

UNIVERSIDAD NACIONAL DEL CAAGUAZÚ

FACULTAD DE CIENCIAS Y TECNOLOGÍA

**IDENTIFICACIÓN DE CARACTERÍSTICAS
EPIDEMIOLOGICAS Y CLIMÁTICAS DE DATOS
RELACIONADOS CON CASOS DENGUE RECOPIRADOS POR
ENTES ESTATALES DEL PARAGUAY ENTRE LOS AÑOS 2009 Y
2013.**

ELABORADO POR:

Marcos Antonio Ortega Velázquez.

TUTOR:

Ing. Héctor Ramiro Estigarribia.

Trabajo presentado a la Facultad de Ciencias y Tecnología de la Universidad Nacional de Caaguazú, como requisito para la obtención del título de Ingeniera en Sistemas Informáticos.

CORONEL OVIEDO – PARAGUAY

2018

PÁGINA DE APROBACIÓN

MESA EXAMINADORA DE SUSTENTACIÓN DE TESIS DE GRADO

Carrera de Ingeniería en Informática

Título de la Tesis: Características Epidemiológicas Y Climáticas Relacionadas
Con El Dengue en base a Recopilados Por Entes Estatales Del Paraguay
Entre los Años 2009 y 2013.

Calificación Obtenida:

Miembro

Miembro

Miembro

Miembro

Presidente

Acta N°: _____

Fecha: _____

Dedicatoria

A **Dios**, que todo me lo hizo posible,

A mis padres:

Corona Velázquez y Angel Ortega,

A mis hermanos:

Daisy, Gloria y Juan

A mis abuelos.

Agradecimientos

A Dios, por colmarme siempre de bendiciones, personas y oportunidades maravillosas.

A mis padres, Corona y Angel y a mis hermanos por apoyarme de manera incondicional, por los sacrificios que han hecho sin dudar en aras a permitirme salir adelante.

A mi tutor, el Ing. Héctor Estigarribia, al Lic. Cristhian Mendieta, al Ing. Fredy Ramirez, al Msc. Santiago Gómez, la Lic. Emilio Soto y a la Lic. Alma Silguero; por brindar su apoyo a este proyecto.

Al Centro de Investigación en Matemática (CIMA); por brindarme la oportunidad de trabajar de manera colaborativa con científicos de renombre a nivel nacional, así como de aprender al lado de otros estudiantes.

Al Programa de Becas de Grado de la Entidad Binacional Itaipú, por haber financiado gran parte de mis estudios de grado por medio de la beca que me fue otorgada en la convocatoria 2013.

A los docentes, funcionarios y directivos de la Facultad de Ciencias y Tecnologías (FCyT), por brindarme siempre su apoyo.

Autor: Marcos ortega

Tutor: Ing. Héctor Estigarribia.

Resumen

En el presente trabajo se realizó el análisis y la reducción de datos innecesarios de tres conjuntos con datos ambientales, sanitarios y epidemiológicos relacionados con casos de Dengue que fueron recopilados por la Dirección General de Vigilancia de la Salud y por la Dirección de Meteorología e Hidrología en el Paraguay durante el periodo de tiempo correspondido entre los años 2009 y 2013.

Fueron aplicadas técnicas selección y limpieza de datos, además de una versión multivariada de la incertidumbre simétrica como técnica para la selección de las características. Como resultado de dicho proceso, se generó un subconjunto mínimo de características, en el cual se mantiene la información realmente útil de los conjuntos originales.

Palabras clave: Dengue, Ciencia de Datos, Descubrimiento del Conocimiento, Selección de Características.

Abstract

In the present work were made an analysis and a data reduction process on three sets with environmental, health and epidemiological data related with Dengue cases that were captured by the General Direction of Health Vigilance and by Meteorology and Hydrology Direction in the interval of time between years 2009 and 2013.

Were applied techniques of data selection, data cleaning and a multivariate version of the symmetric uncertainty. As the result of that process, has generated a minimum subset of features, that maintains the useful information of the original sets.

In the present work an analysis and a data reduction process were made on three sets, with environmental, health and epidemiological data related to Dengue disease cases that were captured by the General Direction of Health Vigilance and by Meteorology and Hydrology Direction in the interval of time between years 2009 and 2013.

Techniques of data selection, data cleaning and a multivariate version of the symmetric uncertainty were applied. As the result of that process, a minimum subset of features has been generated, which maintains the useful information of the original sets.

Keywords: Dengue, Data Science, Knowledge Discovery, Feature Selection.

Indice general

Dedicatoria.....	I
Agradecimientos	II
Resumen.....	III
Abstract	IV
Indice general	V
Indice de figuras.....	X
Índice de fórmulas	XII
CAPITULO I.....	1
1 Marco Introductorio	2
1.1 Introducción	2
1.1.1 Relevancia del trabajo	4
1.1.2 Organización del trabajo	4
1.2 Planteamiento del problema.....	5
1.3 Delimitación y alcance	6
1.4 Pregunta de investigación.....	7
1.5 Objetivos	8
1.5.1 General	8
1.5.2 Específicos:	8
1.6 Justificación	9
CAPITULO II.....	10
2 Marco Teórico	11
2.1 Antecedentes	11

2.1.1 Estado actual de la información con la que cuentan las entidades encargadas de la prevención y lucha contra la enfermedad del dengue..	11
2.1.1.1 Dirección General de Vigilancia de la Salud (D.G.V.S.).....	11
2.1.1.2 Dirección de Meteorología e Hidrología. (D.M.H.)	12
2.1.2 El Dengue en el Paraguay	12
2.1.3 Análisis de variables ambientales, sanitarias y epidemiológicas relacionadas con el Dengue mediante el análisis de datos recopilados. ..	13
2.1.3.1 Proyectos similares desarrollados en el Paraguay	14
Conjuntos de datos establecidos en COMIDENCO	15
2.2 Bases teóricas.	16
2.2.1 El Dengue	16
2.2.1.1 Síntomas.....	16
2.2.1.2 Presencia a nivel internacional.....	17
2.2.2 Vigilancia y valoración de la salud.....	18
2.2.2.1 Salud Pública	18
2.2.2.2 Dirección General de Vigilancia de Salud	19
2.2.3 Dirección de Meteorología e Hidrología	21
2.2.4 Base de datos centralizada.....	22
2.2.5 Proceso de Descubrimiento del Conocimiento.....	24
2.2.5.1 Pasos del P.D.C. según [4]	25
2.2.5.2 Selección de características.....	26
2.2.5.3 ¿Por qué aplicar selección de características?	29
2.2.6 MSU.....	29
2.2.6.1 Cardinalidad univariada y multivariada	30

2.2.6.2 Entropía de Shannon.....	30
Entropía conjunta.....	31
Entropía condicional.....	31
Propiedades de la entropía.....	31
2.2.6.3 Ganancia de información (IG).....	32
Propiedades de la ganancia de información.....	32
2.2.6.4 S.U.....	33
2.2.6.5 Sesgo de MSU.....	35
2.3 Marco Conceptual.....	36
Capítulo III.....	38
3 Marco Metodológico.....	39
3.1 Método O Enfoque De La Investigación.....	39
3.2 Tipos De Estudio.....	39
3.3 Diseño De La Investigación.....	39
3.4 Universo y muestra.....	40
3.4.1 Universo.....	40
3.4.1.1 Las clases:.....	41
3.4.1.2 Las características:.....	41
3.4.2 Muestra.....	47
3.5 Instrumento de recolección de datos.....	57
3.6 Descripción De los Procedimientos De Análisis de Datos.....	57
3.6.1 Tratamiento inicial de los datos.....	57
3.6.1.1 Visualización y conteo de datos.....	57

3.6.1.2 Selección de características.....	58
3.6.1.3 Herramientas software utilizadas para el tratamiento de datos.	58
Python.....	58
3.6.1.4 Plataforma computacional	59
Capítulo IV	60
4 Marco Analítico	61
4.1 Presentación Del Estudio Y De Los Resultados.	61
4.1.1 Algoritmo de MSU en Python	61
4.1.2 Algoritmo propuesto para búsqueda de subconjunto optimo	65
4.1.3 Conjunto de características optimo	67
4.2 Recursos Y Costos.....	69
4.2.1 Costos de inversión.....	69
4.2.2 Recursos Humanos.	69
4.2.3 Otros Recursos.....	70
Capítulo V	71
5 Análisis e interpretación de los resultados	72
5.1 Conjuntos de datos obtenidos.....	72
5.2 Limpieza de datos.	72
5.3 Subconjunto optimo	73
6 Conclusiones.....	75
6.1 Recomendaciones.....	76
6.2 Contribuciones.....	76
6.3 Trabajos Futuros.....	76

7 Bibliografía.....	77
---------------------	----

Indice de figuras

Figura 1: Países en riesgo de dengue endémico y viral, extraída de [7].....	17
Figura 2: Pasos del P.D.C. según [4].....	25
Figura 3: Sesgo [19]	35
Figura 4: Distribución de clases finales por año y cantidad de casos.....	48
Figura 5: Distribución de las categorías pertenecientes a característica lugar_viajo, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final	49
Figura 6: Distribución de las categorías pertenecientes a característica clasificacion_final, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final	50
Figura 7: Distribución de las categorías pertenecientes a característica IgG, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.....	51
Figura 8: Distribución de las categorías pertenecientes a característica NS1_AG, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final	52
Figura 9: Distribución de las categorías pertenecientes a característica IgM en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.....	53
Figura 10: Distribución de las categorías pertenecientes a característica 'PRC' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.....	54
Figura 11: Distribución de las categorías pertenecientes a característica 'serotipo' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final	55
Figura 12: Distribución de las categorías pertenecientes a característica 'alguna_prueba_lab' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final	56
Figura 13: Distribución de las categorías pertenecientes a característica 'otros' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.....	56

Figura 14: Diagrama de flujo del algoritmo de MSU	62
---	----

Índice de fórmulas

Fórmula 1: Número de combinaciones posibles para N características.....	27
Fórmula 2: Entropía de Shannon	30
Fórmula 3: Entropía conjunta.....	31
Fórmula 4: Entropía condicional.....	31
Fórmula 5: Ganancia de información IG	32
Fórmula 6: Definición de SU	33
Fórmula 7: Definición de MSU	34
Fórmula 8: Coeficiente de correlatividad total.....	34
Fórmula 9: Suma de las Entropías Individuales.	34
Fórmula 10: Muestra extrema propuesta en [19].....	36

CAPITULO I

1 Marco Introductorio.

1.1 Introducción.

El Dengue es una enfermedad que constantemente marca presencia en el territorio Paraguayo. Cada año son detectadas diferentes cantidades de casos y algunos de ellos llegan a ser fatales. Han llegado a darse grandes epidemias que afectaron a una cantidad importante de habitantes en el país y en consecuencia el índice de mortandad por la enfermedad incrementó en estos periodos. En vista a estos hechos, en nuestro país se han establecido varias medidas para la prevención y lucha contra el Dengue.

Como parte de los esfuerzos contra el Dengue, la Dirección General de Vigilancia de la Salud¹ recopila semanalmente información relacionada con los casos de Dengue que se dan en nuestro territorio. Sin embargo existe un gran desaprovechamiento del potencial que se encuentra dentro de estos datos u otros que son constantemente capturados o generados por algunas instituciones del Estado, como es en el caso de la Dirección de Meteorología e Hidrología. La correcta utilización de dichos recursos podría constituirse como una importante herramienta para la lucha contra la enfermedad.

Con el auge actual de la Ciencia de Datos y otras ciencias afines, han sido desarrolladas varias técnicas y herramientas enfocadas netamente en el aprovechamiento de los datos capturados en gran cantidad. Las aplicaciones

¹ Dependencia del Ministerio de Salud Pública y Bienestar Social.

varían desde la inteligencia dentro de los negocios para la toma de decisiones² hasta el diagnóstico de enfermedades raras³.

Los datos relacionados con casos de Dengue, también son susceptibles a la aplicación de varios tipos de técnicas y procesos para el descubrimiento del conocimiento, el cuál, según [3] “es un proceso no trivial de identificación de patrones válidos, novedosos, parcialmente útiles; y, en última instancia, comprensibles a partir de los datos”.

En este trabajo se han aplicado tres tareas pertenecientes al proceso de descubrimiento de la información según [4] las cuales son la selección, la limpieza y la reducción de información. Ya que se ha realizado una selección de datos categóricos con relación a un fenómeno sobre los cuales se han aplicado técnicas para la visualización de los mismos, con el fin de apreciar las categorías y la cardinalidad de cada característica⁴, como parte de la limpieza han sido descartadas aquellas que poseían un gran número de registros con valores nulos; y, se ha aplicado una técnica de selección de características conocida como MSU para reducir la dimensionalidad de los datos, siendo este último proceso el objetivo principal de este proyecto. Estas tres tareas del proceso de descubrimiento del conocimiento según [4] deben ser aplicadas antes de las tareas de minería de datos, que es donde se empieza a extraer la información oculta dentro de los datos dependiendo de los objetivos del proceso de descubrimiento de la información mediante la interpretación.

2 Más conocido como: “La Inteligencia de Negocio (BI) es un término genérico que incluye las aplicaciones, la infraestructura y las herramientas, y las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y rendimiento.” [1]

3 Enfermedad rara es aquella cuya prevalencia es inferior a 5 casos por cada 10.000 personas en la Comunidad Europea [2].

4 De aquí en adelante se entenderá como característica a cada variable relacionada con un fenómeno y como categoría a cada valor posible de una característica.

1.1.1 Relevancia del trabajo

El resultado de este trabajo es un conjunto con las características más significativas, al cuál se llega mediante el procesamiento de tres conjuntos de datos que surgen de la unión de los registros proveídos por la Dirección General de Vigilancia de la Salud y por la Dirección de Meteorología e Hidrología del Paraguay. El subconjunto hallado es considerado por este trabajo como optimo para su implementación en técnicas de clasificación de casos de Dengue y esto a su vez puede servir como método de diagnóstico de la enfermedad basado en ciencia de datos.

1.1.2 Organización del trabajo

En el capítulo 2 se presentan el estado actual de los datos relacionados con el Dengue que son capturados y publicados por la Dirección General de Vigilancia de la Salud y los antecedentes de los análisis de estos datos anteriores a este proyecto; también los fundamentos teóricos de la enfermedad del Dengue, de las bases de datos centralizadas, el proceso de descubrimiento del conocimiento y de la Incertidumbre Simétrica Multivariada. En el capítulo 3 se desarrolla el marco metodológico donde se abordan los puntos relacionados con la metodología aplicada en este proyecto, así como los procesos de visualización y limpieza de datos llevados a cabo para la delimitación de las muestras que fueron utilizadas. En el capítulo 4 se trata el marco analítico donde se exponen los algoritmos utilizados para hallar un subconjunto óptimo de características extraído de las muestras establecidas en el capítulo tres. En el capítulo 5 se realiza el análisis y la interpretación de los resultados. Y para terminar en el capítulo 6 son presentadas las conclusiones del trabajo.

1.2 Planteamiento del problema.

Una de las ciencias más novedosas de los últimos años es la Ciencia de Datos, la cuál posee un sin fin de aplicaciones en diferentes campos gracias al proceso de descubrimiento de la información que la caracterizá. Este proceso permite llegar al aprovechamiento de datos capturados mediante la aplicación varias técnicas que han sido desarrolladas para seleccionar, limpiar, reducir, minar e interpretar información hasta llegar a un conocimiento dado, el cuál depende de los objetivos del proceso de descubrimiento del conocimiento [4].

El Estado Paraguay posee instituciones que constantemente almacenan o generan información relacionada con el Dengue; cuenta con un sistema de vigilancia a cargo de la Dirección General de Vigilancia Sanitaria, la cuál es una dependencia del Ministerio de Salud que semanalmente recolectar registros de casos de Dengue, los que poseen datos epidemiológicos y sanitarios de los casos que se dan alrededor del país; también posee una Dirección de Meteorología e Hidrología que día a día genera, captura⁵ y registra datos climáticos.

Actualmente existe un número bastante pequeño de investigaciones realizadas sobre estos datos, lo cuál denota que su potencial no está siendo realmente aprovechado. No existe una identificación clara de los datos más significativos relacionados con la enfermedad, los cuales deberían ser capturados de manera prioritaria cada vez que un caso es notificado, lo cual podría permitir un mejor aprovechamiento y un incremento de valor de estos datos para investigaciones futuras.

⁵ Mediante la utilización de sensores.

1.3 Delimitación y alcance.

El objetivo principal de este proyecto es identificar cuales son las características más significativas relacionadas con el Dengue, cuyos datos hayan sido capturados por Instituciones Estatales del Paraguay.

1.4 Pregunta de investigación.

En base a lo planteado anteriormente, se puede plantear la siguiente pregunta ¿Cuáles son las características más significativas relacionadas con el Dengue entre los datos que son capturados por la Dirección General de Vigilancia de la Salud y por la Dirección de Meteorología e Hidrología?

1.5 Objetivos

1.5.1 General:

Identificar las características más significativas relacionadas con el Dengue dado un conjunto de datos capturados por Dirección General de Vigilancia de la Salud y por la Dirección de Meteorología e Hidrología.

1.5.2 Específicos:

- Establecer conjuntos de datos relacionados con el Dengue con información proporcionada por Instituciones del Estado Paraguayo.
- Aplicar técnicas para la limpieza y el procesamiento de los datos obtenidos para evitar interferencias.
- Seleccionar las características más significativas del conjunto de datos obtenidos.

1.6 Justificación.

De acuerdo con [5] en cuanto al Dengue, el Paraguay es catalogado como un país altamente endémico. Los departamentos de Caaguazú, Concepción, Amambay, Canindeyú, Alto Paraná, Cordillera, Paraguarí y el Área metropolitana de Asunción son constantemente asediados por grandes números de casos y todo esto pesar de la constante labor del sector de la Salud Pública.

Una mayor comprensión de cuan determinantes son las influencias de las variables ambientales, sanitarias y epidemiológicas, y de las interacciones entre estas; podrían brindar un apoyo significativo en la lucha contra la enfermedad, permitiendo a los entes de Salud Pública del Estado Paraguayo analizar e identificar qué datos deben ser capturados de manera prioritaria para realizar estudios posteriores y qué aristas deben ser atacadas para obtener mejores resultados y de en el combate contra la enfermedad.

Para hallar un subconjunto con las características más significativas se deben llevar a cabo las primeras tres etapas de un proceso de descubrimiento del conocimiento, los cuales son la selección, la limpieza y la reducción de dimensionalidad de los datos a lo cuál se llega mediante la aplicación de técnicas de selección de datos.

CAPITULO II

2 Marco Teórico

2.1 Antecedentes

2.1.1 Estado actual de la información con la que cuentan las entidades encargadas de la prevención y lucha contra la enfermedad del dengue.

2.1.1.1 Dirección General de Vigilancia de la Salud (D.G.V.S.)

La D.G.V.S. en el marco de sus funciones emite boletines epidemiológicos semanales que incluyen información sobre Enfermedades de Notificación Obligatoria (E.N.O.) en forma de planillas con casos sospechosos o confirmados, detectados y notificados en todo el sistema de salud y especialmente en los niveles locales.

Estos boletines son alimentados por integrantes del Sistema de vigilancia en salud pública del país, que generan información de interés, los mismos e deben efectuar la notificación de de los eventos de reporte obligatorio definidos en la Resolución S.G. N°190/2013 del M.S.P.B.S, dentro de los términos de estructura de datos, clasificación, periodicidad y destino señalados en los mismos. Debiendo tener en cuenta un conjunto de determinadas características de calidad, veracidad y oportunidad de la información en los datos que son notificados a la D.G.V.S. [6].

Las sub-divisiones que se encargan de realizar las notificaciones de los sucesos que tienen lugar en las diferentes localidades distritales y departamentales se denominan Unidad Epidemiológica Regional (U.E.R.). Las U.E.R. deben efectuar el envío de la notificación a la Unidad de Gestión de Datos (UGD) de la Dirección General de Vigilancia de la Salud (D.G.V.S.) cada día miércoles hasta las 12:00 de la mañana al correo

notifica.D.G.V.S.@gmail.com o al número de fax (021 – 208.217). Una vez que los datos son remitidos a la D.G.V.S., estos son almacenados en un sistema que posee un motor de base de datos Postgres. Es decir, D.G.V.S. no cuenta con un sistema informático distribuido en el cuál los datos podan ser cargados desde las U.E.R. en cualquier parte del país en tiempo real y existe un gran retardo desde la generación de la información hasta su correspondiente publicación en los boletines ya que en lugar de realizar notificaciones diarias de los casos estas son semanales. El gran problema que surge aquí, es que los datos recopilados con este tipo de técnicas difícilmente tendrán todas las ventajas que posee una base de datos centralizada (véase sección 1.8.4.Base de datos centralizada) debido a que los datos no son almacenados desde el origen y esto supone grandes problemas de consistencia y pérdida de datos.

Es bastante difícil acceder a la información detallada que es capturada por la S.G.V.D. debido a problemas burocráticos, en consciencia para este proyecto solo se cuenta con registros de de casos de dengue ocurridos entre los ños 2009 y 2013 (véase la sección 3.4.Población y muestra).

2.1.1.2 Dirección de Meteorología e Hidrología. (D.M.H.)

Según [7] el clima es un factor bastante influyente en la dinámica del dengue. En la D.M.H. los registros de datos climáticos se encuentran bastante estructurados e incluso cumplen con estándares a nivel internacional. Poseen una gran apertura y coordinación con toda institución que necesite de sus servicios y es gracias a esto que es bastante fácil obtener registros históricos del clima.

2.1.2 El Dengue en el Paraguay.

Desde la aparición del Dengue en territorio paraguayo se ha dado un incremento consecutivo de políticas, estrategias y servicios de Salud Pública

que previenen y combaten los brotes de la enfermedad, agilizan la atención eficiente a los enfermos y buscan reducir el número de criaderos de mosquitos. Pese a todos estos esfuerzos, de acuerdo con [5] se registraron epidemias de Dengue en los siguientes periodos 1988-1989; 1999-2000; 2006-2007 y 2012-2013; en los cuales se dio un incremento número de mortalidad en pacientes en comparación a los años en los cuales no se han dado epidemias.

2.1.3 Análisis de variables ambientales, sanitarias y epidemiológicas relacionadas con el Dengue mediante el análisis de datos recopilados.

El Dengue es una enfermedad bastante particular. Su vector, no es nada fácil de eliminar debido a que es un mosquito que es capaz de incubar en un sin fin de ambientes húmedos y con temperaturas relativamente cálidas. Este tipo de escenarios abundan en nuestro país y existen muchas variables que que pueden colaborar o afectar el desarrollo de las larvas. Entre ellas podemos mencionar las temperaturas máximas y mínimas, la lluvia entre otros.

El mosquito adulto también posee varias condicionantes propias que afectan a la propagación de la enfermedad desde el vector. El virus se aloja en el intestino de los mosquitos para luego hacerse paso hasta las glándulas salivales en el periodo de 8 a 12 días⁶, recién pasado este periodo los mosquitos serán capaces de infectar a las personas [8].

La capacidad de vuelo del mosquito se ve fácilmente afectada por las inclemencias del tempo, ahora bien, lo que nos interesa identificar es en qué medida pueden afectar las bajas temperaturas, los vientos fuertes o las lluvias a la propagación de la enfermedad. Si bien la lluvia afecta el vuelo de los mosquitos en ambientes externos, las mismas deberían favorecer la creación de nuevos lugares para la incubación de más larvas.

⁶ Este es tiempo de incubación de la enfermedad en los mosquitos.

El viento y las lluvias normalmente se ven involucrados con las bajadas de temperatura, por culpa de las cuales los mosquitos adultos que se encuentran expuestos tienden a morir, al igual que las larvas que no pueden respirar debido al congelamiento de las superficies de agua de sus lugares de incubación.

El exterminio de los mosquitos por parte del ser humano no puede ser ignorado. Lo ideal es identificar cuales son las acciones más eficientes para la erradicación del vector, teniendo siempre en cuenta la combinación de estas contra-medidas con factores climáticos.

La rápida detección de la enfermedad en las personas y la ejecución de medidas de prevención y saneamiento en el entorno de la persona infectada también puede evitar la propagación desde ese punto.

Existen varios elementos que influyen en la propagación de la enfermedad, cada uno de los cuales posee un número de variables las cuales en muchos casos poseen una interacción entre sí. En estas situaciones la experiencia y la observación son las herramientas más útiles a la hora de determinar cuales son las variables realmente importantes, así como diferenciar todos aquellos datos que en realidad son superfluos o despreciables.

2.1.3.1 Proyectos similares desarrollados en el Paraguay.

Existen dos proyectos financiados por el CONACYT que han trabajado con datos relacionados con el dengue uno de ellos es el proyecto de Datos Abiertos y Alertas Epidemiológicas⁷ llevado a cabo por una empresa llamada

⁷ Este proyecto propone la creación de herramientas de gestión de la información de todas las variables y el estudio de co-variables relacionadas al dengue que permita la normalización de los datos relacionados y favorezca el análisis [9].

Centro de Desarrollo Sostenible (CDS) y el otro es el proyecto lleva el nombre de Construcción de un Modelo de Incidencia de Dengue Aplicado a Comunidades de Paraguay (COMIDENCO)⁸ ejecutado por el Centro de Investigación en Matemática (CIMA) en asociación con la Universidad Nacional de Caaguazú, a través de la Facultad de Ciencias y Tecnologías (FCyT UNCA), la Facultad Politécnica, Laboratorio de Computación Científica y Aplicada (LCCA), el Centro para el Desarrollo de la Investigación Científica (CEDIC) y la Universidad Nacional de Concepción (UNC). Este trabajo en particular fue desarrollado en el marco del proyecto COMIDENCO.

Conjuntos de datos establecidos en COMIDENCO.

Dentro del proyecto COMIDENCO, se han generado tres conjuntos de datos mediante la unión de los datos recopilados por la D.G.V.S. y de los registros del clima de la D.M.H. Estos conjuntos de datos solo cuentan con registros de casos de dengue correspondientes al intervalo [2009, 2013] debido a que los mismos fueron obtenidos de un proyecto anterior publicado por la empresa [11] dicha empresa a su vez obtuvo estos datos del D.G.V.S. (véase 2.2.2.2.D.G.V.S.). Para el proyecto COMIDENCO fue necesario recurrir a este recurso a causa de no haber recibido datos que fuesen lo suficientemente útiles al solicitarlos directamente al D.G.V.S.

Una vez que los datos fueron obtenidos procedieron con tratamiento de los mismos por medio de técnicas de Ciencias de Datos, iniciando con un la unión de los registros del clima y de casos de dengue por medio de las fechas.

⁸ Este proyecto propone construir y verificar un modelo matemático para predecir la tasa de incidencia de dengue, en función a variables ambientales, sanitarias y epidemiológicas y en presencia de acciones de intervención anti-dengue [10].

El trabajo siguiente fue realizar la categorización⁹ de los datos numéricos por medio del software Knime que permitió realizar una codificación de los mismos,

2.2 Bases teóricas.

2.2.1 El Dengue.

El Dengue es una enfermedad clasificada como infección vírica, la cual es transmitida mediante la picadura de las hembras infectadas de mosquitos del género Aedes. La enfermedad posee cuatro serotipos de virus conocidos como (DEN 1, DEN 2, DEN 3 y DEN 4). Los casos de Dengue se dan en las regiones con climas tropicales y subtropicales, especialmente en las áreas urbanas y semiurbanas.

Una vez que una persona ha sido infectada con Dengue, los síntomas empezarán a aparecer pasados los 3 o 14 días (promedio de 4 a 7 días) después de la picadura del mosquito.

2.2.1.1 Síntomas.

Los síntomas son una fiebre elevada (40C°) acompañada de dos de los síntomas siguientes: dolor de cabeza muy intenso, dolor detrás de los globos oculares, dolores musculares y articulares, náuseas, vómitos, agrandamiento de ganglios linfáticos o sarpullido.

El dengue grave es una complicación potencialmente mortal porque cursa con extravasación de plasma, acumulación de líquidos, dificultad respiratoria, hemorragias graves o falla orgánica. No hay tratamiento específico

⁹ Proceso de codificación de grupos de datos numéricos, este proceso es esencial para controlar el sesgo de MSU, ya que al realizar este proceso se disminuye la cardinalidad univariada de las características que originalmente posee valores numéricos.

del dengue ni del dengue grave, pero la detección oportuna y el acceso a la asistencia médica adecuada disminuyen las tasas de mortalidad por debajo del 1% [12].

2.2.1.2 Presencia a nivel internacional.

En los últimos 50 años, su incidencia ha aumentado de manera considerable con su creciente expansión geográfica hacia nuevos países y, en la actual década, desde áreas urbanas a rurales [5]. Cada año se dan aproximadamente 80 millones de casos en más de 100 países, y cerca de 2,5 mil millones de personas viven en países con dengue endémico y el Paraguay forma parte de esta lista de países [12] En la siguiente figura se presenta a los países con riesgo de Dengue endémico y epidémico.

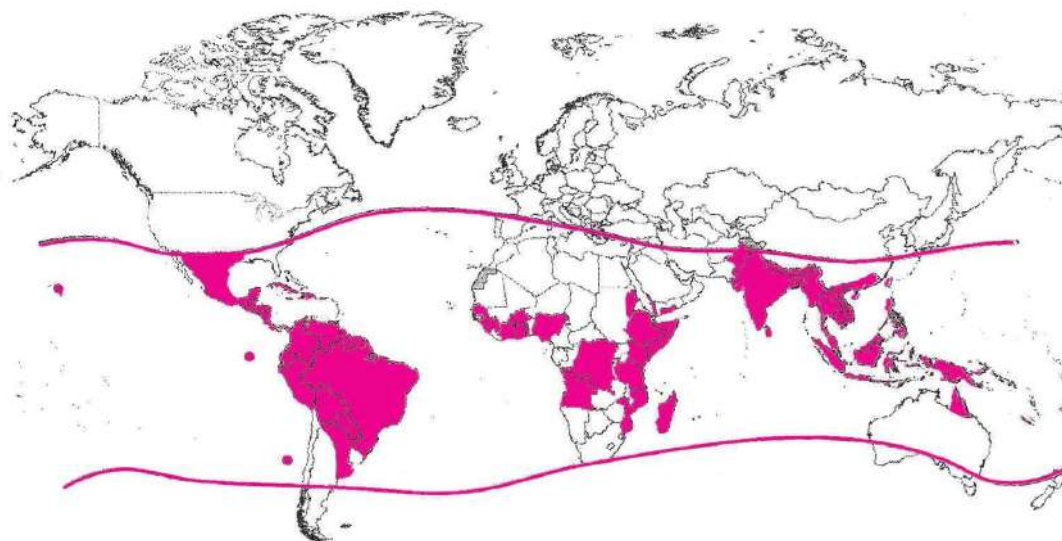


Figura 1: Países en riesgo de dengue endémico y viral, extraída de [7]

2.2.2 Vigilancia y valoración de la salud.

2.2.2.1 Salud Pública.

Las Instituciones de Salud Pública de un Estado tienen como una de sus funciones la realización de la vigilancia y el control constantes del estado de la salud de los pobladores su país. Esta tarea es hecha con la finalidad de identificar problemas de salud en sus comunidades y actuar en consecuencia de manera rápida y eficaz para solucionar dichos problemas. Esto posibilita la detección de nuevos problemas a medida que vayan apareciendo, y el control del número de casos y frecuencia de las enfermedades. Además, la información capturada debe ser usada para determinar si sus estrategias e intervenciones de carácter sanitario¹⁰ y no sanitario¹¹ están rindiendo sus frutos esperados [13].

El estado Paraguayo en particular, cuenta con el Ministerio de Salud Pública y Bienestar Social (M.S.P.B.S.) el cuál según consta en [14] tiene como misión el garantizar el cumplimiento de las funciones de rectoría, conducción, financiamiento y provisión de servicios de salud con el fin de alcanzar la cobertura universal, bajo el enfoque de protección social, en el marco del Sistema Nacional de Salud.

Toda organización que lleva a cabo el control de la salud, debe contar con un conjunto de políticas, las cuales dictan de manera específica los procedimientos a realizar en cada tipo de caso o en cada función determinada.

10 Entiéndase por actividades de carácter sanitario a todas las actividades relacionadas con limpieza, vacunación, chequeo de rutina y otras estrategias utilizadas para combatir enfermedades desde las Instituciones de Salud Pública.

11 Entiéndase a las actividades de carácter no sanitario a aquellas que no tienen incidencia directa sobre la enfermedad mediante intervenciones sanitarias. Ej: Educación de la población y de los profesionales, investigaciones realizadas sobre las enfermedades, etc.

Cuando se detecta algún problema sanitario en el Paraguay, es deber del M.S.P.B.S el establecer las mejores intervenciones y estrategias que puede aplicar para solventar el problema detectado. Debe crear las dependencias necesarias para la gestión puntual de las tareas e identificar cuales son los agentes sanitarios y/o sociales que pueden aplicarlas de la mejor manera.

Para poder ejecutar el control y la respuesta a contingencias de manera eficiente, siempre será necesario contar con información relevante y suficiente, eso suceden en todo tipo de organizaciones y más aún en un estructuras tan grandes como las de los estados.

2.2.2.2 Dirección General de Vigilancia de Salud.

La D.G.V.S es el ente rector del Sistema Nacional de Vigilancia de la Salud. Es una institución estratégica dependiente del Ministerio de Salud Pública y Bienestar Social y tiene como misión el desarrollar los procesos de vigilancia de la salud, con el fin de preservar la salud y reducir los riesgos de enfermar asociados a los determinantes de salud [15] y más cuenta con las siguientes funciones y atribuciones explicitas:

1. Diseñar e implementar políticas de vigilancia de salud a nivel nacional.
2. Desarrollar, coordinar y supervisar la red de Vigilancia de la Salud a nivel nacional.
3. Planificar, organizar, dirigir y supervisar la ejecución de todas las actividades, procesos y proyectos que hacen a la Vigilancia de Salud a nivel nacional.

4. Planificar estratégica y operativamente la Vigilancia de Salud a nivel nacional.
5. Realizar el seguimiento y el control del cumplimiento del Plan Estratégico, el Plan Operativo Anual y el Presupuesto anual de la Dirección; y en su caso aplicar las correcciones y acciones de mejora.
6. Desarrollar, coordinar y supervisar los procesos de alerta epidemiológica y respuesta.
7. Desarrollar, coordinar y supervisar investigación epidemiológica.
8. Coordinar y supervisar el correcto funcionamiento y articulación de sus dependencias.
9. Asesorar en el campo de la vigilancia a otras entidades que lo requieran.
10. Coordinar y gestionar acciones con otras dependencias del Ministerio, así con otras instituciones nacionales e internacionales para la aplicación de medidas de prevención y control.
11. Diseñar y proponer disposiciones legales relacionadas al ámbito de la vigilancia de la salud.
12. Coordinar la implementación de las disposiciones, medidas, recomendaciones y prácticas sugeridas por el Reglamento Sanitario Internacional.

13. Informar de manera oportuna a la población general sobre riesgos para la salud, así como medidas de prevención y control necesarias para minimizar dichos riesgos.

14. Evaluar y cuantificar los recursos disponibles y necesarios para llevar adelante su misión y funciones.

2.2.3 Dirección de Meteorología e Hidrología.

En nuestro país, la Dirección de Meteorología e Hidrología (D.M.H.) es la institución encargada de promover el estudio y desarrollo de la meteorología e hidrología en todo el territorio Paraguayo [8]. Constantemente captura y registra información relacionada con los fenómenos climáticos que se dan sobre nuestro territorio y mantienen un control bastante detallado de las fluctuaciones que llegan a darse con relación a los factores típicos del comportamiento de nuestro clima.

Los servicios específicos ofrecidos por la D.M.H. son:

1. Boletines de Pronósticos Meteorológicos.
2. Boletines de Alertas Meteorológicas.
3. Boletín hidro-meteorológico diario.
4. Boletín Climático Mensual.
5. Boletín de Tendencias !!br0ken!!
6. Provisión de datos Meteorológicos e Hidrológicos al Público en geneal.

2.2.4 Base de datos centralizada.

Los sistemas de tradicionales para procesamiento de archivos físicos¹², la utilización de datos en programas de aplicación y su almacenamiento en archivos separados, son técnicas para el manejo de datos utilizadas en organizaciones grandes.

Por lo general, un departamento que necesita un programa de aplicación¹³ trabajaba con otro departamento que se encarga procesamiento de datos para establecer especificaciones tanto para los programas como para los datos necesarios para él. Normalmente los mismos datos que son capturados se almacenan de manera independiente en varios departamentos dentro de una organización y compartirlos se convierte en una tarea bastante difícil. Cada departamento tiene sus propios archivos de datos que son creados de manera específica para una aplicación, y éstos con estos métodos de trabajo las organizaciones solo logran quedar atrapadas en medio de una maraña de procesos que pueden llegar a ser innecesarios e ineficientes. Para este tipo de escenarios se creó el enfoque de base de datos integrada [16].

Es innegable que el sistema de manejo de la información que posee la D.G.V.S. es bastante coincidente con la descripción anterior. Haciendo una analogía básica podemos decir que las UER son los departamentos que se encargan de la captura de datos y que trabaja con UGD de la D.G.V.S.. En cuanto al DMH es igual ya que los datos que utiliza para proveer sus servicios son recopilados en diferentes puntos geográficos.

El enfoque de base de datos según [16] integrada tiene las siguiente ventajas:

12 Documentos impresos o escritos sobre papel.

13 Entiéndase como programa de aplicación al software que diseñado y utilizado para un conjunto determinado de tareas en un dominio delimitado.

1. Comparación de datos.

La base de datos pertenece a toda la organización. El ABD gestiona los datos, pero éstos no pertenecen a algún individuo o departamento. Por ende, la organización tiene el control sobre los datos que necesita para dirigir su negocio. Muchos usuarios pueden tener autorización para acceder al mismo trozo de información. La autorización para acceder a los datos la otorga el ABD, no otro departamento.

2. Control de redundancia.

Cuando se almacena en una base de datos, la información se integra de modo que múltiples copias de los mismos datos no se almacenan a menos que sea necesario.

3. Consistencia de datos.

Un efecto de eliminar o controlar la redundancia es que los datos son consistentes. Si un ítem de datos aparece solo una vez, cualquier actualización a su valor necesita realizarse solo una vez, y todos los usuarios tendrán acceso al mismo nuevo valor.

4. Estándares de datos mejorados.

El administrador de base de datos (ABD), que es responsable del diseño y mantenimiento de la base de datos para satisfacer las necesidades de todos los usuarios, define y refuerza los estándares de toda la organización para la representación de datos en la base de datos.

5. Mejor seguridad.

Los datos en la base de datos de una organización son un valioso recurso corporativo que se debe proteger de mal uso intencional o accidental. La seguridad de datos es la protección de la base de datos de acceso no autorizado por personas o programas que puedan hacer mal uso o dañar los datos. (pág. 10)

Una buen sistema de manejo de datos basado en el enfoque centralizado agiliza muchas tareas en las organizaciones y permiten producir nuevas informaciones ya que permite a los datos interactuar entre sí.

2.2.5 Proceso de Descubrimiento del Conocimiento.

El Proceso de Descubrimiento del Conocimiento (P.D.C.) hace referencia a la aplicación de un conjunto de pasos sobre datos capturados con anterioridad, con la finalidad de de identificar a aquellos que puedan ser considerados información valiosa, novedosa, potencialmente útil y que posea patrones interpretables. Estos procesos son aplicados especialmente a aquellos datos que son almacenados en bases de datos.

En una gran cantidad de campos de aplicación, se cuenta con información bastante variada y extensa, lo que con normalidad hace que el procesamiento manual de la información sea extremadamente difícil de aplicar, por esto y más es que las alternativas computacionales siempre son las predilectas debido a al potencial de la automatización para llegar a una mayor eficiencia.

2.2.5.1 Pasos del P.D.C. según [4].

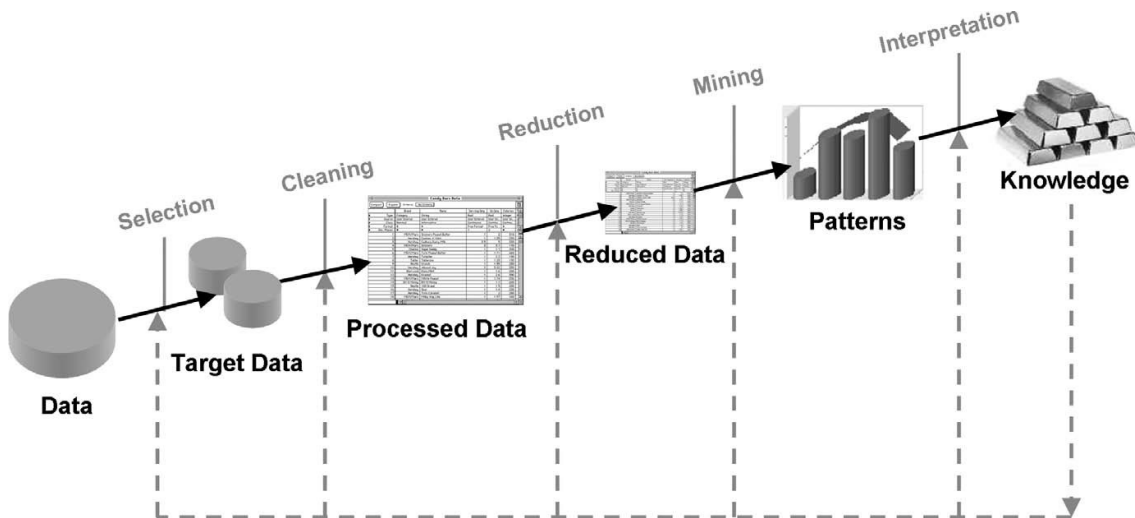


Figura 2: Pasos del P.D.C. según [4].

1. Selección de datos: Se selecciona o crea un conjunto de datos relacionándolos con un objetivo determinado. En caso que existan varios conjuntos de datos, estos pueden unirse para crear el conjunto más apropiado para el objetivo.
2. Limpieza de datos/procesamiento: Esta es una etapa que abarca una basta cantidad de sub-tareas como remoción o reducción de interferencia o datos basura, imputación de variables y discretización de atributos.
3. Reducción de datos: La mayoría de los conjuntos de datos poseen una determinada cantidad de datos redundantes los cuales no poseen valor para el P.D.C. El objetivo principal de este paso es encontrar características útiles para representar a los datos y remover las características irrelevantes. Como resultado de este proceso el tiempo utilizado a la hora de realizar la minería de los datos también se reduce.

4. Minería de datos: Un método de minería de datos (el proceso de extracción de información predictiva oculta alojada dentro de grandes bases de datos) es seleccionado dependiendo de cuales sean los objetivos de la tarea del P.D.C. La elección de algoritmos utilizados depende de muchos factores, incluyendo el entorno al que pertenece el conjunto de datos y los valores que contiene.
5. Interpretación y evaluación: Una vez que un conocimiento ha sido descubierto, es evaluado con respecto a su validez, utilidad, grado de novedad e implicidad. Este paso puede requerir repetir algunos de los pasos.

Las técnicas que son aplicadas dentro de estos pasos se dividen en dos grandes grupos: las técnicas que transforman el significado subyacente de las características de los datos y aquellos que conservan la semántica.

2.2.5.2 Selección de características.

En el ámbito del P.D.C. la selección de características se sitúa como una parte de la tarea de reducción de datos; y es una tarea bastante importante debido a que una vez iniciado el un P.D.C. es normal encontrarse con un gran número de características y con un número aún más grande de combinaciones de estas características.

En ciertas ocasiones una característica por sí sola puede no ser muy informativa o significativa, sin embargo, al combinarla con otras característica o características esta realidad puede cambiar; y, es por esto que deben evaluarse las combinaciones posibles entre las características. El número de subconjuntos de combinaciones posibles con m características pertenecientes a un total de N características está dado por:

$$x = \frac{N!}{m!(N-m)!} \quad (1)$$

Fórmula 1: Número de combinaciones posibles para N características.

El valor resultante puede llegar a ser bastante grande. Muchas veces esto hace parecer que al agregar características de manera consecutiva tarde o temprano uno estará agregando información significativa, lo cuál ayudaría a la identificación de clases; sin embargo esto no es óptimo debido a que se estaría haciendo más daño que bien al agregar una excesiva cantidad de información que también puede ser interferencia, llegando a causar más problemas y confusión. Si bien los conjuntos de datos grandes muchas veces permiten que los algoritmos de Minería de Datos puedan encontrar patrones, es esencial tener en cuenta que esta no es una regla que se aplica a todos los casos. La mayoría de las técnicas trabajan con algún grado de reducción de la información buscando hacer frente a los grandes cúmulos de datos, por lo cual un método eficiente de reducción siempre es necesario.

Función específica:

La función específica de la selección de características se centra en la identificación de los subconjuntos más informativos o significativos pertenecientes a las características de los conjuntos de datos originales y esto es posible gracias a la aplicación de filtros de interferencia. El resultado de este proceso de selección es un conjunto de datos más eficiente con relación a una clase.

Relevancia:

Según [17] la palabra *relevante* hace referencia a aquello que se importante o significativo. [18] mencionan que existen tres tipos diferentes de relevancias en cuanto a categorías de datos: Relevancia fuerte, relevancia débil y irrelevancia. Estas clasificaciones nos ayudan a decidir cuales son las características que deben conservarse y cuales son las que deben eliminarse. Las categorías con relevancias fuertes son en teoría importantes para mantener la estructura en un dominio y deben ser conservadas, las categorías con relevancias débiles pueden o no ser importantes o no, dependiendo de las otras características escogidas mientras que las relevantes deben ser eliminadas.

Redundancia:

Según [17] la redundancia significa sobra o demasiada abundancia de cualquier cosa o en cualquier línea. En el área de la selección de características [18] la redundancia es frecuentemente relacionada con la correlación existente entre las características y se acepta ampliamente que dos características son redundantes cuando sus valores son completamente correlacionales.

Existen dos tipos de medidas extensamente utilizadas para estimar la correlación entre dos variables: lineal y no lineal. En el caso de las relaciones lineales posemos encontrar el coeficiente de correlación de Person y en el caso de las no lineales, existen muchas medidas desarrolladas en base la los conceptos de la entropía. En este trabajo se hablará específicamente de la incertidumbre simétrica o (Symmetrical Uncertainty, S.U.).

2.2.5.3 ¿Por qué aplicar selección de características?

La aplicación de la selección de características es bastante importante como parte del proceso de depuración de la información. En todo trabajo de captura de datos es bastante fácil incluir información no tan importante con relación a un fenómeno ya que en este proceso las instituciones, empresas o personas tienden a capturar una gran cantidad de datos desconociendo cuales son aquellos que realmente poseen valor informativo; lo cuál, es algo normal ya que difícilmente uno podrá reconocer cuales son los datos mas informativos sin realizar un análisis de los datos previo.

Hablando en términos de tecnologías utilizadas en Machine Learning para realizar pronósticos o tareas de clasificación, encontramos como realmente útiles a los trabajos de S.C. ya que la aplicación de las mismas ayudan a evitar ajustes innecesarios en modelos como los de las redes neuronales, lo cual a su vez posibilita disminuir las cargas computacionales generadas por este tipo de técnicas y ayuda a que el modelo final pueda seguir siendo útil ante casos nuevos.

2.2.6 MSU.

Para hablar de MSU (Multivariate Symmetric Uncertainty o Incertidumbre Simétrica Multivariada) primeramente hay que mencionar los conceptos de cardinalidad univariada y multivariada, de Entropía de Shannon y de SU (Symmetric Uncertainty o Incertidumbre Simétrica) como se indica en [19]; ya que MSU es una generalización para varias variables basada en S.U. la cuál a su vez está basada en principios de la entropía de Shannon.

2.2.6.1 Cardinalidad univariada y multivariada.

La palabra cardinalidad posee varias aplicaciones posibles, para este proyecto se utilizará para especificar a la cantidad de valores posibles que puede tener un conjunto determinado.

Dado un conjunto de datos U de datos discretos o categóricos estructurados en filas F y columnas C , donde el nombre de cada columna representa a característica C_i , se entienda como cardinalidad univariada a la cantidad de valores posibles que pertenecen al dominio de cada C_i y a la cardinalidad multivariada como a la cantidad de combinaciones posibles de varias C_i incluyendo la clase a lo largo de las F_n contenidas dentro de U .

2.2.6.2 Entropía de Shannon.

La Entropía de Shannon¹⁴ o H , conocida por muchos como la Medida del Caos, es una medida de incertidumbre en la predicción de una variable discreta X , que posee $\{x_1, \dots, x_n\}$ valores discretos. La misma está definida de la siguiente forma:

$$H(X) = \sum_i P(X_i) \cdot \log_2[P(X_i)] \quad (2)$$

Fórmula 2: Entropía de Shannon.

Donde $H(X)$ es la entropía de una variable X y $P(x_n)$ es la probabilidad de un valor $x_n \in X$.

¹⁴ En honor a Clause Elwood Shannon.

Entropía conjunta:

La entropía conjunta $H(X,Y)$ se define para dos variables (X,Y) independientes que poseen $P(X,Y)$ como su distribución conjunta de probabilidad.

$$H(X,Y) := - \sum_{x \in X} \sum_{y \in Y} P(x,y) \cdot \log_2 [P(x,y)] \quad (3)$$

Fórmula 3: Entropía conjunta.

Entropía condicional:

Dado el valor de una variable discreta Y , es posible cuantificar la cantidad de información necesaria para describir el resultado de una variable X mediante el cálculo de la entropía condicional $H(X|Y)$.

$$H(X|Y) := - \sum_j \{P(y_j) \cdot \sum_i P(X|Y) \cdot \log_2 [P(X|Y)]\} \quad (4)$$

Fórmula 4: Entropía condicional.

Donde $P(x_i|y_i)$ es la probabilidad a posteriori para un valor x_i de la variable X y $P(y_i)$ es la probabilidad a priori de un valor y_i de Y .

Propiedades de la entropía:

1. $H(X) \geq 0$.
2. $H_b(X) = (\log_b a) H_a(X)$.
3. $H(X|Y) \leq H(X)$ y es igual si y solamente si X e Y son independientes.

4. $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$ con igualdad si y solamente si las variables aleatorias de X_i son independientes.
5. $H(X) \leq \log(X)$, y la igualdad se cumple si y solamente si X está uniformemente distribuida y Y .

Teorema (regla de la cadena) Dadas dos variables aleatorias X e Y la entropía conjunta está dada por $H(X, Y) = H(X) + H(X | Y)$.

2.2.6.3 Ganancia de información (IG).

La IG o $IG(X|Y)$ de una variable X dada una variable Y mide la disminución de la incertidumbre acerca del valor de la variable X cuando el valor Y es desconocido, la misma se define de la siguiente forma:

$$IG(X|Y) = H(X) - H(X|Y)$$

Fórmula 5: Ganancia de información IG. (5)

La **IG** permite medir la cantidad de conocimiento de una variable Y que ayuda a que el valor de X sea más fácil de predecir, es por ello que puede ser utilizado como medida de correlación.

Propiedades de la ganancia de información:

1. $I(X; Y) \geq 0$
2. $I(X; Y) = H(X) - H(X|Y)$.
3. $I(X; Y) = H(Y) - H(Y|X)$.

4. $I(X; Y) = H(X) + H(Y) - H(X, Y)$.
5. $I(X; Y) = I(Y; X)$ (*simetría*).
6. $I(X; X) = H(X)$ (*información propia*).

2.2.6.4 S.U.

S.U. es la incertidumbre simétrica de dos variables aleatorias, está formada por la normalización de las entropías de X e Y y se define como sigue:

$$SU(X, Y) = 2 \left[\frac{IG(X|Y)}{H(X) + H(Y)} \right] \quad (6)$$

Fórmula 6: Definición de SU.

Si X e Y son independientes tenemos que $SU(X|Y) = 0$; y si X e Y están completamente correlacionados tenemos que $IG(X|Y) = H(X) = H(Y)$ por lo tanto $SU(X, Y) = 1$; otra observación que podemos hacer en este punto es que los valores de SU se encuentran en el intervalo $(0 \leq SU \leq 1) = SU[0,1]$. Por estas características es que SU es utilizado como medida de correlación y es bastante utilizado como tal para realizar tareas de selección de características ya que permite determinar el nivel de correlación que posee una variable con una clase determinada. También está fuertemente aceptado que SU es una medida de redundancia ya que se considera que dos variables X e Y que poseen una correlación total son mutuamente redundantes.

El SU posee un gran problema y este radica en que solo funciona de manera eficaz sobre dos variables, por lo cual no es práctico a la hora de determinar la redundancia dentro de un sistema con varias variables, ni la significatividad de los mismos [18].

En respuesta a lo anterior tenemos a **MSU**, que es una semimétrica de correlación para varias variables (de ahí su importancia para este trabajo) basada en **SU**, de hecho es una generalización formal del mismo y se define como

$$MSU(X_{1:n}) := \frac{n}{n-1} \cdot \left[\frac{\sum_{i=1}^n H(X_i)}{\sum_{i=1}^n H(X_{1:n})} \right] \quad (7)$$

Fórmula 7: Definición de MSU.

donde

$$C(X_{1:n}) = \sum_{i=1}^n H(X_i) - H(X_{1:n}) \quad (8)$$

Fórmula 8: Coeficiente de correlatividad total.

es el coeficiente de correlatividad total, y

$$H(X_{1:N}) := H(X_1, \dots, X_n) := - \sum_{x_1} \dots \sum_{x_n} P(x_1, \dots, x_n) \log_2 [P(x_1, \dots, x_n)] \quad (9)$$

Fórmula 9: Suma de las Entropías Individuales.

Teniendo en cuenta esto podemos afirmar cuanto sigue:

$$MSU(X_{1:n}) = \frac{n}{n-1} \left[\frac{\sum_{i=1}^n H(X_i)}{\sum_{i=1}^n H(X_i)} - \frac{\sum_{i=1}^n H(X_{1-n})}{\sum_{i=1}^n H(X_i)} \right] = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n H(X_{1-n})}{\sum_{i=1}^n H(X_i)} \right] \quad (10)$$

Esta última afirmación es la que se aplicará en este proyecto y de aquí en más la sumatoria de $i=1$ a n de $H(X1-n)$ se le llamará Suma de Entropías Conjuntas y al denominador de la sumatoria $i=1$ a n de $H(Xi)$ se le conocerá como Suma de las Entropías Individuales.

Los valores posibles de MSU al igual que los del SU se encuentran en el rango $[0,1]$ y ambos pueden ser aplicados sobre datos numéricos discretos y sobre variables aleatorias categóricas.

2.2.6.5 Sesgo de MSU.

En [19] se realiza un análisis experimental del MSU el cuál define de la siguiente manera:

Dado el conjunto $\{A1, A2, \dots, An, C\}$ conformado por n atributos discretos o categóricos junto con la clase C , un tamaño muestral m , una medida de evaluación $M(am1, am2, \dots, amn, cm)$ y una constante θ que indica el valor verdadero del nivel informativo en conjunción de los atributos acerca de la clase, el sesgo de evaluación de atributos está dado por:

$$Sesgo[M] = E_{\theta}(M) - \theta \quad (11)$$

Figura 3: Sesgo [19].

Donde $E_{\theta}(M)$ denota valor esperado de la medida de evaluación M dado el valor verdadero θ que se estima.

En [19] se explica que el sesgo en sí es el error sistemático que tiene lugar en las mediciones de correlaciones entre variables y se determina que los factores implicados en el sesgo son:

- *Cardinalidad Univariada debido a que el MSU es una medida basada en la teoría de la información.*
- *Cardinalidad Multivariada: debido a que el MSU es una medida capaz de evaluar la interacción en conjunto de las variables.*
- *Tamaño Muestral: debido a que el MSU es un estadístico muestral con la propiedad de consistencia, es decir, el mismo converge a su valor verdadero cuando el número de datos de la muestra tiende a infinito o al tamaño total de la población.*

y como conclusión del trabajo desarrollado en [19] se propone que el tamaño muestral estadístico extremo que esté dado en función a un múltiplo de varias veces la cardinal multivariada, de tal forma a facilitar una alta probabilidad de total repetitividad en la muestra, lo cual está dado por la siguiente relación.

$$\text{Tamaño muestral} \approx 10 |clase| \prod_{i=1}^n f_i \quad (12)$$

Fórmula 10: Muestra extrema propuesta en [19]

2.3 Marco Conceptual.

- a) Análisis de datos relacionados con el Dengue capturados en el Paraguay.

- b) Ciencia de datos, aplicación de fases iniciales del proceso de descubrimiento de la información.

Capítulo III

3 Marco Metodológico.

3.1 Método O Enfoque De La Investigación.

El tipo de investigación adoptado para este proyecto es el cualitativo, debido a que se busca determinar el nivel de correlación que existente entre varias variables relacionadas al fenómeno., en este caso la enfermedad del Dengue mediante la aplicación de una medida de correlación basada en la entropía.

3.2 Tipos De Estudio.

El tipo de estudio aplicado en este proyecto es el exploratorio. Los estudios exploratorios se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes [20]. En este caso, el análisis de datos relacionados al Dengue capturados por Instituciones del Estado Paraguayo desde el punto de vista de Ciencia de Datos, es un área muy poco estudiada en la actualidad.

3.3 Diseño De La Investigación.

El diseño de esta investigación es de tipo transeccional correlacional. Estos diseños describen relaciones entre dos o más categorías, conceptos o variables en un momento determinado. A veces, únicamente en términos correlacionales, otras en función de la relación causa efecto [20]. Para el presente proyecto fueron utilizados registros de casos de Dengue capturados por Instituciones Nacionales y se realizó una búsqueda del subconjunto de características más significativo con relación a la variable dependiente conocida como estado estado final.

3.4 Universo y muestra.

3.4.1 Universo.

El universo de este proyecto se establece como base en los tres conjuntos de datos generados en el proyecto COMIDENCO, los mismos poseen variables con valores numéricos que fueron procesadas para convertir sus valores a datos categóricos, este paso es necesario para la aplicación de técnicas basadas en la entropía como MSU. El MSU es escogido para este trabajo debido a que es capaz de trabajar con varias características.

A cada registro de caso de dengue le fueron agregados los datos del clima de 3 días, 5 días y 7 días antes de que se registre el caso, dependiendo del lugar en el cuál este haya sido registrado, lo cuál resulto en un total de 3 conjuntos de datos con un total de 196922, 196920 y 199428 registros de casos de Dengue respectivos a cada conjunto. Las diferencias existentes entre los números de registros de los diferentes conjuntos de datos se debe a que en algunos casos no existían registros climáticos para algunos casos, siendo los registros de clima de 7 días antes los más completos con relación a los casos de Dengue.

Debe tenerse en cuenta que para este trabajo los valores nulos (SD) fueron estimados como una categoría más en todas características, es decir no se realizaron tareas de imputación en este proyecto, a continuación se presentan la las clases y cada una de las características¹⁵ con sus correspondientes dominios:

¹⁵ Las características y la clase son citadas con el nombre original que llevan dentro del conjunto de datos.

3.4.1.1 Las clases:

Las mismas están agrupadas bajo la denominación de 'clasificacion_clinica' y son las siguientes:

- CONFIRMADO: Indica que el caso dio positivo al dengue.
- DESCARTADO: Indica un registro no era sospechoso pero que ya no es un caso de dengue.
- EN ESTUDIO: Indica que el caso está en proceso de estudio de laboratorio, una vez que se cuenta con el resultado .
- SOSPECHOSO: A esta clase pertenecen aquellos casos cuyos estudios no fueron lo suficientemente esclarecedores como para confirmar el caso.

3.4.1.2 Las características:

- 1) 'dia': Número del día en que se dio el caso.
 - Categorías: [1,31].
 - Cardinalidad univariada: 31.
- 2) 'mes': Número del mes dentro del año en el que se dio el caso
 - Categorías: [1,12].
 - Cardinalidad univariada: 12.
- 3) 'semana': Número de la semana dentro del año en la que se dio el caso.
 - Categorías: [1,53].
 - Cardinalidad univariada: 53

- 4) 'institucion_notificacion': Hace referencia al nombre de la institución desde la cuál se notificó un caso determinado. Existe un total de 3689 instituciones que realizaron notificaciones, es decir que posee una cardinalidad bastante alta.
 - Categorías: demasiadas para ser citadas.
 - Cardinalidad univariada: 3689.
- 5) 'edad_discret': Esta característica hace referencia al grupo en el que se encuentra al que pertenece la edad del paciente.
 - Categorías: SD, i0_edad, i1_edad, i2_edad, i3_edad, i4_edad, i5_edad, i6_edad, i7_edad.
 - Cardinalidad univariada: 9.
- 6) 'sexo': Sexo del paciente.
 - Categorías: SD, F y M.
 - Cardinalidad univariada: 3.
- 7) 'viajo': Esta categoría indica si es que el paciente tuvo un viaje o no.
 - Categorías: SD, Urbano, no y sí.
 - Cardinalidad es de 4.
- 8) 'lugar_viajo': Lugares a los cuales viajó el paciente.
 - Categorías: ARGENTINA, BOLIVIA, BRASIL, PARAGUAY, SD.
 - Cardinalidad univariada: 5.
- 9) 'barrio': barrio del que proviene el paciente.
 - Categorías: demasiadas para ser citadas .
 - Cardinalidad univariada: 4470.
- 10)'departamento': Departamento en el cuál se realizó la notificación.
 - Categorías: ALTO PARAGUAY, ALTO PARANA, AMAMBAY, ASUNCION, BOQUERON, CAAGUAZU, CAAZAPA, CANINDEYU,

CENTRAL, CONCEPCION, CORDILLERA, EXTRANJERO, GUAIRA, ITAPUA, MISIONES, PARAGUARI, PRESIDENTE HAYES, SAN PEDRO, SAN PEDRO NORTE, SAN PEDRO SUR, ÑEEMBUCU.

- Cardinalidad univariada: 21.

11)'tuvo_cuadro_similar': Si es que el paciente ha tenido un cuadro similar anteriormente.

- Categorías: no, si, sino, ignora.
- Cardinalidad univariada: 4.

12)'caso_entorno': Si es que hubo un caso de dengue en el entorno del paciente.

- Categorías: si, no.
- Cardinalidad univariada: 2.

13) 'clasificacion_final': Clasificación del tipo de dengue.

- Categorías: CSA, CSA-DG, DG, SD, SSA, SSA-DG
- Cardinalidad univariada: 6

14)'criterio': En base a qué criterio fue asignada la clase final.

- Categorías: CLINICO, EPIDEMIOLOGICO, LABORATORIO, NEXO, PENDIENTE, Pendiente, SD.
- Cardinalidad univariada: 7

15)'fallecido': Señala si es que el paciente falleció o no.

- Categorías: si, no.
- Cardinalidad univariada: 2

16)'síndrome_hemorragico': Si el paciente presentó o no síndrome hemorrágico.

- Categorías: si, no.
- Cardinalidad univariada: 2.

17)'gingivorragia': Si es que se presentó o no el hemorragia en las encías.

- Categorías: si, no.
- Cardinalidad univariada: 2.

18)'derrame_pleural': Si es que se presentó o no este síntoma.

- Categorías: si, no.
- Cardinalidad univariada: 2.

19)'hemorragias_diversas': Si es que se presentó o no este síntoma.

- Categorías: si, no.
- Cardinalidad univariada: 2.

20)'NS1_AG': Presencia de proteínas no estructurales, resultado de hemograma [21].

- Categorías: NEGATIVO, POSITIVO, SD.
- Cardinalidad univariada: 3

21)'IgG': Presencia del anticuerpo específico, fabricado durante las primeras 2 o tres semanas de una infección primaria. el tipo de anticuerpo más abundante en los líquidos corporales. Brinda protección contra las bacterias y las infecciones virales [21].

- Categorías: NEGATIVO, POSITIVO, SD.
- Cardinalidad univariada: 3.

22)'IgM': se encuentra principalmente en la sangre y en el líquido linfático. Es el primer anticuerpo que el cuerpo genera para combatir una infección [21].

- Categorías: NEGATIVO, POSITIVO, SD.
- Cardinalidad univariada: 3.

23)'PCR': Reacción en cadena de la polimerasa, amplifica copias simples de ADN vírico varios millones de veces [21].

- Categorías: NEGATIVO, POSITIVO, SD.
- Cardinalidad univariada: 3.

- 24)'serotipo': Esta característica identificación al serotipo de dengue al que pertenece un caso confirmado.
- Categorías: DENV 1, DENV 2, DENV 3, DENV4, SD
 - Cardinalidad univariada: 5.
- 25)'alguna_prueba_lab': Si es que se realizó alguna prueba laboratorio para detectar el dengue.
- Categorías: POSITIVO, SD.
 - Cardinalidad univariada: 2.
- 26)'tmax_discret': Temperatura máxima diaria en grados Celsius (°C) discretizada [22].
- Categorías: SD, i0_tmax, i1_tmax, i2_tmax, i3_tmax, i4_tmax.
 - Cardinalidad univariada: 6.
- 27)'tmin_discret': Temperatura mínima diaria en grados Celsius (°C) discretizada [22].
- Categorías: SD, i0_tmin, i1_tmin, i2_tmin, i3_tmin, i4_tmin.
 - Cardinalidad univariada: 6.
- 28)'tmed_discret': Temperatura media diaria en grados Celsius (°C) discretizada [22].
- Categorías: SD, i0_tmed, i1_tmed, i2_tmed, i3_tmed, i4_tmed .
 - Cardinalidad univariada: 6.
- 29)'td_discret': Temperatura de rocío en grados Celsius (°C) discretizada [22].
- Categorías: SD, i0_td, i1_td, i2_td, i3_td, i4_td .
 - Cardinalidad univariada: 6.
- 30)'pres_nm_discret': Presión atmosférica reducida al nivel medio del mar discretizada [22].

- Categorías: SD, i0_pres_nm, i1_pres_nm, i2_pres_nm, i3_pres_nm, i4_pres_nm.
 - Cardinalidad univariada: 6.
- 31)'hr_discret': Humedad relativa discretizada (forma original porcentaje %) [22].
- Categorías: SD, i0_hr, i1_hr, i2_hr, i3_hr, i4_hr .
 - Cardinalidad univariada: 6.
- 32)'helio_discret': Horas diarias de sol (forma original horas) [22].
- Categorías: SD, i0_helio, i1_helio, i2_helio, i3_helio, i4_helio.
 - Cardinalidad univariada: 6.
- 33)'nub_discret': Cobertura nubosa discretizada (forma original: octavos) [22].
- Categorías: SD, i0_nub, i1_nub, i2_nub, i3_nub, i4_nub .
 - Cardinalidad univariada: 6.
- 34)'vmax_d_discret': Dirección del viento máximo diario discretizado (forma original: decenas de grados) [22].
- Categorías: SD, i0_vmax_d, i1_vmax_d, i2_vmax_d.
 - Cardinalidad univariada: 4.
- 35)'vmax_f_discret': Velocidad del viento máximo diario (forma original: metros por segundo m s^{-1}) [22].
- Categorías: SD, i0_vmax_f, i1_vmax_f, i2_vmax_f.
 - Cardinalidad univariada: 4.
- 36)'vmed_discret': Velocidad media del viento (forma original: metros por segundo m s^{-1}) [22]
- Categorías: SD, i0_vmed, i1_vmed, i2_vmed.
 - Cardinalidad univariada: 4.
- 37)'distrito_notif': Nombre del distrito al que pertenece el paciente.

- Categorías: demasiado largo para citar.
- Cardinalidad univariada: 239.

38) 'prcp_discret':

- Categorías: i0_prcp, i1_prcp, i2_prcp, i3_prcp, i4_prcp.
- Cardinalidad univariada: 5.

39) 'otros': En esta categoría se especifica cualquier otro tipo de síntoma que el paciente haya tenido.

- Categorías: demasiado largo para citar.
- Cardinalidad univariada: 438.

3.4.2 Muestra.

La muestra utilizada en este trabajo es del tipo no probabilista ya que según lo expresado en [20]

“Las muestras no probabilísticas, la elección de los elementos no depende de la probabilidad, sino de causas relacionadas con las características de la investigación o los propósitos del investigador”.

Para las corridas de MSU fue elaborada una muestra en base lo expresado en 2.2.6.5. Sesgo de MSU, donde según el trabajo hecho en [19] una muestra ideal es aquella que posee todos los casos posibles de cardinalidad multivariada. Para cumplir con dicho requisito se aplicó el proceso de limpieza de datos (véase 2.2.5.1. Pasos del P.D.C.) mediante la visualización de datos hecha con la librería de Python conocida como matplotlib y la análisis de las características de cada característica y clase.

Como resultado del proceso mencionado en el párrafo anterior, del universo presentado en 3.4.1.Universo fueron descartados los registros pertenecientes al año 2009 debido a que muchos de los datos capturados en este año poseen inconsistencia con relación a los registros de los 4 años posteriores con los cuales cuenta este proyecto. Cabe destacar que solo en el 2009 se da la aparición de las clases EN_ESTUDIO y DESCARTADO y ambas con valores muy bajos de 700 y 10 registros respectivamente, dichas cantidades son tan pequeñas que no logran ser apreciadas en la figura 4, por lo cuál la lista de clases finales se reduce a CONFIRMADO y SOSPECHOSO.

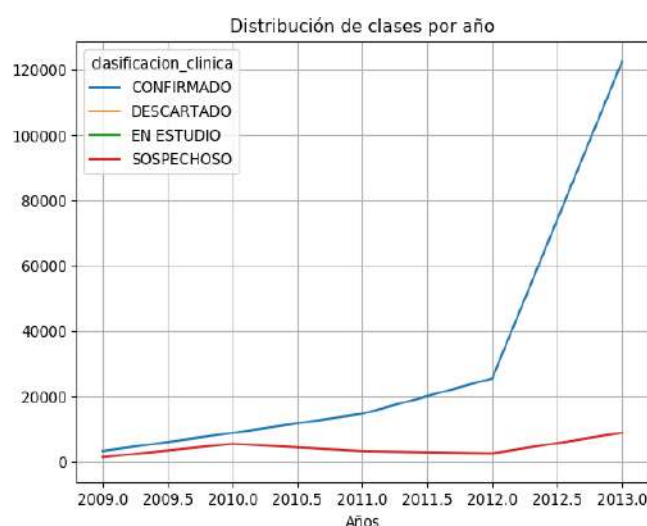


Figura 4: Distribución de clases finales por año y cantidad de casos

En cuanto a las característica, se procedió a hacer un análisis de las carnalidades univariadas de cada una de ellas, para así determinar cuales eran aquellas que se encuentran en un estado aceptable para la aplicación de MSU. En este punto el mayor de los problema detectado para algunas categorías es el alto índice de registros nulos que poseían y en consecuencia esa realidad las siguientes características fueron descartadas del conjunto de características original. Las características descartadas son las siguientes:

1. 'institucion_notificacion': Característica descartada por poseer una alta cardinalidad univariada (un total de 3689 instituciones notificantes) univariada y afectar la cardinalidad multivariada.
2. 'lugar_viajo': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 5 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

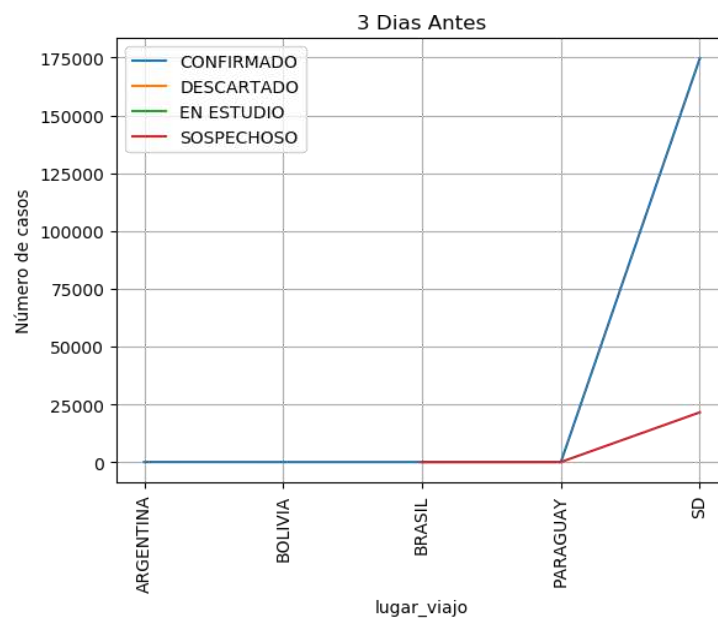


Figura 5: Distribución de las categorías pertenecientes a característica lugar_viajo, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

3. 'barrio': Característica descartada por poseer una alta cardinalidad univariada (un total de 4470 barrios) univariada y afectar la cardinalidad multivariada. Otro problema detectado con esta característica es que existen demasiados barrios en el paraguay que poseen el mismo nombre, lo cuál no es discriminado por MSU a la hora de calcular la suma de la entropía individual de esta característica.
4. 'clasificacion_final': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 6 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

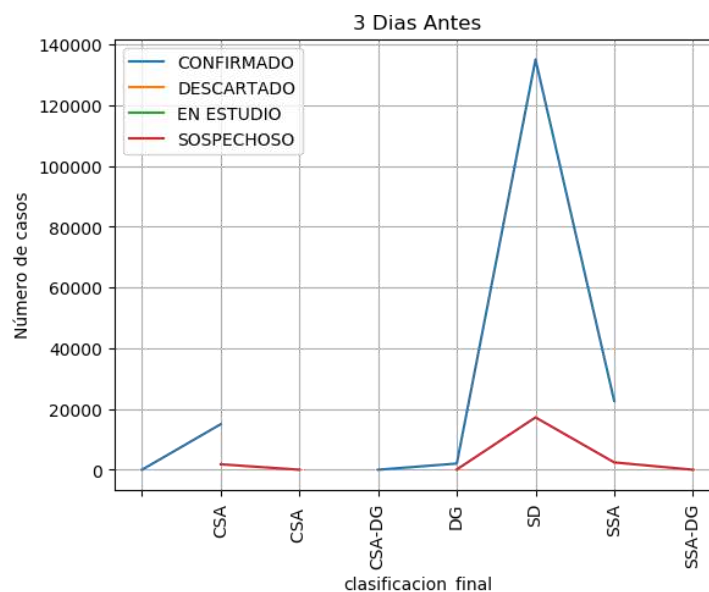


Figura 6: Distribución de las categorías pertenecientes a característica clasificacion_final, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

5. 'NS1_AG': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 7 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.



Figura 7: Distribución de las categorías pertenecientes a característica IgG, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

6. 'IgG': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 8 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

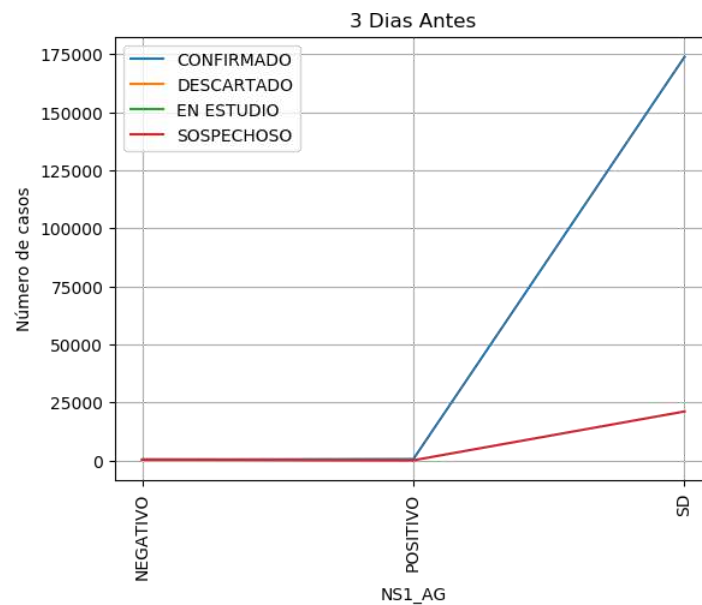


Figura 8: Distribución de las categorías pertenecientes a característica NS1_AG, en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

7. 'IgM': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 9 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

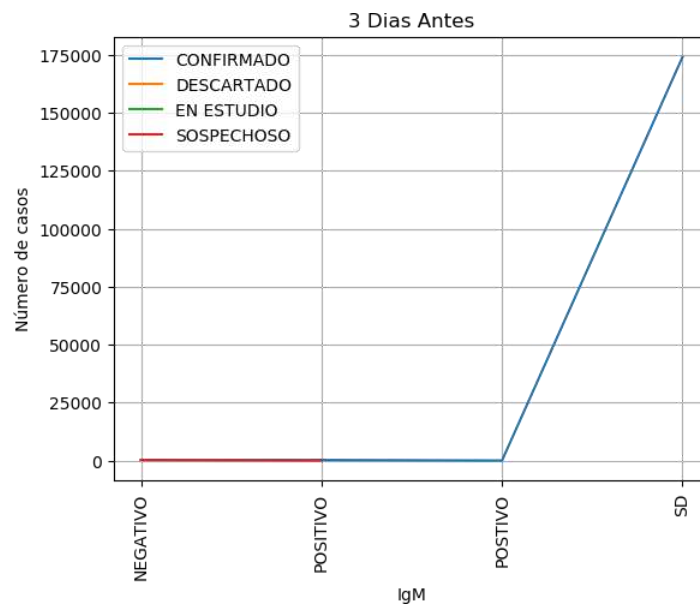


Figura 9: Distribución de las categorías pertenecientes a característica IgM en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

8. 'PRC': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 10 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

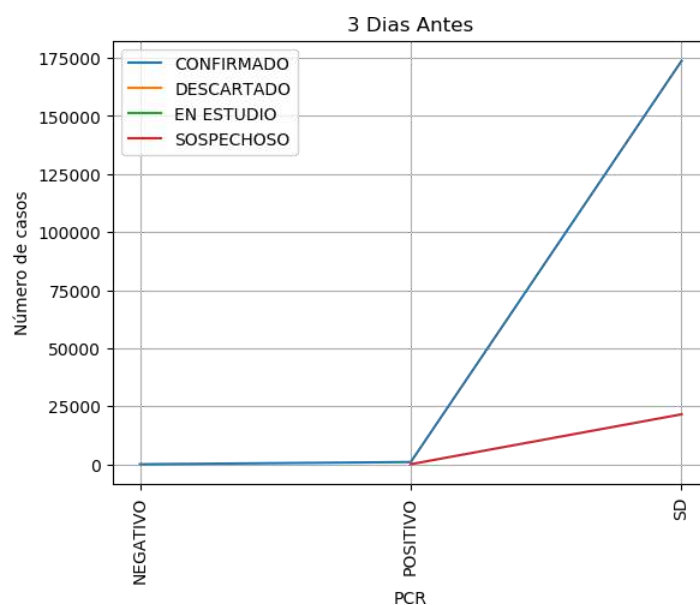


Figura 10: Distribución de las categorías pertenecientes a característica 'PRC' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

9. 'serotipo': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 11 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

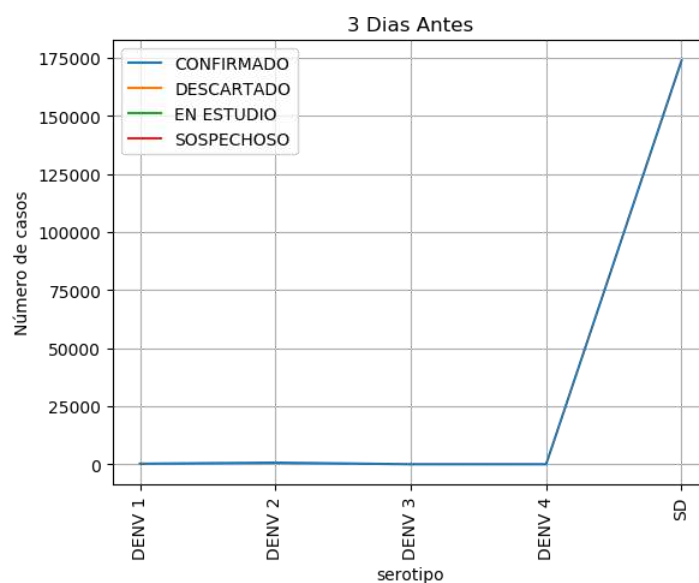


Figura 11: Distribución de las categorías pertenecientes a característica 'serotipo' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

10. 'alguna_prueba_lab': Característica descartada por poseer demasiados datos nulos, como se puede apreciar en la figura 12 donde distribución por número de casos y clase final es mayor en la categoría SD o sin datos.

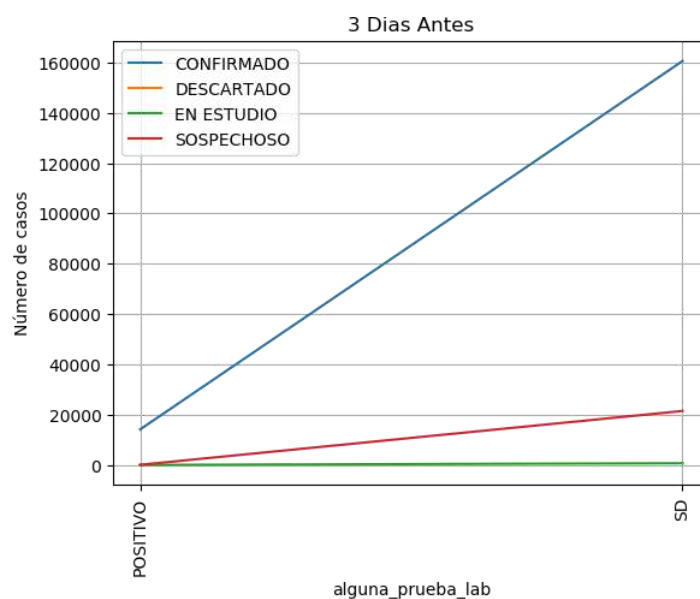


Figura 12: Distribución de las categorías pertenecientes a característica 'alguna_prueba_lab' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

11. 'otros': Característica descartada por poseer demasiados datos nulos.



Figura 13: Distribución de las categorías pertenecientes a característica 'otros' en el conjunto de datos con registros del clima de 3 días de antelación a la notificación, graficadas por clase final.

Como resultado de esta depuración los conjuntos de datos de casos de dengue agrupados con los datos de clima quedaron con el siguiente número de registros: para datos de clima de tres días antes de la notificación con 191551, para datos de clima de cinco días antes de la notificación con 191551 y para datos de clima de siete días antes de la notificación con 194061; mientras que la lista de características final se constituye 28 de características las cuales son: ['dia', 'mes', 'semana', 'edad_discret', 'sexo', 'viajo', 'departamento', 'tuvo_cuadro_similar', 'caso_entorno', 'criterio', 'fallecido', 'síndrome_hemorragico', 'derrame_pleural', 'hemorragias_diversas', 'tmax_discret', 'tmin_discret', 'tmed_discret', 'td_discret', 'pres_nm_discret', 'hr_discret', 'helio_discret', 'nub_discret', 'vmax_d_discret', 'vmax_f_discret', 'vmed_discret', 'distrito_notif', 'prcp_discret', 'gingivorragia'].

3.5 Instrumento de recolección de datos.

Como instrumento recolección de datos según [23] es identificado como **datos secundarios**, debido a que se llevó a cabo una recopilación de archivos electrónicos cuyos registros fueron capturados por la D.G.V.S. y la D.M.H.

3.6 Descripción De los Procedimientos De Análisis de Datos.

En este proyecto se procedió a realizar un análisis con enfoque cuantitativo con la finalidad de identificar cuales son las variables más significativas relacionadas con casos de Dengue que estuvieron siendo capturadas por Instituciones Estatales en el Paraguay.

3.6.1 Tratamiento inicial de los datos.

3.6.1.1 Visualización y conteo de datos.

Superadas la fases de unión y categorización se realizaron tareas de visualización y conteo de datos dentro de cada categoría de manera a poder

apreciar el estado de la cardinalidad univariada de características. Utilizando el lenguaje de programación Python con las librerías Pandas y Matplotlib con la finalidad de identificar la distribución de cada una de las categorías pertenecientes a las distintas características de los registros resultantes del procesamiento previo. Esto permite identificar registros la cantidad de registros nulos así como cualesquier otra anomalía presente dentro de los datos.

3.6.1.2 Selección de características.

Como etapa final se procedió a ejecutar el algoritmo de MSU sobre los datos resultantes de manera a generar un ranking con los valores de MSU, para diferentes subconjuntos de características hasta hallar el subconjunto más significativo.

3.6.1.3 Herramientas software utilizadas para el tratamiento de datos.

Python:

La herramientas principal utilizada fue el lenguaje de programación Python, el cuál es uno de los mejores y más ágiles para este tipo de trabajos. Según [24] Python es uno de los lenguajes más ágiles y eficientes de nuestros tiempos, debido a que posee una sintaxis considerablemente resumida en comparación a sus pares y esto posibilita que el código se claro, fácil de entender y de reescribir. Además se encuentra fuertemente utilizado para investigaciones académicas y en finanzas gracias a que posee una gran variedad de librerías y funciones propias del lenguaje que simplifican las tareas realizadas en ese tipo de trabajos.

Las otra herramientas utilizadas fueron las librerías de Python que llevan los nombres de Pandas y Matplotlib.

3.6.1.4 Plataforma computacional.

- Sistema operativo: Windows 10 Home Single Language.
- Procesador: Intel® Core™ i5-7200U CPU @ 2.50 GHz 2.71 GHz.
- Memoria (RAM): 16,0 GB (15,9 GB utilizable).

Capítulo IV

4 Marco Analítico

4.1 Presentación Del Estudio Y De Los Resultados.

En esta sección se presenta el subconjunto optimo de características hallado y el proceso aplicado sobre la muestra establecida en 3.4.2. Muestra que fue necesario para llegar a él. En primer lugar se aborda un análisis de los procesos que requiere MSU como algoritmo, su traducción al lenguaje de programación Python y su combinación con un algoritmo que permite realizar el análisis consecutivo de diferentes subconjuntos de características hasta encontrar una propuesta optima con el mayor valor de MSU. Este capítulo termina con la exposición de la ejecución del código que será presentado a continuación que permitió encontrar el subconjunto optimo de características propuesto en este proyecto.

4.1.1 Algoritmo de MSU en Python.

Una vez que han sido realizados los procesos de selección y limpieza debe ser realizado el proceso de reducción de los mismos (véase 2.2.5.1.Pasos del P.D.C. según [4]) para reducir los datos, en este proyecto se aplicó la Incertidumbre Simétrica Multivariada o MSU (por su abreviatura en ingles) en combinación con una algoritmo iterativo que será explicado en la siguiente sección.

Para poder aplicar MSU de manera concurrente, su definición expresada en [2.2.6. MSU](#) tuvo que ser analizada y desarrollada desarrollada en código¹⁶. Para mejorar el entendimiento el algoritmo de MSU el siguiente diagrama de flujo:

16 Para este proyecto se optó por el lenguaje de programación interpretado conocido como Python, en vista a la gran fluidez y variedad de herramientas que proporciona para este tipo de trabajos.

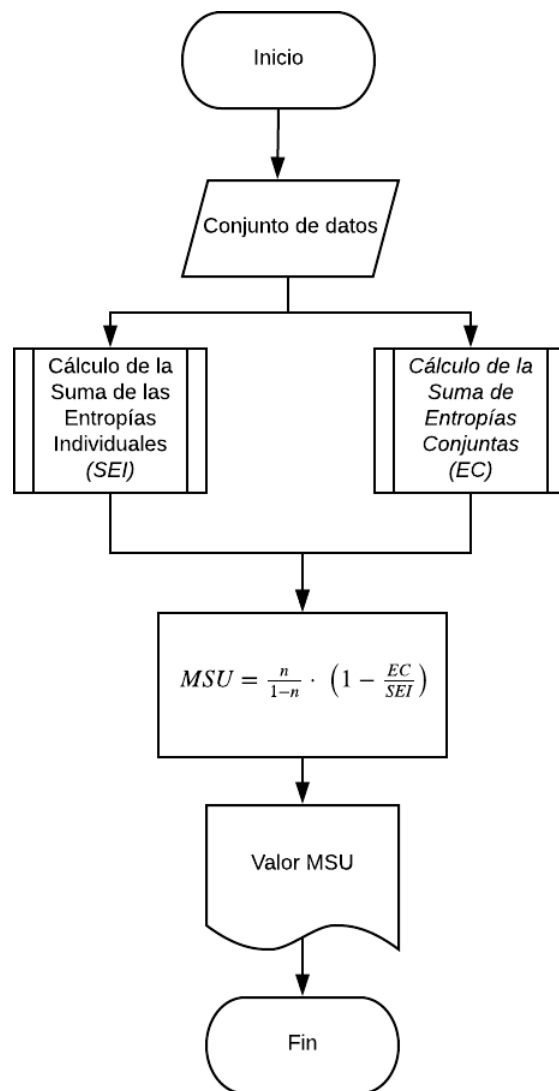


Figura 14: Diagrama de flujo del algoritmo de MSU.

Como puede apreciarse, el proceso del cálculo del MSU inicia desde la provisión de un conjunto de datos del cuál se calculan la Suma de Entropías Conjuntas y la Suma de las Entropías Individuales. En el código desarrollado a partir del diagrama anterior es el siguiente:

```

1. def MSU(df):
2.     n = len(df.columns.tolist())
3.     entropiaConjunta = sumEntropiaConjunta(df)
4.     entropiasIndividuales = sumEntropiaIndividuales(df)
5.     MSU = (n / (n - 1)) * (1 - entropiaConjunta/entropiasIndividuales)
6.     return 'msu', round(MSU, 5)

```

Donde MSU se define como una función en la línea 1, la cuál recibe un conjunto de datos al que le extrae el número de características contenidas como columnas en la línea 2, para posteriormente calcular la Suma de Entropías Conjunta en la línea 3 y la Suma de las Entropías Individuales en la línea 4, dichos cálculos son realizados a en sus funciones correspondientes, una vez estas hayan sido resueltas se procede al calculo del MSU del conjunto en la línea 5 y se regresa el resultado en la línea 6.

Ahora vamos a abordar los algoritmos utilizados para el cálculo de la Suma de Entropías Conjuntas y de la Suma de las Entropías Individuales. Empezando por describir dos las funciones para ejecutadas para cálculo de la probabilidad $P(x)$ y de la entropía $H[P(x)]$.

```

1. def px(subtotal, total):
2.     return px = subtotal / total

```

La función px realiza el cálculo de la distribución de $P(x)$ siendo x una determinada característica¹⁷ o un conjunto de ellas¹⁸, el *subtotal* es el número de veces que apareció x dentro de un conjunto de datos y el *total* representa el número total de registros dentro de dicho conjunto. Este cálculo es un paso previo al cálculo de la entropía de $P(x)$ realizado en la siguiente función.

```

1. def entropy(px):
2.     return (px * math.log(px, 2)) * -1

```

¹⁷ En el caso de las entropías individuales.

¹⁸ Esto se da para las entropías conjuntas.

Estas dos funciones son utilizadas indistintamente para calcular las Sumas de Entropías Conjuntas y la Suma de las Entropías Individuales en las siguientes dos funciones.

```
1. def sumEntropiaConjunta(df):
2.     caracteristicas=df.columns.tolist()
3.     df1 = df.groupby(columnas,
4.         sort=False).size().reset_index().rename(columns={0: 'cantidad'})
5.     df1['px'] = (df1.cantidad / df.shape[0])
6.     df1['entropia'] = df1.px.apply(entropy)
7.     return df1.entropia.sum()
```

Esta función calcula la Entropía Conjunta de un conjunto de datos, en la línea 2 se realiza la lectura de las características del conjunto de datos, en línea 3 se detectan todas las combinaciones posibles entre las características presentes en el conjunto de datos¹⁹ y al mismo tiempo estos registros son agregados con sus correspondientes cantidades de repeticiones a un nuevo conjunto de datos de nombre df1, en la línea 4 se calcula el $P(x)^{20}$ y se agrega a df1 con el nombre 'px', en la línea 5 se calcula la entropía de cada $P(x)$ y esto se agrega a df1 con el nombre 'entropía'; y, al final se retorna la Entropía Conjunta en la línea 6.

```
1. def sumEntropiaIndividuales(df):
2.     entropiasColumnas = {}
3.     for c in df.columns:
4.         df1 = pd.DataFrame(df[c].value_counts(dropna=False, sort=False))
5.         df1['px'] = (df1[c] / df.shape[0])
6.         df1['entropia'] = df1.px.apply(entropy)
7.         entropiasColumnas[c] = df1.entropia.sum()
8.     return sum(entropiasColumnas.values())
```

Con esta función se calcula la Suma de las Entropías Individuales, su llamada se realiza en la línea 4 del algoritmo de MSU descrito anteriormente. Esta función recibe un conjunto de datos del cual columna a columna (es decir una característica a la vez) va calculando el número de repeticiones de cada

¹⁹ Esto también permite determinar la cardinalidad multivariada.

²⁰ Aquí x es una de las combinaciones de características presente en df.

categoría x en la línea 4, con $P(x)$ en la línea 5 y su entropía de $P(x)$ correspondiente en la línea 6; en cada iteración se van almacenando las entropías de cada categoría en una variable cuyo tipo de dato es diccionario, para finalmente retornar la Suma de las Entropías Individuales.

4.1.2 Algoritmo propuesto para búsqueda de subconjunto optimo.

Si bien MSU posee la capacidad de cuantificar la redundancia o dependencia existente entre dos o más características [19], este no es capaz de hallar por sí solo un subconjunto optimo de las mismas ya que para hallar un subconjunto optimo de características es necesario combinar a MSU con la aplicación de algún tipo de estrategia, la cual dado un conjunto de datos permita analizar el MSU diferentes combinaciones de características hasta hallar la combinación ideal con el mayor valor de MSU, el escenario extremo sería evaluar el MSU obtenido por cada combinación posible de las características con las que se cuenta; sin embargo, el costo computacional de un proceso así es demasiado elevado y poco recomendado. Para hacer frente a esta problemática, este proyecto propone el siguiente algoritmo:

```

1. def subConjuntoOptimo(df, caracteristicas, clase):
2.     lenCaracteristicas = len(caracteristicas)
3.     df = df[caracteristicas]
4.     subConjunto = caracteristicas
5.     msuactual = MSU(df)
6.     c=0
7.     print('Original', msuactual, 'Len: ',lenCaracteristicas,
            subConjunto)
8.     for i in caracteristicas:
9.         c+=1
10.        if i != clase:
11.            subConjunto.remove(i)
12.            df2 = df[subConjunto]
13.            msuconjunto = MSU(df2)
14.            if msuconjunto > msuactual:
15.                msuactual = msuconjunto
16.                print('Removiendo: ', i, '...')
17.            else:
18.                subConjunto.append(i)
19.        print(c, ' Actual: ',len(subConjunto), subConjunto, )
20.    lenSubConjuntoN = len(subConjunto)
21.    if lenSubConjuntoN != lenCaracteristicas:
22.        print('Iteracion')
23.        subConjuntoOptimo(df, subConjunto, clase)
24. return 'MSU y subconjunto finales', msuactual, subConjunto

```

La propuesta de arriba básicamente propone encontrar un inicial MSU, el cual se obtiene del conjunto total de características y a continuación ir haciendo cálculos sucesivos de subconjuntos de características extrayendo del mismo una característica a la vez; si el valor de MSU del subconjunto actual es mayor en comparación al valor del MSU calculado previamente, la característica es descartada del subconjunto y el valor del MSU inicial es actualizado por el valor mayor.

Como entradas para este algoritmo definido como una función se deben especificar el conjunto de datos `df`, sus características y la clase²¹. A continuación se procede con la captura de la cantidad de características que

²¹ Es necesario diferenciar la clase, mediante esto es posible evitar que la clase sea descartada como cualquier característica.

han sido recibidas, esto se hace en la variable `lenCaracteristicas`: como puede verse en la línea 2, para `df` se leen solo los registros pertenecientes a las características presentes en la lista `caracteristicas` en la línea 3, los valores de de la lista `caracteristicas` son asignados a una nueva variable de nombre `subConjunto` el cuál será utilizado para establecer el subconjunto óptimo, se calcula un `msuactual` con el MSU del conjunto de datos inicial. En la línea 6 se inicializa un contador `c` que posteriormente permitirá ver el número de característica que han sido analizadas y en la línea 7 se imprimen en pantalla los valores iniciales.

En la línea 8 se especifica un bucle que recorre cada una de las características, dentro de este bucle se calcula el MSU `msuconjunto` de un nuevo subconjunto cada que se extrae una característica y en caso que su valor sea mayor que el contenido en `msuactual` la característica extraída anteriormente es excluida de manera definitiva de `subConjunto` y se actualiza el valor de `msuactual` con el de `msuconjunto`. Finalmente se comprueba si la cantidad de características capturada en `lenCaracteristicas` es diferente a la cantidad de características de `subConjunto` el que sean diferentes indica que hubo reducción del número de características y de ser así se realiza una iteración al algoritmo. El algoritmo se detiene una vez que el valor de MSU deja de mejorar con el descarte sucesivo de características.

4.1.3 Conjunto de características optimo.

Dadas las muestras establecidas en 3.4.2. Muestra una vez aplicados los algoritmos mencionados anteriormente sobre los conjuntos de datos de Dengue con datos climáticos de 3 días, 5 días y 7 días antes del registro de la notificación se obtuvieron los siguientes resultados:

3 Días Antes: ('MSU y subconjunto finales', ('msu', 0.64278084), ['clasificacion_clinica', 'nub_discret', 'vmax_f_discret', 'vmed_discret', 'pres_nm_discret', 'hr_discret', 'tmin_discret', 'dia', 'td_discret', 'semana', 'tmed_discret', 'helio_discret', 'mes', 'tmax_discret', 'prcp_discret', 'vmax_d_discret'])

5 Días Antes: ('MSU y subconjunto finales', ('msu', 0.64470747), ['nub_discret', 'vmax_f_discret', 'vmed_discret', 'pres_nm_discret', 'hr_discret', 'dia', 'helio_discret', 'mes', 'prcp_discret', 'tmax_discret', 'td_discret', 'tmed_discret', 'semana', 'vmax_d_discret'])

7 Días antes: ('MSU y subconjunto finales', ('msu', 0.65261801), ['nub_discret', 'vmax_f_discret', 'vmed_discret', 'pres_nm_discret', 'hr_discret', 'tmin_discret', 'dia', 'td_discret', 'semana', 'tmed_discret', 'helio_discret', 'mes', 'tmax_discret', 'prcp_discret', 'vmax_d_discret'])

Donde las características descartadas fueron:

['fallecido', 'derrame_pleural', 'hemorragias_diversas', 'edad_discret', 'síndrome_hemorragico', 'gingivorragia', 'caso_entorno', 'distrito_notif', 'sexo', 'tuvo_cuadro_similar', 'departamento', 'criterio']

4.2 Recursos Y Costos.

4.2.1 Costos de inversión.

Cantidad	Ocupación	Costo
1	Humanos	18.513.600 Gs.
2	Otros Recursos	7.160.000 Gs.
Total		25.673.600 Gs.

4.2.2 Recursos Humanos.

Cantidad	Ocupación	Costo unitario	Costo
1	Científico de Datos Junior.	1.542.800 Gs.	18.513.600 Gs.
Total			18.513.600 Gs.

4.2.3 Otros Recursos.

Cantidad	Ocupación	Costo Unitario	Costo
1	Computadora de escritorio.	5.000.000 Gs	5.000.000 Gs
1	Impresora	800.000 Gs.	800.000 Gs.
8	Cursos en línea de programación en Python para Ciencia de Datos.	170.000 Gs.	1.360.000 Gs.
Total			7.160.000 Gs

Capítulo V

5 Análisis e interpretación de los resultados

5.1 Conjuntos de datos obtenidos.

Los datos que fueron utilizados en este trabajo provienen del proyecto COMIDENCO (véase la sección 2.1.3.1). Dichos datos están compuestos por registros de casos de dengue capturados por la Dirección General de Vigilancia de la Salud, los cuales fueron unidos con registros del clima provenientes de la Dirección de Meteorología e Hidrología; la unión de esta unión resultó en la generación de un total de tres conjuntos de datos nuevos ya que para cada registro de caso de dengue existían datos del clima con fechas de tres días, cinco días y siete días de antelación a la registro del caso.

Una vez realizada la unión se procedió a realizar la discretización de las características que poseían valores numéricos, como resultado de estos procesos los conjuntos de datos con registros del clima de tres, cinco y siete días antes quedaron con una dimensión total de 196922, 196920 y 199428 filas correspondientes por 63 columnas, existe esta diferencia de números debido que para algunos casos de dengue no existían datos del clima de cinco días antes o de e días antes, siendo así el conjunto con registros de clima de siete días antes el más completo de todos. Entre estos datos aún se encontraban las características numérica originales que habían sido discretizadas y es por ello que se llega a un total de 63 columnas.

5.2 Limpieza de datos.

Esta etapa fue aplicada en 3.4.2. Muestra y para ello solo fueron tomadas las características categóricas, en vista a que la finalidad principal de este proyecto se centraba en la determinación de las variables más significativas y que para ello se aplicaría MSU el cuál solo funciona sobre

variables discretas o categóricas. Por esta razón en este proyecto se dejaron de lado las características numéricas, dejando solo 29 características más la clase como se muestra en 3.4.1. Universo, de las cuales varias características fueron descartadas en 3.4.2. Muestra por contar con carnalidades demasiado altas o por poseer un número bastante elevado de registros nulos; lo cuál no necesariamente significa que la característica como no sean informativas con relación a la case, más bien que la cantidad de registros válidos no es suficiente. También fueron descartados todos los registros que databan del año 2009 ya que la nulidad era prácticamente total en la mayoría de los registros del clima además; al descartar los registros de este año también son descartados dos valores posibles de clase final, los cuales son EN ESTUDIO y DESCARTADO, los cuales poseían un total de 700 y 10 registros, los cuales eran iguales en los tres conjuntos.

Al finalizar la limpieza y el procesamiento de los conjuntos con casos de dengue con registros del clima de tres, cinco y siete días antes quedaron con las dimensiones correspondientes de 191551, 191551 y 194061 registros por 28 características, lo que también resultó en la igualar los registros de casos de dengue entre los primeros dos conjuntos, sin embargo estos seguían siendo diferentes a causa de sus correspondientes registros climáticos. Estos conjuntos de datos finales fueron considerados como el universo en este proyecto

5.3 Subconjunto optimo

Una vez finalizada la limpieza de datos como se especifica en 3.4.2. Muestra y las pruebas de funcionalidad en el algoritmo abordado en la sección 4.1.2. Algoritmo propuesto para búsqueda de subconjunto optimo se procedió a ejecutar dicho algoritmo en los tres conjuntos de datos mencionados en la sección anterior y en todos los casos se obtuvo como resultado un subconjunto optimo de características: ['nub_discret', 'vmax_f_discret', 'vmed_discret',

'pres_nm_discret', 'hr_discret', 'tmin_discret', 'dia', 'td_discret', 'semana', 'tmed_discret', 'helio_discret', 'mes', 'tmax_discret', 'prcp_discret', 'vmax_d_discret']. Con un total de 15 características resultantes, esta coincidencia en los resultados es bastante relevante.

6 Conclusiones

En este trabajo se han establecido exitosamente conjuntos de datos discretos relacionados con casos de Dengue mediante el procesamiento de datos recopilados de la D.G.V.S. y de la D.M.H.

Si bien algunas características no fueron evaluadas mediante MSU a consecuencia de haber sido descartadas durante el proceso de limpieza; y, los registros con los que se contaban eran antiguos y poco numerosos en términos de Ciencias estas, estos fueron útiles para la identificación de sus características más significativas. A lo cuál se llegó mediante la aplicación de MSU en combinación con un algoritmo de búsqueda propuesto por este proyecto.

El subconjunto optimo de características con el mayor valor de MSU que fue identificado, tiene un total de 16 características los cuales provienen de un conjunto total de 28 características depuradas, este hecho demuestra que se ha llegado exitosamente a la fase de reducción dimensional a través de selección de características.

6.1 Recomendaciones.

- Analizar las implicaciones matemáticas de la utilización de MSU con el algoritmo propuesto en 4.1.2. Algoritmo propuesto para búsqueda de subconjunto óptimo.
- Aplicar el mismo proceso establecido en este proyecto con datos más actuales.
- Evaluar el grado de predictibilidad del subconjunto hallado.

6.2 Contribuciones.

Las principales contribuciones de este proyecto son las siguientes: Un conjunto de gráficos y resúmenes de visualización de datos de cada una de las características contenidas en los subconjuntos con datos de clima de tres días, cinco días y siete días antes del registro de cada caso de dengue, así como una muestra de estos datos en la cuál se descartan las características con demasiados registros nulos. También se generó una librería de MSU en Python aplicable a cualquier conjunto de datos categóricos o discretizados la cuál se combina con un algoritmo propuesta que realiza la búsqueda de un subconjunto óptimo de características.

6.3 Trabajos Futuros.

- Aplicación de MSU sobre un subconjunto imputado de datos de la solución propuesta
- Evaluación detallada del sesgo del MSU al aplicarlo sobre datos relacionados con el Dengue.

- Aplicación de técnicas de clasificación sobre el subconjunto optimo resultante de este proyecto para estimar la predictibilidad del mismo con relación a la clase final.

7 Bibliografía

- [1] «What Is Big Data? - Gartner IT Glossary - Big Data». [En línea]. Disponible en: <https://www.gartner.com/it-glossary/big-data/>. [Accedido: 15-nov-2017].
- [2] M. Posada, C. Martín-Arribas, A. Ramírez, A. Villaverde, y I. Abaitua, «Enfermedades raras: Concepto, epidemiología y situación actual en España», *An. Sist. Sanit. Navar.*, vol. 31, pp. 9-20, 2008.
- [3] Usama Fayyad, Gregory Piatetsky-Shapiro, y Padhraic Smyth, «From Data Mining to Knowledge Discovery in Databases», 1996-03-15, vol. 17, n.º 3, pp. 37-54, mar. 1996.
- [4] R. Jensen y Q. Shen, *Computational intelligence and feature selection: rough and fuzzy approaches*. Oxford: Wiley, 2008.
- [5] P. DGVS, «DGVS | Dengue», *Dengue*, 2017. [En línea]. Disponible en: <http://vigisalud.gov.py/paginas/pag/Dengue>. [Accedido: 27-jun-2017].
- [6] P. DGVS, «BOLETÍN EPIDEMIOLÓGICO». Dirección General de Vigilancia de la Salud, 2017.
- [7] Special Programme for Research and Training in Tropical Diseases y World Health Organization, Eds., *Dengue: guidelines for diagnosis, treatment, prevention, and control*, New ed. Geneva: TDR: World Health Organization, 2009.
- [8] «OMS | El mosquito», *WHO*. [En línea]. Disponible en: <http://www.who.int/denguecontrol/mosquito/es/>. [Accedido: 13-ago-2017].
- [9] «Datos Abiertos y Alertas Epidemiológicas», *cds.com.py*, 31-ago-2017. .
- [10] «COMIDENCO - CIMA - Centro de Investigaciones en Matemáticas». [En línea]. Disponible en: <http://cimapy.org/es/investigacion/proyectos/comidenco>. [Accedido: 15-nov-2017].
- [11] «Acerca», *cds.com.py*, 21-sep-2016. .
- [12] O.M.S, «OMS | Dengue», *WHO*. [En línea]. Disponible en: <http://www.who.int/topics/dengue/es/>. [Accedido: 27-jun-2017].
- [13] I. Hernández, Á. Gil de Miguel, y M. Delgado, *Manual de epidemiología y salud pública: para grados en ciencias de la salud*. Madrid: Médica Panamericana, 2013.
- [14] «Mision y Vision - Ministerio de Salud Publica y Bienestar Social». [En línea]. Disponible en: <https://www.mspbs.gov.py/portal/mision.html>. [Accedido: 12-oct-2017].

- [15] «Mision y Vision - D.G.V.S.» [En línea]. Disponible en: <http://www.vigisalud.gov.py/webdgvs/>. [Accedido: 12-oct-2017].
- [16] C. M. Ricardo, *Databases illuminated*. Sudbury, Mass: Jones and Bartlett Publishers, 2004.
- [17] R.- ASALE y R.- ASALE, «Diccionario de la lengua española - Edición del Tricentenario», *Diccionario de la lengua española - Edición del Tricentenario*. [En línea]. Disponible en: <http://dle.rae.es/>. [Accedido: 18-oct-2017].
- [18] H. Liu y H. Motoda, Eds., *Computational methods of feature selection*. Boca Raton: Chapman & Hall/CRC, 2008.
- [19] G. D. S. C. Sosa Gustavo, «REPRESENTATIVIDAD MUESTRAL EN LA INCERTIDUMBRE SIMÉTRICA MULTIVARIADA PARA LA SELECCIÓN DE ATRIBUTOS», Universidad Nacional de Asunción, Asunción, Paraguay, 2017.
- [20] R. Hernández Sampieri, C. Fernández Collado, P. Baptista Lucio, S. Méndez Valencia, y C. P. Mendoza Torres, *Metodología de la investigación*. México, D.F.: McGraw-Hill Education, 2014.
- [21] P. R. Murray, K. S. Rosenthal, y M. A. Pfaller, *Microbiología médica*. Barcelona: Elsevier España, 2014.
- [22] Hernan Veiga, Natalia Herrera, y María de los Milagros Skansi, «Descripcion de controles de calidad de datos climáticos diarios implementados por el Centro regional del Clima para el Sur de América del Sur.», Centro Regional de Clima para el sur de Amércia del Srur, Buenos Aires, 2014-002, 2014.
- [23] R. Hernández Sampieri, C. Fernández Collado, P. Baptista Lucio, S. Méndez Valencia, y C. P. Mendoza Torres, *Metodología de la investigación*. México, D.F.: McGraw-Hill Education, 2014.
- [24] E. Matthes, *Python crash course: a hands-on, project-based introduction to programming*. San Francisco: No Starch Press, 2016.