UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE CIENCIAS QUÍMICAS, FÍSICAS Y MATEMÁTICAS

ESCUELA PROFESIONAL DE MATEMÁTICA CON MENCION EN ESTADÍSTICA

TESIS

COMPARACIÓN DE TÉCNICAS DE IMPUTACIÓN DE DATOS DE ICTERICIA PATOLÓGICA DE NIÑOS ATENDIDOS EN EL HOSPITAL REGIONAL DE CUSCO, 2021

PARA OPTAR AL TITULO PROFESIONAL DE LICENCIADO EN MATEMÁTICA, MENCIÓN ESTADÍSTICA

PRESENTADO POR:

BR. MARIO WILLIAM CALANCHA CUBA

ASESOR:

DR. CLETO DE LA TORRE DUEÑAS

CUSCO – PERÚ

2023

INFORME DE ORIGINALIDAD

(Aprobado por Resolución Nro.CU-303-2020-UNSAAC)

El que suscribe, **Asesor** del trabajo de investigación/tesis titulada: COMPARACION DE TECNICAS DE IMPUTACION DE DATOS DE LETALIDAD PATOLOGICA DE NIÑOS ATENDIDOS EN EL HOSPITAL REGIONAL DE CUSCO, 2021

presentado por: Br. Mario William Calancha Cuba con DNI Nro.: 23937158

presentado por: — con DNI Nro.: —

para optar el título profesional/grado académico de TITULO PROFESIONAL DE LICENCIADO EN MATEMATICAS, MENCION ESTADISTICA

Informo que el trabajo de investigación ha sido sometido a revisión por 03 veces, mediante el Software Antiplagio, conforme al Art. 6° del ***Reglamento para Uso de Sistema Antiplagio de la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de 06 %.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No se considera plagio.	X
Del 11 al 30 %	Devolver al usuario para las correcciones.	—
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, quien a su vez eleva el informe a la autoridad académica para que tome las acciones correspondientes. Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	—

Por tanto, en mi condición de asesor, firmo el presente informe en señal de conformidad y **adjunto** la primera página del reporte del Sistema Antiplagio.

Cusco, 22 de AGOSTO de 2022

Firma

Post firma: DR. CLETO DE LA TORRE BUENAS

Nro. de DNI: 23988416

ORCID del Asesor: 0000-0003-0921-7217

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema Antiplagio: oid-27259:160325427

turnitin

Identificación de reporte de similitud: oid:27259:160325427

NOMBRE DEL TRABAJO
TESIS IMPUTACION OK-22 BACH. MARIO CALANCHA (1).docx

AUTOR
MARIO CALANCHA

RECUENTO DE PALABRAS
13175 Words

RECUENTO DE CARACTERES
70331 Characters

RECUENTO DE PÁGINAS
68 Pages

TAMAÑO DEL ARCHIVO
576.7KB

FECHA DE ENTREGA
Aug 5, 2022 7:41 AM GMT-5

FECHA DEL INFORME
Aug 5, 2022 7:44 AM GMT-5

● 6% de similitud general
El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base d6% Base de

- datos de Internet
- Base de datos de Crossref
- 1% Base de datos de publicaciones
- Base de datos de contenido publicado de Crossr

● Excluir del Reporte de Similitud

- Base de datos de trabajos entregados
- Material citado
- Coincidencia baja (menos de 10 palabras)
- Material bibliográfico
- Material citado
- Bloques de texto excluidos manualmente

DR. CLETO DE LA TORRE DUEÑAS
DNI: 23988416

DEDICATORIA

A Dios y la Virgen del Carmen, por ser un refugio en mi vida y saber que sin ellos nada sería posible.

A mi esposa Youglyn Geanina por su amor incondicional y apoyo por estar siempre a mi lado ser el motor que me guía y ayuda a tomar las mejores decisiones que voy afrontando.

A mis hijos Mauricio Joaquín y Paolo André, quienes son el motivo de todo lo que pueda aspirar.

A mis queridos padres por darme la vida, por su dedicación, esfuerzo y trabajo para que sus hijos salgan adelante, por ser también una constante motivación para el logro de mis objetivos.

Mario William Calancha Cuba

AGRADECIMIENTOS

Mi más sincera gratitud a la Universidad Nacional de san Antonio abad del Cusco, por contribuir grandemente en forjar nuestros ideales, dotándonos de la posibilidad de poder tener esta hermosa escuela profesional de Matemáticas mención estadística que se encuentra a la altura de los requerimientos que la sociedad lo requiere

A toda la plana de docentes de la Universidad, que son guías muy importantes, y quienes con sus conocimientos y experiencias motivaron nuestros ideales, soportes académicos importantes para la culminación de la carrera profesional.

A mis dictaminantes por su paciencia, experiencias y capacidades en la investigación, por sus críticas, las que permitieron el desarrollo y la culminación del presente trabajo de investigación.

Por último, mi agradecimiento principal y gratitud a mi asesor Dr. Cleto De la Torre Dueñas por brindar su apoyo permanente e incondicional, en la ejecución de la presente investigación.

Mario William Calancha Cuba

ÍNDICE

Índice de tablas

Índice de figuras

RESUMEN

Son varias décadas en los que se ha venido estudiando la forma de completar espacios vacíos o datos faltantes, con el fin de obtener un conjunto de datos completos para analizarse por la vía de los métodos estadísticos tradicionales, en los últimos años gracias a los continuos avances de la informática se ha hecho posible el surgimiento y puesta en práctica de nuevas metodologías para la imputación de datos faltantes con métodos más sofisticados, por ello el propósito de la presente investigación fue el de determinar cuál es la técnica de imputación de datos que mejor desempeño presenta en el conjunto de datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021.

La presente investigación es básica descriptiva, con un enfoque cuantitativo con diseño no experimental donde la unidad de análisis son los pacientes infantes del Hospital Regional del Cusco, la muestra estuvo conformado por 656 registros de historias de los pacientes infantes del Hospital Regional del Cusco-2021, donde la técnica de recolección de datos e información fue la revisión documental, en específico revisión de historias clínicas.

Para el análisis se utilizó metodologías de completar datos, como completar con la mediana, completar con regresión utilizando predictores y ruidos aleatorios gaussianos, además de la metodología de KNN k vecinos más cercanos, todos ellos implementados en el software libre R y RStudio; así mismo para ver cuáles son los factores con mayor riesgo en el tipo de ictericia se usará la regresión logística de manera superficial como una complementariedad al trabajo de investigación. Las conclusiones a los que se abordaron con la investigación fueron; la imputación de datos de la variable peso en el alta médica del Recién Nacido en términos de la media la técnica que tiene mayor acercamiento es KNN vecinos más cercanos, mientras que en la variable hematocrito en términos de la media, la técnica que tiene mayor acercamiento es la de Regresión con ruido aleatorio; En general las técnicas basadas en regresión presentan mejor performance a la hora de imputar datos faltantes, como es el caso de la presente investigación.

Palabras claves

Imputación, KNN, regresión, regresión logística, ictericia.

ABSTRACT

There are several decades in which we have been studying how to fill in gaps or missing data, in order to obtain a complete set of data to be analyzed by traditional statistical methods, in recent years thanks to the continuing advances in computer science has made possible the emergence and implementation of new methodologies for the imputation of missing data with more sophisticated methods, Therefore, the purpose of this research was to determine which data imputation technique performs best in the data set of pathological jaundice in children treated at the Regional Hospital of Cusco, 2021.

The present research is basic descriptive, with a quantitative approach with a non-experimental design where the unit of analysis is the infant patients of the Regional Hospital of Cusco, the sample consisted of 656 records of infant patient records of the Regional Hospital of Cusco-2021, where the technique of data collection and information was the documentary review, specifically review of medical records.

For the analysis methodologies were used to complete data, such as completing with the median, completing with regression using predictors and Gaussian random noise, in addition to the methodology of KNN k nearest neighbors, all implemented in the free software R and RStudio; likewise, to see which are the factors with the highest risk in the type of jaundice, logistic regression will be used superficially as a complement to the research work. The conclusions reached by the research were: the imputation of data for the variable weight at medical discharge of the newborn in terms of the mean, the technique that has the closest approach is KNN Nearest Neighbors, while for the variable hematocrit in terms of the mean, the technique that has the closest approach is Regression with random noise; in general, the techniques based on regression present better performance when imputing missing data, as is the case in this research.

Key words

Imputation, KNN, regression, logistic regression, jaundice.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. DESCRIPCIÓN DEL PROBLEMA DE INVESTIGACIÓN

En la mayoría de los estudios muestrales o censales, encontramos múltiples obstáculos y entre los más comunes se encuentra perder una medición u observación, lo que genera espacios vacíos, en la estructura de los datos. De hecho, los datos completos constituyen más una excepción a la regla. Esta situación es una severa limitante, puesto que los métodos estadísticos tradicionales están diseñados para ser aplicados sobre conjuntos de datos completos y además las rutinas de los paquetes estadísticos también asumen que se trabaja con datos completos e incorporan opciones que no siempre son las más adecuadas para imputar observaciones sin que el usuario se dé cuenta de ello. Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio.

Desde hace ya varias décadas, se ha venido estudiando la forma de “completar" estos espacios vacíos, con el fin de obtener un conjunto de datos completos para ser analizados por la vía de los métodos estadísticos tradicionales, en los últimos años gracias a los continuos avances de la informática se ha hecho posible el surgimiento y puesta en práctica de nuevas metodologías para el tratamiento de información con datos faltantes, los cuales, en su mayoría, producen resultados aceptables cuando hay pocos valores perdidos. Aun así, todavía son muchas las falencias que enfrentan las técnicas actuales,

como los sesgos en las estimaciones, alteración de la relación entre las variables, cambios en las varianzas, entre otros y a pesar de la variedad de métodos existentes, el problema permanece abierto, sin que hasta ahora parezca haberse hallado una solución definitiva; además esta situación se complica cuando los datos se presentan en una matriz formada por diversas variables sobre la cual se realizan estudios multivariantes, haciéndose necesario la aplicación de métodos que convenientemente imputen conjuntamente los datos.

La ictericia es un problema muy frecuente en los neonatos; es una patología caracterizada por altos niveles de bilirrubina en la sangre y tejidos corporales, lo cual ocasiona una coloración amarillenta en la piel, mucosas y escleras (parte blanca de los ojos). Se aprecia clínicamente cuando la bilirrubina (B) sérica es superior a 5 mg/dl (85 μmol/L) en neonatos (De la Vera & Montenegro, 2022)

Este trastorno es una de las dos entidades clínicas más frecuentes en la edad neonatal (junto con la dificultad respiratoria) y una de la diez primeras causas de morbilidad neonatal en las unidades de cuidados intermedios, 60% a 70% de los neonatos maduros y 80% o más de neonatos inmaduros llegan a padecer algún grado de ictericia según (Rodriguez, 2001) y (Failachea, 2002) **.**

Su incidencia varia ampliamente entre diversas instituciones y en Norte América es aún la causa más común de readmisiones a unidades de cuidados neonatales**.** Lo cual motiva a identificar los factores de riesgo asociados a esta patología. Las causas son varias, y se han relacionado diversos factores de riesgo con el desarrollo de ictericia neonatal: maternos como incompatibilidad sanguínea, amamantamiento, uso de ciertos fármacos, diabetes gestacional; neonatales como el trauma obstétrico, la mala alimentación, prematuro, género masculino (Trejos & Umanzor, 2018).

A pesar de los recientes avances en el tratamiento de este problema, la toxicidad en el sistema nervioso causada por la bilirrubinemia es aún una importante amenaza y tanto el kernicterus (Ictericia Nuclear) como las alteraciones auditivas son secuelas graves, muchas veces incapacitantes, que aún se siguen observando.

Identificar los factores de riesgo asociados a la ictericia neonatal, nos ayudará a poder crear políticas de prevención y asignar tratamientos oportunos y eficientes para tratar la patología. Para determinar los factores de riesgo asociados a la ictericia neonatal se pueden aplicar distintas metodologías estadísticas una de ellas son los modelos de regresión logística, que son modelos más sofisticados con mejor capacidad para el análisis de este tipo de datos clínicos y epidemiológicos, las aplicaciones que tienen los modelos de regresión logística son diversas en distintos campos, una de estas aplicaciones se realizara en el presente trabajo de investigación.

1.2. FORMULACIÓN DEL PROBLEMA

1.2.1. Problema general

¿Cuál es la técnica de imputación de datos que mejor desempeño presenta en el conjunto de datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021?

1.2.2. Problemas específicos

a) ¿Cuál es la técnica de imputación de datos de medidas de tendencia central más adecuado para el tratamiento de datos ausentes en datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021?

b) ¿Cuál es la técnica de regresión o la técnica de vecinos más cercanos que presente mejor desempeño?

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN

Es muy importante tener los datos completos y mucho más importante de darse el caso de contar con una base de datos que no tiene la información completa contar con técnicas adecuadas para el llenado de los datos, los cuales a su vez conducirán a tener estadísticas más fiables para la toma de decisiones, los datos a trabajar en el presente trabajo son los factores de riesgo que influencian en la presencia de la ictericia, si bien es cierto que en la mayoría de los casos disminuye a medida que pasan los días, pero en otros casos se complica si no se trata a tiempo por ejemplo puede dañar al sistema nervioso central, alteraciones auditivas y entre otros que son irreversibles. La ictericia patológica es una de las razones de hospitalización con mayor frecuencia en neonato. El estudio pretende determinar el mejor método de imputación de datos para información faltante en datos de ictericia patológica; por lo que el presente trabajo de investigación justifica su realización.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. Objetivo general

Determinar la técnica de imputación de datos que mejor desempeño presenta en el conjunto de datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021.

1.4.2. Objetivos específicos

a) Identificar la técnica de imputación de datos de medidas de tendencia central más adecuada para el tratamiento de datos ausentes en datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021.

b) Comparar la técnica de regresión y la técnica de vecinos más cercanos que presente mejor desempeño para la imputación de datos ausentes de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021.

CAPÍTULO II

MARCO TEÓRICO

2.1. MARCO CONCEPTUAL

Imputación: En la ciencia de la estadística, la imputación es la sustitución de valores no informados en una observación; en tanto cabe aclarar que no se trata de solo inventar datos sin ningún criterio para completar, sino que con la imputación tratamos de mantener la distribución de los datos al completar los espacios vacíos. Muchas veces la imputación de datos es un paso previo para poder tratar o analizar los datos con determinadas técnicas estadísticas de análisis.

KNN: *("K-Nearest-Neighbor" o "k-vecinos más cercanos")* es una técnica considerada dentro de los algoritmos de clasificación de aprendizajes supervisados no paramétrico, pero así mismo puede ser utilizado como una técnica de imputación de datos; una nueva muestra se imputa encontrando las muestras en el conjunto de entrenamiento «más cercano» a ella y promedia estos puntos cercanos para completar el valor.

Regresión: es una técnica estadística muy conocida el cual consiste en explicar una de las variables (dependiente) en función de la otra a través de un determinado tipo de función (lineal, parabólica, exponencial, etc.), de forma que la función de regresión se obtiene ajustando las observaciones a la función elegida.

Regresión logística: es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (muchas veces binaria, pero también podría tener más de dos categorías) en función de las variables independientes o predictoras. Es versátil para el modelamiento de la probabilidad de ocurrencia de un evento en función de otros factores.

Esta regresión se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit.

Ictericia: es el color amarillento que se ve en la piel de los recién nacidos. Esto sucede cuando una sustancia química, llamada bilirrubina, se acumula en la sangre del bebé recién nacido. Durante el embarazo, el hígado de la madre elimina la bilirrubina del bebé; pero, una vez que nace, su propio hígado es el que debe realizar esa función. En algunos bebés, el hígado puede no haberse desarrollado lo suficiente como para eliminar la bilirrubina; cuando se acumula demasiada bilirrubina en el cuerpo de un recién nacido, la piel y la parte blanca de los ojos pueden adquirir un color amarillento, a esta coloración amarillenta se la llama ictericia.

2.2. BASES TEÓRICAS RELACIONADAS A IMPUTACIÓN DE DATOS

2.2.1. Datos faltantes

La ausencia de datos es un problema muy recurrente, por ejemplo, cuando los datos no son confiables debido a medidores dañados, la información de la muestra se pierde o se corrompe, o la información no se informa o se informa incorrectamente, entre otros. En estos casos, es necesario evaluar todas las posibles deficiencias y qué soluciones se les ofrecerán antes de comenzar a trabajar con la información (Dagnino, 2014).

Si se tienen demasiados valores faltantes o si los datos omitidos en una variable dependen de otra o más variables se puede optar por descartarlos por completo de la información, pero esto podría conllevar a resultados sesgados o incluso inválidos al momento de realizar el análisis (Castro, 2014).

Existen diversos métodos para rellenar estos espacios vacíos en los datos con la información apropiada. Los métodos más sencillos asignan un valor fijo como la media o la mediana, otros rellenan con un valor existente de manera aleatoria o bien promedian

los datos en una vecindad definida del valor faltante. Estos métodos no proporcionan una solución óptima al problema de la falta de datos y tienden a sesgar la información, por lo que los resultados posteriores al análisis podrían no ser del todo confiables (Castro, 2014).

Para obtener mejores resultados en el análisis de la información se debe hacer un estudio previo de la escala de las variables y su distribución, así como del patrón que siguen los datos faltantes. Es conveniente utilizar distintos métodos para rellenar la información y realizar una evaluación de los mismos para seleccionar el método que mejor se ajusta a los datos y que minimice los posibles errores (Galarza, 2013).

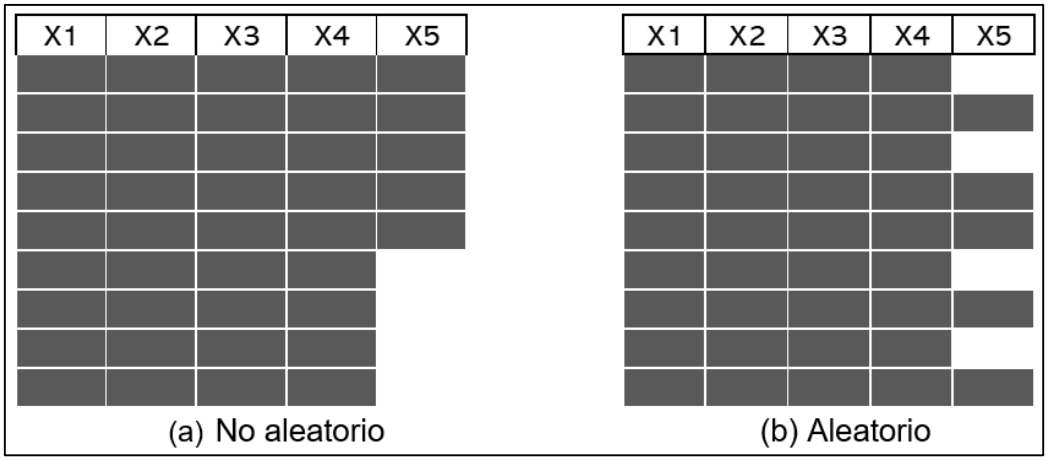
2.2.2. Patrones de Datos Faltantes

Para seleccionar un método adecuado para imputación de datos faltantes es importante encontrar el patrón que sigue la ausencia de datos. En la práctica los conjuntos de datos suelen tener un arreglo rectangular, donde los renglones corresponden a las unidades observadas y las columnas corresponden a las variables o características (Lerdo de Tejada, 2014). Siguiendo esta línea, existen diversas formas de clasificar el patrón de ausencia de datos. Los patrones de ausencia más comunes se definen como sigue:

a) Patrón de datos faltante univariado

El patrón univariado es el caso más simple de presencia de valores perdidos y se identifica cuando se tienen observaciones ausentes únicamente en una variable dentro de un conjunto de datos. La ausencia de registros puede ignorarse si estos presentan un comportamiento aleatorio, es decir, pueden considerarse como una submuestra aleatoria de la población y el análisis puede realizarse con los valores observados. Sin embargo, si la ausencia de registros depende del valor de la misma variable, un análisis sólo con los datos observados que no tome en cuenta este hecho, conduciría a un sesgo en los resultados. (Lerdo de Tejada, 2014)

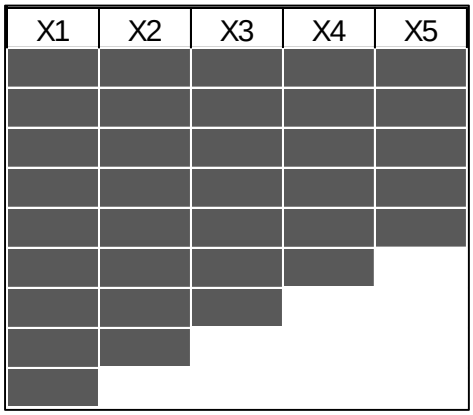
Figura 1. Patrón univariado



b) Patrón de datos faltante Monótono

De acuerdo con (Schafer & Graham, 2002) cuando todas las variables o grupos de variables de un conjunto de datos, por decir $Y_1 \dots Y_p$, se ordenan de tal modo que si Yj es faltante para una observación, entonces las variables $Y_{j+1}, \dots, Y_p$ tampoco son observadas, se dice que los valores ausentes siguen un patrón monótono.

Figura 2. Patrón Monótono



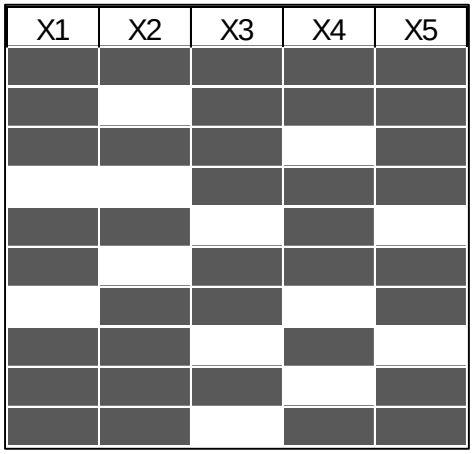
Generalmente este patrón es común en ámbitos como la psicología, la medicina, encuestas de estudios poblacionales o en riesgo de crédito, donde se llevan a cabo estudios longitudinales o de seguimiento de una misma población a lo largo del tiempo y a partir del j-ésimo periodo comienza la perdida de información. Los motivos de esta pérdida

pueden estar asociados al fallecimiento del individuo, al cierre de un crédito o a que el encuestado no se encontraba en el momento de levantamiento de la encuesta. Este problema tradicionalmente se ha resuelto con métodos de imputación simple, que en algunos casos logran resultados aceptables, sin embargo, con el reciente descubrimiento de nuevos métodos de imputación como la imputación múltiple o el algoritmo EM se pueden lograr resultados más confiables en gran parte de las situaciones (Lerdo de Tejada, 2014).

c) Patrón de datos faltante Aleatorio

Cuando la ausencia de datos se presenta en una o más variables sin seguir un orden especifico, entonces se tiene un patrón aleatorio.

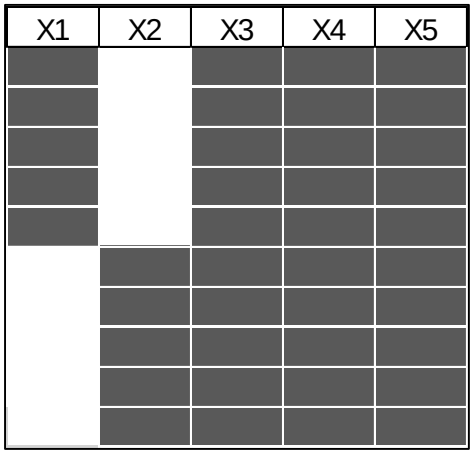
Figura 3. Patrón Aleatorio



d) Patrón de datos faltante de Parámetros no identificados

En este caso se tiene que la ausencia de datos se presenta en dos o más variables para tramos de observaciones excluyentes. Por ejemplo si la ausencia de datos se presenta en dos variables $X1$ y $X2$, las observaciones $1,\cdots,i$ estarán ausentes para la variable $X2$ mientras que las restantes $i+1,\cdots,m$ estarán ausentes para la variable $X1$.

Figura 4: Parámetros no identificados



2.2.3. Distribución de datos perdidos

Para describir las frecuencias y los patrones de los datos faltantes, así como su relación con los datos observados y con ellos mismos, se da al comportamiento de los valores ausentes un enfoque probabilístico. De esta manera, aunque no es posible conocer de manera exacta todas las causas que originan la pérdida de datos, se puede obtener información valiosa que ayude a mitigar este problema (Lerdo de Tejada, 2014). Supongamos que se tiene un conjunto de datos X, de los cuales X_{obs} y X_{miss} corresponden a los valores observados y faltantes respectivamente, de tal manera que X_{obs} - $X_{miss} = X$.

(Rubin, 1976) desarrolló una clasificación para los datos faltantes X_{miss} acuerdo a su estructura de dependencia

- Faltante Completamente Aleatorio (MCAR)
- Faltante Aleatorio (MAR)
- Faltante No Aleatorio (MNAR)

a) Faltante Completamente Aleatorio (MCAR)

En este caso, la probabilidad de que una observación sea faltante no depende de otros datos faltantes o de los datos observados, es decir:

$$P\left(X_{miss} \mid X_{obs}, X_{miss}\right) = P\left(X_{miss}\right)$$

En el mejor de los casos se tienen observaciones perdidas del tipo MCAR. Mientras su representatividad dentro del conjunto de datos sea razonablemente baja se puede ignorar su ausencia, ya que representarían una muestra aleatoria de la población, siendo sus características heterogéneas y similares a las de la población total.

Clasificar los datos faltantes como MCAR es complicado en la práctica, ya que en general los datos faltantes presentan una relación de dependencia con datos observados y no observados. En diversas investigaciones se han desarrollado mecanismos para identificar este patrón de datos, entre los cuales se puede destacar la representación de los datos faltantes en cada variable a través de variables artificiales o dummy (Cohen & Cohen, 1975), en donde se asigna el valor de 0 a la variable artificial cuando el dato es observado y 1 cuando el dato es faltante. Las variables artificiales se utilizan como variables predictivas en un modelo de regresión y se puede evaluar si tienen una relación de dependencia con la variable dependiente a través del coeficiente regresor. Si el coeficiente resulta significativo la ausencia de ciertos datos presenta un comportamiento condicional a la variable explicada y diferente a los datos que sí fueron observados. En cambio, si el coeficiente regresor no resulta significativo, se puede asumir que los datos faltantes presentan un comportamiento aleatorio e independiente a la variable dependiente.

b) Faltante Aleatorio (MAR)

Si los datos presentan el patrón MAR, la probabilidad de que un valor sea faltante depende solo de los datos observados:

$$P\left(X_{miss}\,|X_{obs}, X_{miss}\right) = P\left(X_{miss}\,|X_{obs}\right)$$

Cuando se tienen valores ausentes del tipo MAR es posible encontrar una estructura concreta de su distribución, ya que la probabilidad de que un dato sea faltante se obtiene condicionando sobre los valores observados. En diversos estudios se ha concluido que los datos faltantes del tipo MAR pueden ser ignorados, ya que los resultados obtenidos a través de métodos basados en la verosimilitud no se ven alterados por la ausencia de los MAR. Incluso (Collins, Schafer, & Kam, 2001) demostraron con datos reales que, aunque se asuma erróneamente que los datos faltantes siguen una distribución MAR, este supuesto tiene un impacto poco significativo en los estimadores y errores estándar. Asumir que los datos perdidos son MAR implica no considerar en las estimaciones alguna causa o correlación debida a su ausencia.

Para identificar si los valores ausentes siguen una distribución MAR, (Little R. , 1986) propone un estadístico de prueba con distribución $\chi 2$ con f grados de libertad, donde la hipótesis nula H0 establece que los datos faltantes siguen una distribución MAR. La regla de decisión, como en cualquier prueba de hipótesis, es rechazar Ho conforme a un nivel de significancia α preestablecido.

c) Faltante No Aleatorio (MNAR)

Finalmente, en el caso en que los datos son clasificados como MNAR, la probabilidad de que el dato sea missing o bien $P\left(X_{miss}\,|X_{obs}, X_{miss}\right)$ no puede ser cuantificada puesto que el motivo por el que se tiene el dato faltante depende de los datos faltantes y en algunos casos también de los observados. Por lo tanto, los datos del tipo MNAR no pueden ser ignorados y siempre que se quieran realizar estimaciones con este tipo de datos es necesario incluir un modelo para la probabilidad mencionada anteriormente, que involucre las causas y relaciones de los datos faltantes con la información.

En la práctica, aun cuando los datos faltantes pueden ser ignorados, la meta es reemplazarlos por los valores apropiados con la finalidad de contrarrestar la perdida de información, sobre todo cuando se cuenta con una muestra reducida o la información reportada es escasa.

Cuando los datos son MCAR o MAR, se dice que los motivos que ocasionaron la falta de datos son ignorables, y así es posible simplificar los métodos de estimación. Métodos como el algoritmo EM y la Imputación Múltiple trabajan bajo este supuesto. En general no es fácil obtener evidencia empírica que demuestre que los datos faltantes son MCAR o MAR, sin embargo, se puede justificar la elección del método más conveniente para trabajar con ellos.

d) Cómo probar la existencia de un mecanismo de pérdida de datos en una matriz de datos

Se define como mecanismo de pérdida (proceso de no respuesta) al origen, causas, momento, relaciones, características, que producen la falta de información. Es importante tratar de establecer si las observaciones han sido perdidas al azar o su falta se asocia a causas definibles. Algunas veces el mecanismo está bajo el control del analista, otra no puede controlarlo, pero sí comprenderlo y en muchos casos al no considerarlo explícitamente, se está suponiendo que el mecanismo es ignorable.

La idea de descubrir el mecanismo de pérdida de datos en una matriz de datos es sumamente compleja, por una parte están las matrices que poseen patrones de datos no ignorables, de los cuales el investigador posee información a priori para identificarlos pero en el caso de los mecanismos de pérdida de datos ignorables la identificación de este mecanismo no es tan evidente, solamente existe forma de identificar aquellos que poseen patrones de tipo MCAR pues este es el único mecanismo que produce

proposiciones contrastables. Es claro que habiendo descartado al menos dos de los mecanismos de ausencia de datos el tercero es evidentemente la única opción verificable.

La prueba para definir el mecanismo de pérdida de información se realiza dependiendo si el conjunto de datos es un problema de tipo univariante o multivariante; en el primer caso la prueba es más sencilla pues se emplea una prueba T de Student, pero en el segundo los cálculos son más complejos y debe emplearse la prueba multivariante propuesta por (Little R. , 1986) que no es sino una extensión de la prueba t en la que se coparan simultáneamente las diferencias de medias en todas las variables en el conjunto de datos, a continuación se define cada una de las pruebas mencionadas anteriormente iniciando con el estudio del caso univariante.

e) Prueba t de Student para contrastar el mecanismo de pérdida de información (MCAR)

El método más simple para evaluar la existencia de un mecanismo del tipo MCAR es utilizar una prueba de tipo t de Student para comparar los datos que faltan en subgrupos de datos (Garcia & Palacios, 2013) esta prueba se emplea con una variante en la cual se hace uso de la prueba t de Welch que no asumen varianzas iguales entre los grupos en comparación, esta anotación será importante pues el estadístico de contraste con el que se trabajará, es una variación del estadístico usual de la t de Student empleado usualmente.

Este enfoque separa los datos faltantes y los datos completos para una variable en particular y utiliza una prueba t para examinar las diferencias de medias de grupo con otras variables en el conjunto de datos. El mecanismo MCAR implica que en promedio los casos con datos observados deben comportarse de la misma forma que los datos no observados; por consiguiente, la no significancia en el resultado de la prueba t provee evidencia de que los datos son MCAR, mientras que una prueba t estadísticamente significativa sugiere que los datos son MAR o MNAR.

La hipótesis por probar es:

H0: La media de los datos observados es igual a la media de los datos no observados en la variable de interés. si esta hipótesis se rechaza entonces el mecanismo de pérdida de los datos es de tipo MCAR

H1: La media de los datos observados no es igual a la media de los datos no observados en la variable de interés. El estadístico de contraste empleado en la prueba es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(S_1^2)}{n_1} + \frac{(S_2^2)}{n_2}}}$$

Donde:

t = Estadístico equivalente a t de Student.

$\bar{X}_1$ = Media aritmética del grupo 1.

$\bar{X}_2$ =Media aritmética del grupo 2.

S_1^2 = Varianza del grupo 1.

S_2^2= Varianza del grupo 2.

n_1 = Tamaño de la muestra del grupo 1.

n_2 = Tamaño de la muestra del grupo 2.

f) Prueba de Little MCAR

(Little R. , 1986) propuso una extensión multivariante del enfoque t-test que simultáneamente evalúa las diferencias de medias en cada variable del conjunto de datos. A diferencia de las pruebas t univariado, el procedimiento de Little es una prueba global

de MCAR que se aplica al conjunto de datos. Al igual que el enfoque t-test, prueba de Little evalúa las diferencias de medias entre los subgrupos de casos que comparten el mismo patrón de datos perdidos. La estadística de prueba es una suma ponderada de las diferencias estandarizadas entre las medias de subgrupo y de una gran media global, el estadístico de contraste es el siguiente:

$$d^2 = \sum_{j=1}^{J} n_j \left(\hat{\mu}_j - \hat{\mu}^{MI}\right)^T \hat{\Sigma}_j^{-1} \left(\hat{\mu}_j - \hat{\mu}^{MI}\right)$$

Donde:

n_j : El número de casos en los datos perdidos en el patrón j.

$\hat{\mu}_j$: Contiene la media de las variables para los casos de datos perdidos en el patrón j.

$\hat{\mu}_j^{MI}$: Contiene la estimación por máximo verosimilitud de la gran media calculada para el conjunto de datos con valores completos.

$\hat{\Sigma}_j$: Contiene las estimaciones máximo verosímiles de la matriz de varianzas y covarianzas.

Las Hipótesis a probar son: H_o : Los datos son MCAR y H_1: Los datos son MAR, d^2 se distribuye aproximadamente como el estadístico chi-cuadrado con $\sum k_J - k$ grados de libertad, donde k_j es el número de variables completas del patrón j, y k es el número total de variables.

2.2.4. Técnicas de imputación

Todos los investigadores necesitan depurar los datos que reciben a través de recolecciones de datos antes de proceder a extraer conclusiones, este proceso de depuración consiste en verificar si los valores de cada encuesta satisfacen un conjunto de reglas de consistencias, típicamente conocidas; en el caso que este supuesto no se cumpla

el investigador esta ente un problema que se conoce como Edición e Imputación: "edición" es localizar los campos a modificar e "imputación" es determinar los nuevos valores para tales campos.

Hay una diferencia fundamental entre depuración previa e imputación. Consideremos el conjunto de todas las combinaciones posibles de códigos en un cuestionario, la depuración previa se puede definir como la división del conjunto en dos subconjuntos disjuntos. Las combinaciones que se consideran aceptables y las que se consideran inaceptables, las últimas contienen espacios en blanco no válidos y entradas inconsistentes. Así, la depuración previa es básicamente un diagnóstico y operativamente se puede definir mediante un conjunto de reglas. Por otro lado, la imputación pertenece por naturaleza al tratamiento de datos y es el proceso de asignar valores a datos que falten produciendo así un conjunto de datos completo. No hay un método insesgado conocido de imputación, pero algunos métodos son más adecuados que otros.

Es posible, en lugar de imputar la no-respuesta en el momento en que se preparan las tabulaciones de la encuesta, presentar estas informaciones sobre el tamaño de la no-respuesta. En este caso los usuarios podrían elegir entre diversos métodos de imputación a partir de los datos tabulados.

La situación más sencilla se da cuando hay solo un valor que se puede imputar, en un campo de forma que después de la imputación el valor sea consistente. A este caso se le denomina imputación determinista. Por ejemplo, si la esposa aparece codificada como masculino solo hay un valor posible a imputar al sexo que sea consistente con el resto de la información. A veces hay más de un valor que lo hace consistente. Si es este el caso, se elegir aquel valor particular que es más predominante con relación a la frecuencia total o más recomendable. Un ejemplo de este tipo se encuentra en la encuesta sobre mano de obra. Así, si una persona entre 15 y 16 años no ha rellenado la característica sobre su

actividad laboral en los meses que van de otoño a primavera se le asigna como asistiendo a la escuela, aunque es posible que no asista a la escuela.

Mientras la proporción de tales casos sea pequeña, el efecto de esta imputación será un incremento pequeño en el sesgo, pero habrá reducción en la varianza. En otras situaciones cuando se puede razonablemente imputar un intervalo de valores, necesitamos otros criterios. Uno sería el minimizar el error medio cuadrático de las estimaciones resultantes. La cuestión, es que no se sabe que error cuadrático medio hay que minimizar. Además, no se conocen los diferentes agregados a los que unos datos pueden contribuir y sus diferentes formas de tabulación.

En otras palabras, ¿Cómo se puede predecir el mejor valor de un campo sobre la base de conocer los otros campos del conjunto? Un buen ejemplo de este tipo de imputación es el uso de los datos del mes previo en la encuesta sobre mano de obra: para una determinada persona, difícilmente se encontrará un valor imputado mejor, particularmente en aquellos casos en los que las características demográficas cambien lentamente. Si no disponemos de información pasada se tiene que recurrir a otros métodos de imputación.

2.2.4.1. Clasificación de las Técnicas de Imputación

Existen diversas clasificaciones para los distintos métodos de imputación pues cada uno de ellos se aplica para diferentes patrones de pérdida de respuesta, atendiendo a la clasificación de (Goicochea, 2002); y complementando con algunas otras técnicas estudiadas se presenta la siguiente clasificación:

a) Técnicas Determinísticas

Este tipo de técnicas se emplea cuando al repetir la imputación en varias unidades bajo las mismas condiciones, producirán las mismas respuestas. Algunas de las técnicas que conducen a estos resultados son:

Fichero Caliente (Hot Deck)

Es un método usual de ajustar conjuntos de datos para valores no observados y admite diversas variantes. Generalmente el fichero caliente es un procedimiento de duplicación. Cuando falta un valor, se duplica un valor ya existente en la muestra para reemplazarlo. El principal propósito del fichero caliente es reducir el sesgo debido a la no respuesta. Para reducir este sesgo, el procedimiento de fichero caliente incorpora un método de clasificación. Todas las unidades muéstrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. De modo que el supuesto implícito que se está utilizando es que dentro de cada clasificación la no respuesta sigue la misma distribución que los que responden. Tal supuesto impone una fuerte restricción para las variables de clasificación. Estas variables han de estar correlacionadas con los valores que falten y con los valores de los que contestan. Si esto no se mantiene el fichero-caliente reduce solo en parte el sesgo debido a la no-respuesta i) produce un conjunto de datos limpios, esto es, un conjunto de datos completo y claro; ii) reduce el sesgo mientras preservemos las distribuciones conjuntas y marginales. Por ejemplo, si sustituyéramos un valor que falte por la media, la distribución de los valores muéstrales resultaría afectada. Y si escogiéramos aleatoriamente un valor entre los datos se reduciría la distorsión de la distribución, pero no el sesgo.

Como método de imputación los procedimientos de fichero caliente tienen ciertos rasgos atractivos entre los que se encuentran los siguientes:

1. Los procedimientos conducen a una post-estratificación sencilla;

2. No se presentan problemas especiales de encajar conjuntos de datos;

3. No se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que falten.

Imputación Haciendo uso de la Media

Este método, propuesto por primera vez por Wilks (1932), es posiblemente uno de los procedimientos de imputación más antiguo y sencillo. Existen dos variantes las medias incondicionadas y las medias condicionadas.

a.- Imputación haciendo uso de las Medias incondicionadas

La forma más simple de imputación no aleatoria de un valor desconocido consiste en asignar el valor promedio de la variable que lo contiene, calculado en los casos que tienen valor. Si se trata de una variable categórica se imputa la moda de la distribución. Consiste en estimar los valores perdidos de la j-ésima variable mediante la media de sus valores observados (Little & Rubin, 1987); la expresión clásica para el cálculo es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde:

x_i: $Valores\ observados\ de\ la\ variable\ x$

n: $Cantidad\ de\ individuos$

El uso de imputación por la media no es una técnica recomendable cuando posteriormente se desea realizar un análisis estadístico mediante técnicas de regresión. Bajo este procedimiento de imputación, el valor medio de la variable se preserva, pero otros

estadísticos que definen la forma de la distribución varianza, covarianza, cuartiles, sesgo, curtosis, entre otros, pueden ser afectados.

El uso de este método afectará la correlación entre la variable imputada y cualquiera otra, reduciendo su variabilidad. Esto es, la sustitución de la media en una variable puede llevar a perjudicar estimaciones de los efectos de otra o todas las variables en un análisis de regresión, porque el perjuicio en una correlación puede afectar los pesos de todas las variables. Adicionalmente, si se imputa un gran número de valores usando la media, la distribución de frecuencias de la variable imputada puede ser engañosa debido a demasiados valores localizados centralmente creando una distribución más alargada o leptocurtica (Rovine & Delaney, 1990)

b.- Imputación de medias condicionadas

Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases ajustadas e imputar los valores faltantes de los valores observados en la misma clase.

Una variante del procedimiento anterior se presenta cuando las respuestas de cada variable son agrupadas en clases disjuntas con diferentes medias, y a cada registro faltante se le imputara con la media respectiva de su grupo. La sustitución de los datos faltantes por la media reduce la amplitud del intervalo de confianza debido a la disminución de la varianza del estimador. Al igual que el procedimiento de medias, en este caso se asume que los datos faltantes siguen un patrón MCAR y existirán tantos promedios como categorías se formen, lo cual contribuye a atenuar los sesgos en cada celda, pero de ninguna manera los elimina. Este procedimiento tiene las mismas desventajas que el caso anterior, pero en menor proporción por estar agrupadas. Igualmente es de fácil aplicación.

En la medida que la falta de información por categoría sea baja, los sesgos disminuyen, pero no desaparecen. No obstante, no se sugiere utilizar este procedimiento en la medida de que se disponga de una mejor alternativa para sustituir la información omitida.

Imputación usando la mediana

Dado que la media es afectada por la presencia de valores extremos, parece natural usar la mediana en vez de la media con el fin de asegurar robustez. En este caso el valor faltante de una característica dada es reemplazado por la mediana de todos los valores conocidos de ese atributo. Este método es también una opción recomendada cuando la distribución de los valores de una característica es sesgada (Acuña & Rodriguez, 2004).

Obviamente técnicas como la imputación de la media y la mediana, sólo son aplicables a variables cuantitativas y no pueden usarse con valores faltantes en una característica categórica, en cuyo caso puede usarse la imputación de la moda. Estos métodos de imputación son aplicados separadamente en cada característica que contiene valores faltantes. Nótese que la estructura de correlación de los datos no está siendo considerada en los métodos anteriores.

Por lo tanto, una medida alternativa de tendencia central representa mejor la distribución subyacente y por tanto una mejor estimación para los valores faltantes. La mediana, en particular, frecuentemente funciona bien como una medida de tendencia central cuando las distribuciones se desvían considerablemente de la distribución normal estándar. El procedimiento para sustituir la mediana para los valores faltantes para una variable particular sigue la misma lógica y protocolo que la sustitución de la media.

Imputación por Regresión

Este método, propuesto por (Buck, 1960), supone que las filas de la matriz de datos constituyen una muestra aleatoria de una población normal multivalente. El vector de medias y la matriz de varianzas y covarianzas de los datos completos son utilizados como estimaciones de los parámetros poblacionales, con los cuales se ajustan ecuaciones en regresión para cada una de las variables con datos perdidos, en término de las restantes. Ante la presencia de un patrón de datos faltantes MAR es posible utilizar modelos de regresión para imputar información en la variable Y, a partir de un grupo de covariables $(X_1, X_2, \ldots, X_p)$ correlacionadas.

Se considera una variable Y_i que presenta n_{per} valores perdidos y $n_i = n - n_{per}$ valores observados. Se supone que las $k - 1$ restantes variables X_j , con $i \neq j$, no presentan valores perdidos. Con este método se estima la regresión de la variable Y_i sobre las variables X_j ,$\forall j \neq i$, a partir de los n casos completos y se imputa cada valor perdido con la predicción dada por la ecuación de regresión estimada. Esto es, si para el caso I el valor y_{li} no se observa, entonces se imputa mediante:

$$\hat{y}_{li} = \hat{\beta}_{0.obs} + \sum_{j \neq i} \hat{\beta}_{j.obs}\, x_{lj}$$

Donde $\hat{\beta}_{0.obs}$ y $\hat{\beta}_{j.obs}$, $j \neq i$ representan los coeficientes de la regresión de X_i, $\forall i \neq i$, basadas en las n_i observaciones completas.

2.3. BASES TEÓRICAS RELACIONADAS A LA ICTERICIA NEONATAL

2.3.1. Ictericia neonatal.

Es una afección que ocasiona una pigmentación amarillenta en la piel, mucosas y fluidos del cuerpo. La pigmentación más notoria es en la piel y en escleras (parte

blanca de los ojos). Esta pigmentación resulta de la acumulación de la bilirrubina en los tejidos, un pigmento producto del metabolismo de la hemoglobina. (Ortiz, 2010)

2.3.2. Hiperbilirrubinemia Neonatal.

La hiperbilirrubinemia neonatal se manifiesta como la coloración amarillenta de la piel y mucosas que refleja un desequilibrio temporal entre la producción y la eliminación de bilirrubina. Es clínicamente evidente cuando existe una concentración de bilirrubina mayor de 5mg/dl en suero. Las causas de ictericia neonatal son múltiples y producen hiperbilirrubinemia directa, indirecta o combinada, de severidad variable (Manotas, 2005):

1. No conjugada: Es la elevación de la bilirrubina sérica no conjugada a niveles superiores a 1,3 – 1.5 mg/dl.
2. Conjugada: Es la elevación de la bilirrubina sérica mayor de 1,5 mg/dl y más de 10% de la concentración sérica total.

2.3.3. Ictericia fisiológica.

Ictericia monosintomática de inicio a partir del segundo día de vida, con un pico máximo de bilirrubina de 12- 15 mg/dl en el 3º -5º día, no persistiendo más allá del 7ª día. No requiere tratamiento, pero si observación y seguimiento por si se tratase de una ictericia patológica. Se debe a una limitación del hígado para metabolizar el exceso de bilirrubina producida en los primeros días de vida (Ortiz, 2010).

2.3.4. Ictericia patológica.

Se considera que la ictericia es patológica cuando aparece es en las primeras 24 horas de vida y cuando la bilirrubina sérica es mayor a 15 mg/dl, la ictericia persiste después del 8º día y el incremento de la bilirrubina es en más de 5 mg/dl (Ortiz, 2010).

2.3.5. Ictericia por lactancia materna.

Ictericia asintomática de inicio tardía entre el 4º y 7º día con cifras de bilirrubina hasta 20 mg/dl en la 2º 3º semana que puede prolongarse hasta la 4º - 12º semana de vida.

El diagnóstico es clínico tras la exclusión de otras causas. El tratamiento es aumentar el número de tomas, buena hidratación y, si es preciso por la cifra de bilirrubina es la fototerapia. Es debida principalmente a un incremento de la circulación enterohepática con aumento de la reabsorción de bilirrubina.

2.3.6. Tratamientos de la Ictericia

Existen actualmente 3 tipos de tratamientos:

•**Terapia farmacológica**, en la cual se utilizan diferentes fármacos como Mesoporfirina, Fenobarbital y Administración oral de sustancias no absorbibles.

•**Fototerapia,** en la cual se hace uso de la luz visible para descomponer la bilirrubina en productos polarizados hidrosolubles (24-30 %).

•**Exanguinotransfusión**, mecanismo de acción que se basa en la remoción mecánica de sangre del neonato por sangre de un donador.

2.4. ANTECEDENTES DE ESTUDIO

2.4.1. Antecedentes Internacionales

(Quesada, 2011) En su estudio "*observacional de hiperbilirrubinemia neonatal*", en un hospital de tercer nivel Julio 2010 a junio 2011 Ecuador. Llega a la conclusión que la hiperbilirrubinemia neonatal se presentó en el 43.5% de los pacientes ingresados en el servicio de neonatología y fue una de las principales causas de hospitalización.

(Parodi & colaborades, 2005) Realizaron la revisión bibliográfica "*Ictericia Neonatal*", en Uruguay, teniendo como base que la Ictericia en el recién nacido es la mayor parte de las veces es un hecho fisiológico, se llegó a las siguientes conclusiones: 1) Ha disminuido los casos de Ictericia por incompatibilidad Rh debido a la utilización profiláctica de inmunoglobulinas Anti-D. 2) La administración de fototerapia ha disminuido la práctica de exanguineotransfusión. 3) Tanto la fototerapia como la exanguineotransfusión siguen siendo los pilares del tratamiento, aunque no están exentos de riesgos.4) El egreso precoz del hospital de los recién nacidos puede incrementar el riesgo de complicaciones debidas a ictericia temprana no detectada.

(Manotas, 2005) En su estudio "*Ictericia Neonatal*", en Montevideo**,** tuvo las siguientes conclusiones: las enfermedades hematológicas neonatales, especialmente las del tipo hemolítico, son menos frecuentes en la actualidad, pero algunas son tan graves que pueden afectar de manera irreversible al sistema nervioso central; las enfermedades hematológicas propias de la niñez tampoco son usuales en el neonato, pero cuando se presentan deben interpretarse como situaciones que requieren solución rápida, por tanto, la tendencia actual es mejorar las medidas preventivas con terapias efectivas y de aplicación temprana que permitan disminuir las repercusiones de dichas enfermedades a corto y largo plazo. En esta revisión se consideran, básicamente, los trastornos hemolíticos que afectan al neonato más a menudo y que producen elevaciones de la bilirrubina de diferente magnitud e importancia. Es preciso iniciar el estudio de los problemas hemolíticos revisando el metabolismo fetal y neonatal de la bilirrubina, ya que es precisamente allí donde se encuentra la explicación fisiopatológica de dichos problemas.

(Caiza & Colaboradores, 2006) Condujeron en Ecuador un "*estudio descriptivo*" realizado con 1406 registros neonatales de niños con hiperbilirrubinemia que necesitaron de cuidados intermedios o intensivos en el servicio de Neonatología del Hospital Gineco Obstétrico Isidro Ayora de la ciudad de Quito, entre los años 2001 al 2005. Los factores de riesgo encontrados para recién nacidos prematuros al compararlos con neonatos a término fueron: cesárea (p<0,00001, OR 2,13 IC 1,76-2,679), infección neonatal (p<0,00001 OR 2,16 IC 1,73-2,69); la enfermedad de membrana hialina asociada a la prematurez fue el factor de riesgo más importante encontrado en este estudio (p<0,00000001 OR 29,3 IC 12,15-70,8). Otros síndromes de dificultad respiratorios también constituyeron factores de riesgo (p<0,00001 OR 2,86 IC 2,24-3,65)

2.4.2. Antecedentes nacionales.

(Bazalar, 2014) En su estudio de "*prevalencia y causas de ictericia neonatal*" en el Hospital Nacional Ramiro Prialé Huancayo en el periodo 2010-2011 en el cual se llegó a la conclusión que la prevalencia de ictericia neonatal es de 4,2% de recién nacidos con ictericia siendo las principales causas de ictericia neonatal, la ictericia fisiológica, hipo alimentación, incompatibilidad de grupo sanguíneo ABO y la frecuencia de ictericia según sexo fue masculino en 51.665% y según la edad gestacional fue a término en un 92.20%.

(Alvear, 2011) En su trabajo de investigación sobre el estudio de *"Ictericia Fisiológica en recién nacidos a término en el 2011"* concluyo que la incidencia total de la ictericia neonatal fisiología neonatal en este estudio fue de 5.2% nacidos vivos encontrando relación directa entre el sexo masculino y la ictericia fisiológica, 57.1% de casos fueron de sexo masculino.

(Perez, 2006) En un "*estudio prospectivo*", en el que registró 1,327 casos de hiperbilirrubinemia, en neonatos nacidos y atendidos en el Hospital Daniel Alcides Carrión de la ciudad de Huancayo, de un total de 3280 recién nacidos vivos. La tasa de morbilidad por hiperbilirrubinemia neonatal fue de 405 por cada 1000 recién nacidos vivos, encontró además discreto predominio del sexo masculino sobre el femenino, sin diferencia significativa. Además, más de la mitad de los casos de hiperbilirrubinemia fueron para los recién nacidos a término, con peso adecuado para la edad gestacional. Se registró que el mayor índice de hiperbilirrubinemia se presentó con valores de hematocrito de 45 a 49%. En relación a la edad materna, la mayor incidencia de Hiperbilirrubinemia, se presentó en el grupo etáreo comprendido entre los 15 a 29 años. Las principales causas de hiperbilirrubinemia neonatal fueron la prematuridad, el sufrimiento fetal agudo, la hipoxia neonatal y el parto por cesárea; una gran parte de las madres de hijos hiperbilirrubinémicos no presentaron ningún control prenatal. El índice de hiperbilirrubinemia fue de 40.46%; existiendo predominio del sexo masculino sobre el femenino.

(Morachino, 2011) desarrollo la tesis intitulada, "*correlación entre bilirrubina sérica y bilirrubinemia transcutánea en neonatos ictéricos*" donde la población estuvo conformada por 259 neonatos de los cuales 157 fueron varones 102 mujeres concluyendo que el icterómetro transcutáneo tiene una relación positiva con los valores de bilirrubina sérica con una alta correlación siendo esta diferencia estadísticamente significativa.

CAPÍTULO III

HIPÓTESIS Y VARIABLES

3.1. HIPÓTESIS

3.1.1. Hipótesis general

La técnica de imputación de datos que mejor desempeño presenta en el conjunto de datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021 es la imputación por regresión.

3.1.2. Hipótesis especificas

a) La técnica de imputación de datos de medidas de tendencia central más adecuada para el tratamiento de datos ausentes en datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021 es por la mediana.

b) La técnica de regresión es la técnica más adecuada a comparación de los vecinos más cercanos como técnica de imputación de datos.

3.2. IDENTIFICACIÓN DE VARIABLES E INDICADORES

Variable de estudio: Técnicas de imputación de datos

Indicadores:

Imputación haciendo uso de las Medias

Imputación usando la mediana

Imputación usando modelos de regresión considerando variables predictoras.

Imputación usando modelos de regresión adicionando un residuo aleatorio.

Imputación usando k vecinos más cercanos.

OPERACIONALIZACIÓN DE LA VARIABLE

VARIABLE	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	DIMENSIONES	INDICADORES
Técnicas de imputación de datos	La imputación de datos implica estimar o "rellenar" los valores faltantes con valores estimados basados en la información disponible en el conjunto de datos. El objetivo de la imputación de datos es minimizar el impacto de los valores faltantes en los análisis subsiguientes y en las conclusiones que se extraen de los datos. (Galarza, 2013)	En lugar de simplemente eliminar las observaciones con valores faltantes, lo que podría resultar en una pérdida significativa de información, las técnicas de imputación buscan preservar la integridad del conjunto de datos al sustituir los valores faltantes por estimaciones razonables.	-Imputación haciendo uso de las Medias -Imputación usando la mediana -Imputación usando modelos de regresión considerando variables predictoras. -Imputación usando modelos de regresión adicionando un residuo aleatorio. -Imputación usando k vecinos más cercanos.	Para la verificación de la calidad de la imputación se utilizará dos métodos de comparación: a. Por medidas estadísticas de resumen. -Media -Desviación estándar -Mediana -Mínimo -Máximo -Rango -Asimetría -Curtosis -Se b. Por gráficos -Gráfico de densidades. -Grafico de barras

CAPÍTULO IV

METODOLOGÍA

4.1. TIPO Y DISEÑO DE INVESTIGACIÓN.

El presente estudio está enmarcado en el tipo de investigación básica descriptiva, con un enfoque cuantitativo. (Sanchez, Reyes, & Mejía, 2018).

Diseño.

La presente investigación está enmarcada en el diseño no experimental; (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014) la "investigación no experimental es la que se realiza sin manipular deliberadamente variables. Es decir, se trata de estudios donde no se varía en forma intencional las variables independientes para ver su efecto sobre otras variables".

4.2. UNIDAD DE ANÁLISIS

La unidad de análisis, pacientes infantes del Hospital Regional del Cusco, 2021.

4.3. POBLACIÓN DE ESTUDIO

Todos los registros de historias de los pacientes infantes del Hospital Regional del Cusco, 2021.

4.4. SELECCIÓN DE MUESTRA

Es un subconjunto representativo de la población; una muestra puede ser probabilística (aleatoria) o no probabilística, que Jiménez C. (1983) precisa que la muestra "es una parte o subconjunto de una población, que pone de manifiesto las propiedades de la población".

La muestra es una parte representativa de la población y se obtiene a partir de la siguiente relación:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} * p * q}{\varepsilon^2}$$

Donde

P=0.50: Probabilidad de éxito

$\varepsilon = 0.0383 = 3.83\%$: Error del estudio.

Reemplazando los valores, se tiene:

$$n = \frac{(1,96)^2 * 0,50 * 0,50}{(0,0383)^2} \approx 656$$

Se consideró las 656 historias que se encontraron hasta este periodo.

4.5. TÉCNICAS DE RECOLECCIÓN DE DATOS E INFORMACIÓN

Se hará uso de la técnica de la revisión documental, en específico revisión de historias clínicas.

El instrumento por utilizar para el recojo de datos en el presente trabajo de investigación es la ficha de recolección de datos ver anexo.

4.6. ANÁLISIS E INTERPRETACIÓN DE LA INFORMACION.

Para el análisis se utilizará metodologías de completar datos, como por ejemplo completar con la mediana, completar con regresión utilizando predictores y ruidos aleatorios gaussianos, además de la metodología de KNN k vecinos más cercanos, todos ellos implementados en el software libre R y Rstudio; así mismo para ver cuáles son los factores con mayor riesgo en el tipo de ictericia se usará la regresión logística de manera superficial como una complementariedad al trabajo de investigación

CAPÍTULO V

RESULTADOS Y DISCUSIÓN

Para el presente capitulo se cuenta con una base de datos con 656 historias de los pacientes infantes del Hospital Regional del Cusco quienes desarrollaron o nacieron con ictericia, la cual tiene algunos registros en que no están completos, hubo perdida de información primaria, por lo tanto ahí tenemos se tiene un propósito, en vez de eliminar registros con datos no completos se decide determinar cuál es la técnica de imputación de datos que mejor desempeño presenta en el conjunto de datos de ictericia patológica de niños atendidos en el Hospital Regional de Cusco, 2021.

5.1. ANÁLISIS DESCRIPTIVO Y EXPLORATORIO DE LA BASE DE DATOS DE ICTERICIA

La presente investigación y la base de datos que se obtuvo está conformada por 32 variables, las que se presentan a continuación en la *Tabla 1*.

Luego se realizó un análisis exploratorio de los datos, se pudo observar con los siguientes códigos implementados en el R, que las variables que presentan datos perdidos o faltantes en la data son el peso en el alta del Recién Nacido (RN) el 4.87% de datos en esa variable, Periodo de embarazo (1.82% de datos perdidos en esta variable) y la cantidad de hematocritos del RN presenta 5.03% de datos perdidos en dicha variable.

Tabla 1: Variables analizadas en la data de Ictericia

Nro.	Variable	Descripción de la variable
1	Edad	Edad de la madre
2	gestas	Número de gestas
3	hijos	Número de hijos
4	edages	Edad gestacional

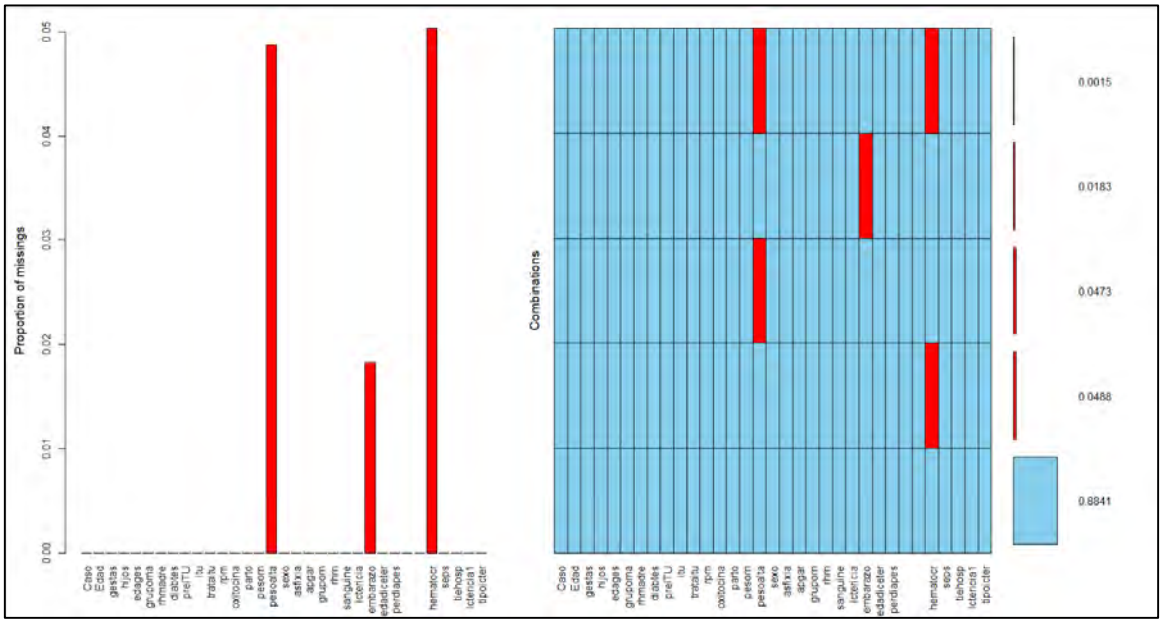
5	grupoma	Grupo
6	rhmadre	Factor Rh
7	diabtes	Presenta diabetes
8	preITU	Presencia de ITU
9	itu	Número de ITU
10	trataitu	Recibió tratamiento para ITU
11	rpm	Tiempo de RPM
12	oxitocina	Aplicación oxitocina
13	parto	Tipo de parto
14	pesorn	Peso del RN
15	pesoalta	Peso de alta de RN
16	sexo	Sexo del RN
17	asfixia	Asfixia
18	apgar	Apgar
19	gruporn	Grupo
20	rhrn	Factor Rh
21	sanguine	Incompatibilidad sanguínea
22	ictericia	Hermano con antecedentes de ictericia
23	embarazo	Periodo de embarazo
24	edadiceter	Periodo de aparición de ictericia
25	perdiapes	Perdida del mas del 10% de peso en la primera semana
26	cefalolohema	Presencia de cefalohematoma
27	TIPOhematocrito	Hematocrito
28	hematocr	Número de hematocrito
29	seps	Presencia de sepsis
30	tiehosp	Tiempo de hospitalización
31	Ictericia1	Presencia de Ictericia
32	tipoicter	Tipo de Ictericia

```r
#Para ver que columnas tienen valores perdidos
which(colSums(is.na(datos))!=0)
## pesoalta embarazo hematocr
##       16       24       29

#Para ver el porcentaje de valores perdidos en las columnas
colmiss=c(16,24,29)
per.miss.col=100*colSums(is.na(datos[,colmiss]))/dim(datos)[1]
per.miss.col

## pesoalta embarazo hematocr
## 4.878049 1.829268 5.030488
```

Para visualizar gráficamente el comportamiento de los datos perdidos se tiene el siguiente grafico implementado con la siguiente función escrita en R

```
# Aggregation plot
x11()
a=aggr(datos,numbers=T)
A
```

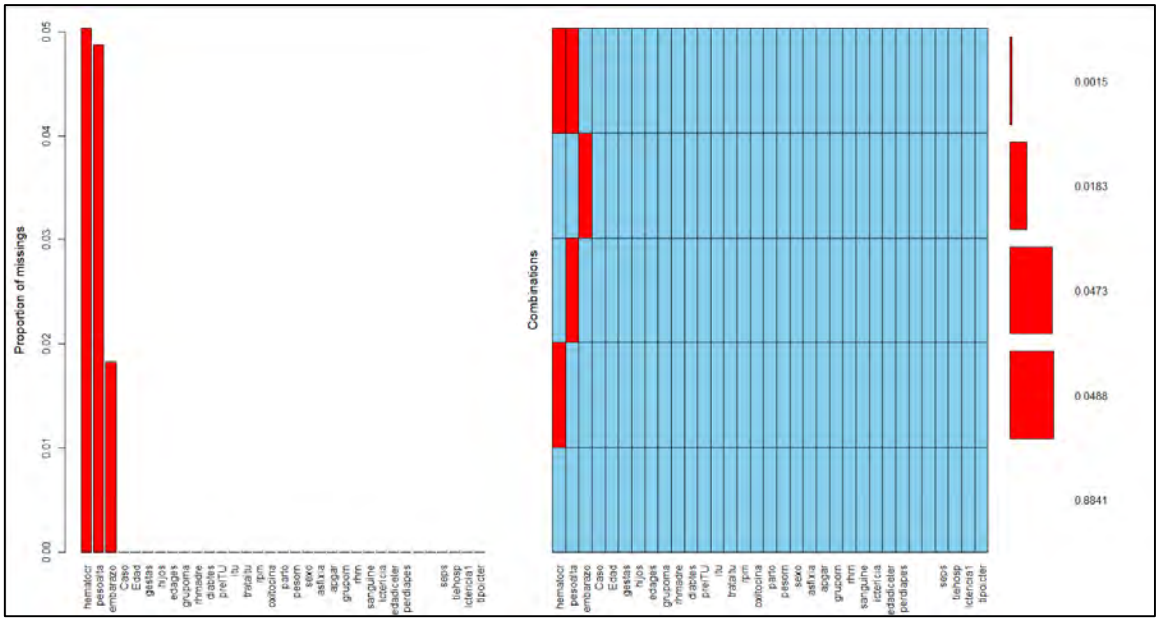
Figura 5: Distribución de los datos perdidos en la data.



Ordenando el gráfico con algunas funciones se tiene el gráfico con una distribución ordenada de la variable.

```
a=aggr(datos,numbers=T, sortComb=TRUE,
     sortVar=TRUE, only.miss=TRUE)
```

Figura 6: Distribución de los datos perdidos ordenado de manera descendente.



5.2. PRUEBA DE MEDIAS PARA VERIFICAR MECANISMO DE PERDIDA DE DATOS

Con la intención de verificar si alguna variable tiene relación con el mecanismo de perdida de datos de la variable hematocritos, encontrándose que por ejemplo la variable edad gestacional de la madre en promedio es similar tanto para el grupo que tiene datos completos, como con el grupo que no tiene datos. Así se concluye con un pvalor de 0.6262 que no existe diferencias significativas entre la edad promedio gestacional considerando el grupo con datos completos y faltantes, podríamos asumir que el mecanismo de la perdida de datos es MAR (Mecanismo Faltante Aleatorio)

```
# Prueba t de medias
t.test(edages ~ is.na(hematocr), data=datos)  #

##
##  Welch Two Sample t-test
##
## data:  edages by is.na(hematocr)
## t = 0.49134, df = 35.204, p-value = 0.6262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5135205  0.8415520
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            38.70947            38.54545
```

Así mismo para verificar si alguna variable tiene relación con el mecanismo de perdida de datos de la variable hematocritos, encontrándose que por ejemplo la variable edad gestacional de la madre en promedio es similar tanto para el grupo que tiene datos completos, como con el grupo que no tiene datos. Así se concluye con un p_valor de 0.1237 que no existe diferencias significativas entre la edad promedio gestacional considerando el grupo con datos completos y faltantes de la variable peso de alta del RN, podríamos asumir que el mecanismo de la perdida de datos es MAR (Mecanismo Faltante Aleatorio)

```
t.test(edages ~ is.na(pesoalta), data=datos)  #

##
##  Welch Two Sample t-test
##
## data:  edages by is.na(pesoalta)
## t = 1.5746, df = 37.71, p-value = 0.1237
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1074748  0.8590774
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            38.71955            38.34375
```

5.3. IMPUTACIÓN DE DATOS FALTANTES

5.3.1. Imputación usando medidas de tendencia central

Las variables que tenían datos faltantes son el peso en el alta del Recién Nacido (RN), Periodo de embarazo y la cantidad de hematocritos del RN estas variables, para ello se utilizó en variables cuantitativas reemplazar por la mediana y en variables cualitativas reemplazar con el valor de la moda; con esta finalidad se usa la librería DMwR la cual tiene una función denominada centralImputation.

```
library(DMwR)
datos.c<-centralImputation(datos)
summary(datos.c)
##  oxitocina          parto         pesorn        pesoalta          sexo
##  Si:458    Parto normal:381   Min.   :1050   Min.   :1545   Masculino:322
##  No:198    Cesarea     :275   1st Qu.:2950   1st Qu.:2850   Femenino :334
##                               Median :3230   Median :3100
##                               Mean   :3201   Mean   :3089
##                               3rd Qu.:3490   3rd Qu.:3345
##                               Max.   :4590   Max.   :4490
##
##  asfixia               apgar     gruporn  rhrn    sanguine ictericia
##  Si: 29   Normal          :571   O :521   +:651   Si:116   Si:255
##  No:627   Depresion leve  : 75   A : 96   -:  5   No:540   No:401
##           Depresion severa: 10   B : 39
##                                  AB:  0
##
##         embarazo        edadiceter  perdiapes cefalolohema   TIPOhematocrit
##  Pre termino: 54   12 Hrs    : 83   No:636    Si: 20       Poliglobulia: 71
##  A termino  :602   24 Hrs    :473   Si: 20    No:636       Normal      :585
```

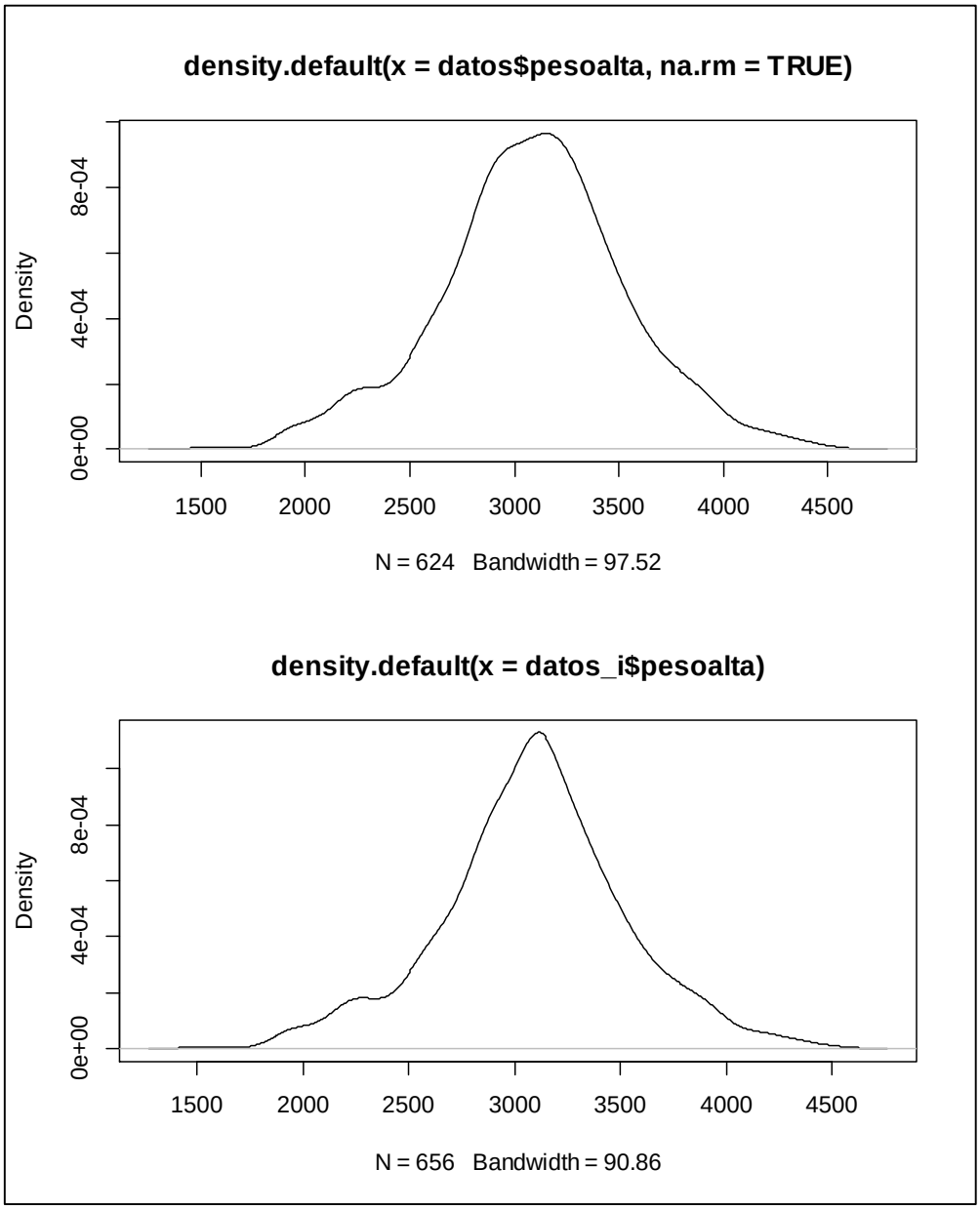
Así como la librería DMwR existe otra librería VIM la cual tiene una función initialise la cual puede completar los datos faltantes se decidió reemplazarla por la mediana.

```
datos_i<-initialise(datos,method="median")
summary(datos_i)

##  oxitocina          parto         pesorn        pesoalta           sexo
##  No:198    Cesarea     :275   Min.   :1050   Min.   :1545   Femenino :334
##  Si:458    Parto normal:381   1st Qu.:2950   1st Qu.:2850   Masculino:322
##                               Median :3230   Median :3100
##                               Mean   :3201   Mean   :3089
##                               3rd Qu.:3490   3rd Qu.:3345
##                               Max.   :4590   Max.   :4490

##         embarazo        edadiceter  perdiapes cefalolohema
TIPOhematocrito
##  A termino  :602   12 Hrs    : 83   No:636    No:636       Normal      :585
##  Pre termino: 54   24 Hrs    :473   Si: 20    Si: 20       Poliglobulia: 71
##                    48 Hrs    : 64
##                    72 H rs   : 28
##                    96 Hrs    :  4
##                    Quinto dia:  3
##                    Setimo dia:  1
```

Luego se decidió comparar la densidad de las variables que han sido imputadas para ver cuál de ellos mantiene las características originales de la data sin imputar.

```
x11()
par(mfrow=c(2,1))
plot(density(datos$pesoalta, na.rm =TRUE))
plot(density(datos_i$pesoalta))
```

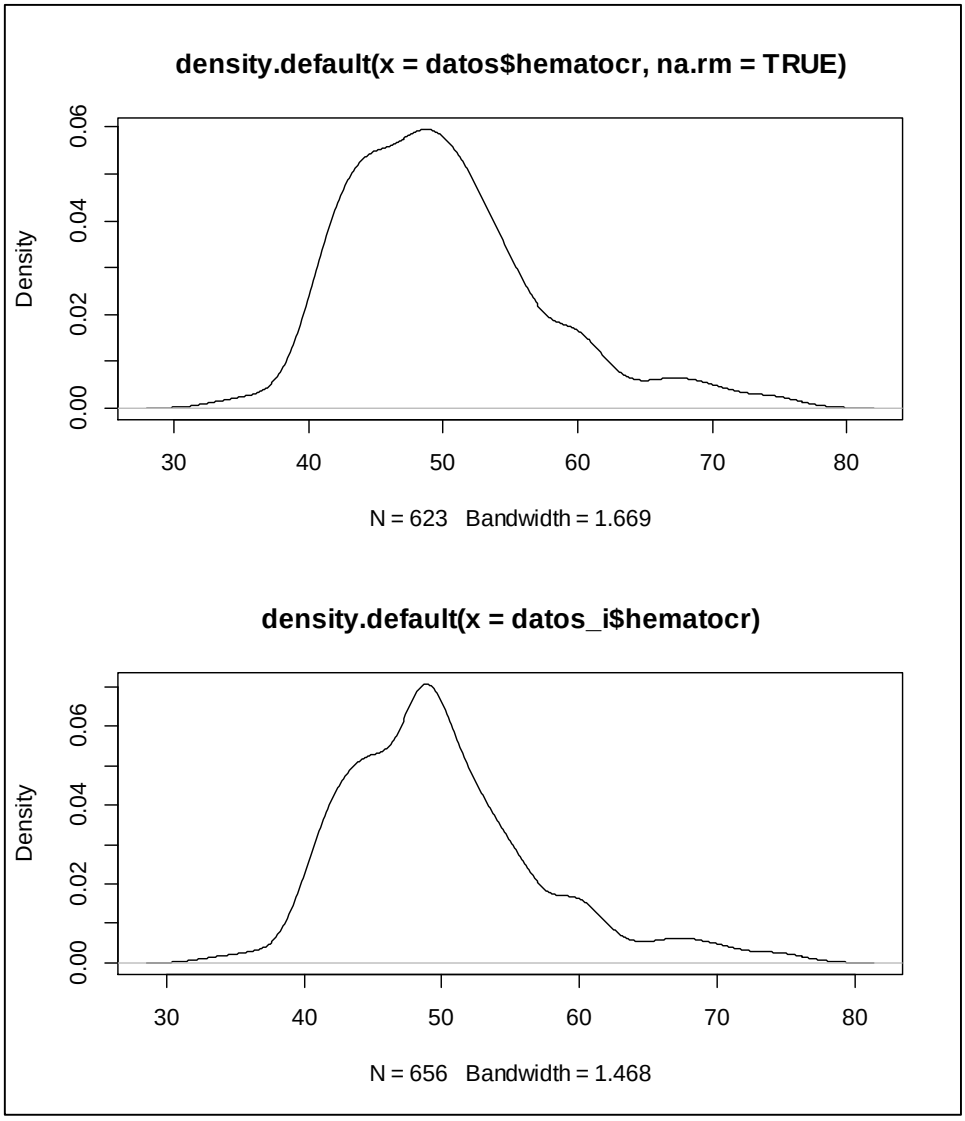
Figura 7: Variable Peso del Alta del RN antes y después de la imputación con la mediana



Se pudo observar en la *Figura 7*, que la distribución se volvió más puntiaguda debido a que se incrementó solo un valor que es la mediana. De la misma manera se pudo hallar para la variable número de hematocritos observándose en la *Figura 8*, existe el mismo efecto que con la anterior variable.

```
x11()
par(mfrow=c(2,1))
plot(density(datos$hematocr, na.rm =TRUE))
plot(density(datos_i$hematocr))
```

Figura 8: Variable hematocrito antes y después de la imputación con la mediana

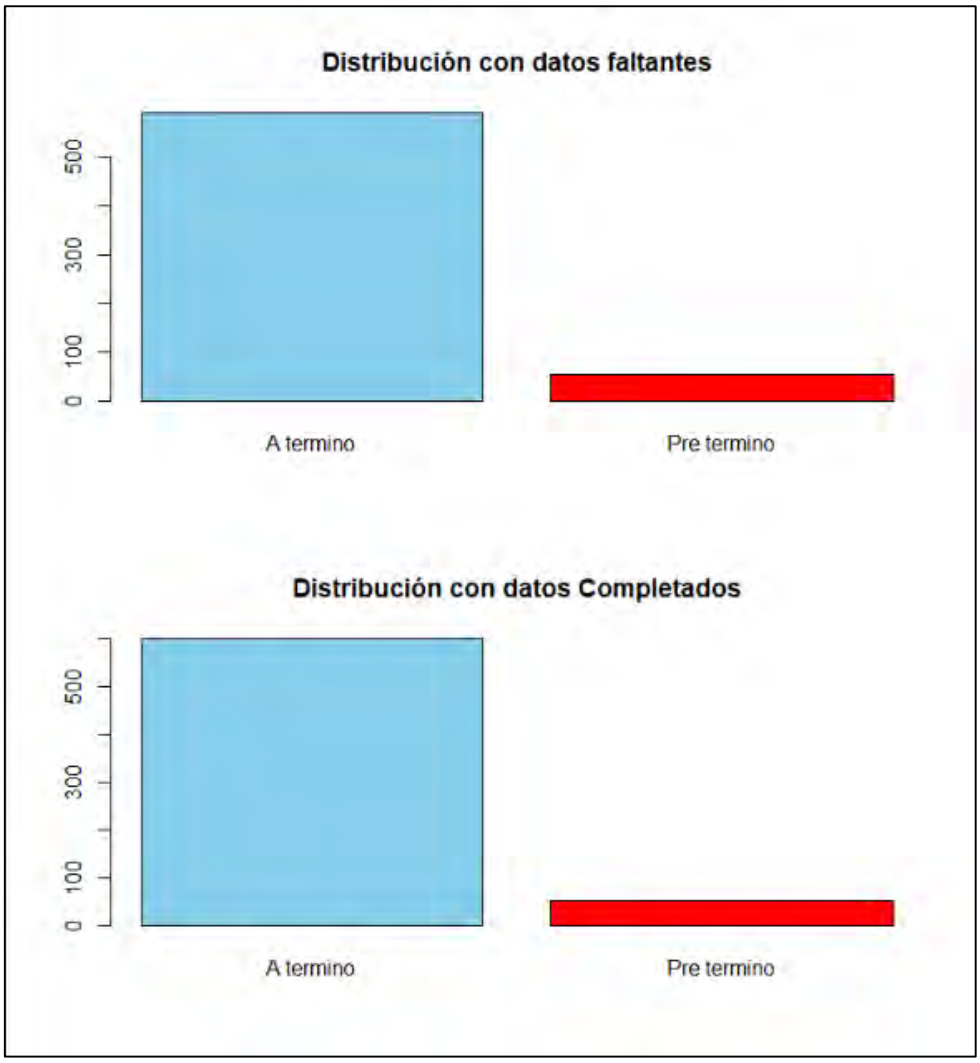


Ahora para la variable embarazo que es un variable categórica, se puede observar que la imputación que se realizó lo hizo con la moda, reemplazándose los valores faltantes con la moda.

```
# variable embarazo
table(datos$embarazo)

##
## Pre termino   A termino
##          54         590

table(datos_i$embarazo)

 ##
##   A termino Pre termino
##         602          54
```

```
par(mfrow=c(2,1))
barplot(table(datos$embarazo), col=c("red","skyblue"))
barplot(table(datos_i$embarazo),col=c("skyblue", "red"))
```

Figura 9: Variable embarazo antes y después de completar la data



5.3.2. Imputación usando modelos de regresión con la media por grupo de ictericia

Con la librería *simputacion* se puede utilizar modelos de regresión para la imputación de los datos, en este caso particular se reemplazó a las variables cuantitativas *pesoalta* y *hematocritos* con su media, pero esta media es obtenida dentro de cada grupo o tipo de ictericia, lo cual ayuda acercar o mejorar un poco más a la situación real, a continuación se muestra el código y salida de una porción de la data real e imputada para ver la diferencia de los valores con los cuales han sido imputados.

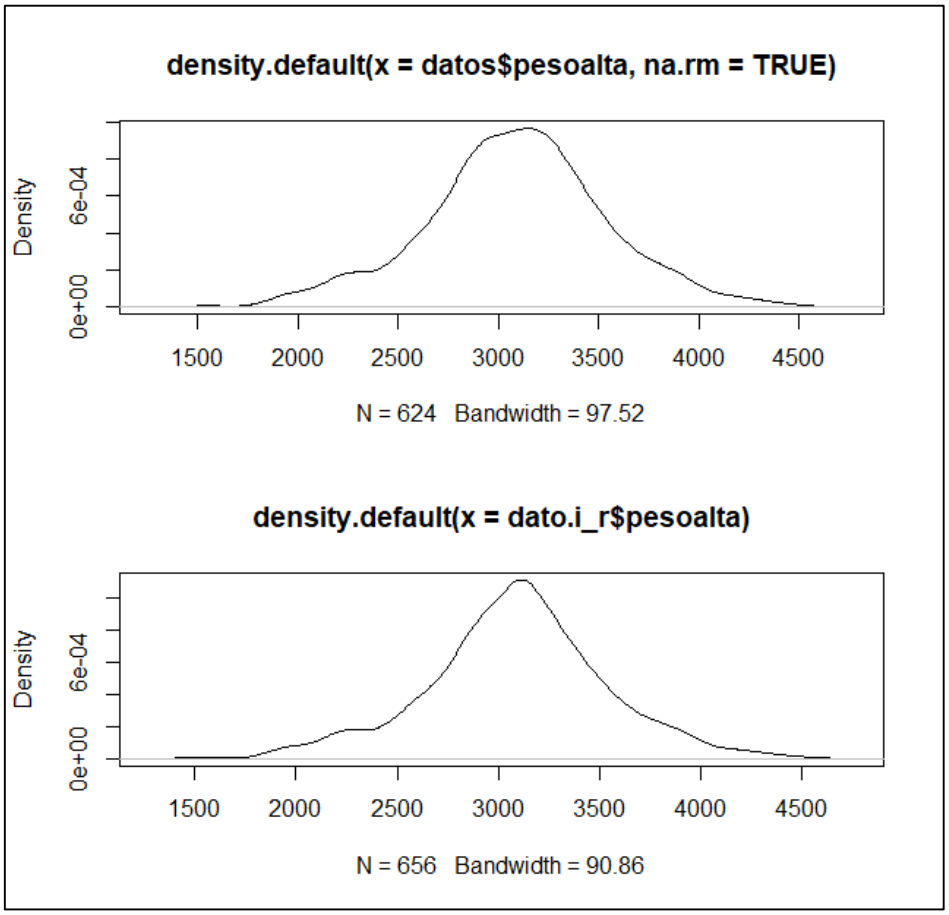
```r
library(simputation)

# Reemplazando por la media de cada tipo ictericia
dato.i_r <- impute_lm(datos, pesoalta + hematocr ~ 1 | tipoicter)  # s
i tengo pocos datos perdidos
datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285     2060 Pre termino       NA Fisiologico
## 10    2950     2950   A termino       NA Fisiologico
## 11    3310     3110   A termino       NA Fisiologico
## 12    3655     3350   A termino       NA Fisiologico
## 13    3120     2875   A termino       NA  Patologico
## 14    2950     2725   A termino       NA Fisiologico
## 59    3620       NA   A termino       68 Fisiologico
## 60    3440       NA   A termino       56 Fisiologico
## 61    3030       NA   A termino       60 Fisiologico
## 62    3140       NA   A termino       51 Fisiologico
## 63    3350       NA   A termino       52 Fisiologico
## 64    3030       NA   A termino       52 Fisiologico
## 307   3230       NA   A termino       44 Fisiologico
## 308   3440       NA   A termino       49  Patologico
## 309   1900       NA   A termino       58  Patologico
## 310   2560       NA   A termino       45 Fisiologico
## 311   3090     3050   A termino       54 Fisiologico
```

```r
dato.i_r[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

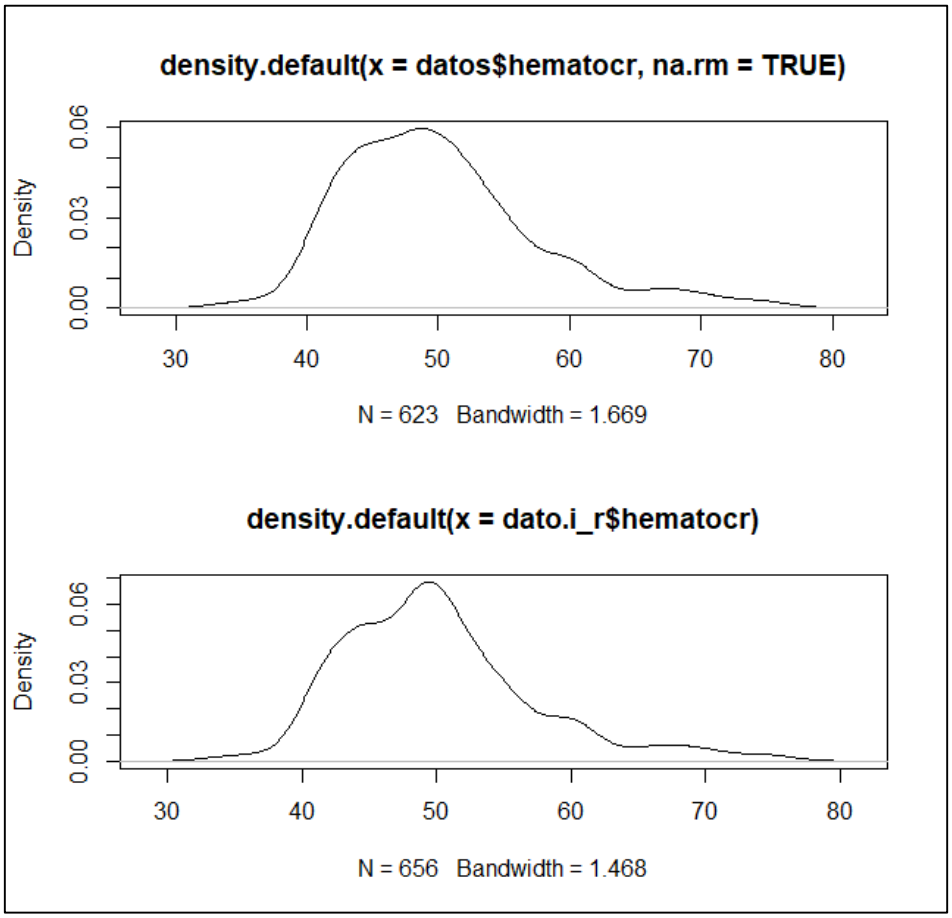
```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285 2060.000 Pre termino 49.46135 Fisiologico
## 10    2950 2950.000   A termino 49.46135 Fisiologico
## 11    3310 3110.000   A termino 49.46135 Fisiologico
## 12    3655 3350.000   A termino 49.46135 Fisiologico
## 13    3120 2875.000   A termino 51.37321  Patologico
## 14    2950 2725.000   A termino 49.46135 Fisiologico
## 59    3620 3114.116   A termino 68.00000 Fisiologico
## 60    3440 3114.116   A termino 56.00000 Fisiologico
## 61    3030 3114.116   A termino 60.00000 Fisiologico
## 62    3140 3114.116   A termino 51.00000 Fisiologico
## 63    3350 3114.116   A termino 52.00000 Fisiologico
## 64    3030 3114.116   A termino 52.00000 Fisiologico
## 307   3230 3114.116   A termino 44.00000 Fisiologico
## 308   3440 3036.349   A termino 49.00000  Patologico
## 309   1900 3036.349   A termino 58.00000  Patologico
## 310   2560 3114.116   A termino 45.00000 Fisiologico
## 311   3090 3050.000   A termino 54.00000 Fisiologico
```

Figura 10:Variable peso del alta del RN antes y después de completar datos con el método de la media por grupo de ictericia



Obsérvese en las *Figuras 10 y 11* donde se imputo y se completó valores perdidos que la distribución de las variables en el antes y después, existe una pequeña variación de la distribución de los datos, modificando la cúspide de estas variables, se podría decir que no es tan recomendado esta técnica en el momento de la praxis.

Figura 11: Variable hematocrito antes y después de completar datos con el método de la media por grupo de ictericia



5.3.3. Imputación usando modelos de regresión considerando variables predictoras extras.

```
# Considerando otras variables como predictoras
dato.i_rp <- impute_lm(datos, pesoalta + hematocr ~ Edad + edages + pe
sorn + tiehosp | tipoicter)  # talvez deberia ser solo sea.surface.Tem
p
datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285     2060 Pre termino       NA Fisiologico
## 10    2950     2950   A termino       NA Fisiologico
## 11    3310     3110   A termino       NA Fisiologico
## 12    3655     3350   A termino       NA Fisiologico
## 13    3120     2875   A termino       NA  Patologico
## 14    2950     2725   A termino       NA Fisiologico
## 59    3620       NA   A termino       68 Fisiologico
## 60    3440       NA   A termino       56 Fisiologico
## 61    3030       NA   A termino       60 Fisiologico
## 62    3140       NA   A termino       51 Fisiologico
## 63    3350       NA   A termino       52 Fisiologico
## 64    3030       NA   A termino       52 Fisiologico
## 307   3230       NA   A termino       44 Fisiologico
## 308   3440       NA   A termino       49  Patologico
```

```
## 309   1900       NA   A termino       58  Patologico
## 310   2560       NA   A termino       45 Fisiologico
## 311   3090     3050   A termino       54 Fisiologico
```

```
dato.i_rp[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285 2060.000 Pre termino 51.49282 Fisiologico
## 10    2950 2950.000   A termino 49.43825 Fisiologico
## 11    3310 3110.000   A termino 49.99756 Fisiologico
## 12    3655 3350.000   A termino 48.29898 Fisiologico
## 13    3120 2875.000   A termino 53.77667  Patologico
## 14    2950 2725.000   A termino 48.74166 Fisiologico
## 59    3620 3436.252   A termino 68.00000 Fisiologico
## 60    3440 3258.837   A termino 56.00000 Fisiologico
## 61    3030 2924.118   A termino 60.00000 Fisiologico
## 62    3140 2995.414   A termino 51.00000 Fisiologico
## 63    3350 3287.238   A termino 52.00000 Fisiologico
## 64    3030 2896.869   A termino 52.00000 Fisiologico
## 307   3230 3097.885   A termino 44.00000 Fisiologico
## 308   3440 3246.696   A termino 49.00000  Patologico
## 309   1900 1895.749   A termino 58.00000  Patologico
## 310   2560 2434.404   A termino 45.00000 Fisiologico
## 311   3090 3050.000   A termino 54.00000 Fisiologico
```

Figura 12: Variable peso del alta del RN antes y después de completar datos usando modelos de regresión considerando variables predictoras extras.

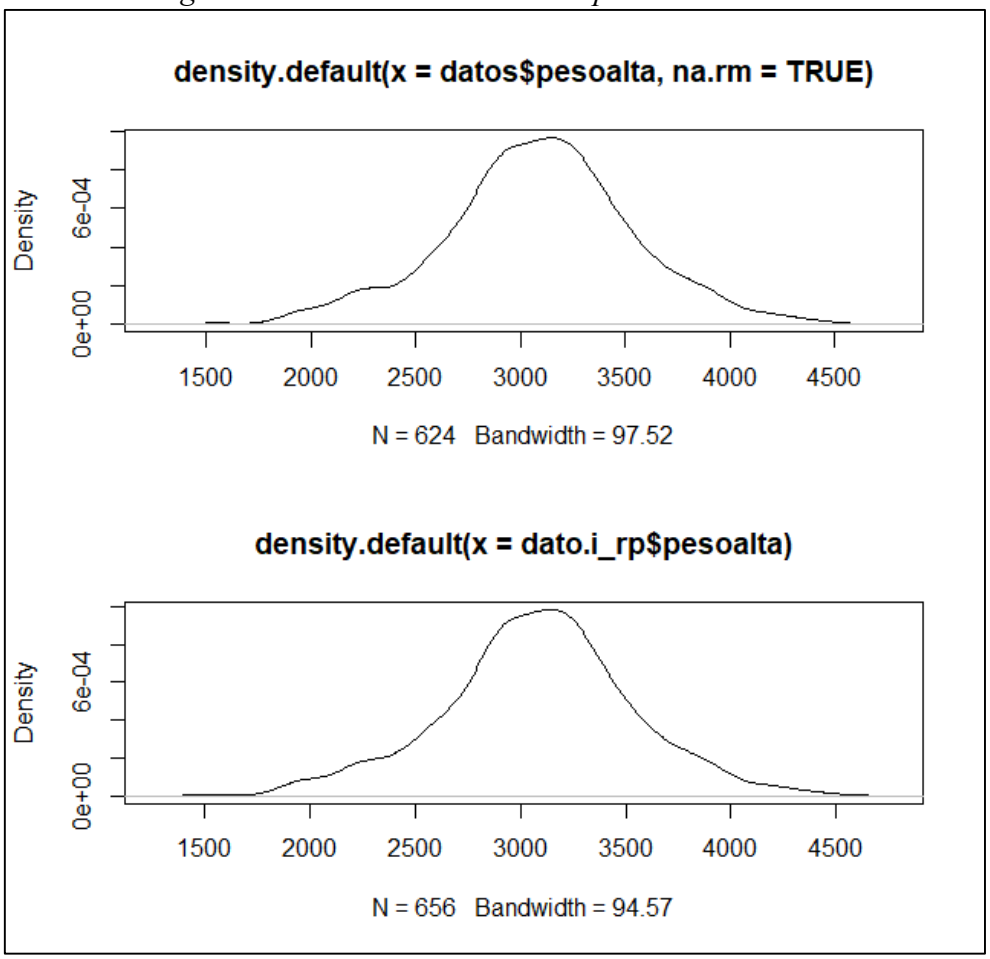
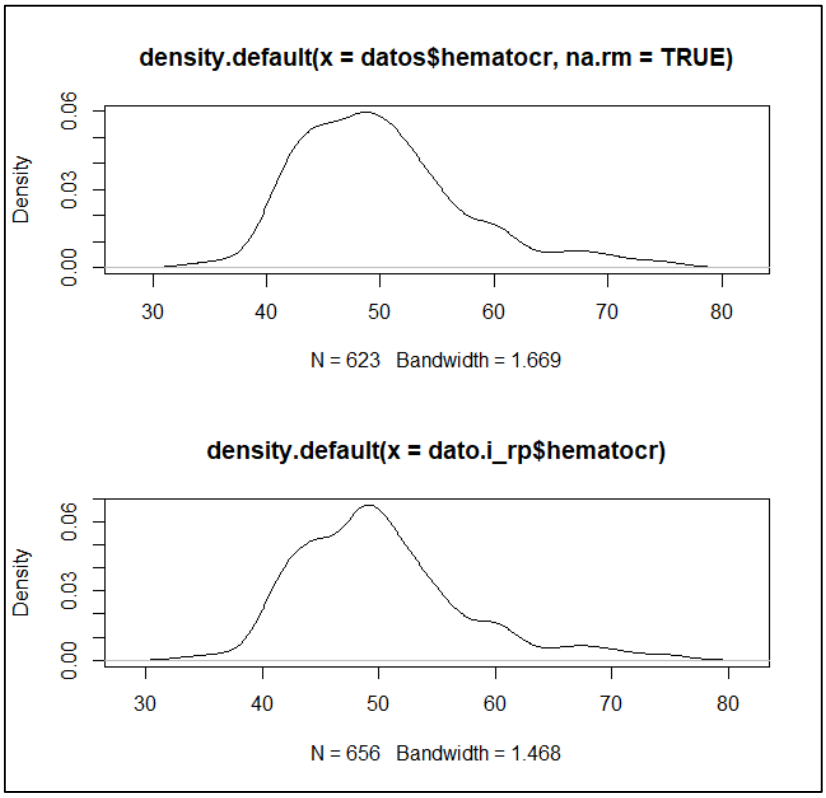


Figura 13:Variable hematocritos antes y después de completar datos usando modelos de regresión considerando variables predictoras extras.



En las *Figuras 12 y 13* donde se imputo y se completó valores perdidos, la distribución de la variable peso del alta del RN en el antes y después presentan semejanzas, entonces se pudo observar que la variable está muy bien representada, mientras que la otra variable hematocritos en la parte central presenta una elevación distinta del comportamiento original de la data no considerando los datos perdidos.

5.3.4. Imputación usando modelos de regresión adicionando un residuo aleatorio.

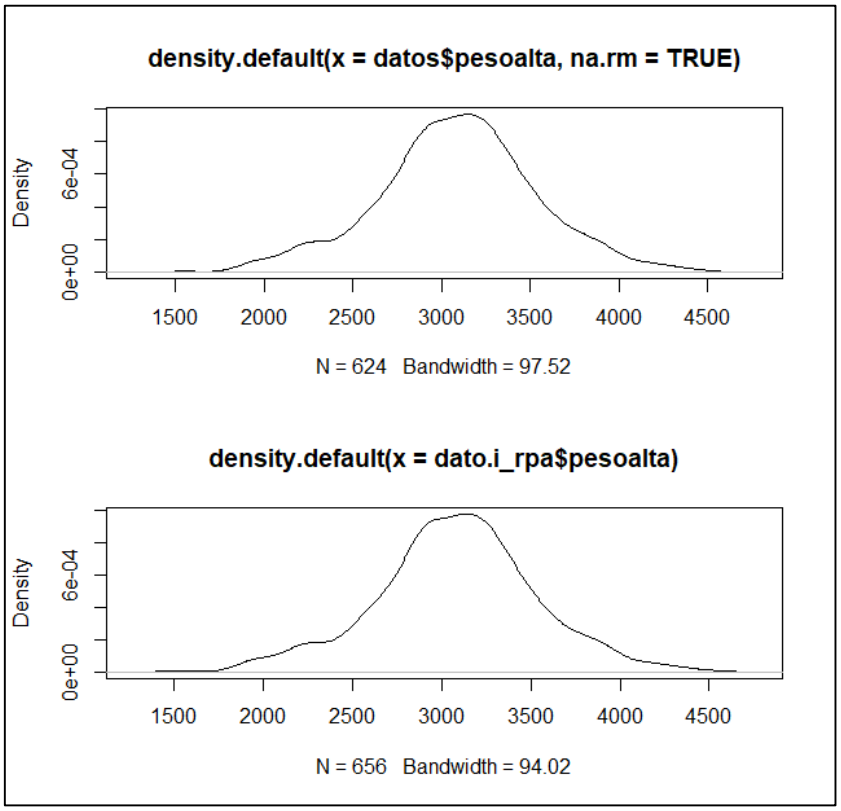
```
# Adicionando un residuo aleatorio
dato.i_rpa <- impute_lm(datos, pesoalta + hematocr ~ Edad + edages + p
esorn + tiehosp, add_residual = "normal") ## es como poner prediction
en regresion
datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]

##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285     2060 Pre termino       NA Fisiologico
## 10    2950     2950   A termino       NA Fisiologico
## 11    3310     3110   A termino       NA Fisiologico
## 12    3655     3350   A termino       NA Fisiologico
## 13    3120     2875   A termino       NA  Patologico
```

```
## 14    2950     2725   A termino       NA Fisiologico
## 59    3620       NA   A termino       68 Fisiologico
## 60    3440       NA   A termino       56 Fisiologico
## 61    3030       NA   A termino       60 Fisiologico
## 62    3140       NA   A termino       51 Fisiologico
## 63    3350       NA   A termino       52 Fisiologico
## 64    3030       NA   A termino       52 Fisiologico
## 307   3230       NA   A termino       44 Fisiologico
## 308   3440       NA   A termino       49  Patologico
## 309   1900       NA   A termino       58  Patologico
## 310   2560       NA   A termino       45 Fisiologico
## 311   3090     3050   A termino       54 Fisiologico
```

```
dato.i_rpa[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

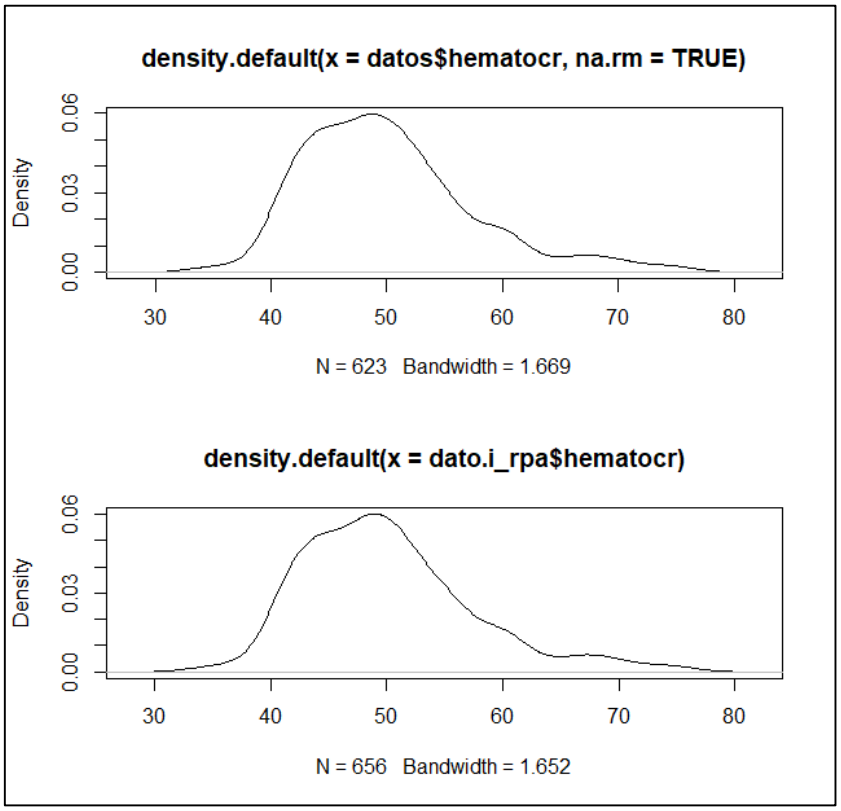
```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285 2060.000 Pre termino 53.61396 Fisiologico
## 10    2950 2950.000   A termino 46.32558 Fisiologico
## 11    3310 3110.000   A termino 55.37611 Fisiologico
## 12    3655 3350.000   A termino 52.04203 Fisiologico
## 13    3120 2875.000   A termino 61.88478  Patologico
## 14    2950 2725.000   A termino 44.50021 Fisiologico
## 59    3620 3497.637   A termino 68.00000 Fisiologico
## 60    3440 3084.031   A termino 56.00000 Fisiologico
## 61    3030 2888.894   A termino 60.00000 Fisiologico
## 62    3140 3092.591   A termino 51.00000 Fisiologico
## 63    3350 3394.034   A termino 52.00000 Fisiologico
## 64    3030 2736.444   A termino 52.00000 Fisiologico
## 307   3230 2786.341   A termino 44.00000 Fisiologico
## 308   3440 3389.106   A termino 49.00000  Patologico
## 309   1900 1460.003   A termino 58.00000  Patologico
```

Figura 14: Variable peso alta antes y después de completar datos usando modelos de regresión adicionando un residuo aleatorio.



En la *Figura 14*, obsérvese que la distribución de la variable peso del alta del RN en el antes y después presentan similaridades, entonces se pudo observar que la variable está muy bien representada, además en la *Figura 15* también la variable hematocritos en el después tiene mayor similaridad en comparación con anteriores metodologías de imputación de datos.

Figura 15: Variable hematocrito antes y después de completar datos usando modelos de regresión adicionando un residuo aleatorio.



5.3.5. Imputación usando k vecinos más cercanos

```
# Usando la libreria VIM
dato_vars <- c("pesoalta","hematocr","embarazo")
dato_i_knn <- VIM::kNN(data=datos, variable=dato_vars)

datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285     2060 Pre termino       NA Fisiologico
## 10    2950     2950   A termino       NA Fisiologico
## 11    3310     3110   A termino       NA Fisiologico
## 12    3655     3350   A termino       NA Fisiologico
## 13    3120     2875   A termino       NA  Patologico
## 14    2950     2725   A termino       NA Fisiologico
## 59    3620       NA   A termino       68 Fisiologico
## 60    3440       NA   A termino       56 Fisiologico
## 61    3030       NA   A termino       60 Fisiologico
## 62    3140       NA   A termino       51 Fisiologico
## 63    3350       NA   A termino       52 Fisiologico
## 64    3030       NA   A termino       52 Fisiologico
## 307   3230       NA   A termino       44 Fisiologico
## 308   3440       NA   A termino       49  Patologico
## 309   1900       NA   A termino       58  Patologico
## 310   2560       NA   A termino       45 Fisiologico
## 311   3090     3050   A termino       54 Fisiologico
```

```
dato.i_rpa[c(9:14,59:64,307:311),c(15:16,24,29,33)]
```

```
##     pesorn pesoalta    embarazo hematocr   tipoicter
## 9     2285 2060.000 Pre termino 53.61396 Fisiologico
## 10    2950 2950.000   A termino 46.32558 Fisiologico
## 11    3310 3110.000   A termino 55.37611 Fisiologico
## 12    3655 3350.000   A termino 52.04203 Fisiologico
## 13    3120 2875.000   A termino 61.88478  Patologico
## 14    2950 2725.000   A termino 44.50021 Fisiologico
## 59    3620 3497.637   A termino 68.00000 Fisiologico
## 60    3440 3084.031   A termino 56.00000 Fisiologico
## 61    3030 2888.894   A termino 60.00000 Fisiologico
## 62    3140 3092.591   A termino 51.00000 Fisiologico
## 63    3350 3394.034   A termino 52.00000 Fisiologico
## 64    3030 2736.444   A termino  52.00000 Fisiologico
## 307   3230 2786.341   A termino 44.00000 Fisiologico
## 308   3440 3389.106   A termino 49.00000  Patologico
## 309   1900 1460.003   A termino 58.00000  Patologico
## 310   2560 2349.866   A termino 45.00000 Fisiologico
## 311   3090 3050.000   A termino 54.00000 Fisiologico
```

En la *Figura 16*, obsérvese que la distribución de la variable peso del alta del RN en el antes y después presentan similaridades entonces se pudo observar que la variable está muy bien representada con el algoritmo de los vecinos más cercanos, además en la *Figura 17* también la variable hematocritos en el después tiene mayor similaridad lo cual permite a esta metodología competir con los métodos de regresión quienes eran los que mejor comportamiento de completar datos tienen.

Figura 16: Variable pesoalta antes y después de completar datos usando modelos usando k vecinos más cercanos

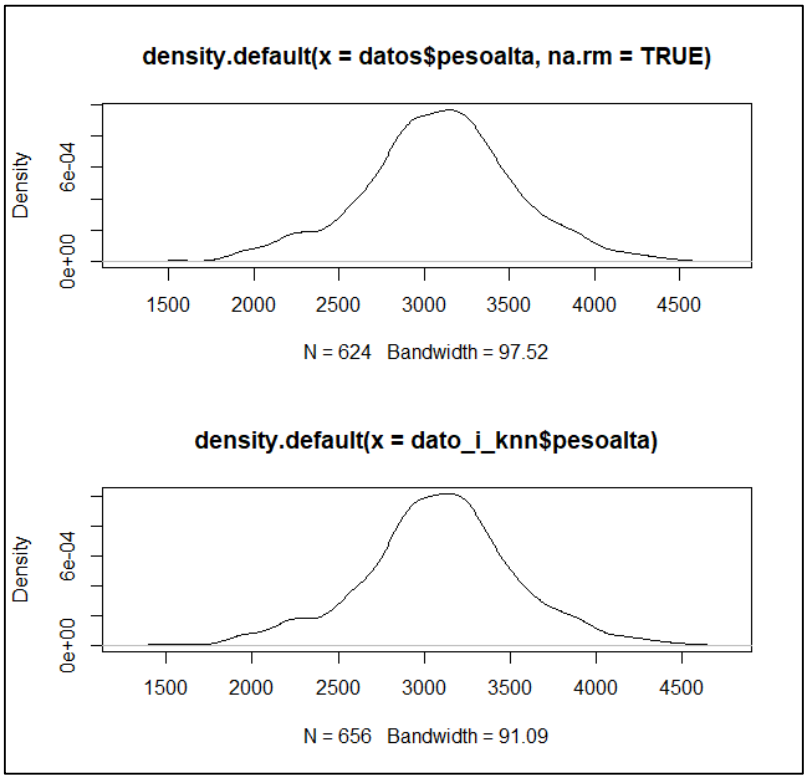
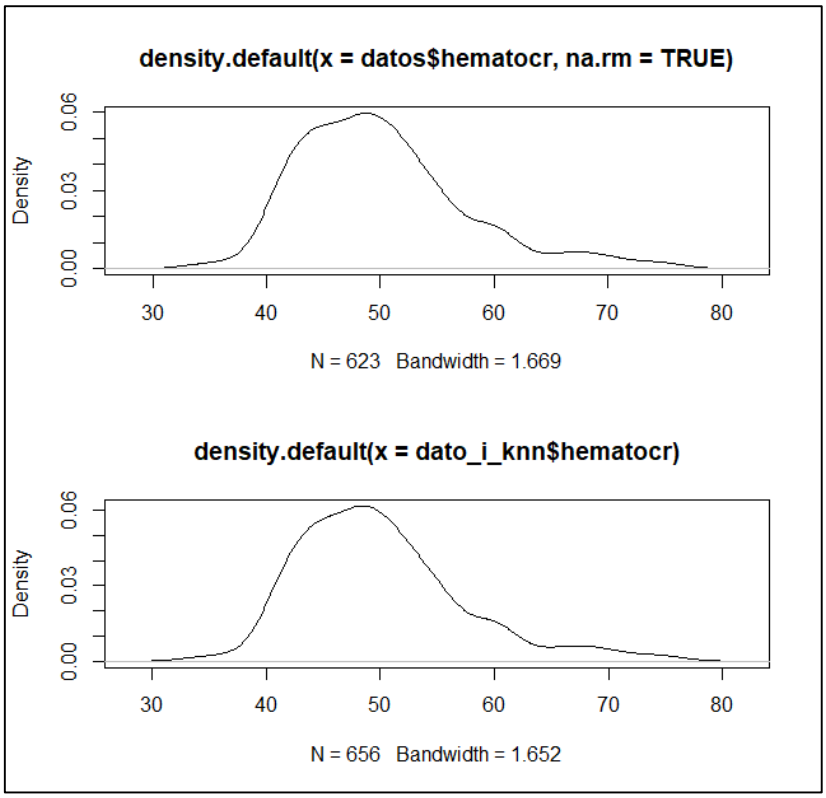


Figura 17: Variable hematocrito antes y después de completar datos usando modelos usando k vecinos más cercanos



5.4. COMPARACIÓN DE LAS TÉCNICAS UTILIZADAS PARA LA IMPUTACIÓN DE DATOS

Tabla 2: Comparación de técnicas de imputación de datos faltantes en la variable Peso del alta del Recién nacido.

	Variable peso de la alta médica del Recién Nacido					
Técnica	**Original**	**Reemplazo con la Mediana**	**Regresión media grupo ictericia**	**Regresión con predictores**	**Regresión con ruido aleatorio**	**KNN Vecinos más cercanos**
n	624	656	656	656	656	656
Media	3088.07	3088.65	3088.51	3080.62	3081.94	3087.79
Desviación estándar	451.81	440.64	440.7	451.79	449.52	442.17
Mediana	3100	3100	3111.5	3098.32	3098.5	3100
Mínimo	1545	1545	1545	1545	1545	1545
Máximo	4490	4490	4490	4490	4490	4490
Rango	2945	2945	2945	2945	2945	2945
Asimetría	-0.08	-0.08	-0.08	-0.09	-0.08	-0.08
Curtosis	0.36	0.53	0.53	0.37	0.40	0.49
Se	18.09	17.2	17.21	17.64	17.55	17.26

Se puede observar en la *Tabla 2*, que las comparaciones en términos de la media, la técnica que tiene mayor acercamiento es la de KNN Vecinos más cercanos, seguido de Regresión media grupo ictericia y por último la técnica reemplazando con la mediana, bueno en términos de la desviación estándar esta la técnica de la regresión con predictores, seguido de la regresión con ruido aleatorio.

Mientras que en términos de la asimetría todos conservan un valor muy cercano o igual a -0.08, por tanto, en base a este indicador todas las técnicas van bien, ahora con respecto a la curtosis la técnica que obtuvo un valor más cercano al de la data sin imputar fue la técnica de la regresión con predictores, seguido de la regresión con ruido aleatorio.

Tabla 3: Comparación de técnicas de imputación de datos faltantes en la variable hematocritos

	Variable hematocritos					
Técnica	**Original**	**Reemplazo con la Mediana**	**Regresión media grupo ictericia**	**Regresión con predictores**	**Regresión con ruido aleatorio**	**KNN Vecinos más cercanos**
N	623	656	656	656	656	656
Media	50.1	50.05	50.09	50.09	50.1	50.02
Desviación estándar	7.41	7.23	7.23	7.23	7.39	7.27
Mediana	49	49	49.46	49	49	49
Mínimo	33	33	33	33	33	33
Máximo	77	77	77	77	77	77
Rango	44	44	44	44	44	44
Asimetría	0.94	0.99	0.97	0.97	0.91	0.97
Curtosis	1.04	1.27	1.25	1.24	0.99	1.19
Se	0.3	0.28	0.28	0.28	0.29	0.28

Se puede observar en la *Tabla 3*, en la imputación de datos de la variable hematocrito que las comparaciones en términos de la media la técnica que tiene mayor acercamiento es la de Regresión con ruido aleatorio seguido de KNN Vecinos más cercanos, bueno en términos de la desviación estándar esta la técnica de Regresión con ruido aleatorio seguido de KNN Vecinos más cercanos también.

Mientras que en términos de la asimetría todos conservan un valor muy cercano o igual a 0.94, por tanto, en base a este indicador todas las técnicas van bien, ahora con respecto a la curtosis la técnica que obtuvo un valor más cercano al de la data sin imputar fue la Regresión con ruido aleatorio.

5.5. ANÁLISIS DE LA ICTERICIA NEONATAL

```
table(dato.i_rp$tipoicter)
## Fisiologico  Patologico
##         440         216
prop.table(table(dato.i_rp$tipoicter))*100
## Fisiologico  Patologico
##    67.07317    32.92683
```

Se puede observar que del total de niños que presentaron ictericia neonatal el 67.07% presentaba una ictericia del tipo fisiológico mientras que el 32.93% de los recién nacidos tenía una ictericia neonatal del tipo patológico.

Factores asociados al tipo de ictericia mediante regresión logística

```
modelo1 = glm(tipoicter~Edad+gestas+hijos+edages+rhmadre+itu+trataitu+
pesorn+pesoalta+gruporn,data = dato.i_rp, family = "binomial")
summary(modelo1)
```

```
##
## Call:
## glm(formula = tipoicter ~ Edad + gestas + hijos + edages + rhmadre +
##     itu + trataitu + pesorn + pesoalta + gruporn, family = "binomial",
##     data = dato.i_rp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0075  -0.8939  -0.7715   1.3132   1.8092
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.839e+00  2.204e+00   0.834   0.4042
## Edad        -3.077e-03  1.685e-02  -0.183   0.8551
## gestas       2.038e-01  1.378e-01   1.478   0.1393
## hijos       -2.121e-01  1.652e-01  -1.284   0.1992
## edages      -6.161e-02  6.177e-02  -0.997   0.3186
## rhmadre-    -1.368e+01  4.365e+02  -0.031   0.9750
## itu          3.385e-01  1.588e-01   2.131   0.0331 *
## trataituNo   3.219e-01  1.903e-01   1.691   0.0907 .
## pesorn      -7.777e-04  5.234e-04  -1.486   0.1373
## pesoalta     5.137e-04  5.270e-04   0.975   0.3296
## grupornA     3.868e-01  2.353e-01   1.644   0.1002
## grupornB     7.143e-01  3.396e-01   2.103   0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

Se puede observar del modelo de regresión logística que las variables que se encuentran relacionadas con el tipo de ictericia son el ITU (Número de Infecciones de tracto unitario) y el grupo B sanguíneo del recién nacido. Y hasta podría ser la variable los que no tuvieron un tratamiento para la ITU (trataituno)

```
confint(modelo1)
```

```
##                     2.5 %       97.5 %
## (Intercept) -2.4841448508 6.174477e+00
## Edad        -0.0361810858 2.998174e-02
## gestas      -0.0689602734 4.735566e-01
## hijos       -0.5366138282 1.131735e-01
## edages      -0.1832933635 5.928196e-02
## rhmadre-               NA 3.635829e+01
## itu          0.0265921018 6.502770e-01
## trataituNo  -0.0516551808 6.951411e-01
## pesorn      -0.0018256811 2.337333e-04
## pesoalta    -0.0005067543 1.569242e-03
## grupornA    -0.0800942074 8.447018e-01
## grupornB     0.0400288223 1.380381e+00
```

```
exp(confint(modelo1))
```

```
##                   2.5 %       97.5 %
## (Intercept) 0.08339684 4.803318e+02
## Edad        0.96446563 1.030436e+00
## gestas      0.93336376 1.605695e+00
## hijos       0.58472488 1.119826e+00
## edages      0.83252389 1.061074e+00
## rhmadre-            NA 6.168866e+15
## itu         1.02694883 1.916072e+00
## trataituNo  0.94965627 2.003992e+00
## pesorn      0.99817598 1.000234e+00
## pesoalta    0.99949337 1.001570e+00
## grupornA    0.92302939 2.327284e+00
## grupornB    1.04084077 3.976415e+00
```

El factor de riesgo para tener una ictericia fisiológica esta fluctuando de 1.04 a 3.97 con más probabilidad de presentar este tipo de ictericia para aquellos recién nacidos que tienen grupo sanguíneo B, en comparación al grupo sanguíneo O, además podemos afirmar que existe de 1.02 a 1.91 más probabilidad de tener ictericia fisiológica por cada incremento en el número de ITU (infecciones de tracto urinario)

5.6. DISCUSIÓN DE RESULTADOS

(Manotas, 2005) En su estudio “Ictericia Neonatal”, llega a conclusiones, las enfermedades hematológicas neonatales, especialmente las del tipo hemolítico, son menos frecuentes en la actualidad, pero algunas son tan graves que pueden afectar de

manera irreversible al sistema nervioso central; por tanto, la tendencia actual es mejorar las medidas preventivas con terapias efectivas y de aplicación temprana que permitan disminuir las repercusiones de dichas enfermedades a corto y largo plazo. Además, con el presente estudio se pudo observar que el grupo B sanguíneo del recién nacido tiene relación con la ictericia fisiológica confirmando lo concluido por Manotas.

(Caiza & Colaboradores, 2006) Los factores de riesgo encontrados para recién nacidos prematuros al compararlos con neonatos a término fueron: cesárea, infección neonatal; la enfermedad de membrana hialina asociada a la prematurez fue el factor de riesgo más importante encontrado en este estudio. Nosotros en el presente estudio encontramos que el factor de riesgo para tener una ictericia fisiológica esta fluctuando de 1.04 a 3.97 con más probabilidad de presentar este tipo de ictericia para aquellos recién nacidos que tienen grupo sanguíneo B, en comparación al grupo sanguíneo O, además podemos afirmar que existe de 1.02 a 1.91 más probabilidad de tener ictericia fisiológica por cada incremento en el número de ITU (infecciones de tracto urinario), recordemos que a diferencia de los otros estudios nosotros teníamos una base de datos de recién nacidos que ya presentaban ictericia, y lo que se está investigando es cuales son los factores de mayor riesgo asociado al tipo de ictericia (Fisiológico y patológico)

(Bazalar, 2014) En su estudio de "prevalencia y causas de ictericia neonatal" llegó a la conclusión que la prevalencia de ictericia neonatal es de 4,2% de recién nacidos con ictericia siendo las principales causas de ictericia neonatal, la ictericia fisiológica, hipo alimentación, incompatibilidad de grupo sanguíneo ABO las cuales se podrían afirmar de alguna manera con nuestro estudio que el grupo sanguíneo es un factor importante para el desarrollo de la ictericia en los neonatos.

CONCLUSIONES

1. La imputación de datos de la variable peso en el alta médica del Recién Nacido en términos de la media la técnica que tiene mayor acercamiento es KNN Vecinos más cercanos, en términos de la desviación estándar esta la técnica de la regresión con predictores, en términos de la asimetría todos conservan un valor muy cercano o igual a -0.08, por tanto, en base a este indicador todas las técnicas presentan mejor desempeño, ahora con respecto a la curtosis la técnica que obtuvo un valor más cercano fue con la técnica de la regresión con predictores.

2. La imputación de datos de la variable hematocrito en términos de la media la técnica que tiene mayor acercamiento es la de Regresión con ruido aleatorio, en términos de la desviación estándar esta la técnica de Regresión con ruido aleatorio, mientras que en términos de la asimetría todos conservan un valor muy cercano o igual a 0.94, por tanto, en base a este indicador todas las técnicas van bien, ahora con respecto a la curtosis la técnica que obtuvo un valor más cercano fue la Regresión con ruido aleatorio.

3. En general las técnicas basadas en regresión presentan una adecuada performance a la hora de imputar datos faltantes, como se ha podido determinar en el caso de la presente investigación.

RECOMENDACIONES

- Se recomienda a las personas que trabajan con datos, utilizar las técnicas basadas en regresión con predictores y con ruido aleatorio para la imputación de valores faltantes en una data.
- Antes de imputar los datos se recomienda observar si existe algún patrón de perdida de datos como mecanismo no aleatorio de datos de ser así las técnicas utilizadas en la presente investigación tendrían que ser ajustadas al patrón de datos que se encontró.

REFERENCIAS BIBLIOGRÁFICAS

Abraira, V. (1996). *metodos multivariados en bioestadistica .* Madrid, España: Centro de estudios Ramon Areces.

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. En D. Banks, F. McMorris, P. Arabie, & W. Gaul, *Classification, Clustering, and Data Mining Applications* (págs. 639-647). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-17103-1_60

Alvarez, R. (1995). *Estadística multivariante y no paramétrica con SPSS. Aplicación a las ciencias de la salud.* Madrid, España: Diaz de Santos S.A.

Alvear, C. (2011). Ictericia fisiologica en recien nacidos a termino en el 2011.

Andersen, E. (1990). *Introduction to the statistical analysis of categorical data.* New York, U.S.A.

Ari, E. (2016). Using Multinomial Logistic Regression to Examine the Relationship Between Children's Work Status and Demographic Characteristics. *Research Journal of Politics, Economics and Management*, 77-93.

Bazalar. (2014). Prevalencia y causas de ictericia neonatal. Huancayo.

Buck, S. (1960). Method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of he Royal Statistics Society, 22*, 302-307.

Caiza, & Colaboradores. (2006). Estudio Descriptivo. Ecuador.

Castro, M. (2014). *Imputación de datos faltantes en un modelo de tiempo de fallo acelerado.* España: Universidad de Coruña, Santiago Compostela y Vigo.

Cohen, J., & Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L., Schafer, J., & Kam, C. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods, 6*, 330-351. Obtenido de https://doi.org/10.1037/1082-989X.6.4.330

Dagnino, J. (2014). Datos faltantes (Missing values). *Revista Chilena de Anestesia*, 332-334.

De la Vera, R., & Montenegro, G. (2022). *Caracterización clínica neurológica de ictericia neonatal asociado a patologías perinatales.* Guayaquil: Universidad de Guayaquil. Obtenido de http://repositorio.ug.edu.ec/bitstream/redug/67853/1/CD%203706-

%20DE%20LA%20VERA%20CARRION%2c%20RONNY%20DAMIAN%3b%20MONTENEGRO%20AVATA%2c%20GREGORY%20MIGUEL.pdf

Dueñas, M. (s.f.). *Modelos de respuesta discreta en R y aplicación con datos reales.* España: Universidad de Granada.

Failachea, o. (2002). Ictericia neonatal. Uruguay.

Galarza, L. (2013). *Comparación mediante simulación de los métodos em e imputación múltiple para datos faltantes.* Lima: Universidad Nacional Mayor de San Marcos.

Garcia, R., & Palacios, D. (2013). *Analisis y algoritmo de selección de técnicas determinísticas y estocásticas de imputación de datos.* El salvador: UNIVERSIDAD DE EL SALVADOR.

Goicochea, P. (2002). *Imputación basada en árboles de clasificación.* Eustat. Obtenido de https://www.eustat.eus/document/datos/ct_04_c.pdf

Gómez Hernández, S., & Palacios Arias, D. (2013). *Modelación logística multinomial para clasificar los hogares de El Salvador por nivel de pobreza.* El salvador: Universidad de El salvador.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación.* México D.F.: McGraw-Hill.

Hosmer, D., & Lemeshow, S. (1989). *Applied logistic Regression.* New York, U.S.A.: John Wiley & Sons.

Lerdo de Tejada, M. (2014). *Estimación de datos faltantes con el Algoritmo EM.* México: Universidad Nacional Autónoma de México. Obtenido de https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000708980/3/0708980.pdf

Li, J., Bioucas-Dias, J., & Plaza, A. (2012). Spectral–Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields. *in IEEE Transactions on Geoscience and Remote Sensing*, 809-823.

Little, R. (1986). A test of Missing Completely at Random for multivariate data with missing values. *Sociological Methods and Research*.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics.* New York: John Wiley & Sons, Inc.

Manotas. (2005). Ictericia Neonatal. Montevideo.

Morachino, G. (2011). Correlación entre bilirrubina sérica y bilirrubinemia transcutanea en neonatos ictéricos. Trujillo, Perú.

Ortiz, P. (2010). *Ictericia clinica en neonatos y correlación con valores sericos de bilirrubina. Hospital José María Velasco Ibarra. Tena 2010.* RioBamba-Ecuador: Escuela Superior Politecnica de Chimborazo.

Osorio, D., Ospina, J., & Lenis, D. (2009). Planteamiento del modelo logístico multinomial a través de la funcion canónica de enlace de la familia exponencial. *Heurística, 16*, 105-115.

Parodi, & colaborades. (2005). Ictericia neonatal. Uruguay.

Perez. (2006). Estudio prospectivo. Huancayo.

Quesada. (2011). Observacional de hiperbilirrubinemia neonatal. Ecuador.

Rodriguez López, J. (2017). *Metodos numericos para la aproximacion ed reaices multiples de ecuaciones no lineales.* España: Universidad de Salamanca.

Rodriguez, B. R. (2001). *Hiperbilirrubinemia neonatal.* Mc Graw.Hill Interamericana.

Roque, M. (2018). *Modelos de regresión logística multinomial de la calidad de la fibra de alpaca Huacaya en función de sus caracteristicas:sexo y edad -Corani, Carabaya, Puno-2017.* Puno: Universidad Nacional del Altiplano.

Rovine, M., & Delaney, M. (1990). Missing Data Estimation in Developmental Research. (A. V. ed., Ed.) *Statistical Methods in Longitudinal Research: Principles and Structuring Change, 1*, 35-79.

Rubin, D. (1976). Inference and Missing Data. *Journal of the Royal Statistical Society. Biometrika, 63*(3).

Sanchez, H., Reyes, C., & Mejía, K. (2018). *Manual de términos en investigación científica, tecnológica y humanística.* Lima, Perú: Universidad Ricardo Palma.

Schafer, J., & Graham, J. (2002). Missing Data, Our View of the State of the Art. *Psychological Methods, 7*(2).

Silva, A. (1990). *Excursion a la regresion logistica en ciencias de la salud.* Madrid, España.

Trejos, M., & Umanzor, G. (2018). *Factores de riesgo de ictericia en recién nacidos del Hospital Escuela "Dr. Oscar Danilo Rosales Argüello", León. Noviembre 2017 - Abril 2018.* Nicaragua: Universidad Nacional Autónoma de Nicaragua. Obtenido de http://riul.unanleon.edu.ni:8080/jspui/bitstream/123456789/7337/1/241494.pdf

INSTRUMENTO

FICHA DE RECOLECCIÓN DE DATOS

FACTORES DE RIESGO E ICTERICIA NEONATAL EN EL HOSPITAL REGIONAL DEL CUSCO

1.- MADRE

Edad:…….

Procedencia: Urbano ( ) Rural ( )

Estado Socioeconómico: A( ) B( ) C( ) D( )

Edad gestacional (FUR):……..

Controles prenatales:…….

La lactancia materna es exclusiva: Si ( ) No ( )

Incompatibilidad Sanguínea ABO: Si ( ) No ( )

Tipo de parto: eutócico o vaginal ( ) Distócico o cesárea ( )

Infecciones del tracto urinario: Si presento ( ) No presento ( )

Oxitocina: Si( ) No ( )

NEONATO:

Sexo: Masculino ( ) Femenino ( )

Policitemia: Presenta ( ) No presenta ( )

Asfixia: Presenta ( ) No presenta ( )

Cefalohematoma Presenta ( ) No presenta ( )

Sepsis neonatal: Presenta ( ) No presenta ( )

CODIGO DE R

```
library(foreign)
datos <-read.spss("tesis final solo ictericia.sav",
                  use.value.labels=TRUE,
                  to.data.frame=TRUE)
# Aggregation plot
x11()
a=aggr(datos,numbers=T)
a
summary(a)
aggr(datos,numbers=T, sortComb=TRUE,
     sortVar=TRUE, only.miss=TRUE)
# Prueba t de medias
t.test(edages ~ is.na(hematocr), data=datos)  #
t.test(edages ~ is.na(pesoalta), data=datos)  #
#  Imputación
# Usando una medida de Tendencia Central  1% al 5%
library(DMwR2)
datos.c<-centralImputation(datos)
summary(datos.c)
datos_i<-initialise(datos,method="median")
summary(datos_i)
# variable peso alta
mean(datos$pesoalta, na.rm =TRUE)
median(datos$pesoalta, na.rm =TRUE)
x11()
par(mfrow=c(2,1))
plot(density(datos$pesoalta, na.rm =TRUE))
plot(density(datos_i$pesoalta))
```

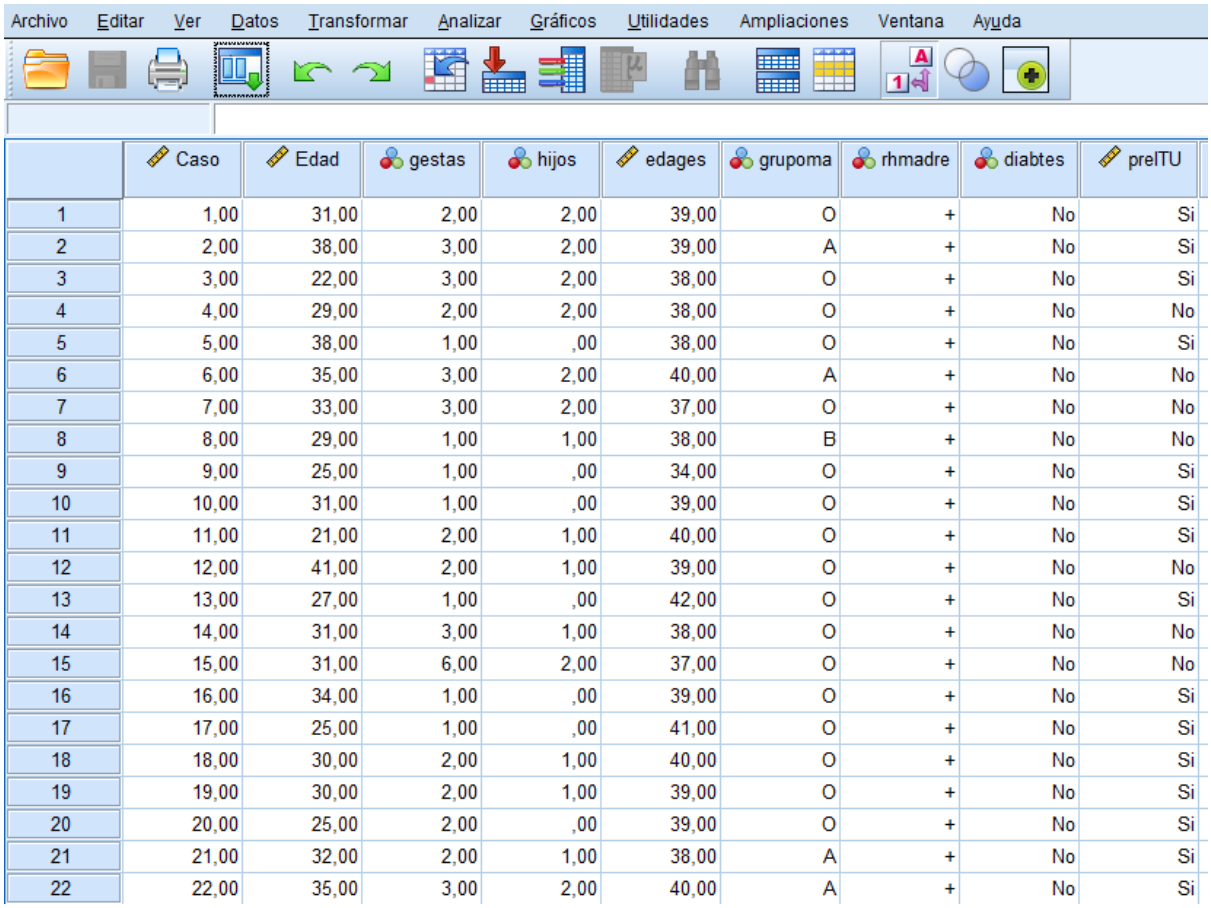
```
# variable hematocritos

mean(datos$hematocr, na.rm =TRUE)

median(datos$hematocr, na.rm =TRUE)

x11()

par(mfrow=c(2,1))

plot(density(datos$hematocr, na.rm =TRUE))

plot(density(datos_i$hematocr))


# variable hematocritos

table(datos$embarazo)

table(datos_i$embarazo)


par(mfrow=c(2,1))

barplot(table(factor(datos$embarazo,levels=c("A termino","Pre
termino"))), col=c("skyblue", "red"), main= "Distribución con datos
faltantes")

barplot(table(datos_i$embarazo),col=c("skyblue", "red"), main=
"Distribución con datos Completados")


#

# Usando Modelos de Regresión

# Reemplazando por la media de cada tipo ictericia

dato.i_r <- impute_lm(datos, pesoalta + hematocr ~ 1 | tipoicter)  #
si tengo pocos datos perdidos

datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]

dato.i_r[c(9:14,59:64,307:311),c(15:16,24,29,33)]

par(mfrow=c(2,1))

plot(density(datos$pesoalta, na.rm =TRUE))

plot(density(dato.i_r$pesoalta))
```

```
par(mfrow=c(2,1))

plot(density(datos$hematocr, na.rm =TRUE))

plot(density(dato.i_r$hematocr))


# Considerando otras variables como predictoras

dato.i_rp <- impute_lm(datos, pesoalta + hematocr ~ Edad + edages +
pesorn + tiehosp | tipoicter)  # talvez deberia ser solo
sea.surface.Temp

datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]

dato.i_rp[c(9:14,59:64,307:311),c(15:16,24,29,33)]

par(mfrow=c(2,1))

plot(density(datos$pesoalta, na.rm =TRUE))

plot(density(dato.i_rp$pesoalta))


par(mfrow=c(2,1))

plot(density(datos$hematocr, na.rm =TRUE))

plot(density(dato.i_rp$hematocr))


# Adicionando un residuo aleatorio

dato.i_rpa <- impute_lm(datos, pesoalta + hematocr ~ Edad + edages +
pesorn + tiehosp, add_residual = "normal") ## es como poner
prediction en regresion

datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]

dato.i_rpa[c(9:14,59:64,307:311),c(15:16,24,29,33)]

par(mfrow=c(2,1))

plot(density(datos$pesoalta, na.rm =TRUE))

plot(density(dato.i_rpa$pesoalta))


par(mfrow=c(2,1))

plot(density(datos$hematocr, na.rm =TRUE))

plot(density(dato.i_rpa$hematocr))
```

```
# K-Vecinos mas cercanos
dato_vars <- c("pesoalta","hematocr","embarazo")
dato_i_knn <- VIM::kNN(data=datos, variable=dato_vars)

datos[c(9:14,59:64,307:311),c(15:16,24,29,33)]
dato_i_knn[c(9:14,59:64,307:311),c(15:16,24,29,33)]
par(mfrow=c(2,1))
plot(density(datos$pesoalta, na.rm =TRUE))
plot(density(dato_i_knn$pesoalta))

par(mfrow=c(2,1))
plot(density(datos$hematocr, na.rm =TRUE))
plot(density(dato_i_knn$hematocr))
#---------------------------------------
# Comparacion
#---------------------------------------
library(psych)
describe(datos[,c(16,24,29)])
describe(datos.c[,c(16,24,29)])
describe(datos_i[,c(16,24,29)])
describe(dato.i_r[,c(16,24,29)])
describe(dato.i_rp[,c(16,24,29)])
describe(dato.i_rpa[,c(16,24,29)])
describe(dato_i_knn[,c(16,24,29)])

## ANÁLISIS DESCRIPTIVO Y REGRESIÓN LOGÍSTICA
table(dato.i_rp$tipoicter)
prop.table(table(dato.i_rp$tipoicter))*100
```

```
# dato.i_rp$tipoicter<-relevel(dato.i_rp$tipoicter, ref =
"Patologico")

modelo1=glm(tipoicter~Edad+gestas+hijos+edages+rhmadre+itu+trataitu+
pesorn+pesoalta+gruporn,data = dato.i_rp,family = "binomial")

summary(modelo1)
```

DATOS-A



	Caso	Edad	gestas	hijos	edages	grupoma	rhmadre	diabtes	prelTU
1	1,00	31,00	2,00	2,00	39,00	O	+	No	Si
2	2,00	38,00	3,00	2,00	39,00	A	+	No	Si
3	3,00	22,00	3,00	2,00	38,00	O	+	No	Si
4	4,00	29,00	2,00	2,00	38,00	O	+	No	No
5	5,00	38,00	1,00	,00	38,00	O	+	No	Si
6	6,00	35,00	3,00	2,00	40,00	A	+	No	No
7	7,00	33,00	3,00	2,00	37,00	O	+	No	No
8	8,00	29,00	1,00	1,00	38,00	B	+	No	No
9	9,00	25,00	1,00	,00	34,00	O	+	No	Si
10	10,00	31,00	1,00	,00	39,00	O	+	No	Si
11	11,00	21,00	2,00	1,00	40,00	O	+	No	Si
12	12,00	41,00	2,00	1,00	39,00	O	+	No	No
13	13,00	27,00	1,00	,00	42,00	O	+	No	Si
14	14,00	31,00	3,00	1,00	38,00	O	+	No	No
15	15,00	31,00	6,00	2,00	37,00	O	+	No	No
16	16,00	34,00	1,00	,00	39,00	O	+	No	Si
17	17,00	25,00	1,00	,00	41,00	O	+	No	Si
18	18,00	30,00	2,00	1,00	40,00	O	+	No	Si
19	19,00	30,00	2,00	1,00	39,00	O	+	No	Si
20	20,00	25,00	2,00	,00	39,00	O	+	No	Si
21	21,00	32,00	2,00	1,00	38,00	A	+	No	Si
22	22,00	35,00	3,00	2,00	40,00	A	+	No	Si

DATOS-B

aciones Ventana Ayuda

	sanguine	ictericia	embarazo	edadicete r	perdiapes	cefalolohe ma	TIPOhematocrito	hematocr	seps	tiehosp
+	Si	No	A termino	12 Hrs	No	No	Normal	55,00	No	4,00
+	No	Si	A termino	48 Hrs	No	No	Poliglobulia	67,00	Si	5,00
+	No	No	A termino	24 Hrs	No	No	Normal	.	No	3,00
+	Si	Si	A termino	48 Hrs	Si	No	Normal	.	No	3,00
+	No	No	A termino	24 Hrs	No	No	Normal	.	Si	4,00
+	No	No	A termino	24 Hrs	No	No	Normal	.	No	2,00
+	No	No	Pre termino	12 Hrs	No	No	Normal	.	No	28,00
+	No	Si	A termino	24 Hrs	No	No	Normal	.	No	2,00
+	No	Si	A termino	24 Hrs	No	No	Normal	.	No	2,00
+	No	No	A termino	24 Hrs	No	No	Normal	.	Si	2,00
+	No	No	A termino	24 Hrs	No	No	Normal	.	No	3,00
+	No	Si	A termino	12 Hrs	No	No	Normal	49,00	No	3,00
+	No	Si	A termino	48 Hrs	No	No	Normal	45,00	No	10,00
+	No	No	A termino	24 Hrs	No	No	Normal	56,00	No	2,00
+	No	No	A termino	24 Hrs	No	No	Normal	45,00	No	3,00
+	No	No	A termino	24 Hrs	No	No	Normal	51,00	No	3,00
+	Si	Si	A termino	12 Hrs	No	No	Normal	47,00	No	8,00
+	Si	Si	A termino	12 Hrs	Si	No	Poliglobulia	62,00	No	4,00
+	No	No	.	24 Hrs	No	No	Normal	55,00	No	2,00
+	No	No	.	24 Hrs	No	No	Normal	45,00	No	3,00
+	No	No	.	48 Hrs	No	No	Normal	56,00	Si	9,00
+	No	Si	.	24 Hrs	No	No	Normal	52,00	No	7,00
+	No	No	.	24 Hrs	No	No	Normal	48,00	No	2,00
+	Si	Si	.	12 Hrs	No	No	Normal	48,00	No	5,00
+	No	No	A termino	48 Hrs	No	Si	Normal	40,00	Si	8,00
+	No	No	A termino	24 Hrs	No	No	Normal	48,00	No	2,00
+	No	No	A termino	24 Hrs	No	No	Normal	42,00	No	2,00
+	No	No	A termino	24 Hrs	No	No	Normal	54,00	No	6,00
+	No	Si	A termino	24 Hrs	No	No	Poliglobulia	61,00	No	3,00
+	Si	No	A termino	24 Hrs	No	No	Normal	53,00	No	5,00
+	No	Si	A termino	24 Hrs	No	No	Poliglobulia	63,00	No	2,00
+	Si	Si	Pre termino	48 Hrs	No	No	Normal	42,00	No	5,00
+	No	No	A termino	24 Hrs	No	No	Normal	47,00	No	4,00
+	No	Si	A termino	24 Hrs	No	No	Normal	56,00	No	3,00
+	Si	No	A termino	24 Hrs	No	No	Normal	54,00	No	8,00
+	No	No	A termino	24 Hrs	No	No	Poliglobulia	60,00	No	5,00