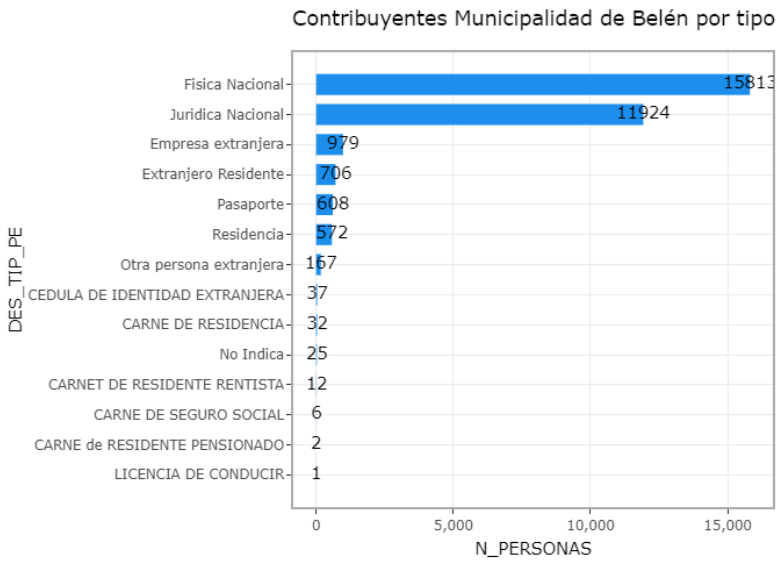
```
# Se realiza el acceso a la base de datos

con <- odbcConnect("cnx", uid = "mineria", pwd="gbsystem01")
```

Fuente: Elaboración propia.

2.3.7 Selección de datos

Para efectos del presente proyecto se toman en cuenta las personas físicas, en las cuales los registros de cédula indicados en la municipalidad y los datos del Tribunal Supremo de Elecciones coinciden, en totalidad la municipalidad tiene 30884 registros de personas, cabe indicar que estas personas pueden ser físicas, jurídicas, nacionales y extranjeros, la distribución se puede observar en el siguiente gráfico:

Gráfico 9. **Contribuyentes Municipalidad por tipo**



Fuente: Elaboración propia.

Se obtiene que un total de 13,590 registros coinciden entre el padrón electoral y la base de datos de personas de la Municipalidad:

Figura 34. **Conteo de personas.**

```
# conteo de personas db municipal y padrón
registros <-sqlQuery(con,"SELECT COUNT(*) FROM DEC.COM_PERDET PER INNER JOIN MINERIA.PADRON PAD ON PER.CEDULA=
PAD.CEDULA;")
registros
```

	COUNT(*) <int>
1	13590

1 row

Fuente: Elaboración propia.

2.3.8 Generar reporte en formato CSV

Se crea una consulta SQL con el fin de generar los datos desde la base de datos, posteriormente estos datos serán utilizados desde R:

Para ver la consulta completa, ver anexos.

Figura 35. **Consulta SQL que obtiene información de los contribuyentes**

```
SELECT INF.CEDULA,
       INF.TIP_PERSON,
       INF.COD_PROVIN,
       INF.COD_CANTON,
       INF.COD_DISTRI,
       CASE NVL (PAD.CODELEC, 'NV')
          WHEN 'NV' THEN 'NV'
          WHEN '407001' THEN 'SAN ANTONIO'
          WHEN '407002' THEN 'LA RIBERA'
          WHEN '407003' THEN 'LA ASUNCION'
          WHEN '408003' THEN 'LLORENTE'
          WHEN '408001' THEN 'SAN JOAQUIN'
          ELSE 'OTRO'
       END
          AS VOTO_DISTRITO,
       CASE NVL (PAD.PROVINCIA, 'NV')
          WHEN 'NV' THEN 'NV'
          WHEN 'HEREDIA' THEN PAD.PROVINCIA
          WHEN 'ALAJUELA' THEN PAD.PROVINCIA
          WHEN 'SAN JOSE' THEN PAD.PROVINCIA
          WHEN 'PUNTARENAS' THEN PAD.PROVINCIA
          ELSE 'OTRO'
       END
          AS V_PROVINCIA,
       CASE NVL (PAD.CANTON, 'NV')
          WHEN 'NV' THEN 'NV'
          WHEN 'BELEN' THEN PAD.CANTON
          WHEN 'CENTRAL' THEN PAD.CANTON
          WHEN 'FLORES' THEN PAD.CANTON
          WHEN 'SANTA ANA' THEN PAD.CANTON
          WHEN 'ESTADOS UNIDOS' THEN PAD.CANTON
          ELSE 'OTRO'
```

Fuente: Elaboración propia.

2.3.9 Individuos con registros duplicados

Se observa que hay personas que tienen duplicados los registros en el maestro de matrimonios.

A continuación, se pueden observar los individuos:

Figura 36. **Registros duplicados en el patrón de matrimonios**

```
duplicados<-(datos[duplicated(datos$CEDULA, .keep_all= TRUE),])
duplicados
```

	CEDULA <int>	TIP_PERSON <fct>	COD_PRO... <int>	COD_CAN... <fct>	COD_DISTRI <fct>	VOTO_DISTRITO <fct>	V_PROVINCIA <fct>
324	105270395	F	4	07	1	SAN ANTONIO	HEREDIA
3666	400930586	F	4	07	2	LA RIBERA	HEREDIA
3959	401050438	F	4	07	1	SAN ANTONIO	HEREDIA
4261	401160329	F	4	07	1	SAN ANTONIO	HEREDIA
5912	600790316	F	4	07	2	LLORENTE	HEREDIA

5 rows | 1-8 of 46 columns

Fuente: Elaboración propia.

Validación mediante consulta SQL en la base de datos se obtiene que los registros en el patrón de matrimonios se encuentran duplicados:

Figura 37. **Registros duplicados en el patrón de matrimonios**

```
con <- odbcConnect("cnx", uid = "mineria", pwd="gbsystem01")
matrimnios_spc <-sqlQuery(con,"SELECT *
  FROM MINERIA.TSE_MATRIMONIOS
 WHERE CEDULA_HOMBRE IN ('400930586',
                  '105270395',
                  '401050438',
                  '600790316',
                  '401160329')
                 or CEDULA_MUJER IN ('400930586',
                  '105270395',
                  '401050438',
                  '600790316',
                  '401160329');")
matrimnios_spc
```

	CITA_MATRIMONIO <dbl>	CEDULA_HOM... <int>	NOMBRE_HOMBRE <chr>	APP1_HOM... <chr>	APP2_HOM... <chr>
1	2.008846e+12	600790316	FERMIN DANILO DEL CARMEN	JIMENEZ	BADILLA
2	2.008846e+12	600790316	DANILO	JIMENEZ	BADILLA
3	2.008846e+12	600790316	FERMIN DANILO DEL CARMEN	JIMENEZ	BADILLA
4	4.003737e+12	401050438	JULIO ALFREDO DE JESUS	CHAVES	MURILLO
5	4.003544e+12	400930586	ALEXIS GERARDO DE LAS PIEDADES	CHAVES	DELGADO
6	4.003618e+12	400930586	ALEXIS	CHAVES	DELGADO
7	4.003741e+12	401050438	JULIO	CHAVES	MURILLO
8	4.003922e+12	401160329	LUIS ALONSO	VENEGAS	PEREIRA
9	4.003923e+12	401160329	LUIS ALONSO DE JESUS	VENEGAS	PEREIRA

9 rows | 1-6 of 38 columns

Fuente: Elaboración propia.

Validación en el portal de consulta de civiles, se observa que estas personas tienen matrimonios:

Figura 38. **Portal de consulta de personas por cédula, Tribunal Supremo de Elecciones.**

Fuente: Consulta de civiles, TSE.

Se eliminan los registros que no están correctos:

Figura 39. **Eliminación de registros incorrectos**

```
datos<- (datos[-(datos$CEDULA=='400930586'& datos$ESTADO_CIVIL=='1'),])
datos<- (datos[-(datos$CEDULA=='600790316' & datos$ESTADO_CIVIL=='4'),])
datos<- (datos[-(datos$CEDULA=='105270395'& datos$ESTADO_CIVIL=='1'),])
datos<- (datos[-(datos$CEDULA=='401160329'& datos$ESTADO_CIVIL=='1'),])
datos<- (datos[-(datos$CEDULA=='401050438'& datos$ESTADO_CIVIL=='1'),])
dim(datos)
```

```
## [1] 6489   45
```

Fuente: Elaboración propia.

2.3.10 Lectura del reporte desde R

Figura 40. **Función para lectura de reporte de individuos en R**

```
datos <- read.table("../datos/reporte_2021.csv",stringsAsFactors = T,  header=TRUE, sep=';', dec =
','')
dim(datos)
```

Fuente: Elaboración propia.

Este conjunto de datos contiene 6489 individuos y 38 variables:

Figura 41. **Variables seleccionadas.**

```
## 'data.frame':    6489 obs. of  38 variables:
##  $ COD_PROVIN         : Factor w/ 4 levels "1","2","4","8": 3 3 3 3 3 3 3 3 3 3 ...
##  $ COD_CANTON         : Factor w/ 7 levels "BEL","EX","HE",..: 1 1 3 1 1 1 1 1 1 1 ...
##  $ COD_DISTRI         : Factor w/ 7 levels "EX","HE","LA",..: 3 7 2 7 7 3 7 7 4 7 ...
##  $ VOTO_DISTRITO      : Factor w/ 7 levels "LA_ASUNCION",..: 2 2 1 5 6 3 6 5 1 6 ...
##  $ V_PROVINCIA        : Factor w/ 6 levels "ALAJUELA","HEREDIA",..: 2 2 2 6 2 2 2 1 2 2 ...
##  $ V_CANTON           : Factor w/ 7 levels "BELEN","CENTRAL",..: 1 1 1 2 1 3 1 2 1 1 ...
##  $ COD_TARIFA_REP     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_TARIFA_RESID   : Factor w/ 2 levels "N","S": 2 2 1 2 2 2 2 2 2 2 ...
##  $ COD_TARIFA_DOM     : Factor w/ 2 levels "N","S": 2 2 2 2 2 2 2 2 2 1 ...
##  $ COD_TARIFA_SOCIAL  : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_TARIFA_COMER1  : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_TARIFA_SOC     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_TARIFA_ORD     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_SERVIC_MPO     : Factor w/ 2 levels "N","S": 2 2 1 2 2 2 2 2 2 2 ...
##  $ COD_SERVIC_LVP     : Factor w/ 2 levels "N","S": 2 2 1 2 2 2 2 2 2 2 ...
##  $ COD_SERVIC_AGU     : Factor w/ 2 levels "N","S": 2 2 2 2 2 2 2 2 2 1 ...
##  $ COD_SERVIC_IBI     : Factor w/ 2 levels "N","S": 2 2 2 2 2 2 2 2 2 2 ...
##  $ COD_SERVIC_BAS     : Factor w/ 2 levels "N","S": 2 2 1 2 2 2 1 2 2 1 ...
##  $ COD_SERVIC_PZV     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_SERVIC_PAT     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_SERVIC_LIC     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COD_SERVIC_CEM     : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 2 2 1 ...
##  $ SUM_MONTO_FINCA    : num  1.83e+08 1.25e+07 0.00 1.28e+08 7.32e+07 ...
##  $ SUM_MONTO_IMPONIBLE: num  91600000 12476250 0 64000000 18287662 ...
##  $ EDAD               : int  90 90 88 87 86 85 86 85 84 84 ...
##  $ SEXO               : Factor w/ 2 levels "F","M": 2 1 1 1 2 1 1 2 1 1 ...
##  $ N_HIJOS            : int  0 0 0 0 0 0 2 0 1 0 ...
##  $ ESTADO_CIVIL       : Factor w/ 6 levels "0","1","2","3",..: 2 6 2 1 3 1 6 6 3 6 ...
##  $ TIPO_RELACION      : Factor w/ 5 levels "0","2","3","4",..: 2 5 2 2 2 5 5 5 2 5 ...
##  $ CANT_CUENTAS       : int  2 2 2 2 2 3 4 3 5 1 ...
##  $ N_PROPIEDADES      : int  1 1 1 1 1 1 3 1 2 1 ...
##  $ N_PAT_COMER        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ N_PAT_LIC          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ IND_AFILIADO       : Factor w/ 2 levels "N","S": 2 1 2 2 1 2 1 2 1 1 ...
##  $ PROP_SENAS         : Factor w/ 2 levels "N","S": 2 2 2 2 2 2 1 2 1 2 ...
##  $ CONSTRUC_FINCA     : Factor w/ 2 levels "N","S": 2 2 1 2 2 2 2 2 2 1 ...
##  $ CONTRUCCIONES_FINCA: int  1 1 0 1 1 2 2 1 2 0 ...
##  $ MOROSO             : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 2 ...
```

Fuente: Elaboración propia.

2.3.11 Resumen de los datos

A continuación, el resumen de los datos a utilizar:

Figura 42. **Resumen de datos a utilizar**

```
##   COD_PROVIN_1       COD_PROVIN_2      COD_PROVIN_4     COD_PROVIN_8
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :1.0000   Median :0.00000
##  Mean   :0.005856   Mean   :0.02527   Mean   :0.9655   Mean   :0.00339
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##  COD_CANTON_BEL   COD_CANTON_EX      COD_CANTON_HE      COD_CANTON_NV
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:1.0000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :1.0000   Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.9755   Mean   :0.003699   Mean   :0.004623   Mean   :0.002003
##  3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##  COD_CANTON_OTRO    COD_CANTON_SANA    COD_CANTON_SJ      COD_DISTRI_EX
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.00000
##  Mean   :0.007551   Mean   :0.001233   Mean   :0.005394   Mean   :0.00339
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.00000
##  COD_DISTRI_HE     COD_DISTRI_LA    COD_DISTRI_LR    COD_DISTRI_NV
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.00262   Mean   :0.2523   Mean   :0.2047   Mean   :0.002003
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##  COD_DISTRI_OTRO   COD_DISTRI_SA    VOTO_DISTRITO_LA_ASUNCION
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.01479   Mean   :0.5203   Mean   :0.1573
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##  VOTO_DISTRITO_LA_RIBERA VOTO_DISTRITO_LLORENTE VOTO_DISTRITO_NV
##  Min.   :0.0000          Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:0.00000        1st Qu.:0.0000
##  Median :0.0000          Median :0.00000        Median :0.0000
##  Mean   :0.2589          Mean   :0.02142        Mean   :0.0131
##  3rd Qu.:1.0000          3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.0000          Max.   :1.00000        Max.   :1.0000
```

```
##  VOTO_DISTRITO_OTRO VOTO_DISTRITO_SAN_ANTONIO VOTO_DISTRITO_SAN_JOAQUIN
##  Min.   :0.0000     Min.   :0.0000            Min.   :0.000000
##  1st Qu.:0.0000     1st Qu.:0.0000            1st Qu.:0.000000
##  Median :0.0000     Median :0.0000            Median :0.000000
##  Mean   :0.1697     Mean   :0.3708            Mean   :0.008784
##  3rd Qu.:0.0000     3rd Qu.:1.0000            3rd Qu.:0.000000
##  Max.   :1.0000     Max.   :1.0000            Max.   :1.000000
##  V_PROVINCIA_ALAJUELA V_PROVINCIA_HEREDIA V_PROVINCIA_NV   V_PROVINCIA_OTRO
##  Min.   :0.00000      Min.   :0.0000      Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000      1st Qu.:1.0000      1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000      Median :1.0000      Median :0.0000   Median :0.00000
##  Mean   :0.05317      Mean   :0.8525      Mean   :0.0131   Mean   :0.01942
##  3rd Qu.:0.00000      3rd Qu.:1.0000      3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00000      Max.   :1.0000      Max.   :1.0000   Max.   :1.00000
##  V_PROVINCIA_PUNTARENAS V_PROVINCIA_SAN_JOSE V_CANTON_BELEN  V_CANTON_CENTRAL
##  Min.   :0.000000       Min.   :0.00000      Min.   :0.000   Min.   :0.00000
##  1st Qu.:0.000000       1st Qu.:0.00000      1st Qu.:1.000   1st Qu.:0.00000
##  Median :0.000000       Median :0.00000      Median :1.000   Median :0.00000
##  Mean   :0.007705       Mean   :0.05409      Mean   :0.787   Mean   :0.08091
##  3rd Qu.:0.000000       3rd Qu.:0.00000      3rd Qu.:1.000   3rd Qu.:0.00000
##  Max.   :1.000000       Max.   :1.00000      Max.   :1.000   Max.   :1.00000
##  V_CANTON_FLORES   V_CANTON_NV     V_CANTON_OTRO     V_CANTON_SANTA_ANA
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.000000
##  Mean   :0.0319   Mean   :0.0131   Mean   :0.07551   Mean   :0.007397
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000
##   V_CANTON_USA      COD_TARIFA_REP_N COD_TARIFA_REP_S   COD_TARIFA_RESID_N
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :0.000000   Median :1.0000   Median :0.000000   Median :0.0000
##  Mean   :0.004161   Mean   :0.9961   Mean   :0.003853   Mean   :0.1785
##  3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.000000   Max.   :1.0000
##  COD_TARIFA_RESID_S COD_TARIFA_DOM_N COD_TARIFA_DOM_S COD_TARIFA_SOCIAL_N
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :1.0000     Median :0.0000   Median :1.0000   Median :1.0000
##  Mean   :0.8215     Mean   :0.4127   Mean   :0.5873   Mean   :0.9932
##  3rd Qu.:1.0000     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  COD_TARIFA_SOCIAL_S COD_TARIFA_COMER1_N COD_TARIFA_COMER1_S COD_TARIFA_SOC_N
##  Min.   :0.000000    Min.   :0.0000      Min.   :0.00000     Min.   :0.0000
##  1st Qu.:0.000000    1st Qu.:1.0000      1st Qu.:0.00000     1st Qu.:1.0000
##  Median :0.000000    Median :1.0000      Median :0.00000     Median :1.0000
##  Mean   :0.006781    Mean   :0.9736      Mean   :0.02635     Mean   :0.9952
##  3rd Qu.:0.000000    3rd Qu.:1.0000      3rd Qu.:0.00000     3rd Qu.:1.0000
##  Max.   :1.000000    Max.   :1.0000      Max.   :1.00000     Max.   :1.0000
```

```
##  COD_TARIFA_SOC_S   COD_TARIFA_ORD_N COD_TARIFA_ORD_S  COD_SERVIC_MPO_N
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.000000   Median :1.0000   Median :0.00000   Median :0.0000
##  Mean   :0.004777   Mean   :0.9817   Mean   :0.01834   Mean   :0.2194
##  3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##  COD_SERVIC_MPO_S COD_SERVIC_LVP_N COD_SERVIC_LVP_S COD_SERVIC_AGU_N
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.7806   Mean   :0.2196   Mean   :0.7804   Mean   :0.3994
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  COD_SERVIC_AGU_S COD_SERVIC_IBI_N   COD_SERVIC_IBI_S COD_SERVIC_BAS_N
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.000000   Median :1.0000   Median :0.0000
##  Mean   :0.6006   Mean   :0.001541   Mean   :0.9985   Mean   :0.4267
##  3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.0000   Max.   :1.0000
##  COD_SERVIC_BAS_S COD_SERVIC_PZV_N COD_SERVIC_PZV_S  COD_SERVIC_PAT_N
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :1.0000   Median :1.0000   Median :0.00000   Median :1.0000
##  Mean   :0.5733   Mean   :0.9633   Mean   :0.03668   Mean   :0.9792
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##  COD_SERVIC_PAT_S COD_SERVIC_LIC_N COD_SERVIC_LIC_S   COD_SERVIC_CEM_N
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.000000   1st Qu.:1.00
##  Median :0.0000   Median :1.0000   Median :0.000000   Median :1.00
##  Mean   :0.0208   Mean   :0.9977   Mean   :0.002312   Mean   :0.89
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:1.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :1.00
##  COD_SERVIC_CEM_S     SEXO_F          SEXO_M      IND_AFILIADO_N
##  Min.   :0.00     Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.00     1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.00     Median :1.000   Median :0.000   Median :1.000
##  Mean   :0.11     Mean   :0.528   Mean   :0.472   Mean   :0.615
##  3rd Qu.:0.00     3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1.00     Max.   :1.000   Max.   :1.000   Max.   :1.000
##  IND_AFILIADO_S   PROP_SENAS_N      PROP_SENAS_S    CONSTRUC_FINCA_N
##  Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.385   Mean   :0.03668   Mean   :0.9633   Mean   :0.1436
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

```
##  CONSTRUC_FINCA_S SUM_MONTO_FINCA     SUM_MONTO_IMPONIBLE      EDAD
##  Min.   :0.0000   Min.   :0.000e+00   Min.   :        0   Min.   :  3.0
##  1st Qu.:1.0000   1st Qu.:3.010e+07   1st Qu.:  2520503   1st Qu.: 44.0
##  Median :1.0000   Median :6.000e+07   Median : 16443288   Median : 56.0
##  Mean   :0.8564   Mean   :1.060e+08   Mean   : 31801970   Mean   : 54.8
##  3rd Qu.:1.0000   3rd Qu.:1.220e+08   3rd Qu.: 40437599   3rd Qu.: 65.0
##  Max.   :1.0000   Max.   :4.457e+09   Max.   :682186490   Max.   :133.0
##     N_HIJOS       ESTADO_CIVIL TIPO_RELACION  CANT_CUENTAS    N_PROPIEDADES
##  Min.   :0.0000   0:1705       0:1623        Min.   : 1.000   Min.   : 1.000
##  1st Qu.:0.0000   1:3352       2:3781        1st Qu.: 1.000   1st Qu.: 1.000
##  Median :1.0000   2: 424       3:   3        Median : 2.000   Median : 1.000
##  Mean   :0.9886   3:   2       4: 649        Mean   : 2.162   Mean   : 1.371
##  3rd Qu.:2.0000   4: 647       5: 433        3rd Qu.: 3.000   3rd Qu.: 1.000
##  Max.   :8.0000   5: 359                     Max.   :20.000   Max.   :15.000
##   N_PAT_COMER        N_PAT_LIC        CONTRUCCIONES_FINCA MOROSO
##  Min.   :0.00000   Min.   :0.000000   Min.   : 0.000      N:4797
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 1.000      S:1692
##  Median :0.00000   Median :0.000000   Median : 1.000
##  Mean   :0.02219   Mean   :0.002466   Mean   : 1.832
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.: 2.000
##  Max.   :3.00000   Max.   :2.000000   Max.   :44.000
```

Fuente: Elaboración propia.

Del resumen de los datos, se obtiene lo siguiente:

- La edad mínima es de un individuo de 3 años, la media está en 56 años y una edad máxima de 133 años.
- Se identifican 3429 personas femeninas y 3065 masculinas.
- Se observa que la variable TIPO_PERSONA solo tiene 3 individuos de tipo X, esto corresponde a personas físicas que al momento de registro no se validó la correcta asignación y el sistema le asigno el valor por defecto, por lo cual se quita esta variable para futuros análisis.

Figura 43. **Individuos categorizados con el tipo "X"**

```
table(datos$TIP_PERSON)
```

```
##
##    E    F    X
##    1 6490    3
```

Fuente: Elaboración propia.

- Para efectos de este análisis se puede omitir la cedula del individuo.

- Se observa que la variable de código de tarifa: COD_TARIFA_IND solo posee un valor que es N, por lo cual se elimina.

Figura 44. **Eliminación de registro N en COD_TARIFA_IND**

```
table(datos$COD_TARIFA_IND)
```

```
##
##    N
## 6494
```

Fuente: Elaboración propia.

- Se observa que la variable de código de tarifa: COD_TARIFA_INDUST solo posee un valor que es N, por lo cual se elimina.

Figura 45. **Eliminación de registro N en COD_TARIFA_INDUST**

```
table(datos$COD_TARIFA_INDUST)
```

```
##
##    N
## 6494
```

Fuente: Elaboración propia.

- Se observa que la variable: COD_TARIFA_PRE tiene solo un individuo con valor N, por lo cual se elimina esta variable.

Figura 46. **Eliminación de registro N en COD_TARIFA_PRE**

```
table(datos$COD_TARIFA_PRE)
```

```
##
##    N    S
## 6493    1
```

Fuente: Elaboración propia.

- Se observa que la variable: COD_TARIFA_COMER3 tiene solo 4 individuos con valor N, por lo cual se elimina esta variable.

Figura 47. **Eliminación de registro N en COD_TARIFA_COMER3**

```
table(datos$COD_TARIFA_COMER3)
```

```
##
##    N    S
## 6490    4
```

Fuente: Elaboración propia.

- Se observa que la variable: COD_TARIFA_COMER2 tiene solo 5 individuos con valor N, por lo cual se elimina esta variable.

Figura 48. **Eliminación de registros N en COD_TARIFA_COMER2**

```
table(datos$COD_TARIFA_COMER2)
```

```
##
##    N    S
## 6489    5
```

Fuente: Elaboración propia.

- Se observa que la variable: COD_TARIFA_IND no tiene individuos con valor S, por lo cual se elimina esta variable.

Figura 49. **Eliminación de la variable COD_TARIFA_IND**

```
table(datos$COD_TARIFA_IND)
```

```
##
##    N
## 6494
```

Fuente: Elaboración propia.

Se desestiman dichas variables para futuros análisis:

Figura 50. **Desestimación de variables**

```
datos$TIP_PERSON<-NULL
datos$CEDULA<-NULL
datos$COD_TARIFA_IND<-NULL
datos$COD_TARIFA_INDUST<-NULL
datos$COD_TARIFA_PRE<-NULL
datos$COD_TARIFA_COMER3<-NULL
datos$COD_TARIFA_COMER2<-NULL
datos$COD_TARIFA_IND<-NULL
```

2.3.12 Imputación de datos

Se identifican los NA's en el set de datos, un total de 277 registros, estos son individuos que no tienen indicado la provincia, cantón ni el distrito por parte de la municipalidad, para estos individuos se les asigna el respectivo lugar de votación:

Figura 51. **Individuos sin provincia, cantón y distrito**

	COD_PRO... <int>	COD_CAN... <fct>	COD_DISTRI <fct>	VOTO_DISTRITO <fct>	V_PROVINCIA <fct>	V_CAN... <fct>	COD_TARIFA_REP <fct>
57	NA			SAN ANTONIO	HEREDIA	BELEN	N
173	NA			OTRO	SAN JOSE	CENTRAL	N
198	NA			OTRO	SAN JOSE	CENTRAL	N
205	NA			LA ASUNCION	HEREDIA	BELEN	N
239	NA			OTRO	ALAJUELA	CENTRAL	N
242	NA			SAN ANTONIO	HEREDIA	BELEN	N

Fuente: Elaboración propia.

2.3.13 Recodificación de variables

- Se recodifica la variable tipo de relación ya que esta es una variable categórica:

```
datos[, 'TIPO_RELACION'] <- as.factor(datos[, 'TIPO_RELACION'])
```

- Se recodifica la variable COD_PROVIN a factor, además se unifican los códigos de las provincias de la siguiente forma:

Figura 52. **Codificación de variable COD_PROVIN a factor**

```
datos[, 'COD_PROVIN'] <- as.character(datos[, 'COD_PROVIN'])
datos$COD_PROVIN[datos$COD_PROVIN=="ALAJUELA"]<-"2"
datos$COD_PROVIN[datos$COD_PROVIN=="HEREDIA"]<-"4"
datos$COD_PROVIN[datos$COD_PROVIN=="PUNTARENAS"]<-"8"
datos$COD_PROVIN[datos$COD_PROVIN=="SAN JOSE"]<-"1"
datos$COD_PROVIN[datos$COD_PROVIN=="NV"]<-"8"
datos$COD_PROVIN[datos$COD_PROVIN=="OTRO"]<-"8"
datos[, 'COD_PROVIN'] <- as.factor(datos[, 'COD_PROVIN'])
table(datos$COD_PROVIN)
```

```
##
##    1    2    4    8
##   38  164 6270   22
```

Fuente: Elaboración propia.

- Se recodifica la variable COD_CANTON a factor, además se unifican los códigos de los cantones de la siguiente forma:

Figura 53. **Codificación de variable COD_CANTON a factor**

```
datos[, 'COD_CANTON'] <- as.character(datos[, 'COD_CANTON'])
# codigos utilizados
datos$COD_CANTON[datos$COD_CANTON=="H1"]<-"HE"
datos$COD_CANTON[datos$COD_CANTON=="F1"]<-"EX"
datos$COD_CANTON[datos$COD_CANTON=="A1"]<-"BEL"
datos$COD_CANTON[datos$COD_CANTON=="07"]<-"BEL"
datos$COD_CANTON[datos$COD_CANTON=="7"]<-"BEL"
datos$COD_CANTON[datos$COD_CANTON=="BELEN"]<-"BEL"
datos$COD_CANTON[datos$COD_CANTON=="CENTRAL"]<-"SJ"
datos$COD_CANTON[datos$COD_CANTON=="ESTADOS UNIDOS"]<-"EX"
datos$COD_CANTON[datos$COD_CANTON=="FLORES"]<-"HE"
datos$COD_CANTON[datos$COD_CANTON=="SANTA ANA"]<-"SANA"
datos[, 'COD_CANTON'] <- as.factor(datos[, 'COD_CANTON'])
table(datos$COD_CANTON)
```

```
##
##  BEL   EX   HE   NV OTRO SANA   SJ
## 6335   24   30   13   49    8   35
```

Fuente: Elaboración propia.

- Se recodifica la variable COD_DISTRI a factor, además se unifican los códigos de los distritos de la siguiente forma:

Figura 54. **Codificación de la variable COD_DISTRI a factor**

```
datos[, 'COD_DISTRI'] <- as.character(datos[, 'COD_DISTRI'])
# codigos utilizados
datos$COD_DISTRI[datos$COD_DISTRI=="0"]<-"SA"
datos$COD_DISTRI[datos$COD_DISTRI=="1"]<-"SA"
datos$COD_DISTRI[datos$COD_DISTRI=="02"]<-"LA"
datos$COD_DISTRI[datos$COD_DISTRI=="2"]<-"LA"
datos$COD_DISTRI[datos$COD_DISTRI=="03"]<-"LR"
datos$COD_DISTRI[datos$COD_DISTRI=="3"]<-"LR"
datos$COD_DISTRI[datos$COD_DISTRI=="4"]<-"HE"
datos$COD_DISTRI[datos$COD_DISTRI=="8"]<-"SA"
datos$COD_DISTRI[datos$COD_DISTRI=="L"]<-"EX"
datos$COD_DISTRI[datos$COD_DISTRI=="LA ASUNCION"]<-"LA"
datos$COD_DISTRI[datos$COD_DISTRI=="LA RIBERA"]<-"LR"
datos$COD_DISTRI[datos$COD_DISTRI=="LLORENTE"]<-"HE"
datos$COD_DISTRI[datos$COD_DISTRI=="SAN ANTONIO"]<-"SA"
datos$COD_DISTRI[datos$COD_DISTRI=="SAN JOAQUIN"]<-"HE"
datos[, 'COD_DISTRI'] <- as.factor(datos[, 'COD_DISTRI'])
table(datos$COD_DISTRI)
```

```
##
##   EX   HE   LA   LR   NV OTRO   SA
##   22   17 1639 1328   13   96 3379
```

Fuente: Elaboración propia.

- Se recodifica la variable de la provincia de votación a factor, además se unifican los códigos de la siguiente forma:

Figura 55. **Codificación de la variable V_PROVINCIA a factor**

```
datos[, 'V_PROVINCIA'] <- as.character(datos[, 'V_PROVINCIA'])
datos$V_PROVINCIA[datos$V_PROVINCIA=="SAN JOSE"]<-"SAN_JOSE"
datos[, 'V_PROVINCIA'] <- as.factor(datos[, 'V_PROVINCIA'])
table(datos$V_PROVINCIA)
```

```
##
##   ALAJUELA    HEREDIA         NV       OTRO PUNTARENAS   SAN_JOSE
##        345       5537         85        126         50        351
```

Fuente: Elaboración propia.

- Se recodifica la variable del cantón de votación a factor, además se unifican los códigos de la siguiente forma:

Figura 56. **Codificación de la variable V_CANTON a factor.**

```
datos[, 'V_CANTON'] <- as.character(datos[, 'V_CANTON'])
datos$V_CANTON[datos$V_CANTON=="ESTADOS UNIDOS"]<-"USA"
datos$V_CANTON[datos$V_CANTON=="SANTA ANA"]<-"SANTA_ANA"
datos[, 'V_CANTON'] <- as.factor(datos[, 'V_CANTON'])
table(datos$V_CANTON)
```

```
##
##     BELEN   CENTRAL    FLORES        NV      OTRO SANTA_ANA       USA
##      5112       525       207        85       490        48        27
```

Fuente: Elaboración propia.

- Se recodifica la variable del distrito de votación a factor, además se unifican los códigos de la siguiente forma:

Figura 57. **Codificación de la variable VOTO_DISTRITO a factor**

```
datos[, 'VOTO_DISTRITO'] <- as.character(datos[, 'VOTO_DISTRITO'])
datos$VOTO_DISTRITO[datos$VOTO_DISTRITO=="LA ASUNCION"]<-"LA_ASUNCION"
datos$VOTO_DISTRITO[datos$VOTO_DISTRITO=="LA RIBERA"]<-"LA_RIBERA"
datos$VOTO_DISTRITO[datos$VOTO_DISTRITO=="SAN ANTONIO"]<-"SAN_ANTONIO"
datos$VOTO_DISTRITO[datos$VOTO_DISTRITO=="SAN JOAQUIN"]<-"SAN_JOAQUIN"
datos[, 'VOTO_DISTRITO'] <- as.factor(datos[, 'VOTO_DISTRITO'])
table(datos$VOTO_DISTRITO)
```

```
##
## LA_ASUNCION   LA_RIBERA    LLORENTE          NV        OTRO SAN_ANTONIO
##        1021        1682         139          85        1101        2409
## SAN_JOAQUIN
##          57
```

Fuente: Elaboración propia.

- Se recodifica la variable del tipo de relación a factor, además se unifican los códigos de la siguiente forma:

Figura 58. **Codificación de la variable TIPO_RELACION a factor**

```
datos[, 'TIPO_RELACION'] <- as.character(datos[, 'TIPO_RELACION'])
datos$TIPO_RELACION[datos$TIPO_RELACION=="7"]<-"2"
datos[, 'TIPO_RELACION'] <- as.factor(datos[, 'TIPO_RELACION'])
table(datos$TIPO_RELACION)
```

```
##
##    0    2    3    4    5
## 1623 3783    3  649  436
```

Fuente: Elaboración propia.

- Se recodifica la variable del estado civil a número, además se unifican los códigos de la siguiente forma:

Figura 59. **Codificación de la variable ESTADO_CIVIL a número.**

```
datos[, 'ESTADO_CIVIL'] <- as.factor(datos[, 'ESTADO_CIVIL'])
```

Fuente: Elaboración propia.

Variable del monto de la finca: Se observa que hay 180 registros que tienen un valor en esta variable igual a 0. Además, 274 registros con un monto menor a 5 millones de colones. Estos valores no son imputados.

Figura 60. **Registros por valor de la finca**

```
mean(datos$SUM_MONTO_FINCA)
```

```
## [1] 105967194
```

```
dim(datos[datos$SUM_MONTO_FINCA==0,])
```

```
## [1] 180  38
```

```
dim(datos[datos$SUM_MONTO_FINCA<5000000,])
```

```
## [1] 274  38
```

Figura 61. **Vista general de los datos**

COD_PROVIN	COD_CANTON	COD_DISTRI	VOTO_DISTRITO	V_PROVINCIA	V_CANTON	COD_TARIFA_F
4	BEL	SA	SAN_ANTONIO	HEREDIA	BELEN	N
4	BEL	SA	SAN_ANTONIO	HEREDIA	BELEN	N
4	BEL	SA	SAN_ANTONIO	HEREDIA	BELEN	N
4	BEL	LA	LA_RIBERA	HEREDIA	BELEN	N
4	BEL	LA	LA_RIBERA	HEREDIA	BELEN	N
4	BEL	LA	LA_RIBERA	HEREDIA	BELEN	N
4	BEL	SA	LA_RIBERA	HEREDIA	BELEN	N
4	HE	HE	LA_ASUNCION	HEREDIA	BELEN	N
4	BEL	SA	OTRO	SAN_JOSE	CENTRAL	N
4	BEL	SA	SAN_ANTONIO	HEREDIA	BELEN	N

Fuente: Elaboración propia.

2.3.14 Análisis exploratorio

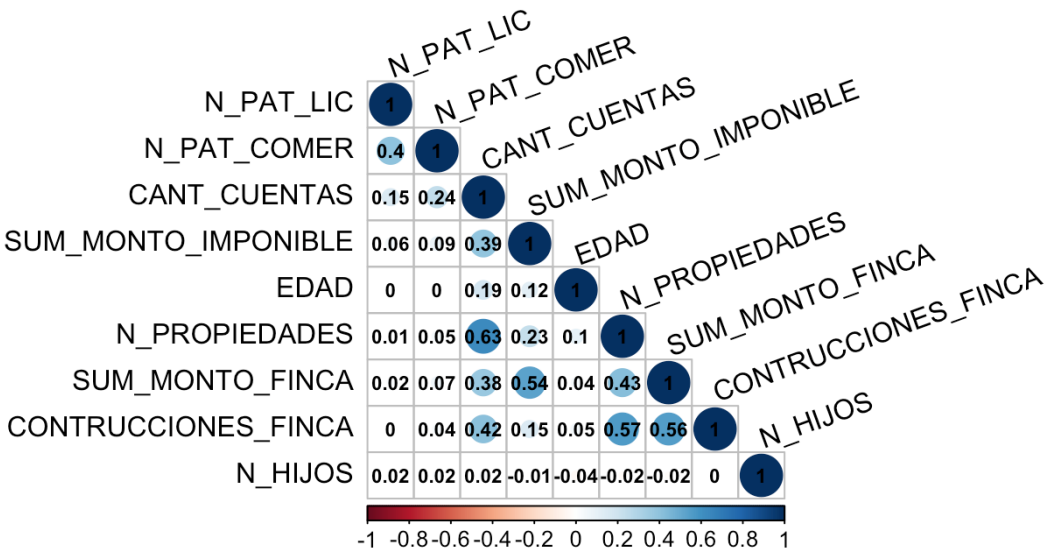
Interpretación de la correlación de las variables numéricas:

- Se puede observar que existe una correlación fuerte y positiva entre la suma del monto de las fincas y la cantidad de construcciones en la finca, es decir entre más

construcciones hay en las fincas, mayor es el monto del valor de la finca y viceversa.

- Se observa una correlación fuerte y positiva entre la cantidad de propiedades y la cantidad de cuentas, indicando que entre mayor es la cantidad de propiedades más cuentas de servicios requiere la persona.

Gráfico 10. **Matriz de correlaciones**



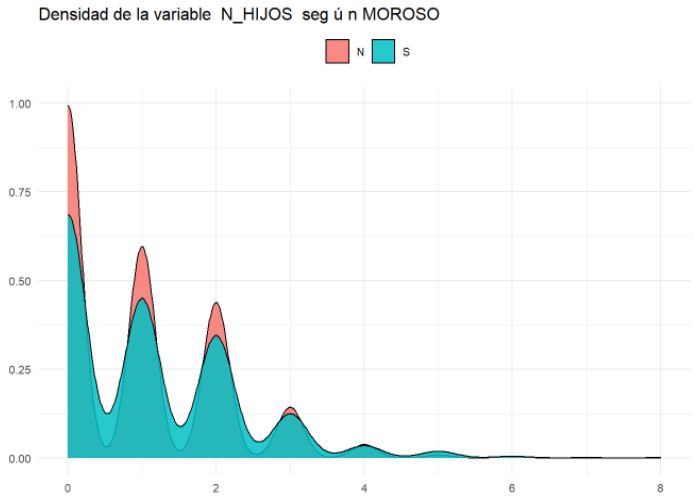
Fuente: Elaboración propia.

2.3.15 Análisis de densidad

Variables numéricas:

Variable número de hijos: Se observa que la variable número de hijos presenta una densidad que permite una diferenciación mayormente en para los individuos no morosos:

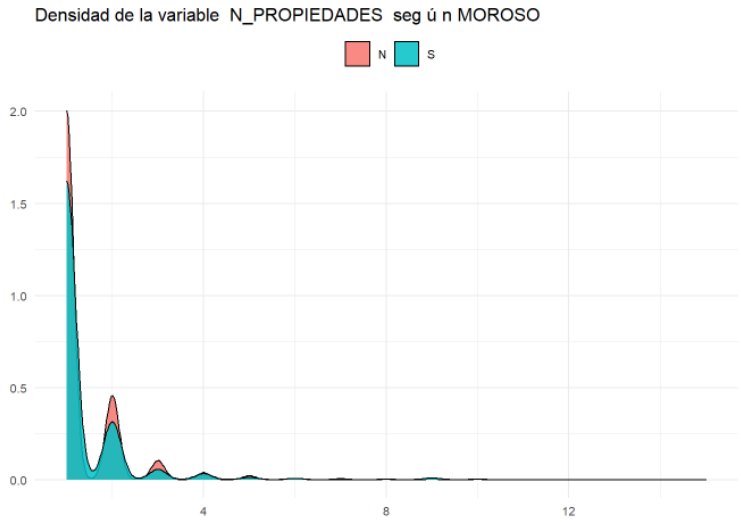
Gráfico 11. **Densidad de la variable N_Hijos según Morosos**



Fuente: Elaboración propia.

Variable número de propiedades: Se observa que la variable de número de propiedades no presenta mucha diferenciación para los individuos con respecto a la variable a predecir:

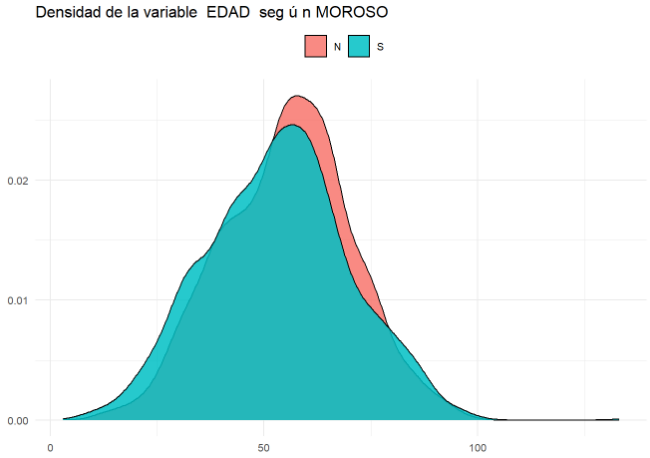
Gráfico 12. **Densidad de la variable N_PROPIEDADES según Morosos**



Fuente: Elaboración propia.

Variable edad: Se observa que para la variable edad, principalmente superior a los 50 años estos individuos son no morosos:

Gráfico 13. **Densidad de la variable EDAD según Moroso**



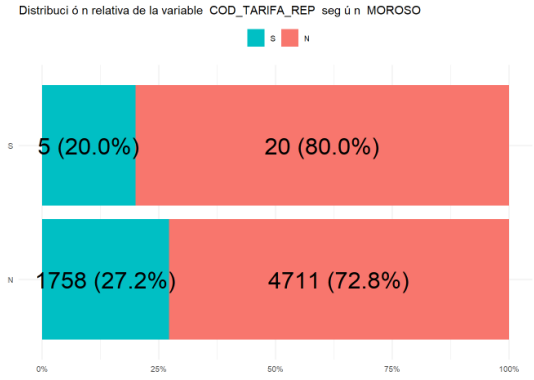
Fuente: Elaboración propia.

Variables categóricas:

Variable código tarifa para REP:

Esta variable se observa que no presenta una buena distribución que pueda separar los individuos con respecto a la variable a predecir:

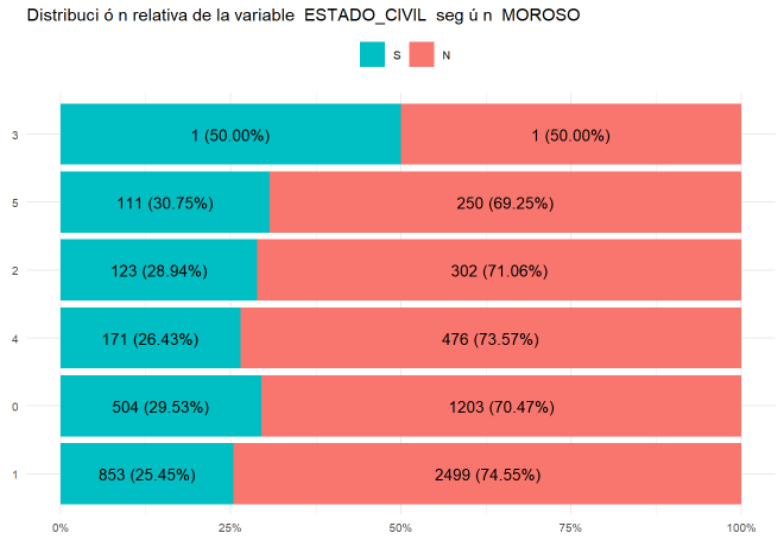
Gráfico 14. **Distribución relativa de la variable COD_TARIFA_REP según Moroso**



Fuente: Elaboración propia.

Variable estado civil: Se observa que esta variable se puede considerar como una buena variable predictiva ya que presenta relativamente buena separación de individuos morosos y no morosos:

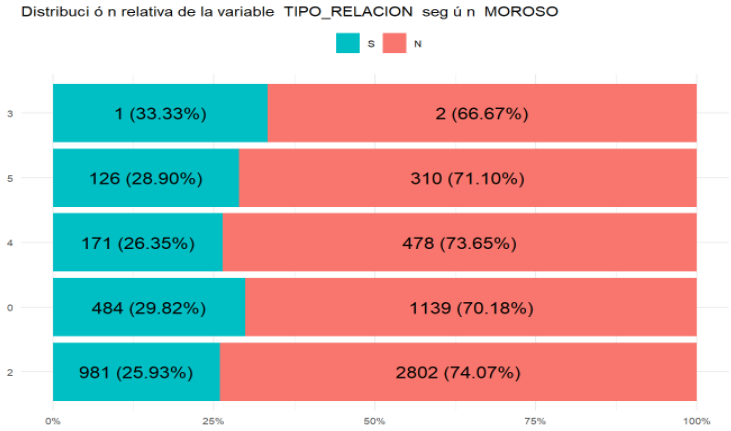
Gráfico 15. **Distribución relativa de la variable ESTADO_CIVIL según Moroso.**



Fuente: Elaboración propia.

Variable tipo de relación: Esta variable presenta una distribución que permite evidenciar una mayor cantidad de individuos no morosos para cada una de las categorías de esta variable:

Gráfico 16. **Distribución relativa de la variable TIPO_RELACION según Moroso**

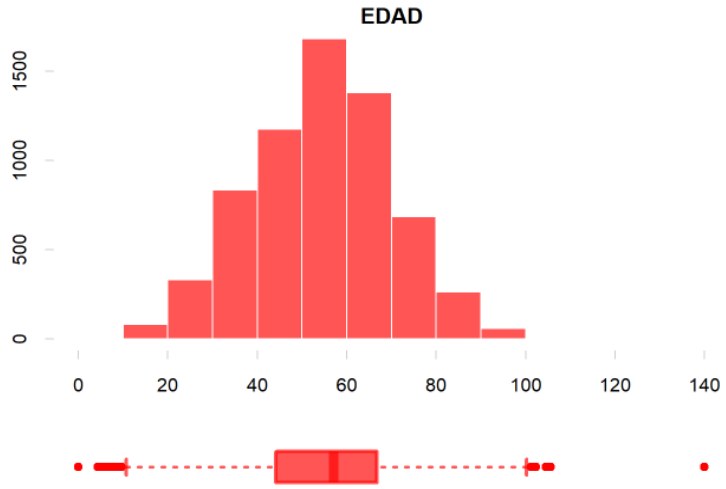


Fuente: Elaboración propia.

Análisis numérico:

Variable edad: Se observa que la edad sigue una distribución normal, la media se encuentra en 56 años, se puede ver un individuo con una edad cercana a los 140 años:

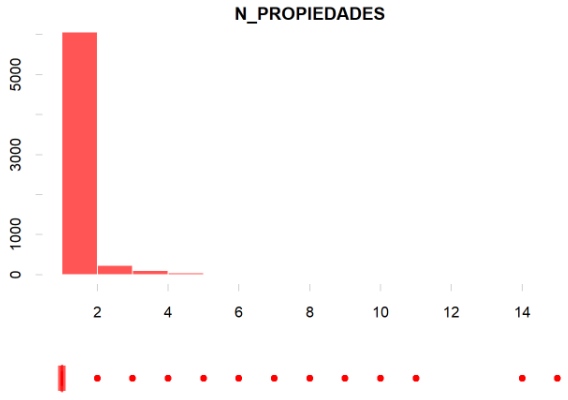
Gráfico 17. **Distribución de la variable edad**



Fuente: Elaboración propia.

Variable de número de propiedades: Se observa que la mayor cantidad de individuos tienen de 0 a 2 propiedades:

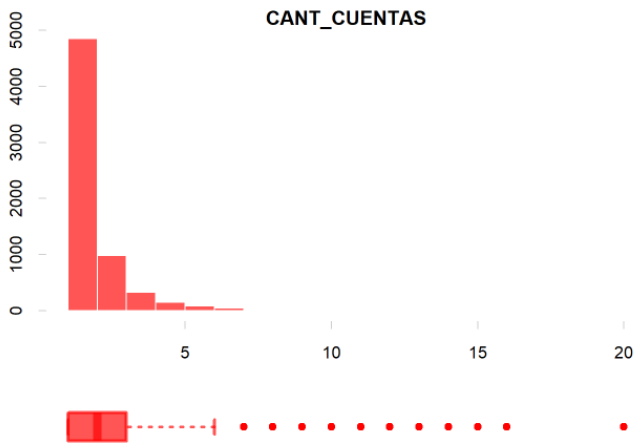
Gráfico 18. **Distribución de número de propiedades por individuo**



Fuente: Elaboración propia.

Variable cantidad de cuentas: Se obtiene que la mayor cantidad de individuos tienen menos de 5 cuentas:

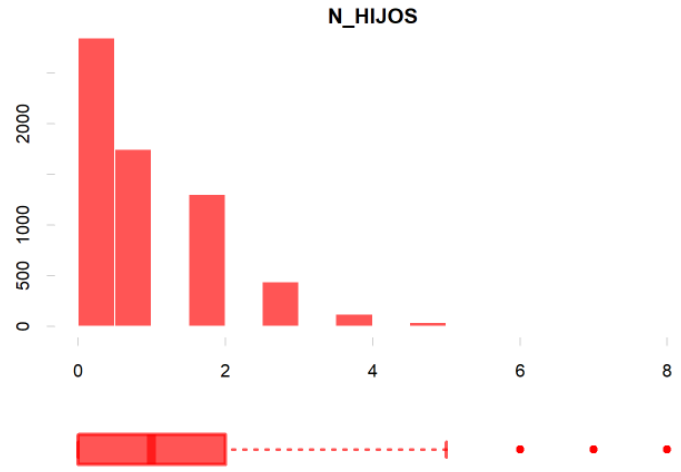
Gráfico 19. **Distribución de cuentas por individuo**



Fuente: Elaboración propia.

Variable de número de hijos: Se obtiene que los individuos tienen en su mayoría menos de 2 hijos:

Gráfico 20. **Distribución del número de hijos por individuo**



Fuente: Elaboración propia.

2.3.16 Códigos disyuntivos completos

Se procede a transformar las variables categóricas a códigos disyuntivos completos:

Figura 62. **Transformación de variables categóricas a códigos disyuntivos.**

```
datos_aux<-var.categoricas(datos)
names<-colnames(datos_aux)
names<-names[names !='MOROSO']
names<-names[names !='TIPO_RELACION']
names<-names[names !='ESTADO_CIVIL']

datos1<- select(datos,all_of(names))

datos<-select(datos,-all_of(names))

datos_dummy <- dummy.data.frame(datos1,sep='_',verbose=F)
```

Fuente: Elaboración propia

Así, el set de datos queda conformado por las siguientes variables:

Tabla 13. Nombres de variables

COD_PROVIN_1	COD_PROVIN_2	COD_PROVIN_4
COD_PROVIN_8	COD_CANTON_BEL	COD_CANTON_EX
COD_CANTON_HE	COD_CANTON_NV	COD_CANTON_OTRO
COD_CANTON_SANA	COD_CANTON_SJ	COD_DISTRI_EX
COD_DISTRI_HE	COD_DISTRI_LA	COD_DISTRI_LR
COD_DISTRI_NV	COD_DISTRI_OTRO	COD_DISTRI_SA
V_CANTON_CENTRAL	V_PROVINCIA_OTRO	VOTO_DISTRITO_LA_ASUNCION
V_CANTON_FLORES	V_CANTON_BELEN	VOTO_DISTRITO_LA_RIBERA
V_CANTON_OTRO	V_CANTON_NV	VOTO_DISTRITO_LLORENTE
V_CANTON_USA	V_CANTON_SANTA_ANA	VOTO_DISTRITO_NV
COD_TARIFA_REP_N	COD_TARIFA_REP_S	VOTO_DISTRITO_OTRO
COD_TARIFA_RESID_S	COD_TARIFA_RESID_N	VOTO_DISTRITO_SAN_ANTONIO
COD_TARIFA_DOM_N	COD_TARIFA_DOM_S	VOTO_DISTRITO_SAN_JOAQUIN
COD_TARIFA_SOCIAL_N	COD_TARIFA_SOCIAL_S	V_PROVINCIA_ALAJUELA
COD_TARIFA_COMER1_N	COD_TARIFA_COMER1_S	V_PROVINCIA_HEREDIA
COD_TARIFA_SOC_N	COD_TARIFA_SOC_S	V_PROVINCIA_NV
COD_TARIFA_ORD_N	COD_TARIFA_ORD_S	V_PROVINCIA_PUNTARENAS
COD_SERVIC_MPO_N	COD_SERVIC_MPO_S	V_PROVINCIA_SAN_JOSE

COD_SERVIC_LVP_N	COD_SERVIC_LVP_S	COD_SERVIC_AGU_N
COD_SERVIC_AGU_S	COD_SERVIC_IBI_N	COD_SERVIC_IBI_S
COD_SERVIC_BAS_N	COD_SERVIC_BAS_S	COD_SERVIC_PZV_N
COD_SERVIC_PAT_S	COD_SERVIC_PZV_S	COD_SERVIC_PAT_N
COD_SERVIC_CEM_S	SEXO_F	SEXO_M
IND_AFILIADO_N	IND_AFILIADO_S	PROP_SENAS_N
PROP_SENAS_S	CONSTRUC_FINCA_N	CONSTRUC_FINCA_S
SUM_MONTO_FINCA	SUM_MONTO_IMPONIBLE	EDAD
N_HIJOS	ESTADO_CIVIL	TIPO_RELACION
CANT_CUENTAS	N_PROPIEDADES	N_PAT_COMER
N_PAT_LIC	CONTRUCCIONES_FINCA	MOROSO

Una vez concluido el proceso de limpieza y selección de variables, se procede a guardar los datos limpios:

Figura 63. **Almacenamiento del set de datos final**

```
write.table(datos,file="../datos/datos_limpios.csv", sep=";", dec = ',', row.names = TRUE)
saveRDS(datos, file = "../datos/datos_limpios.rds")
```

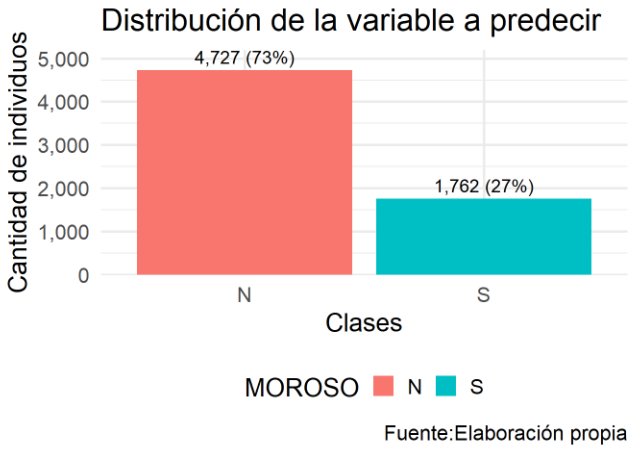
Fuente: Elaboración propia.

2.4 Modelado

2.4.1 Distribución de la variable a predecir

El presente proyecto consiste en un problema desbalanceado, es decir, hay más individuos de una categoría, en este caso, se tiene que el conjunto de datos presenta un 27% de individuos que pertenecen a la categoría de moroso, y 73% que no son morosos.

Gráfico 21. **Distribución de la variable a predecir**



Fuente: Elaboración propia.

2.4.2 Algoritmos utilizados

Para efectos del presente proyecto se utilizarán los siguientes algoritmos para elegir el mejor modelo:

- Redes neuronales
- Arboles de decisión
- Máquinas de soporte vectorial
- Bosques aleatorios
- Método de Naive Bayes
- Regresión lineal
- K-vecinos más cercanos
- Xtreme Gradient Boosting
- Potenciación

2.4.3 Estructuras del proyecto en RStudio

Se crea un proyecto en RStudio con la siguiente estructura:

- Directorio calibrar: Directorio utilizado para calibrar los modelos.
- Directorio datos: En este directorio se guardan los datos tanto limpios como datos originales.
- Directorio de imágenes: Se guardan las imágenes requeridas para generar el documento auto reproducible.
- Directorio: modelos: En este directorio se guardan las matrices de confusión de los modelos.
- Directorio de probabilidad de corte: En este directorio se generan los modelos utilizando la técnica de probabilidad de corte.
- Directorio de reportes: En este directorio se guardan los documentos generados como resultado del proyecto.
- Directorio de utilidades: En este directorio se guardan las herramientas utilizadas en el proyecto como lo son funciones, gráficos, paquetes:

Figura 64. **Directorio de utilidades**

2.4.4 Procesamiento en paralelo

Para efectos de este proyecto y con el fin de disminuir el tiempo de procesamiento se trabaja en la mayor parte con procesamiento en paralelo utilizando los paquetes de R: Snow y parallel.

Figura 65. **Procesamiento en paralelo.**

```
peones <- parallel::detectCores()
peones
```

```
## [1] 4
```

Fuente: Elaboración propia.

2.4.5 Validación cruzada

Los modelos son evaluados utilizando set de datos de pruebas y entrenamiento generados mediante validación cruzada, se utilizan 5 validaciones cruzadas con una cantidad de 10 grupos:

Figura 66. **Modelos de pruebas y entrenamiento generados con validación cruzada.**

```
numero.filas <- nrow(datos)
cantidad.validacion.cruzada <- 5
cantidad.grupos <- 10

numero.filas
```

```
## [1] 6494
```

```
cantidad.validacion.cruzada
```

```
## [1] 5
```

```
cantidad.grupos
```

```
## [1] 10
```

Fuente: Elaboración propia.

2.4.6 Calibración de modelos

Se procede a calibrar los siguientes modelos:

2.4.6.1 Calibración modelo potenciación:

Se calibran los siguientes algoritmos:

- Discrete
- Real
- Gentle

Los algoritmos son calibrados en paralelo utilizando validación cruzada, finalmente se guardan sus matrices de confusiones:

Figura 67. **Almacenamiento de matrices de confusión**

```
saveRDS(MCs.discrete, file = "./calibrar/ada/m_cal_MCs_discrete.rds")
saveRDS(MCs.real, file = "./calibrar/ada/m_cal_MCs_real.rds")
saveRDS(MCs.gentle, file = "./calibrar/ada/m_cal_MCs_gentle.rds")

stopCluster(clp) #cerrar el proceso
```

Fuente: Elaboración propia.

Se evalúan los resultados:

Figura 68. **Resultado de los modelos**

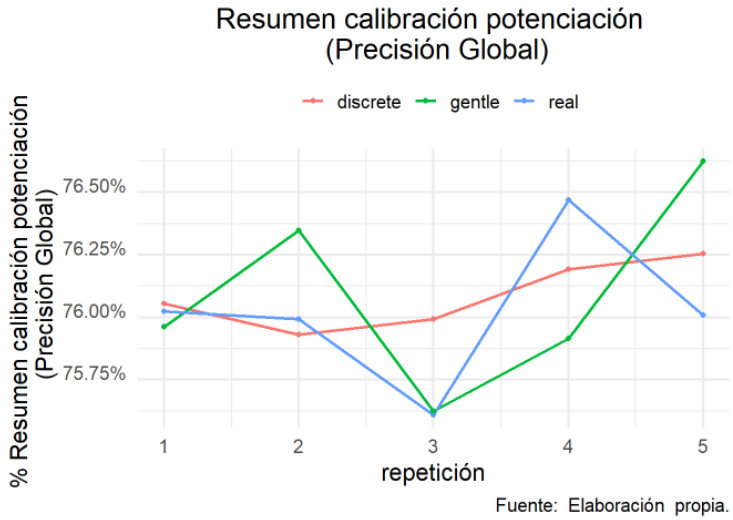
rep <int>	discrete <dbl>	real <dbl>	gentle <dbl>
1	0.7605482	0.7602402	0.7596243
2	0.7593163	0.7599322	0.7634740
3	0.7599322	0.7560825	0.7562365
4	0.7619341	0.7647059	0.7591623
5	0.7625500	0.7600862	0.7662458

5 rows

Fuente: Elaboración propia.

Evaluación tomando en cuenta la precisión global: Se observa que el algoritmo que obtiene una mejor precisión global es el algoritmo gentle.

Gráfico 22. **Resumen calibración potenciación (Precisión Global)**

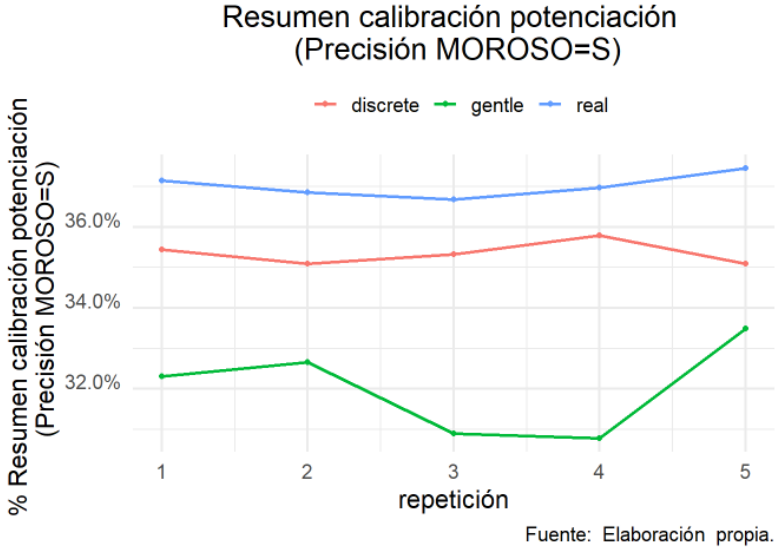


Fuente: Elaboración propia.

Evaluación tomando en cuenta la categoría del Sí:

Se obtiene que el algoritmo que mejor predice cuando un individuo va a ser moroso es el algoritmo gente:

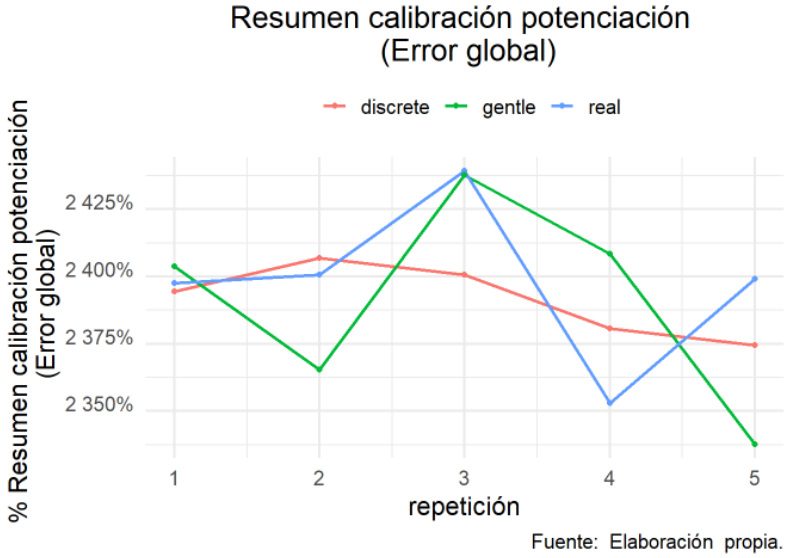
Gráfico 23. **Resumen calibración potenciación (Precisión MOROSO=S)**



Fuente: Elaboración propia.

Evaluación teniendo en cuenta el error global: Se obtiene que el algoritmo que presenta el menor error es el algoritmo gentle:

Gráfico 24. **Resumen calibración potenciación (Error global)**



Fuente: Elaboración propia.

Conclusión calibración potenciación: Se obtiene que para el modelo de potenciación el algoritmo que tiene una mejor predicción es el algoritmo gentle.

2.4.6.2 Calibración KNN

Se calibran los siguientes algoritmos:

- rectangular
- triangular
- epanechnikov
- biweight
- triweight
- cos
- inv
- gaussian
- optimal

Se cargan las matrices de confusión como parte del resultado de la calibración en paralelo:

Figura 69. **Carga de matrices de confusión**

```
MCs.rectangular <- readRDS(file = "../calibrar/knn/m_cal_MCsrectangular.rds")
MCs.triangular <- readRDS(file = "../calibrar/knn/m_cal_MCstriangular.rds")
MCs.epanechnikov <- readRDS(file = "../calibrar/knn/m_cal_MCsepanechnikov.rds")

MCs.biweight <- readRDS(file = "../calibrar/knn/m_cal_MCsbiweight.rds")
MCs.triweight <- readRDS(file = "../calibrar/knn/m_cal_MCstriweight.rds")
MCs.cos <- readRDS(file = "../calibrar/knn/m_cal_MCscos.rds")

MCs.inv <- readRDS(file = "../calibrar/knn/m_cal_MCs.inv.rds")
MCs.gaussian <- readRDS(file = "../calibrar/knn/m_cal_MCsgaussian.rds")
MCs.optimal <- readRDS(file = "../calibrar/knn/m_cal_MCsoptimal.rds")
```

Fuente: Elaboración propia.

Figura 70. **Resultados numéricos de la calibración en paralelo.**

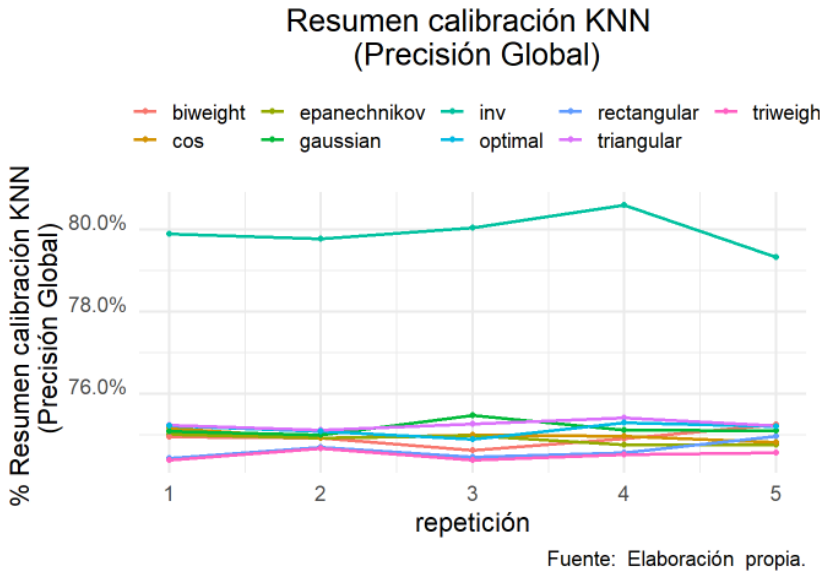
rep <int>	rectangular <dbl>	triangular <dbl>	epanechnikov <dbl>	biweight <dbl>	triweight <dbl>	cos <dbl>	inv <dbl>
1	0.2669817	0.3041937	0.2894271	0.3201418	0.3366804	0.3018311	0.5858156
2	0.2604843	0.3047844	0.2911991	0.3154164	0.3425871	0.2941524	0.5796690
3	0.2539870	0.2906084	0.2835204	0.3047844	0.3295924	0.2953337	0.5900709
4	0.2409923	0.2988777	0.2835204	0.3112817	0.3313644	0.3036031	0.5938534
5	0.2569403	0.2959244	0.2906084	0.3236858	0.3449498	0.2923804	0.5867612

5 rows | 1-8 of 10 columns

Fuente: Elaboración propia.

Evaluación tomando la precisión global: Se observa que el algoritmo que tiene mejor desempeño tomando la precisión global es el algoritmo inv:

Gráfico 25. **Resumen calibración KNN (Precisión Global)**

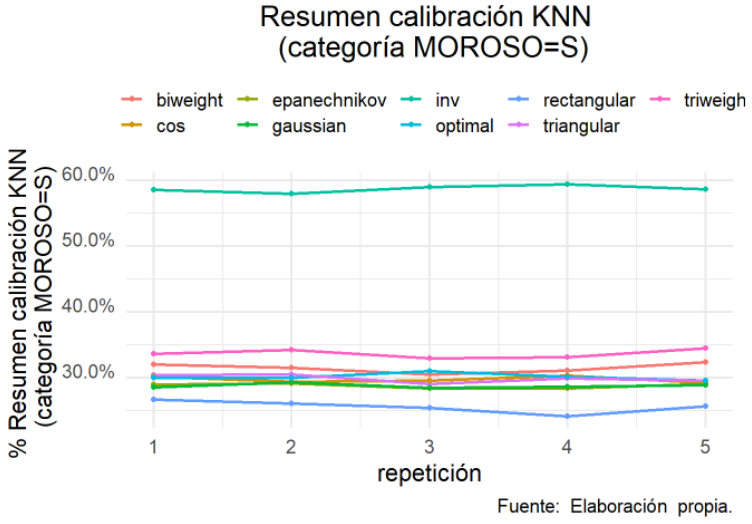


Fuente: Elaboración propia.

Evaluación tomando la categoría Sí:

El algoritmo que tiene mejor predicción para la categoría del sí, es el algoritmo inv:

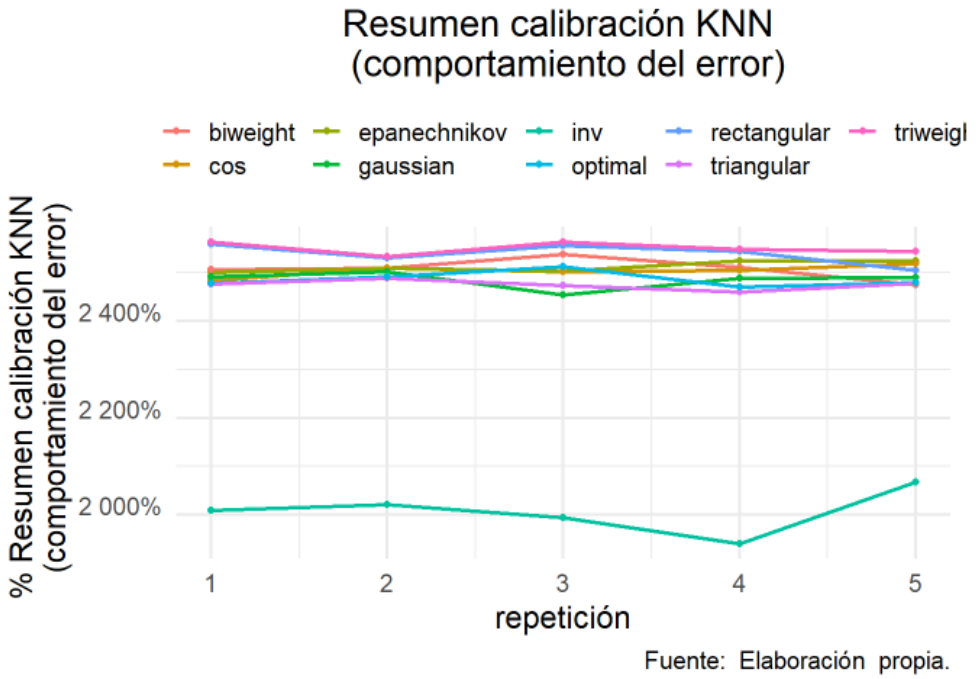
Gráfico 26. **Resumen calibración KNN (categoría MOROSO=S)**



Fuente: Elaboración propia.

Evaluación teniendo en cuenta el error: Se observa que el algoritmo que tiene un error menor es el algoritmo inv.

Gráfico 27. **Resumen calibración KNN (Comportamiento del error)**



Fuente: Elaboración propia.

Conclusión calibración KNN: Según los resultados obtenidos de la calibración del modelo knn, se obtiene que el algoritmo que tiene un mejor desempeño en la predicción es el algoritmo inv.

2.4.6.3 Calibración máquinas de soporte vectorial

Se calibran los siguientes algoritmos:

- linear
- polynomial
- radial basis
- sigmoid

A continuación, se muestra la lectura de las matrices de confusión:

Figura 71. **Lectura de las matrices de confusión**

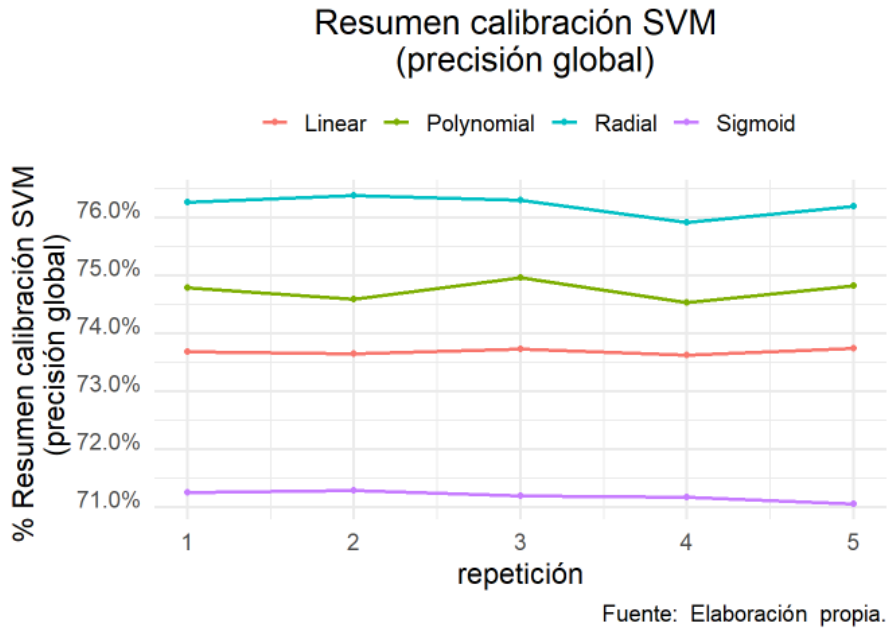
```
MCs.linear <- readRDS(file = "../calibrar/svm/m_cal_MCslinear.rds")
MCs.polynomial <- readRDS(file = "../calibrar/svm/m_cal_MCspolynomial.rds")
MCs.sigmoid <- readRDS(file = "../calibrar/svm/m_cal_MCssigmoid.rds")
MCs.radial <- readRDS(file = "../calibrar/svm/m_cal_MCsradial.rds")
```

Fuente: Elaboración propia.

Evaluación tomando la precisión global como referencia:

Se obtiene que el algoritmo que obtiene mejores resultados para la precisión global es el algoritmo radial.

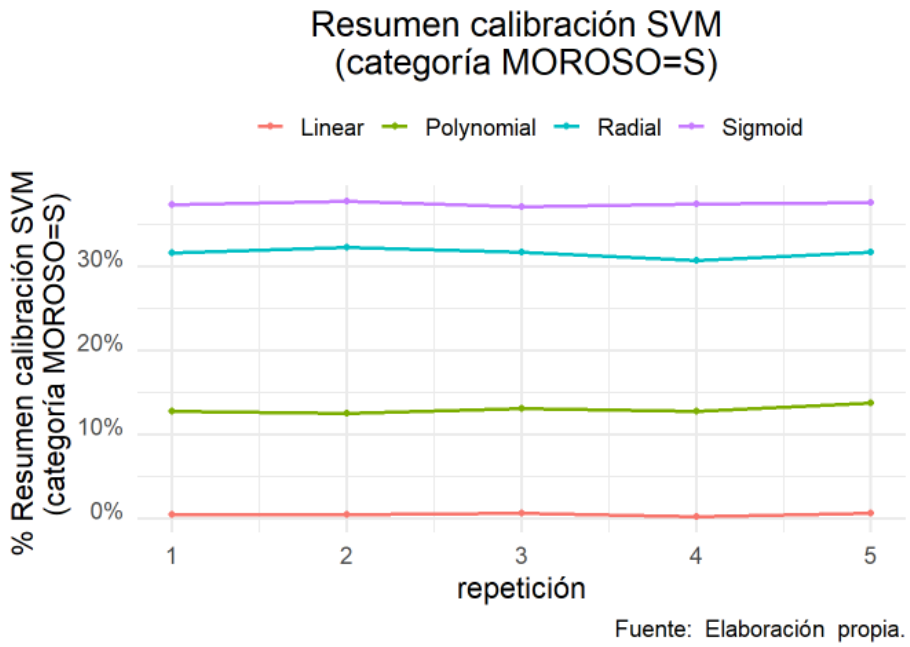
Gráfico 28. **Resumen calibración SVM (Presión global)**



Fuente: Elaboración propia.

Evaluación tomando en cuenta la categoría del sí: Se obtiene que el algoritmo que mejor predice el sí, es el algoritmo sigmoid:

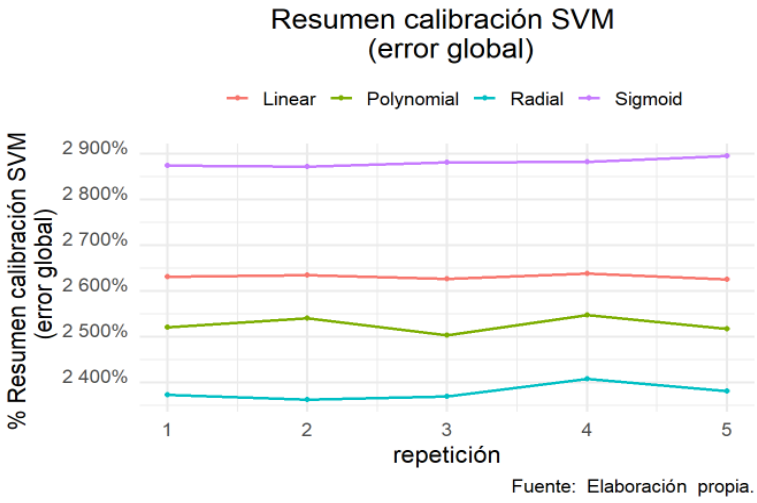
Gráfico 29. **Resumen calibración SVM (Categoría MOROSO=S)**



Fuente: Elaboración propia.

Evaluación tomando el error global como referencia: Se obtiene que el algoritmo que tiene un menor porcentaje de error es radial:

Gráfico 30. **Resumen calibración SVM (error global)**



Fuente: Elaboración propia.

Conclusión máquinas de soporte vectorial: Se identifica que tomando en cuenta la categoría de moroso sí, el algortimo que tiene un mejor resultado de predicción es el sigmoid.

2.4.6.4 Calibración árboles de decisión

Se calibran:

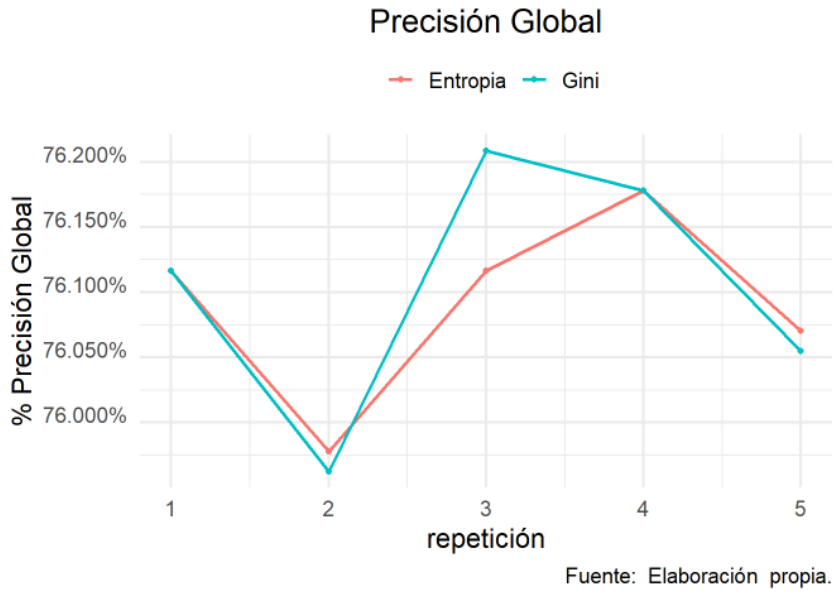
- Gini
- Información ganada o entropía

Figura 72. **Lectura de las matrices de confusión**

```
MCs.information <- readRDS(file = "../calibrar/rpart/m_cal_MCsinformation.rds")
MCs.gini <- readRDS(file = "../calibrar/rpart/m_cal_MCsgini.rds")
```

Evaluación tomando en cuenta la precisión global: Se observa que no existe una diferencia entre ambos algoritmos:

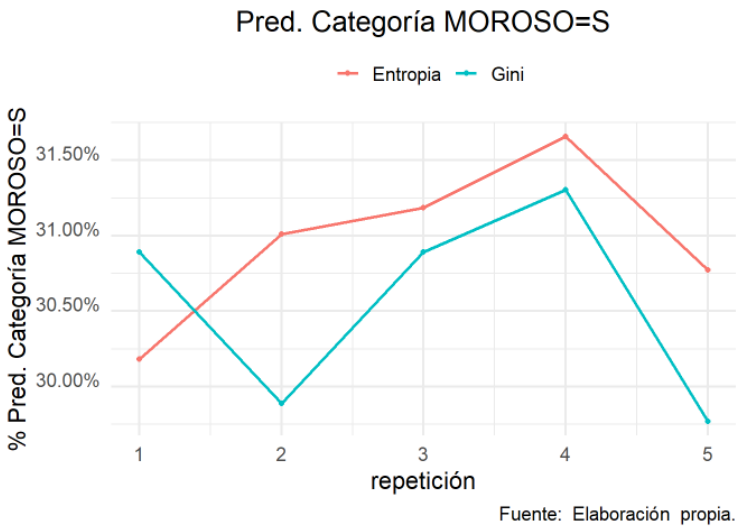
Gráfico 31. **Precisión Global**



Fuente: Elaboración propia.

Evaluación tomando la categoría morosa como referencia: Se observa que el algoritmo que obtiene mejores resultados para este caso es entropía:

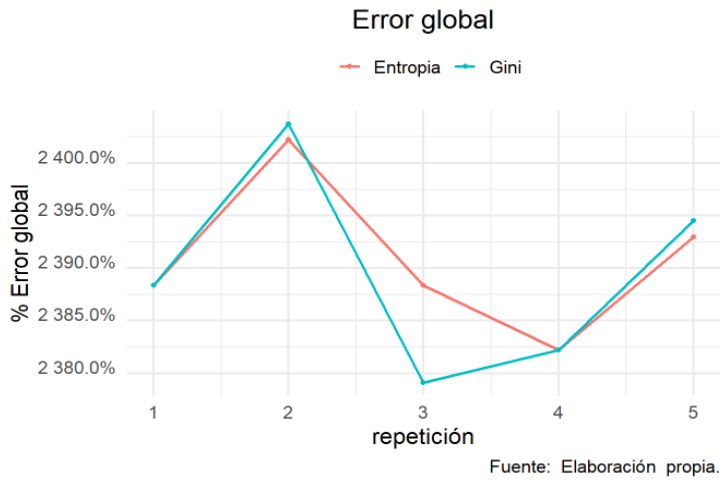
Gráfico 32. **Pred. Categoría moroso**



Fuente: Elaboración propia.

Evaluación tomando el error global como referencia: Se observa que no existe mayor diferencia entre ambos algoritmos.

Gráfico 33. **Error Global**



Fuente: Elaboración propia.

Conclusión árboles de decisión: Se concluye que los resultados son muy similares, por lo cual se puede utilizar cualquier algoritmo para este conjunto de datos.

2.4.7 Generación de modelos

Una vez se han calibrado los algoritmos y se han generado los modelos, se precede a evaluar los modelos con el fin de determinar cuál obtiene la mejor predicción para el conjunto de datos:

Los modelos han sido generados en paralelo, se procede a la lectura de las matrices de confusión:

Figura 73. **Lectura de modelos generados en paralelo.**

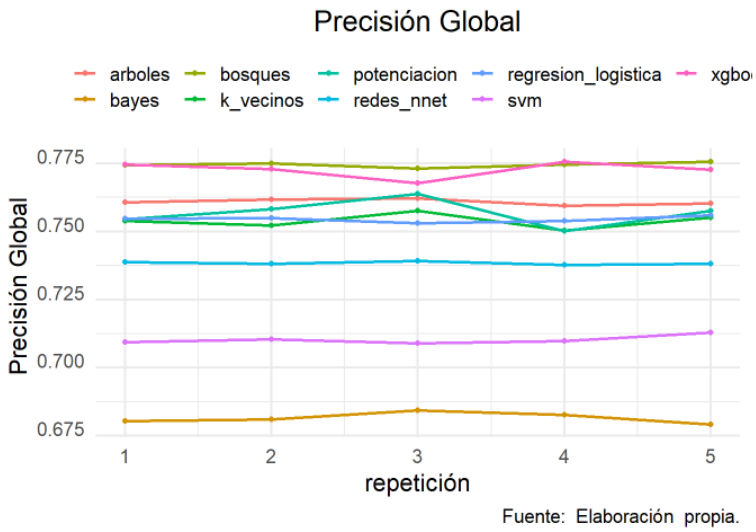
```
MCs.svm <- readRDS(file = "../modelos/mc_models/mc_svm.rds")
MCs.knn <- readRDS(file = "../modelos/mc_models/mc_knn.rds")

MCs.bayes <- readRDS(file = "../modelos/mc_models/mc_bayes.rds")
MCs.arbol <- readRDS(file = "../modelos/mc_models/mc_arbol.rds")

MCs.bosque <- readRDS(file = "../modelos/mc_models/mc_bosque.rds")
MCs.potenciacion <- readRDS(file = "../modelos/mc_models/mc_poten.rds")

MCs.red <- readRDS(file = "../modelos/mc_models/mc_red.rds")
MCs.xgboost <- readRDS(file = "../modelos/mc_models/mc_xgb.rds")

#MCs.red.neu <- readRDS(file = "../modelos/mc_models/mc_redneu.rds")
MCs.glm <- readRDS(file = "../modelos/mc_models/mc_glm.rds")
```

Fuente: Elaboración propia.

2.4.8 Análisis tomando en cuenta la precisión global

Se obtiene que el modelo de XGBoosting y bosques son los algoritmos que tienen mejores índices en cuanto a la precisión global.

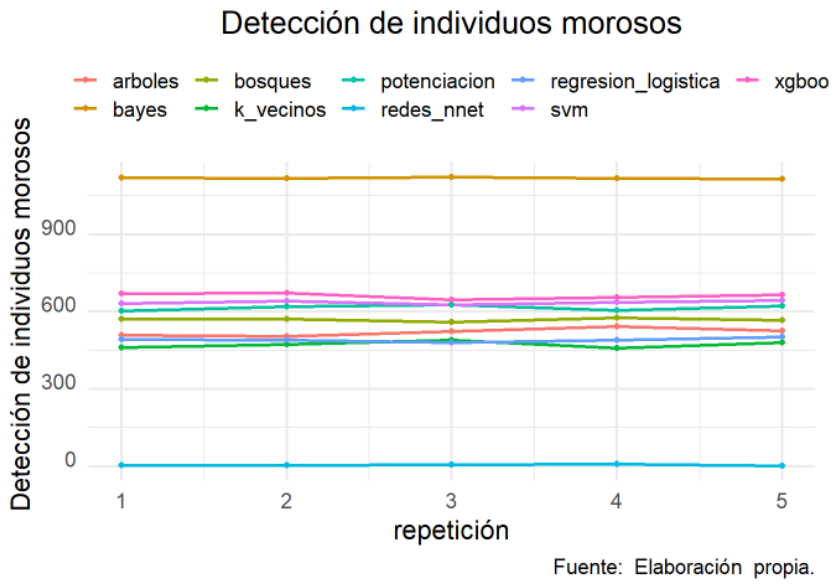
Gráfico 34. **Precisión Global**



Fuente: Elaboración propia.

Análisis de algoritmos tomando en cuenta la suma de predicciones del sí: Se observa que el modelo que obtiene mejores resultados es el modelo de Bayes:

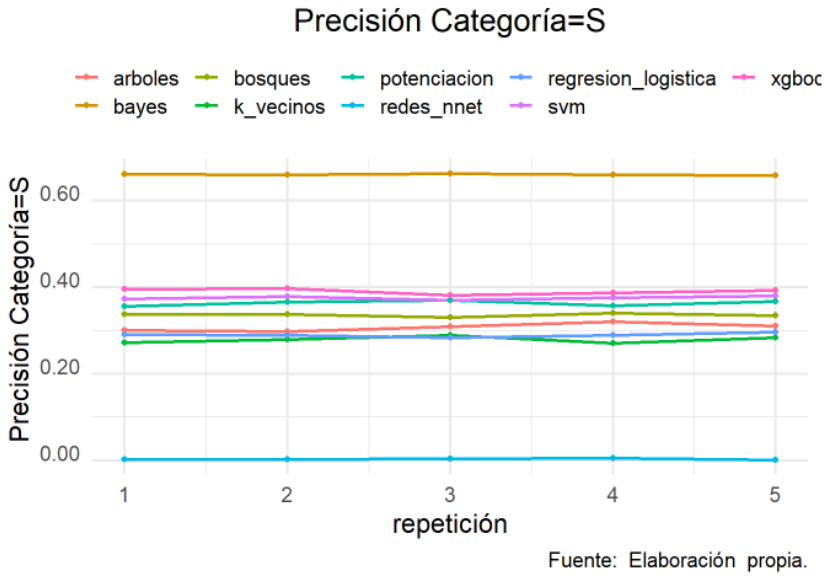
Gráfico 35. **Detección de individuos morosos**



Fuente: Elaboración propia.

Análisis tomando en cuenta la precisión de la categoría del sí: Se observa que de igual manera el modelo de bayes es el que tiene mejores resultados:

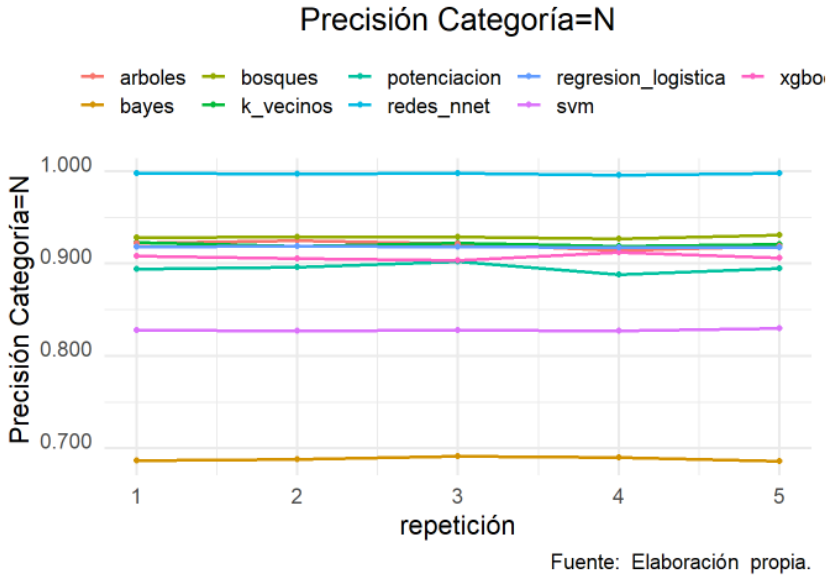
Gráfico 36. **Precisión Categoría = S**



Fuente: Elaboración propia.

Análisis tomando en cuenta la categoría del no: Se obtiene que el modelo de redes neuronales obtiene un resultado del 100%, sin embargo, este modelo no es capaz de predecir de forma correcta la categoría del sí. Además, se observa que el modelo bayesiano obtiene valores cercanos al 70%.

Gráfico 37. **Evaluación de modelos, precisión de la categoría del no**



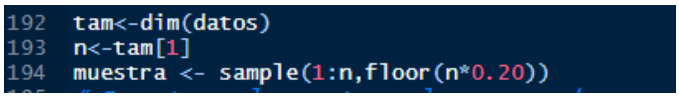
Conclusión de elección de modelo: En este caso y por las características del proyecto, la categoría que resulta importante predecir es cuando un individuo va a presentar problemas de morosidad. De esta forma, se obtiene que el modelo que obtiene una mejor predicción para la categoría de interés es el modelo bayesiano.

2.4.9 Probabilidad de corte

Se sugiere además el uso de la técnica de probabilidad de corte, en este caso los algoritmos pueden funcionar de una mejor manera si se sacrifica la categoría del no:

Se define el tamaño de la muestra: se usa el 20% de los datos para pruebas y el 80% restante para entrenamiento:

Figura 74. **Tamaño de la muestra en probabilidad de corte**



```
192  tam<-dim(datos)
193  n<-tam[1]
194  muestra <- sample(1:n,floor(n*0.20))
```

Fuente: Elaboración propia.

Probabilidad de corte bayes: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 75. **Lectura de matrices de confusón, modelo bayesiano**

```
prediccion<-readRDS( file = "../prob_corte/resultados/pred_train.bayes.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.bayes.rds")
```

Fuente: Elaboración propia.

Se observa que utilizando una probabilidad de 0.6, este algoritmo obtiene índices predictivos de un 69% de predicción para ambas categorías de interés:

Figura 76. **Matrices de confusión utilizando probabilidad de corte, modelo bayesiano**

```
## Corte usado para la Probabilidad = 0.6
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 669 292
##     S 104 233
##
## Overall Accuracy: 0.6949
## Overall Error:    0.3051
##
## Category Accuracy:
##
##             N            S
##      0.696150     0.691395
```

Fuente: Elaboración propia.

Probabilidad de corte potenciación: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 77. **Lectura de las matrices de confusión modelo de potenciación**

```
prediccion<-readRDS(file = "../prob_corte/resultados/pred_train.ada.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.ada.rds")

Clase <- ttesting
head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo, se obtiene que la probabilidad de corte 0.25 permite predecir un 69% de individuos no morosos y un 68% de individuos morosos.

Figura 78. **Matrices de confusión usando probabilidad de corte, modelo potenciación**

```
##
## Corte usado para la Probabilidad = 0.25
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 665 296
##     S 105 232
##
## Overall Accuracy: 0.6911
## Overall Error:    0.3089
##
## Category Accuracy:
##
##             N            S
##      0.691988     0.688427
## ========================================
```

Fuente: Elaboración propia.

Probabilidad de corte k vecinos: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 79. **Lectura de matrices de confusión k vecinos**

```
prediccion<-readRDS( file = "../prob_corte/resultados/pred_train.knn.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.knn.rds")

Clase <- ttesting
head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para K vecinos, se sugiere la probabilidad de corte de 0.25, esta permite predecir un 68.15% de los individuos no morosos, y un 72.70% de individuos morosos.

Figura 80. **Matrices de confusión usando probabilidad de corte, modelo k vecinos**

```
## Corte usado para la Probabilidad = 0.25
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 655 306
##     S  92 245
##
## Overall Accuracy: 0.6934
## Overall Error:    0.3066
##
## Category Accuracy:
##
##             N            S
##      0.681582     0.727003
```

Fuente: Elaboración propia.

Probabilidad de corte XGBoosting: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 81. **Lectura de matrices de confusión, modelo XGBoosting**

```
prediccion<-readRDS( file = "../prob_corte/resultados/pred_train.xgboost.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.xgboost.rds")

Clase <- ttesting
head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo, se sugiere utilizar una probabilidad de corte de 0.2, la cual permite predecir un 66.38% de los individuos no morosos, y un 70.91 de individuos morosos.

Figura 82. **Matrices de confusión modelo XGBoosting**

```
## Corte usado para la Probabilidad = 0.2
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 638 323
##     S  98 239
##
## Overall Accuracy: 0.6757
## Overall Error:    0.3243
##
## Category Accuracy:
##
##             N           S
##      0.663892    0.709199
## ----------------------------------------
```

Fuente: Elaboración propia.

Probabilidad de corte con redes neuronales (nnet): Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 83. **Lectura de matrices de confusión modelo redes neuronales**

```
prediccion<-readRDS( file = "../prob_corte/resultados/pred_train.nnet.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.nnet.rds")

Clase <- ttesting
head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo no es posible determinar una probabilidad de corte que se puede utilizar:

Figura 84. **Matrices de confusión utilizando probabilidad de corte, modelo de redes neuronales**

```
## Corte usado para la Probabilidad = 0.35
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 961   0
##     S 337   0
##
## Overall Accuracy: 0.7404
## Overall Error:    0.2596
##
## Category Accuracy:
##
##             N            S
##      1.000000     0.000000
## ========================================
## Corte usado para la Probabilidad = 0.3
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 961   0
##     S 337   0
##
## Overall Accuracy: 0.7404
## Overall Error:    0.2596
##
## Category Accuracy:
##
##             N            S
##      1.000000     0.000000
## ========================================
```

Fuente: Elaboración propia.

Probabilidad de corte con bosques aleatorios: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 85. **Lectura de matrices de confusión, modelo bosques aleatorios**

```
prediccion<-readRDS( file = "../prob_corte/resultados/pred_train.randomForest.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.randomForest.rds")
Clase <- ttesting

head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo se sugiere utilizar una probabilidad de corte de 0.25, esta permite predecir un 74.50 % de los individuos no morosos, y un 70.62% de los individuos morosos:

Figura 86. **Matríz de confusión usando probabilidad de corte, modelo bosques aleatorios**

```
## Corte usado para la Probabilidad = 0.25
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 716 245
##     S  99 238
##
## Overall Accuracy: 0.7350
## Overall Error:    0.2650
##
## Category Accuracy:
##
##             N            S
##      0.745057     0.706231
```

Fuente: Elaboración propia.

Probabilidad de corte con árboles de decisión: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 87. **Lectura de las matrices de confusión árboles de decisión**

```
prediccion<-readRDS(file = "../prob_corte/resultados/pred_train.rpart.rds")
ttesting<-readRDS(file = "../prob_corte/resultados/clase_train.rpart.rds")
Clase <- ttesting

head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo, se sugiere utilizar una probabilidad de corte de 0.3, esta permite predecir un 72.11% de individuos no morosos, y un 70.02% de los individuos morosos.

Figura 88. **Matriz de confusión modelo de árboles de decisión**

```
## Corte usado para la Probabilidad = 0.3
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 693 268
##     S 101 236
##
## Overall Accuracy: 0.7157
## Overall Error:    0.2843
##
## Category Accuracy:
##
##              N            S
##       0.721124     0.700297
##
```

Fuente: Elaboración propia.

Probabilidad de corte con regresión logística: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 89. **Lectura de las matrices de confusión modelo de regresión logistica.**

```
prediccion<-readRDS(file = "../prob_corte/resultados/pred_train.glm.rds")
ttesting<-readRDS(file = "../prob_corte/resultados/clase_train.glm.rds")
Clase <- ttesting

head(prediccion$prediction)
```

Fuente: Elaboración propia.

Para este modelo, se sugiere utilizar una probabilidad de corte de 0.25, el modelo es capaz de predecir un 70.13% de los individuos no morosos y un 70.32% de los individuos morosos.

Figura 90. **Matriz de confusion utilizando probabilidad de corte, modelo regresión logística**

```
## Corte usado para la Probabilidad = 0.25
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 674 287
##     S 100 237
##
## Overall Accuracy: 0.7018
## Overall Error:    0.2982
##
## Category Accuracy:
##
##             N            S
##      0.701353     0.703264
```

Fuente: Elaboración propia.

Probabilidad de corte con máquinas de soporte vectorial: Una vez generado el modelo se cargan las predicciones y los datos de pruebas para realizar la evaluación:

Figura 91. **Lectura de las matrices de confusión, modelo svm**

```
prediccion<-readRDS(file = "../prob_corte/resultados/pred_train.svm.rds")
ttesting<-readRDS( file = "../prob_corte/resultados/clase_train.svm.rds")
Clase <- ttesting

head(prediccion$prediction)
```

Fuente: Elaboración propia.

Este modelo no presenta índices que sean recomendables tal como se evidencia en la siguiente imagen, la probabilidad de corte de 0.25 permite predecir un 57.64% de los individuos no morosos y un 69.43% de los individuos morosos:

Figura 92. **Matriz de confusión usando probabilidad de corte, modelo svm**

```
## Corte usado para la Probabilidad = 0.25
##
## Confusion Matrix:
##       Pred
## Clase   N   S
##     N 554 407
##     S 103 234
##
## Overall Accuracy: 0.6071
## Overall Error:    0.3929
##
## Category Accuracy:
##
##             N           S
##      0.576483    0.694362
```

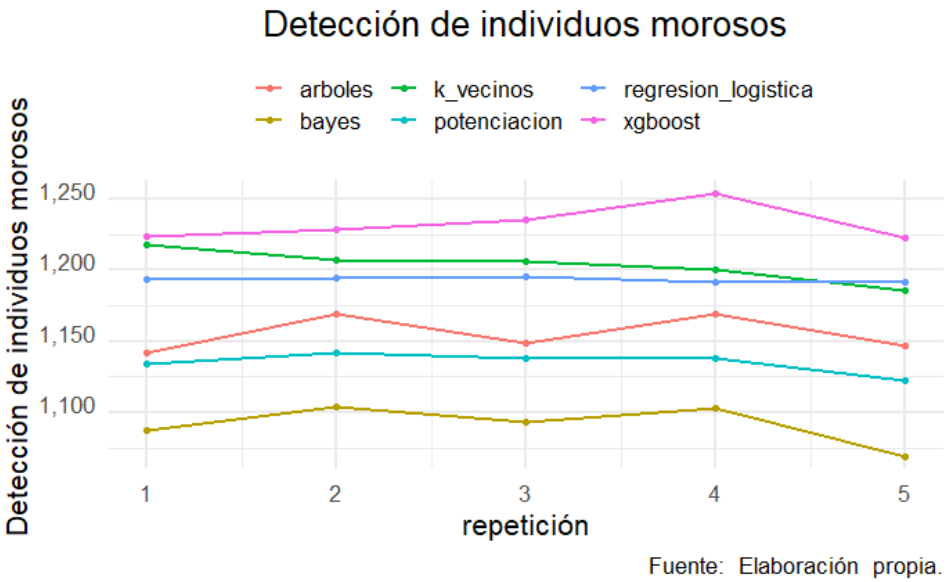
Fuente: Elaboración propia.

2.4.10 Probabilidad de corte y validación cruzada

Con el fin de conocer el verdadero error de los modelos, se generan pruebas con validación cruzada, cuyos resultados se muestran a continuación:

Con respecto a la suma de los individuos que fueron identificados como morosos, se obtiene que el modelo de XGBoosting es el que predice mayor cantidad:

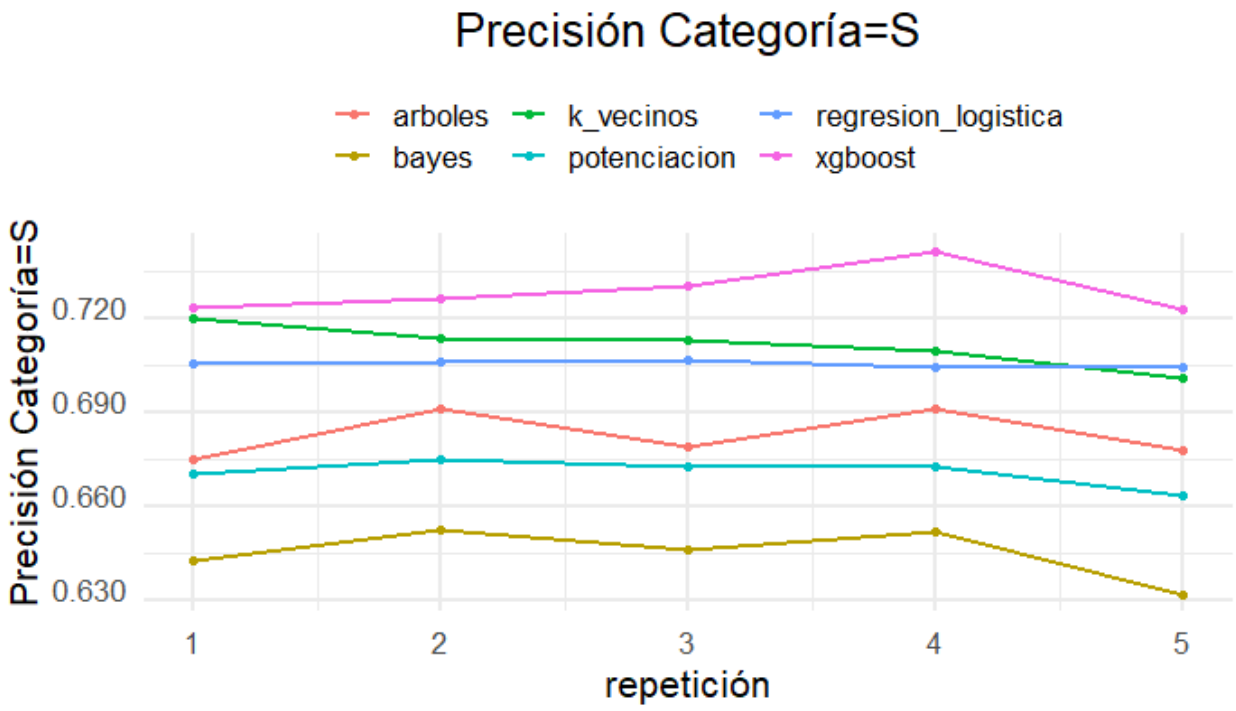
Gráfico 38. **Detección de individuos morosos, evaluación de modelos**



Fuente: Elaboración propia.

Tomando en cuenta la categoría del Sí, se obtiene que el modelo que predice más como es de esperar es XGBoosting y K vecinos.

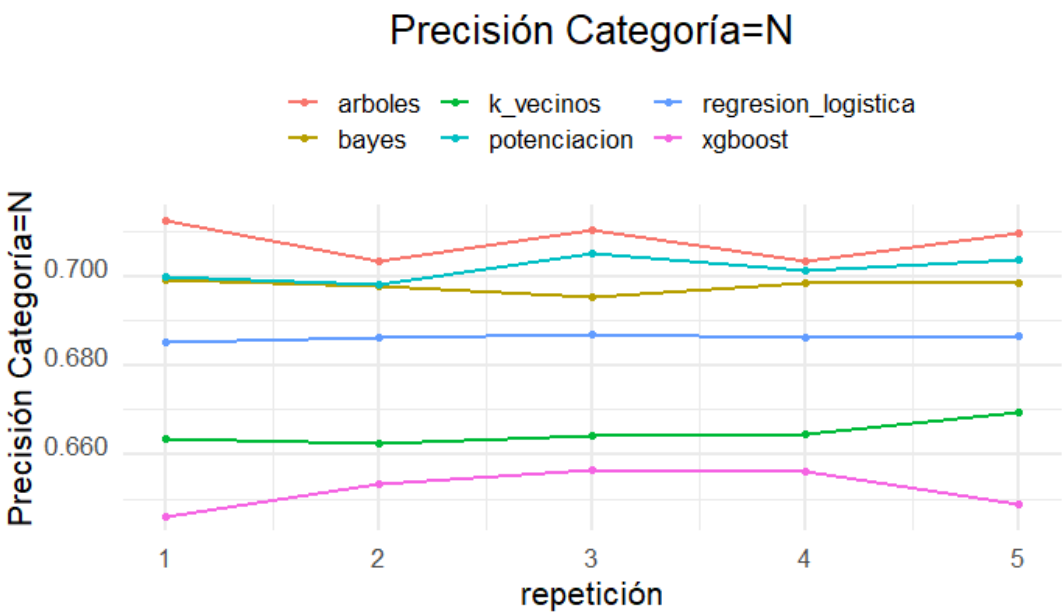
Gráfico 39. **Precisión de la categoría del sí, evaluación de modelos**



Fuente: Elaboración propia.

Si se toma en cuenta la precisión de predicción de la categoría del No, se obtiene que el modelo de árboles de decisión es el que mejor predice esta categoría:

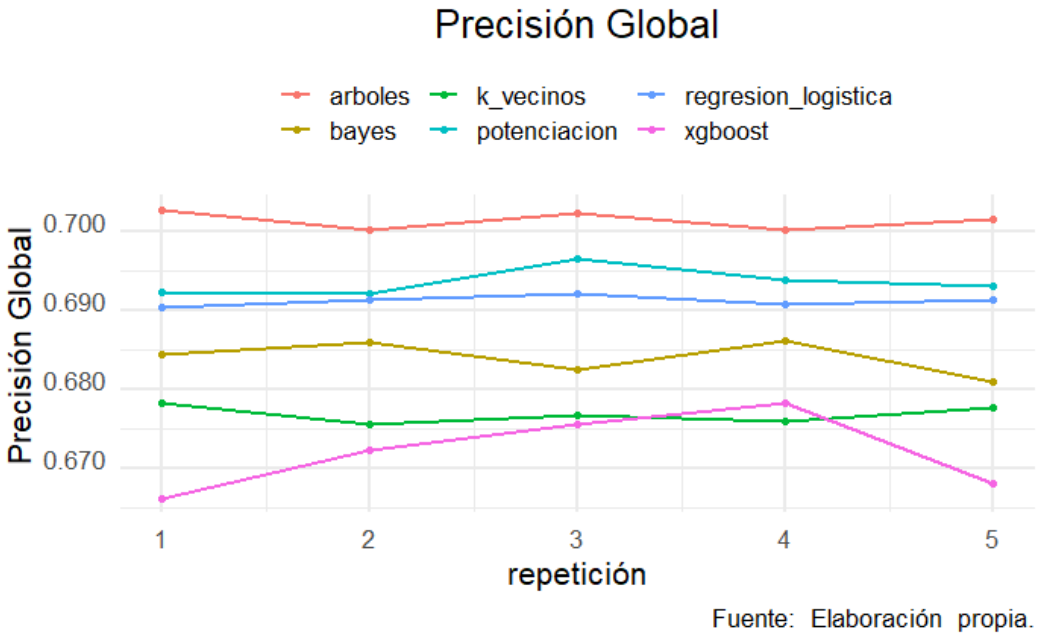
Gráfico 40. **Precisión categoría del no, evaluación de modelos**



Fuente: Elaboración propia.

Con respecto a la precisión global, se obtiene que el modelo de árboles de decisión es el modelo que mejor precisión global presenta:

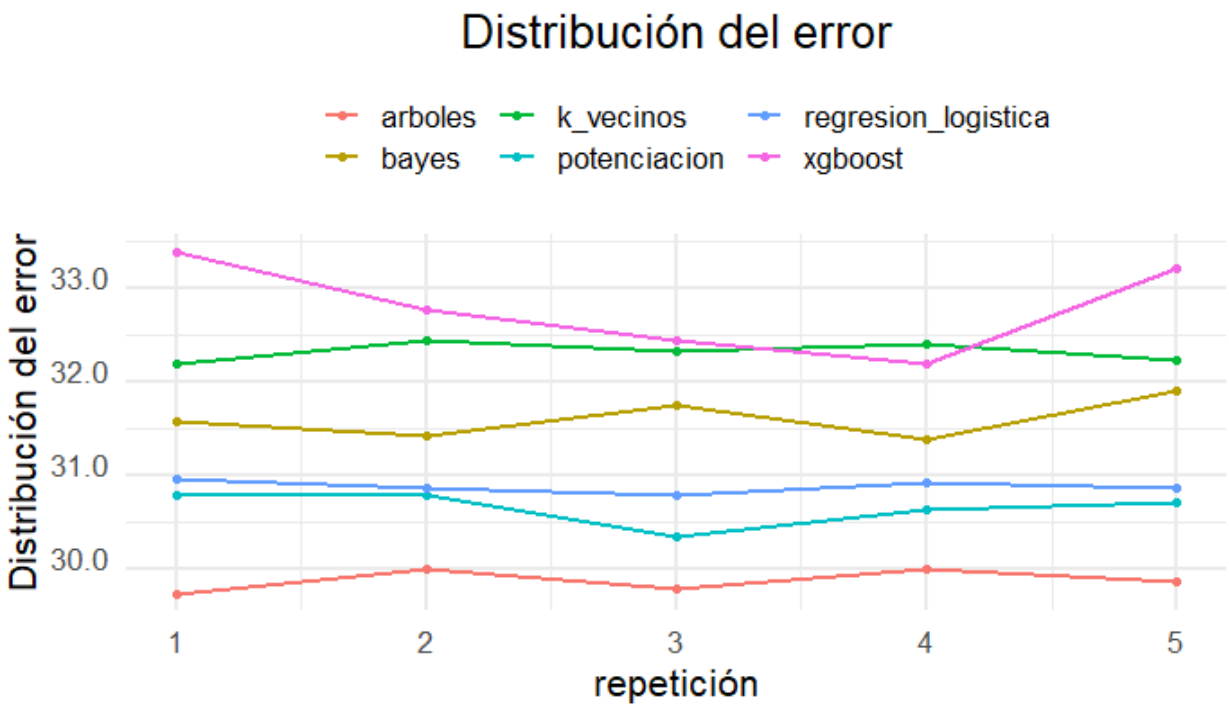
Gráfico 41. **Precisión global, evaluación de modelos**



Fuente: Elaboración propia.

En consideración del error, se obtiene que el modelo que menor error general presenta es el de árboles de decisión:

Gráfico 42. **Distribución del error, evaluación de modelos**



Fuente: Elaboración propia.

2.4.11 Elección del mejor modelo

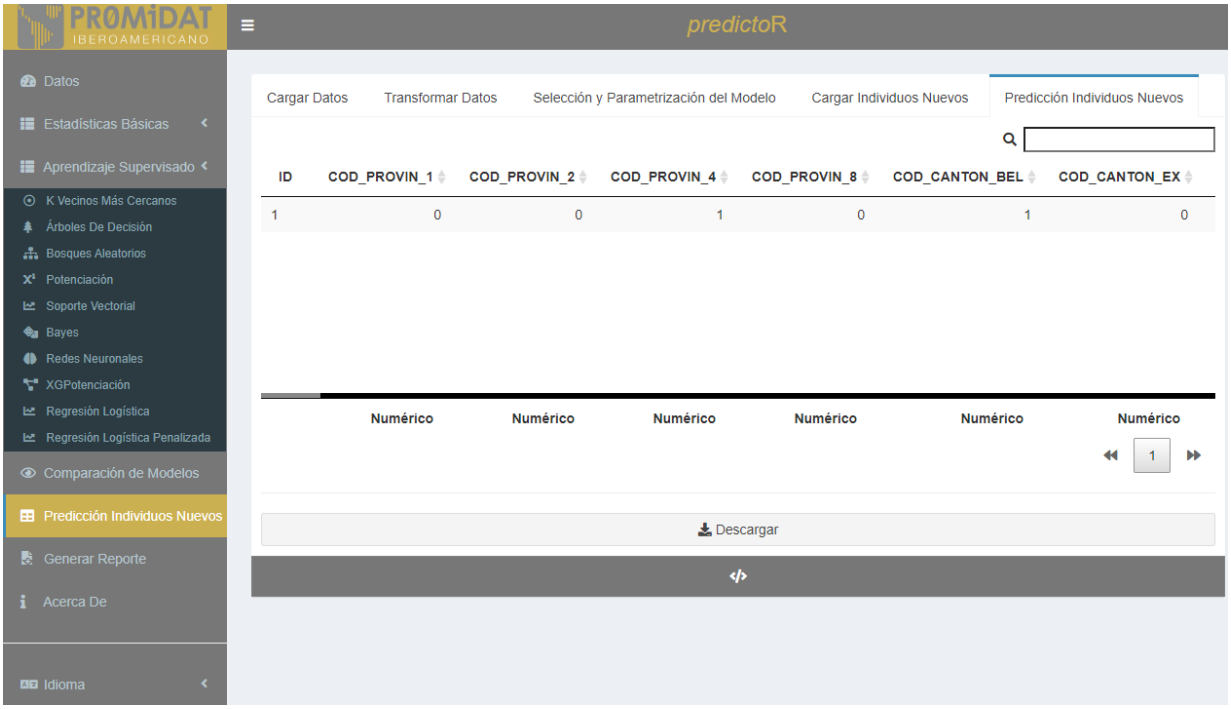
Según los datos obtenidos tanto en las etapas previas como en la presente, se sugiere utilizar el modelo de Bayes, este modelo obtiene la siguiente predicción: utilizando un corte de 0.6, este modelo predice un 69.13% de los individuos que son morosos, y un 69.61% de los individuos que no son morosos, además se propone el modelo de árboles de decisión que obtiene resultados de un 70% de predicción para ambas categorías.

2.4.12 Predicción de nuevos individuos

Por facilidad del negocio, se recomienda predecir nuevos individuos utilizando el paquete predictoR disponible en el CRAN de R; para instalar el paquete se debe ejecutar la siguiente instrucción:

Install.packages(‘predictoR’, dependecies=TRUE)

Figura 93. **Predicción de nuevos individuos en predictoR**



Fuente: Elaboración propia.

Como se ha podido observar, en el paquete no se utiliza la probabilidad de corte, por lo cual se facilita el código para predecir nuevos individuos utilizando el lenguaje R:

Figura 94. **Script de predicción de nuevos indiduos**

```
suppressMessages(library(rattleR))
suppressMessages(library(rpart))
#Lectura del set de datos limpio
datos <- readRDS(file = "./datos/datos_limpios.rds")
datos$MOROSO <- factor(datos$MOROSO, levels = c("N", "S"))
str(datos)
# Se genera el modelo con todos los datos
modelo<-train.bayes(datos,formula=MOROSO~.)
modeloarboles<-train.rpart(datos,formula=MOROSO~.,control = rpart.control(minsplit = 2, maxdepth = 30),
                           parms = list(split = "gini"))

# Se lee el set de individuos nuevos y se recodifican las variables requeridas
datos_ind_nuevos <- read.table("./datos/nuevos-ind_predecir.csv",stringsAsFactors = T,
                               header=TRUE, sep=';', dec = ',')
datos_ind_nuevos$ESTADO_CIVIL<-as.factor(datos_ind_nuevos$ESTADO_CIVIL)
datos_ind_nuevos$TIPO_RELACION<-as.factor(datos_ind_nuevos$TIPO_RELACION)
# Estadísticas básicas
dim(datos_ind_nuevos)
str(datos_ind_nuevos)
datos_ind_nuevos
# Se predicen los nuevos individuos
prediccion_prob <- predict(modelo, datos_ind_nuevos, type = "prob")
prediccion_prob_arboles <- predict(modeloarboles, datos_ind_nuevos, type = "prob")

prediccion_prob
prediccion_prob_arboles
# Se obtienen los indices de predicción de la segunda categoria: S
Score <- prediccion_prob$prediction[, 2]
Score

Scorearb <- prediccion_prob_arboles$prediction[, 2]
Scorearb
```

Fuente: Elaboración propia.

3. Validación de la propuesta

En la presente propuesta de solución los diferentes algoritmos de aprendizaje supervisado se han evaluado utilizando técnicas de validación cruzada o cross-validation, calibración de modelos, procesamiento en paralelo y evaluando la predicción por clase y predicción por probabilidad, de esta forma se ha logrado obtener los modelos que obtienen mejor predicción para el problema descrito a lo largo del proyecto los cuales son el modelo de Bayes t el modelo de árboles de decisión, el sumarizado de estos resultados se pueden observar en la siguiente tabla:

Tabla 14. Resultado de los modelos según la probabilidad de corte

Modelo	Predicción SI	% Predicción No	Probabilidad de corte
Bayes	69.13%	69.61%	0.6
Regresión logística	70.32%	70.13%	0.25
Árboles de decisión	70.02%	72.11%	0.3
XGBoosting	70.91%	66.38%	0.2
K vecinos	72.70%	68.15%	0.25
Potenciación	68.84%	69.19%	0.25

Fuente: Elaboración propia.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

1. Conclusiones

Siendo la minería de datos un área de conocimiento muy importante en el proceso de darle un significado y valor agregado a los datos, se ha brindado en este proyecto un modelo de aprendizage supervisado que permite la predicción de individuos morosos, de esta manera se ha logrado brindar una propuesta de mejora en el afan de solucionar los problemas de morosidad que se enfrentan en la municipalidad.

Cabe mencionar que la fase de obtención de datos, así como el proceso de comprensión de estos, han sido de los puntos que requirieron mayor tiempo en la elaboración y conclusión del presente proyecto. En la etapa de compresión y mediante el análisis de los datos se identificaron debilidades en los datos suministrados por la municipalidad entre los que se pueden mencionar la existencia de campos que no fueron asignados de forma correcta, existe además una deficiencia en la calidad de datos ya que hay valores faltantes conocidos comúnmente como datos nulos. A pesar de estos inconvenientes, fue posible obtener la información valiosa de las personas como edad, sexo, lugar de nacimiento, entre otras, esto gracias a la cooperación de entidades como el Tribunal Supremo de Elecciones y Registro Civil lo cual aportó un gran valor y significancia a los datos, además de facilitar la obtención de valores para variables predictoras con campos incompletos. Por otro lado, se realizó una implementación de conocimiento muy importante en cuanto a técnicas y uso de herramientas de manipulación y análisis de datos.

Resulta importante mencionar que, como parte del proceso de minería de datos, y al requerir procesamientos y cálculos computacionales complejos, fue necesario la implementación de paquetes en el lenguaje R, además del uso de técnicas de programación como lo es la programación en paralelo esto con el fin de disminuir los tiempos de procesamiento y agilizar la generación y evaluación de los modelos evaluados.

Finalmente, siendo el valor principal de este proyecto, se logró obtener un modelo de minería de datos que es capaz de predecir hasta un 70% en cada categoría de la variable

a predecir haciendo notar que la debida implementación y uso del mismo generaría un valor agregado a la muncipalidad, ya que permite mejorar la recaudación de impuestos municipales ayudando de esta forma tener una mayor disposición de recursos monetarios para afrontar los proyectos actuales y futuros, logrando de tal forma mejorar la gestión municipal que propicia una mejor calidad de vida a los habitantes del cantón, siendo este uno de los mayores objetivos trazados por las entidades municipales, proveyendo lo anterior un valor agregado a la presente solución; además, esta solución podría representar una solución innovadora en el modelo de negocio de la municipalidad, representando un punto de partida en la apertura en la ciencia de datos y aplicación de ésta en los procesos de toma de decisiones.

2. Limitaciones

A continuación, se definen las limitaciones encontradas en este proyecto:

- En este proyecto se utilizaron únicamente las siguientes bases de datos: base de datos del Sistema Integrado Municipal, base de datos de afiliados de la municipalidad, padrón electoral del Tribunal Supremo de Elecciones, consulta de personas por cédula del Tribunal Supremo de Elecciones (para validar información de personas), base de datos de nacimientos, defunciones y matrimonios del Registro Civil.
- El presente modelo toma en consideración solamente personas físicas contribuyentes de la municipalidad cuyos datos fueron posible validar con la base de datos del Tribunal Supremo de Elecciones.
- Como parte del proceso de minería de datos se descartaron algunas variables que no aportaban valor o significancia al modelo propuesto.
- No se recomendará equipo de software ni hardware a la organización patrocinadora para la futura implementación del modelo.
- La implementación del presente modelo será brindado a la municipalidad con el fin de que ellos sean los encargados de ajustarlo a sus modelos de desarrollos e intereses organizacionales.

3. Trabajos futuros

Consideramos que este proyecto brinda la posibilidad de incursionar en el área de datos y además permite visualizar el uso que se le puede dar a la minería de datos en un ambiente organizacional, se considera además que con la solución propuesta se facilita la apertura en el ámbito tecnológico tanto a la municipalidad como en las propuestas de aprendizaje educativo y profesional en el país, también se genera un gran aporte a la comunidad científica que desee basar sus investigaciones tomando como referencia esta solución.

En las organizaciones gubernamentales de nuestro país existen pocos estudios relacionados a la morosidad en donde se implemente la minería de datos como eje principal de solución, por lo que se abren las puertas a desarrollar e incursionar en estos apartados y mejorar así los procesos internos, externos y de recaudación monetaria en los diferentes entes organizacionales.

REFERENCIAS

Antezana Bustamante, D. A. (2018). Impacto de la implementación de minería de datos en el mantenimiento y análisis de la información catastral en una municipalidad distrital.

Asensio Romero, P. (2012). El libro de la gestión municipal. Editorial Díaz de Santos, S.A.

Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Brachfield, P. J. (2000). Las Leyes Europeas y Españolas Contra La Morosidad Descifradas y su Aplicación Práctica. Lucha Contra La Morosidad.

Bryla, B. (2015). Oracle database 12c DBA handbook. New York: McGraw-Hill Education, p. [5-6].

Cabello, M. V. N. (2010). Introducción a las bases de datos relacionales. Editorial Visión Libros.

Camacho Portillo, I. (1 de marzo de 2015). Técnicas de negociación con clientes morosos. Madrid, Madrid, España.

Contraloría General de la República. (2017). Índice de Gestión Municipal, Resultados del periodo 2017. San José.

Contraloría General de la República. (2018). Índice de Gestión Municipal, Resultados del periodo 2018. San José.

Contraloría General de la República. (2015). Informe de la Auditoría de Carácter Especial Acerca de la Gestión de Cobro de los Tributos Municipales en la Municipalidad de Santa Cruz. San José: Editorial Costa Rica.

DECSA Costa Rica. (09 de marzo de 2019). Yaipan. Obtenido de Yaipan Especialistas en Tecnologías de información: https://www.yaipan.com.

Field, A., Miles, J., & Field, Z. Discovering Statistics Using R (2012).

Gala, A. (2008). "Consejos para evitar la morosidad". Revista La Gaceta de los Negocios. N° 9.

Grus, J. (2019). Data science from scratch: first principles with python. O'Reilly Media.

López, C. P. (2007). Minería de datos: técnicas y herramientas. Editorial Paraninfo.

Machine Learning, Data Science, Big Data, Analytics, AI. (12 de junio de 2021). KD nuggets. Obtenido de: https://www.kdnuggets.com.

Matich, D. J. (2001). Redes Neuronales: Conceptos básicos y aplicaciones. Universidad Tecnológica Nacional, México.

Medina Merino, R. F., & Ñique Chacón, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, (010), 165-189.

Municipalidad de Belén, Costa Rica. (29 de mayo de 2021). Informe de labores 2019, Municipalidad de Belén. Obtenido de página oficial de la municipalidad: https://www.belen.go.cr.

Oldemar, R. [Oldemar Rodriguez]. (2013, agosto 13). Clase No .1 Minería de Datos [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=pReo00nAG4c&t=3586s.

Picazo Giménez, G. D. (1 de enero de 1994). La Mora del Deudor. Madrid, Madrid, España. Recuperado el 06 de abril de 2019.

Procuraduría General de la república. (2019). Sistema Costarricense de Información Jurídica. [En línea] Disponible en: http://www.pgrweb.go.cr [Obtenido 08 marzo. 2019].

Procuraduría General de la República. (23 de diciembre de 2011). Sistema Costarricense de información Jurídica. Obtenido de Procuraduría General de la Républica web site: http://www.pgrweb.go.cr/.

Rodríguez, Ó. (21 de Julio de 2014). Penurias financieras aquejan a 38 municipalidades. La Nación, p.1.

Salas, R. (2004). Redes neuronales artificiales. Universidad de Valparaıso. Departamento de Computación, 1.

Solomon, S., Nguyen, H., Liebowitz, J., & Agresti, W. (2006). Using data mining to improve traffic safety programs. Industrial Management & Data Systems, 106(5), 621-643.

Tribunal Supremo de Elecciones. (16 de marzo de 2021). Consulta de personas. Obtenido del Tribunal Supremo de Elecciones, sitio web: https://www.consulta.tse.go.cr/consulta_persona/consulta_cedula.aspx.

Han, J. and Kamber, M. (2012). Data mining. 3rd ed. Haryana, India: Elsevier.

Bates, D., Bengtsson, H., & Bivand, R. (s.f.). El Proyecto R para Computación Estadística. Obtenido de El Proyecto R para Computación Estadística: https://www.r-project.org/

Parmer, J., Parmer , C., & Johnson, A. (2013). Plotly. Obtenido de Plotly: https://plotly.com/about-us/

Wickham, H. (s.f.). Tidyverse. Obtenido de Tidyverse: https://tidyverse.tidyverse.org/index.html

ANEXOS

Anexo 1

Lista de servicios Municipalidad

AUX_CONTAB	COD_SERVIC	DES_SERVIC
PAT	COM	PATENTE COMERCIAL
CUF	AMN	AMNISTIA
PAT	LIE	PAT.LIC.EXT
PAT	TLB	TIMB.PRO.PARQUE
PAT	LIC	PAT.LICORES
PAT	MPT	MULTA POR DECLARACION TARDIA
PAT	NCI	NOTA DE CREDITO INTERNA
CUF	SPC	IMPUESTO PERMISO CONSTRUCCION
CUF	INT	INTERESES SERVICIOS CUF
PAT	INT	INTERESES IMPUESTO PAT
CEM	INT	INTERESES SERVICIOS
LIC	LIC	LICENCIA PARA EXPENDIO DE BEBIDAS ALCOHOLICAS
LIC	T77	TIMBRE PRO PARQUES NACIONALES
CUF	MIC	MULTA IMPUESTO CONSTRUCCION
CUF	AGF	SERV. AGUA FIJO
PAT	P21	carga de datos
PAT	P17	ALQUILER DE LOCALES
PAT	P19	MULTA POR DECLARACION TARDIA
CEM	INH	SERVICIO DE INHUMACIÓN
CEM	EXH	SERVICIO DE EXHUMACIÓN
PAT	RCO	AJ. POS PATENTE COMERCIAL
PAT	RTL	REAJUSTE + TIMB.PRO.PARQUE
CUF	GES	SERVICIOS AMBIENTALES

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	BAI	RECOLECCION INFECTO CONTAGIOSO
CUF	IBI	IMPUESTO DE BIENES INMUEBLES
CUF	SRA	SERVICIO RECAUDACION APORTES P-CONVENIO
CUF	AMA	ALQUILER DE MAQUINARIA Y EQUIPOS
PAT	NCO	AJ. NEG. PATENTE COMERCIAL
PAT	NTL	REAJUSTE - TIMB.PRO.PARQUE
CUF	S15	SERV.AMBIENTAL DOMICILIARIO-VIEJO
PAT	ABO	APORTE 10% ENTRADAS BALNEARIO
CUF	SIA	SERVICIO INST Y DERIVACION DE AGUA
CUF	OSC	OTROS SERVICIOS COMUNITARIOS
CUF	SPI	SERVICIO DE PUBLICIDAD E IMPRESION
CUF	VOS	VENTA DE OTROS SERVICIOS
PAT	P03	IMP. ESPECTACULOS PUBLICOS
PAT	P06	PATENTES LICORES NAC. Y EXT.
PAT	P07	PATENTES DE LICORES NACIONALES
PAT	P08	PATENTES LICORES EXTRANJEROS
PAT	P12	PATENTE TIENDA LICORES
CEM	DER	DERECHO CEMENTERIO
CUF	LVP	LIMPIEZA DE VIAS Y SITIOS PUBLICOS
CUF	MPO	MANT. PARQ. Y OBRAS DE ORNATO
PAT	T77	TIMBRE PRO PARQUES NACIONALES

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	CPA	CONT.CUIDADOS PALIATIVOS BELEN
CUF	CCR	CONTRIBUCION CRUZ ROJA
CUF	RBR	RECOLECCION DE BASURA RECICLABLE
CUF	MEI	MEDIDOR INDUSTRIAL
CUF	ALO	ALQUILER DE LOCALES
CUF	AGU	SERV. AGUA POTABLE
CUF	REC	MULTA POR RECONEXION DE HIDROMETRO
CEM	ALQ	ALQUILER CEMENTERIO
PQT	BE2	BOLETA ESTACIONAMIENTO 1 HORA
PQT	BE1	BOLETA ESTACIONAMIENTO 1/2 HORA
PQT	DE1	DESC. ESTACIONAMIENTO 1/2 HORA
PQT	DE2	DESC. ESTACIONAMIENTO 1 HORA
CEM	CEM	DERECHO CEMENTERIO
CEM	MAN	MANTENIMIENTO CEMENTERIO
CUF	BAS	RECOLECCION RESIDUOS SOL. Y VALORIZABLES
PAT	LIM	PATENTES LICORES NAC. Y EXT.
PAT	LIN	PATENTES DE LICORES NACIONALES
CUF	ALC	ALCANTARILLADO SANITARIO Y PTAR
PAT	ROT	IMP. ROTULOS ANUNC. Y VALLAS
CUF	REM	REMATES Y CONFISCACIONES
CUF	OIV	OTROS INGRESOS VARIOS NO ESPECIF.
CUF	RAG	AJ. POS. SERV. AGUA POTABLE

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	NAG	AJ. NEG. SERV. AGUA POTABLE
CUF	NCI	NOTA DE CREDITO INTERNA
PAT	IND	PATENTE INDUSTRIAL
CUF	TCO	TIMBRE CONSTRUCCION DE SOCIEDADES
CUF	AIF	APORTE IFAM IMP. RUEDO LEY 6909-83
CUF	OEP	OTROS IMPUESTOS SOBRE E.P.
CUF	IAP	INTERESES POR ARREGLO DE PAGO
CUF	RPA	AJ. POS. CONT.CUID. PALIATIV. BELEN
CUF	NPA	AJ. NEG. CONT.CUID. PALIATIV. BELEN
LIC	RLI	AJ. POS PATENTE LICORES
LIC	NLI	AJ. NEG PATENTE LICORES
CUF	RSP	AJ. POS. IMP. PERMISO CONSTRUCCION
PAT	IAP	INTERESES POR ARREGLO DE PAGO
CEM	CCJ	COSTAS P/COBRO JUDICIAL
CUF	APF	APORTE INST. PUBLICAS FINANCIERAS
CUF	PDC	DOCUM.POR PERDIDA Y DETERIORO BIENES CON
PAT	CTL	CANON ARRENDAMIENTO AREAS PUBLICAS
PAT	AMN	AMNISTIA
CEM	AMN	AMNISTIA
LIC	AMN	AMNISTIA

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	RBA	AJ. POS. REC. RESID. SOL. Y VALORIZABLES
CUF	NBA	AJ. NEG. REC. RESID. SOL. Y VALORIZABLES
CEM	RDE	AJ. POS. DERECHO CEMENTERIO
CEM	NDE	AJ. NEG. DERECHO CEMENTERIO
CEM	RMA	AJ. POS. MANTENIMIENTO CEMENTERIO
CEM	NMA	AJ. NEG. MANTENIMIENTO CEMENTERIO
CUF	RAL	AJ. POS. ALCANT. SANIT. Y PTAR
CUF	NAL	AJ. NEG. ALCANT. SANIT. Y PTAR
CUF	RLV	AJ. POS. LIMPIEZA DE VIAS Y SITIOS PUBL.
CUF	NLV	AJ. NEG. LIMPIEZA DE VIAS Y SITIOS PUBL.
CUF	RGE	AJ. POS. SERVICIOS AMBIENTALES
CUF	NGE	AJ. NEG. SERVICIOS AMBIENTALES
CUF	EBA	APORTE VOLUNTARIO EBAIS
HID	NCI	NOTA DE CREDITO INTERNA
CUF	MID	MULTA POR IMCUMP. ART.85 CM
CUF	TPP	TIMBRE PRO PARQUES INCISO B
CUF	TTI	TIMBRE TRASPASO IBI
CUF	TMU	TIMBRES MUNICIPALES
CUF	MLT	MULTAS ART.234 INC.D LEY TRANSITO
CUF	MPQ	MULTAS POR INFRACCION LEY PARQUIM

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	PZV	DEFINIR SERVICIO PZV
CEM	IAP	INTERESES POR ARREGLO DE PAGO
PAT	OEP	IMP. ESPECTACULOS PUBLICOS
PAT	CCJ	COSTAS P/COBRO JUDICIAL
PQT	INT	INTERESES SERVICIOS PARQUIMETROS
PAT	CPP	CONT.CUIDADOS PALIATIVOS BELEN
LIC	RTL	REAJUSTE + TIMB.PRO.PARQUE
LIC	NTL	REAJUSTE - TIMB.PRO.PARQUE
CUF	IVA	IMPUESTO VALOR AGREGADO
CUF	RIV	AJ. POS. IMPUESTO VALOR AGREGADO
CUF	NIV	AJ. NEG. IMPUESTO VALOR AGREGADO
CUF	RIB	AJ. POS. IMPUESTO DE BIENES INMUEBLES
CUF	NIB	AJ. NEG. IMPUESTO DE BIENES INMUEBLES
CUF	RMP	AJ. POS. MANT. PARQ. Y OBRAS DE ORNATO
CUF	NMP	AJ. NEG. MANT. PARQ. Y OBRAS DE ORNATO
CUF	RAF	AJ. POS. SERV. AGUA FIJO
CUF	NAF	AJ. NEG. SERV. AGUA FIJO
LIC	NCI	NOTA DE CREDITO INTERNA
CEM	NCI	NOTA DE CREDITO INTERNA
PAT	EXP	IMP. EXPLOTACION DE CANTERAS
LIC	INT	INTERESES IMPUESTO LIC

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	THI	TIMBRE HIPOTECAS Y CEDULAS HIPOTECA
PAT	RIR	AJ. POS IMP. ROTULOS
PAT	NIR	AJ. NEG IMP. ROTULOS
CUF	MRA	MULTA SEGUN ART 41 REGL.ACUEDUCTO
CUF	IST	INTERESES S/ TITULOS INST.PUB. FINANC
CUF	RST	REC. SIMPLIFICACION TRIBUTARIA LEY 8114
CUF	SRT	SANCION RESARCITORIA A TERCEROS
CUF	ASP	APORTE SECTOR PRIVADO (TRANS.CORRI)
CUF	NSP	AJ. NEG. IMP. PERMISO CONSTRUCCION
PQT	COM	COMISIÓN RECAUDACION INS SICSOA
CUF	APJ	APORTE CONS. NAC.PERSONA JOVEN
CUF	AMH	APORTE MINISTERIO HACIENDA PART. ESPEC.
CUF	AIL	APORTE IFAM. LICORES LEY 6796
PAT	NMP	AJ. NEG. DECLARACION TARDIA
PAT	RMP	AJ. POS. DECLARACION TARDIA
CUF	IN2	INTERESES IMPUESTOS CUF
CUF	IOP	INCUMPLIMIENTO ART. 84 Y 85 CM
CUF	DET	DERECHO ENTRADAS TEATRO NACIMIENTO

AUX_CONTAB	COD_SERVIC	DES_SERVIC
CUF	DES	DERECHO Y ESTACIONAMIENTO Y TERMIN
CUF	RCR	AJ. POS. CONT. CRUZ ROJA
CUF	NCR	AJ. NEG. CONT. CRUZ ROJA
PQT	INF	MULTAS POR INFRACCION LEY PARQUIM
PAT	RLI	AJ. POS PATENTE LICORES
PAT	NLI	AJ. NEG PATENTE LICORES
CUF	CCJ	COSTAS P/COBRO JUDICIAL
PAT	NPP	AJ. NEG. CONT.CUID. PALIATIV. BELEN
PAT	RPP	AJ. POS. CONT.CUID. PALIATIV. BELEN
CUF	ASC	APORTE SECTOR PRIVADO (TRANS.CAPITA)
CUF	NIN	AJ. NEG. INTERESES SERVICIOS CUF
CUF	RIN	AJ. POS. INTERESES SERVICIOS CUF

Anexo 2

Comando de importación del respaldo de base de datos Oracle:

impdp \'/ as sysdba\' directory=BACKUPS DUMPFILE=ORCL_28-01-21.dmp LOGFILE=IMP-ORCL-DEC-06-02-21.log schemas=DEC remap_tablespace=TAB_01:SUB,TAB_02:SUB,TAB_03:SUB,TAB_04:SUB,TAB_05:SUB

Anexo 3

Utilidad de conteo de filas de las tablas utilizadas.

```
-- crea la tabla requerida
CREATE TABLE MINERIA.CONTEODATOS AS SELECT OWNER, OBJECT_NAME, OBJECT_TYPE, 0 AS REGISTROS
    FROM DBA_OBJECTS WHERE 1=2;

-- query para obtener la sentencia de insert

 SELECT 'INSERT INTO MINERIA.CONTEODATOS VALUES('||CHR(39)||OWNER||CHR(39)||','||CHR(39)||OBJECT_NAME||CHR(39)||','|
 CHR(39)||OBJECT_TYPE||CHR(39)||','||'(SELECT COUNT(*) FROM '||OWNER||'.'||OBJECT_NAME||'));'
    FROM DBA_OBJECTS
   WHERE OBJECT_NAME IN ('COM_PERSON','COM_INFPER','CUF_TTRAIN','CUF_PREPTO',
                         'CUF_CERTIF','COM_PERDET','CUF_HIDROM','CUF_PROPIE',
                         'CUF_CUENTAS','CUF_CTAAUX','CUF_TARIFA','CUF_TARMED',
                         'CUF_PERCOS','CUF_PERCOB','CUF_PERCOH','CUF_PATLIC',
                         'CUF_PATENT','CUF_PREPTN','PADRON','AFILIADOS',
                         'TSE_NACIMIENTOS','TSE_MATRIMONIOS','TSE_DEFUNCIONES','DIST_ELECT',
                         'CUF_INMALF','CUF_AVALUO','CUF_CONSFI','CUF_CEMENT',
                         'CUF_DECBIE','CUF_SEROCU','CUF_CUENTA','V_SERVICIO','V_CUF_CERTIF','V_MOROSOS',
                         'V_CUENTA_SERVICIO','V_SERVICIOS_X_CUENTA','V_TARIFA',
                         'V_PERSONA_X_TARIFA','V_AFILIADOS','V_CANTIDAD_CUENTAS','V_NACIMIENTO_TSE',
                         'V_VALORES_PROPIES','V_CANT_PROPIEDADES','MATRIMONIO_X_CED','MATRIMONIOS_HOMBRES',
                         'MATRIMONIO_X_CED_MUJER','MATRIMONIOS_MUJERES','V_HIJOS_MUJERES',
                         'V_HIJOS_HOMBRES','V_HIJOS','V_MATRIMONIOS','V_PATENTE_COMER','V_PATENTE_LIC',
                         'V_CANT_PAT_COMER','V_CANT_PAT_LIC','PROPIE_SIN_SENAS','V_SERVICIO_CEM')
                         AND STATUS='VALID'
ORDER BY OBJECT_TYPE;
```

Anexo 4

Sentencia SQL para extraer los individuos que conforman el conjunto de datos utilizado.

```
SELECT INF.CEDULA,
    INF.TIP_PERSON,
    INF.COD_PROVIN,
    INF.COD_CANTON,
    INF.COD_DISTRI,
        CASE NVL (PAD.CODELEC, 'NV') WHEN 'NV' THEN 'NV'
                WHEN '407001' THEN 'SAN ANTONIO'
                WHEN '407002' THEN 'LA RIBERA'
                WHEN '407003' THEN 'LA ASUNCION'
                WHEN '408003' THEN 'LLORENTE'
```

```
                    WHEN '408001' THEN 'SAN JOAQUIN' ELSE 'OTRO'
END AS VOTO_DISTRITO,
          CASE NVL (PAD.PROVINCIA, 'NV') WHEN 'NV' THEN 'NV'
                    WHEN 'HEREDIA' THEN PAD.PROVINCIA
                    WHEN 'ALAJUELA' THEN PAD.PROVINCIA
                    WHEN 'SAN JOSE' THEN PAD.PROVINCIA
                    WHEN 'PUNTARENAS' THEN PAD.PROVINCIA
                    ELSE 'OTRO' END AS V_PROVINCIA,
         CASE NVL (PAD.CANTON, 'NV') WHEN 'NV' THEN 'NV'
                    WHEN 'BELEN' THEN PAD.CANTON
                    WHEN 'CENTRAL' THEN PAD.CANTON
                    WHEN 'FLORES' THEN PAD.CANTON
                    WHEN 'SANTA ANA' THEN PAD.CANTON
                    WHEN 'ESTADOS UNIDOS' THEN PAD.CANTON
                    ELSE 'OTRO' END AS V_CANTON,
    CASE NVL (VPXT.COD_TARIFA_REP, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
      AS COD_TARIFA_REP,
    CASE NVL (VPXT.COD_TARIFA_COMER2, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
      AS COD_TARIFA_COMER2,
    CASE NVL (VPXT.COD_TARIFA_INDUST, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
      AS COD_TARIFA_INDUST,
    CASE NVL (VPXT.COD_TARIFA_RESID, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
      AS COD_TARIFA_RESID,
    CASE NVL (VPXT.COD_TARIFA_DOM, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
      AS COD_TARIFA_DOM,
```

```
      --CASE NVL (VPXT.COD_TARIFA_RES, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
       --  AS COD_TARIFA_RES,
      CASE NVL (VPXT.COD_TARIFA_COMER3, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_COMER3,
      CASE NVL (VPXT.COD_TARIFA_SOCIAL, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_SOCIAL,
      CASE NVL (VPXT.COD_TARIFA_IND, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_IND,
      --CASE NVL (VPXT.COD_TARIFA_T1, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
       --  AS COD_TARIFA_T1,
      CASE NVL (VPXT.COD_TARIFA_COMER1, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_COMER1,
      CASE NVL (VPXT.COD_TARIFA_SOC, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_SOC,
      --CASE NVL (VPXT.COD_TARIFA_FRANCA, 'NA') WHEN 'NA' THEN
'N' ELSE 'S' END
      --   AS COD_TARIFA_FRANCA,
      CASE NVL (VPXT.COD_TARIFA_PRE, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_PRE,
      CASE NVL (VPXT.COD_TARIFA_ORD, 'NA') WHEN 'NA' THEN 'N'
ELSE 'S' END
         AS COD_TARIFA_ORD,
```

```
    CASE NVL (COD_SERVIC_MPO, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_MPO,
    CASE NVL (COD_SERVIC_LVP, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_LVP,
    CASE NVL (COD_SERVIC_AGU, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_AGU,
    CASE NVL (COD_SERVIC_IBI, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_IBI,
    CASE NVL (COD_SERVIC_BAS, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_BAS,
    CASE NVL (COD_SERVIC_PZV, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_PZV,
            CASE NVL (VCPC.N_PAT_COMER,0) WHEN 0 THEN 'N'
ELSE 'S' END
      AS COD_SERVIC_PAT,
            CASE NVL (VCPL.N_PAT_LIC,0) WHEN 0 THEN 'N' ELSE 'S'
END
      AS COD_SERVIC_LIC,
            CASE NVL (vsc.NUM_PERSON,-1) WHEN -1 THEN 'N' ELSE
'S' END
      AS COD_SERVIC_CEM,
    VPRO.SUM_MONTO_FINCA,
    VPRO.SUM_MONTO_IMPONIBLE,
    NAC.EDAD,
```

```
    CASE NVL (NAC.SEXO , 'NI') WHEN '1' THEN 'M' WHEN '2' THEN 'F'
ELSE 'NI' END as SEXO,
    NVL (VH.N_HIJOS, 0) AS N_HIJOS,
    NVL (VM.ESTADO_CIVIL, 0) AS ESTADO_CIVIL,
    NVL (VM.TIPO_RELACION, 0) AS TIPO_RELACION,
    CUE.CANTIDAD AS CANT_CUENTAS,
    PROPS.N_PROPIEDADES,
        NVL (VCPC.N_PAT_COMER,0) AS N_PAT_COMER,
        NVL (VCPL.N_PAT_LIC,0) AS N_PAT_LIC,
    CASE NVL (AF.IDENTIFICACION, 'NA') WHEN 'NA' THEN 'N' ELSE 'S'
END
      AS IND_AFILIADO,
        CASE NVL (PSS.NUM_PERSON, 0) WHEN 0 THEN 'S' ELSE 'N'
END AS PROP_SENAS,
        CASE NVL (vcc.c_construc, 0) WHEN 0 THEN 'N' ELSE 'S' END AS
CONSTRUC_FINCA,
        CASE NVL (vcc.c_construc, 0) WHEN 0 THEN 0 ELSE vcc.c_construc
END AS CONTRUCCIONES_FINCA,
    CASE NVL (CERT.NUM_PERSON, 0) WHEN 0 THEN 'N' ELSE 'S' END
AS MOROSO
 FROM DEC.COM_PERDET INF
        LEFT JOIN MINERIA.PROPIE_SIN_SENAS PSS ON
INF.NUM_PERSON=PSS.NUM_PERSON
    LEFT JOIN MINERIA.V_PERSONA_X_TARIFA VPXT
      ON (INF.NUM_PERSON = VPXT.NUM_PERSON)
    LEFT JOIN MINERIA.V_SERVICIOS_X_CUENTA VSXC
      ON INF.NUM_PERSON = VSXC.NUM_PERSON
    INNER JOIN MINERIA.V_VALORES_PROPIES VPRO ON INF.CEDULA
= VPRO.CEDULA
```

```
    INNER JOIN MINERIA.V_NACIMIENTO_TSE NAC ON INF.CEDULA =
NAC.CEDULA
    INNER JOIN MINERIA.V_CANTIDAD_CUENTAS CUE
      ON CUE.NUM_PERSON = INF.NUM_PERSON
    INNER JOIN MINERIA.V_CANT_PROPIEDADES PROPS
      ON INF.NUM_PERSON = PROPS.NUM_PERSON
    LEFT JOIN MINERIA.V_AFILIADOS AF ON INF.CEDULA =
AF.IDENTIFICACION
   LEFT JOIN MINERIA.V_MOROSOS CERT
      ON INF.NUM_PERSON = CERT.NUM_PERSON
    LEFT JOIN MINERIA.V_PADRON PAD ON INF.CEDULA =
PAD.CEDULA
    LEFT JOIN MINERIA.V_HIJOS VH ON INF.CEDULA = VH.CEDULA
    LEFT JOIN MINERIA.V_MATRIMONIOS VM ON INF.CEDULA =
VM.CEDULA
        LEFT JOIN MINERIA.V_CANT_PAT_COMER VCPC ON
INF.NUM_PERSON=VCPC.NUM_PERSON
        left join mineria.V_SERVICIO_CEM vsc on
INF.NUM_PERSON=vsc.NUM_PERSON
        LEFT JOIN MINERIA.V_CANT_PAT_LIC VCPL ON
INF.NUM_PERSON=VCPL.NUM_PERSON
        LEFT JOIN mineria.v_cant_constru vcc on
INF.NUM_PERSON=vcc.NUM_PERSON
```