



UNIVERSIDAD TECNOLÓGICA METROPOLITANA  
FACULTAD DE INGENIERÍA  
ESCUELA DE INFORMÁTICA

**MODELO PREDICTIVO DE CONTACTABILIDAD CAWI,  
MEDIANTE REDES NEURONALES**

TRABAJO DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN  
COMPUTACIÓN MENCIÓN INFORMÁTICA

AUTOR:

GARRIDO VALENZUELA, SEBASTIÁN ALEJANDRO

PROFESOR GUÍA:

FIGUEROA PLAZA, PABLO ANDRÉS

SANTIAGO – CHILE

20 de Diciembre del 2021

## **DEDICATORIA**

El presente trabajo está dedicado a la Facultad de Ingeniería, a la Escuela de Informática, a todos los profesores por ayudarme en mi formación profesional y académica, y aportar conocimientos e ideas al área de ciencia de datos. También lo dedico a mi familia, por estar siempre apoyándome, sea como sea, en las diferentes etapas de este proceso universitario.

## AGRADECIMIENTOS

Mi profundo agradecimiento para el profesor guía, Ing. Pablo Figueroa Plaza, y a su vez subgerente de empresa “*Activa Research*”, por confiar en mí, abrirme las puertas, acompañarme en este proceso de titulación, y permitirme realizar el proceso investigativo. De igual manera, durante el transcurso de mi carrera “Ingeniería Civil en Computación Mención Informática”, mis agradecimientos a la Universidad Tecnológica Metropolitana, a toda la Facultad de Ingeniería, a todos los profesores quienes, con la enseñanza de sus valiosos conocimientos, hicieron crecer a sus alumnos, hasta convertirse en un profesional como ingeniero(a). Gracias a cada una de estas personas por su paciencia y dedicación, y también a mis compañeros(as) quienes hemos estudiado, trabajado y compartido en conjunto, durante nuestro proceso de aprendizaje y desarrollo, no cabe duda de que hay mucho por aprender en varios ámbitos.

## TABLA DE CONTENIDOS

<b>Capítulo 1.</b>	<b>Introducción.....</b>	<b>1</b>
<b>Capítulo 2.</b>	<b>Marco Teórico.....</b>	<b>3</b>
2.1.	CAWI (Computer Assisted Web Interviewing) .....	3
2.2.	Modelos.....	4
2.2.1.	Análisis Factorial: .....	5
2.2.2.	Redes Neuronales Artificiales (RNA): .....	7
<b>Capítulo 3.</b>	<b>Marco Metodológico .....</b>	<b>18</b>
3.1.	Metodología.....	18
3.1.1.	Etapa I, Comprensión del negocio: .....	20
3.1.2.	Etapa II, Comprensión de los datos: .....	21
3.1.3.	Etapa III, Preparación de los datos: .....	23
3.1.4.	Etapa IV, Modelado:.....	24
3.1.5.	Etapa V, Evaluación de proyecto: .....	25
3.2.	Herramientas.....	28
<b>Capítulo 4.</b>	<b>Comprensión de Negocio.....</b>	<b>31</b>
4.1.	Definición del Problema .....	31
4.2.	Solución Propuesta .....	33
4.3.	Objetivos .....	34
4.3.1.	Objetivo general: .....	34
4.3.2.	Objetivos específicos: .....	35
4.4.	Alcances y Limitaciones .....	35
4.4.1.	Alcances: .....	35
4.4.2.	Limitaciones: .....	36
4.5.	Hipótesis.....	36
<b>Capítulo 5.</b>	<b>Comprensión de los Datos.....</b>	<b>37</b>
5.1.	Descripción del Dataset .....	37
5.1.1.	Sendinblue: .....	38
5.1.2.	Base de datos de respuestas:.....	41
5.2.	Descripción de Variables .....	42
<b>Capítulo 6.</b>	<b>Preparación de los Datos .....</b>	<b>47</b>
6.1.	Construcción y estructuración del conjunto de datos de prueba .....	47

6.1.1.	Extracción de muestra y estructuración de Sendinblue:.....	51
6.2.	Análisis Descriptivo de los Datos .....	61
6.2.1.	Estadística con relación al horario de envíos: .....	63
6.2.2.	Estadísticas con relación a envíos con respuesta: .....	71
6.2.3.	Estadística con relación al perfil de encuestados: .....	75
6.3.	Detección y tratamiento de outliers:.....	95
6.4.	Transformación de Valores en los Datos .....	102
6.5.	Análisis Factorial .....	104
6.6.	Definición de Variables Input y Target .....	109
<b>Capítulo 7.</b>	<b>Modelado y Evaluación .....</b>	<b>111</b>
7.1.	Segmentación Data en Grupo de Entrenamiento y Evaluación .....	111
7.2.	Diseño de Arquitectura y Definición de Parámetros .....	113
7.3.	Resultados de Evaluación de Modelos.....	118
7.3.1.	Análisis y observaciones en los resultados: .....	122
7.4.	Validación de Modelos, Matriz de Confusión.....	125
7.5.	Prototipo Elaborado .....	129
<b>Capítulo 8.</b>	<b>Conclusiones .....</b>	<b>135</b>
<b>Capítulo 9.</b>	<b>Referencias Bibliográficas .....</b>	<b>138</b>
<b>Capítulo 10.</b>	<b>Anexos .....</b>	<b>143</b>
10.1.	Estadísticas CAWI, Estados de Encuestas Enviadas:.....	143
10.1.1.	Enero del año 2020.....	143
10.1.2.	Febrero del año 2020 .....	144
10.1.3.	Marzo del año 2020.....	145
10.1.4.	Abril del año 2020 .....	146
10.1.5.	Mayo del año 2020.....	147
10.1.6.	Junio del año 2020 .....	148
10.1.7.	Julio del año 2020 .....	149
10.1.8.	Agosto del año 2020 .....	150
10.1.9.	Septiembre del año 2020 .....	151
10.1.10.	Octubre del año 2020 .....	152
10.1.11.	Noviembre del año 2020 .....	153
10.1.12.	Diciembre del año 2020.....	154
10.2.	Listado de Diseño de Arquitecturas de Redes Neuronales: .....	155

10.3.	Código de Fuente Utilizado, Redes Neuronales Artificiales .....	171
10.4.	Resultados de Evaluación en RNA por Arquitectura .....	173
10.5.	Código de Fuente Utilizado, Matriz de Confusión.....	182
10.6.	Validación de Archivos de Entrada en Interfaz Gráfica .....	183
10.7.	Resultados de Predicciones Horaria de Respuestas.....	187

## ÍNDICE DE TABLAS

Tabla 5-1: Cantidad de encuestas enviadas con su respectivo estado 2020 .....	40
Tabla 5-2: Cantidad de encuestas respondidas desde la BD de respuesta 2020 .....	41
Tabla 6-1: Extracción de muestra entre Sendinblue y BD de cliente .....	53
Tabla 6-2: Pérdida de datos en estructuración de Sendinblue .....	62
Tabla 6-3: Cantidad de envíos por mes .....	62
Tabla 6-4: Frecuencia con relación a horas de envíos .....	64
Tabla 6-5: Comparativa frecuencia con relación a horas de envíos con y sin respuesta .....	65
Tabla 6-6: Frecuencia de envíos con respecto a los días de semana .....	67
Tabla 6-7: Comparativa sobre los envíos en los días de semana, con y sin respuesta.....	68
Tabla 6-8: Porcentaje de frecuencia de respuestas en un determinado horario.....	70
Tabla 6-9: Cantidad de envíos con y sin respuesta 2020 .....	71
Tabla 6-10: Cantidad de respuestas estimadas y comparación con la BD de respuesta .....	72
Tabla 6-11: Cantidad de veces que se ha respondido una encuesta por personas .....	73
Tabla 6-12: Frecuencia sobre días transcurridos hasta recibir una respuesta .....	74
Tabla 6-13: Frecuencia con relación a N° de aperturas por envío.....	74
Tabla 6-14: Comparativa N° de aperturas en los envíos con y sin respuesta .....	75
Tabla 6-15: Frecuencia con relación a sexo de los encuestados 2020 .....	76
Tabla 6-16: Frecuencia con relación a sexo de los encuestados con respuesta 2020 .....	77
Tabla 6-17: Frecuencia con relación a sexo de los encuestados S/R 2020 .....	77
Tabla 6-18: Frecuencia con relación a la edad de los encuestados 2020 .....	78
Tabla 6-19: Frecuencia con relación a la edad de los encuestados C/R 2020 .....	79
Tabla 6-20: Frecuencia con relación a la edad de los encuestados S/R 2020 .....	79
Tabla 6-21: Frecuencia con relación al rango de edades 2020 .....	80
Tabla 6-22: Frecuencia con relación al segmento de los encuestados 2020 .....	82
Tabla 6-23: Frecuencia con relación al segmento de los encuestados C/R 2020 .....	83
Tabla 6-24: Frecuencia con relación al segmento de los encuestados S/R 2020 .....	83
Tabla 6-25: Frecuencia con relación al subsegmento de los encuestados 2020.....	84
Tabla 6-26: Frecuencia con relación al subsegmento de los encuestados C/R 2020.....	86
Tabla 6-27: Frecuencia con relación al subsegmento de los encuestados S/R 2020.....	87
Tabla 6-28: Frecuencia con relación al segmento agrupado de los encuestados 2020 .....	89
Tabla 6-29: Frecuencia con relación al segmento agrupado de los encuestados C/R 2020 .....	89
Tabla 6-30: Frecuencia con relación al segmento agrupado de los encuestados S/R 2020 .....	90
Tabla 6-31: Frecuencia con relación a la carterización de los encuestados 2020 .....	91
Tabla 6-32: Frecuencia con relación a la carterización de los encuestados C/R 2020.....	92
Tabla 6-33: Frecuencia con relación a la carterización de los encuestados S/R 2020 .....	93
Tabla 6-34: Dataset N°1, frecuencia de envíos con y sin respuesta .....	100

Tabla 6-35: Dataset N°2, frecuencia de envíos con y sin respuesta .....	101
Tabla 6-36: Codificación de datos del dataset estructurado .....	103
Tabla 7-1: División de los datasets para entrenamiento y de prueba .....	112
Tabla 7-2: Resultados de evaluación arquitectura de RNA N°23 .....	121
Tabla 7-3: Resultados de evaluación arquitectura de RNA N°27 .....	121
Tabla 7-4: Resultados de evaluación arquitectura de RNA N°36 .....	122
Tabla 7-5: Resultados estimados de las respuestas reales dentro de un horario determinado.....	131
Tabla 7-6: Resultados estimados general de las respuestas predichas .....	132
Tabla 7-7: Resultado de predicción horaria del Modelo Red36_D1S2, dataset N°1 usado.....	133
Tabla 7-8: Resultado de predicción horaria del Modelo Red36_D1S2, dataset N°2 usado.....	134
Tabla 10-1: Cantidad de encuestas enviadas con su respectivo estado Enero del 2020 .....	143
Tabla 10-2: Cantidad de encuestas enviadas con su respectivo estado Febrero del 2020 .....	144
Tabla 10-3: Cantidad de encuestas enviadas con su respectivo estado Marzo del 2020.....	145
Tabla 10-4: Cantidad de encuestas enviadas con su respectivo estado Abril del 2020 .....	146
Tabla 10-5: Cantidad de encuestas enviadas con su respectivo estado Mayo del 2020.....	147
Tabla 10-6: Cantidad de encuestas enviadas con su respectivo estado Junio del 2020 .....	148
Tabla 10-7: Cantidad de encuestas enviadas con su respectivo estado Julio del 2020 .....	149
Tabla 10-8: Cantidad de encuestas enviadas con su respectivo estado Agosto del 2020 .....	150
Tabla 10-9: Cantidad de encuestas enviadas con su respectivo estado Septiembre del 2020 .....	151
Tabla 10-10: Cantidad de encuestas enviadas con su respectivo estado Octubre del 2020.....	152
Tabla 10-11: Cantidad de encuestas enviadas con su respectivo estado Noviembre del 2020 .....	153
Tabla 10-12: Cantidad de encuestas enviadas con su respectivo estado Diciembre del 2020 .....	154
Tabla 10-13: Diseño y parámetros de arquitectura RNA N°1 .....	156
Tabla 10-14: Diseño y parámetros de arquitectura RNA N°2 .....	156
Tabla 10-15: Diseño y parámetros de arquitectura RNA N°3 .....	156
Tabla 10-16: Diseño y parámetros de arquitectura RNA N°4 .....	157
Tabla 10-17: Diseño y parámetros de arquitectura RNA N°5 .....	157
Tabla 10-18: Diseño y parámetros de arquitectura RNA N°6 .....	157
Tabla 10-19: Diseño y parámetros de arquitectura RNA N°7 .....	158
Tabla 10-20: Diseño y parámetros de arquitectura RNA N°8 .....	158
Tabla 10-21: Diseño y parámetros de arquitectura RNA N°9 .....	158
Tabla 10-22: Diseño y parámetros de arquitectura RNA N°10 .....	159
Tabla 10-23: Diseño y parámetros de arquitectura RNA N°11 .....	159
Tabla 10-24: Diseño y parámetros de arquitectura RNA N°12 .....	160
Tabla 10-25: Diseño y parámetros de arquitectura RNA N°13 .....	160
Tabla 10-26: Diseño y parámetros de arquitectura RNA N°14 .....	160
Tabla 10-27: Diseño y parámetros de arquitectura RNA N°15 .....	161
Tabla 10-28: Diseño y parámetros de arquitectura RNA N°16 .....	161
Tabla 10-29: Diseño y parámetros de arquitectura RNA N°17 .....	162

Tabla 10-30: Diseño y parámetros de arquitectura RNA N°18 .....	162
Tabla 10-31: Diseño y parámetros de arquitectura RNA N°19 .....	162
Tabla 10-32: Diseño y parámetros de arquitectura RNA N°20 .....	163
Tabla 10-33: Diseño y parámetros de arquitectura RNA N°21 .....	163
Tabla 10-34: Diseño y parámetros de arquitectura RNA N°22 .....	163
Tabla 10-35: Diseño y parámetros de arquitectura RNA N°23 .....	164
Tabla 10-36: Diseño y parámetros de arquitectura RNA N°24 .....	164
Tabla 10-37: Diseño y parámetros de arquitectura RNA N°25 .....	164
Tabla 10-38: Diseño y parámetros de arquitectura RNA N°26 .....	165
Tabla 10-39: Diseño y parámetros de arquitectura RNA N°27 .....	165
Tabla 10-40: Diseño y parámetros de arquitectura RNA N°28 .....	165
Tabla 10-41: Diseño y parámetros de arquitectura RNA N°29 .....	166
Tabla 10-42: Diseño y parámetros de arquitectura RNA N°30 .....	166
Tabla 10-43: Diseño y parámetros de arquitectura RNA N°31 .....	167
Tabla 10-44: Diseño y parámetros de arquitectura RNA N°32 .....	167
Tabla 10-45: Diseño y parámetros de arquitectura RNA N°33 .....	167
Tabla 10-46: Diseño y parámetros de arquitectura RNA N°34 .....	168
Tabla 10-47: Diseño y parámetros de arquitectura RNA N°35 .....	168
Tabla 10-48: Diseño y parámetros de arquitectura RNA N°36 .....	169
Tabla 10-49: Diseño y parámetros de arquitectura RNA N°37 .....	169
Tabla 10-50: Diseño y parámetros de arquitectura RNA N°38 .....	170
Tabla 10-51: Diseño y parámetros de arquitectura RNA N°39 .....	170
Tabla 10-52: Diseño y parámetros de arquitectura RNA N°40 .....	170
Tabla 10-53: Resultados de evaluación arquitectura de RNA N°1 a 20 .....	173
Tabla 10-54: Resultados de evaluación arquitectura de RNA N°21 .....	174
Tabla 10-55: Resultados de evaluación arquitectura de RNA N°22 .....	174
Tabla 10-56: Resultados de evaluación arquitectura de RNA N°24 .....	175
Tabla 10-57: Resultados de evaluación arquitectura de RNA N°25 .....	175
Tabla 10-58: Resultados de evaluación arquitectura de RNA N°26 .....	176
Tabla 10-59: Resultados de evaluación arquitectura de RNA N°28 .....	176
Tabla 10-60: Resultados de evaluación arquitectura de RNA N°29 .....	177
Tabla 10-61: Resultados de evaluación arquitectura de RNA N°30 .....	177
Tabla 10-62: Resultados de evaluación arquitectura de RNA N°31 .....	178
Tabla 10-63: Resultados de evaluación arquitectura de RNA N°32 .....	178
Tabla 10-64: Resultados de evaluación arquitectura de RNA N°33 .....	179
Tabla 10-65: Resultados de evaluación arquitectura de RNA N°34 .....	179
Tabla 10-66: Resultados de evaluación arquitectura de RNA N°35 .....	180
Tabla 10-67: Resultados de evaluación arquitectura de RNA N°37 .....	180
Tabla 10-68: Resultados de evaluación arquitectura de RNA N°38 .....	181

Tabla 10-69: Resultados de evaluación arquitectura de RNA N°39 .....	181
Tabla 10-70: Resultados de evaluación arquitectura de RNA N°40 .....	182
Tabla 10-71: Resultado de predicción horaria del Modelo Red23_D1S2, dataset N°1 usado.....	187
Tabla 10-72: Resultado de predicción horaria del Modelo Red23_D1S2, dataset N°2 usado.....	188
Tabla 10-73: Resultado de predicción horaria del Modelo Red23_D2S2, dataset N°1 usado.....	188
Tabla 10-74: Resultado de predicción horaria del Modelo Red23_D2S2, dataset N°2 usado.....	189
Tabla 10-75: Resultado de predicción horaria del Modelo Red27_D1S2, dataset N°1 usado.....	189
Tabla 10-76: Resultado de predicción horaria del Modelo Red27_D1S2, dataset N°2 usado.....	190
Tabla 10-77: Resultado de predicción horaria del Modelo Red27_D2S2, dataset N°1 usado.....	190
Tabla 10-78: Resultado de predicción horaria del Modelo Red27_D2S2, dataset N°2 usado.....	191
Tabla 10-79: Resultado de predicción horaria del Modelo Red36_D2S2, dataset N°1 usado.....	191
Tabla 10-80: Resultado de predicción horaria del Modelo Red36_D2S2, dataset N°2 usado.....	192

# ÍNDICE DE ILUSTRACIONES

Ilustración 2-1: Esquema de un Análisis Factorial .....	6
Ilustración 2-2: Estructura general de una red neuronal.....	8
Ilustración 2-3: Gráfico de función de activación, Sígmoide .....	11
Ilustración 2-4: Gráfico de función de activación, rectificador Lineal Unitario (RELU) .....	12
Ilustración 2-5: Gráfico de función de activación, Tangente Hiperbólica .....	13
Ilustración 2-6: Gráfico de función de activación, ELU .....	14
Ilustración 2-7: Gradiente descendente estocástico y su dirección de descenso de gradiente .....	15
Ilustración 3-1: Ciclo de vida CRISP-DM para el desarrollo del proyecto .....	19
Ilustración 3-2: Etapa I de CRISP-DM, Comprensión de negocio .....	20
Ilustración 3-3: Etapa II de CRISP-DM, Comprensión de los datos.....	21
Ilustración 3-4: Etapa III de CRISP-DM, Preparación de los datos.....	23
Ilustración 3-5: Etapa IV de CRISP-DM, Modelado .....	24
Ilustración 3-6: Etapa V de CRISP-DM, Evaluación de Proyecto .....	26
Ilustración 5-1: Porcentajes de estados en los correos enviados 2020 .....	40
Ilustración 5-2: Cantidad de respuestas vs mes, encuestas del año 2020 .....	42
Ilustración 6-1: Flujos de datos y su procesamiento para construcción de datos CAWI.....	50
Ilustración 6-2: Proceso de estructuración de datos .....	60
Ilustración 6-3: Porcentaje de envíos por mes 2020.....	63
Ilustración 6-4: Gráfico de barras sobre frecuencia de horas de envíos 2020.....	64
Ilustración 6-5: Gráfico de barras sobre frecuencia de horas de envíos con respuesta 2020 .....	66
Ilustración 6-6: Gráfico de barras sobre frecuencia de horas de envíos sin respuesta 2020.....	66
Ilustración 6-7: Porcentaje de envíos realizados dentro del día de la semana 2020 .....	67
Ilustración 6-8: Porcentaje de envíos realizados dentro del día de la semana 2020 con respuesta.....	69
Ilustración 6-9: Porcentaje de envíos realizados dentro del día de la semana 2020 sin respuesta.....	69
Ilustración 6-10: Porcentaje de envíos con y sin respuesta 2020 .....	71
Ilustración 6-11: Porcentaje de sexo de encuestados 2020 .....	76
Ilustración 6-12: Porcentaje de sexo de encuestados C/R 2020 .....	77
Ilustración 6-13: Porcentaje de sexo de encuestados S/R 2020 .....	78
Ilustración 6-14: Porcentaje respecto al rango de edades de encuestados 2020.....	80
Ilustración 6-15: Porcentaje respecto al rango de edades de encuestados 2020 C/R .....	81
Ilustración 6-16: Porcentaje respecto al rango de edades de encuestados 2020 S/R.....	81
Ilustración 6-17: Porcentaje de segmentos de encuestados 2020 .....	82
Ilustración 6-18: Porcentaje de segmentos de encuestados C/R 2020 .....	83
Ilustración 6-19: Porcentaje de segmentos de encuestados S/R 2020.....	84
Ilustración 6-20: Porcentaje de subsegmentos de encuestados 2020.....	85
Ilustración 6-21: Porcentaje de subsegmentos de encuestados C/R 2020.....	87

Ilustración 6-22: Porcentaje de subsegmentos de encuestados S/R 2020.....	88
Ilustración 6-23: Porcentaje de segmentos agrupados de encuestados 2020.....	89
Ilustración 6-24: Porcentaje de segmentos agrupados de encuestados C/R 2020 .....	90
Ilustración 6-25: Porcentaje de segmentos agrupados de encuestados S/R 2020.....	91
Ilustración 6-26: Porcentaje de encuestados carterizados 2020 .....	92
Ilustración 6-27: Porcentaje de encuestados carterizados C/R 2020 .....	93
Ilustración 6-28: Porcentaje de encuestados carterizados S/R 2020.....	94
Ilustración 6-29: Comparativa de BoxPlot, resultado de mitigación de outliers en campo edad.....	96
Ilustración 6-30: Comparativa de histograma, mitigación de outliers en Segmento .....	97
Ilustración 6-31: Comparativa de histograma, mitigación de outliers en Subsegmento.....	98
Ilustración 6-32: Comparativa de histograma, mitigación de outliers en Agrupación de Segmentos.....	99
Ilustración 6-33: Comparativa de BoxPlot, resultado de mitigación de outliers en campo apertura .....	99
Ilustración 6-34: Dataset N°1, porcentaje de envíos con y sin respuesta .....	101
Ilustración 6-35: Dataset N°2, porcentaje de envíos con y sin respuesta .....	102
Ilustración 6-36: Gráfico de Codo para obtener número de factores, dataset N°1 .....	105
Ilustración 6-37: Dataset N°1, valores de los factores para cada dato campo.....	106
Ilustración 6-38: Dataset N°1, estadística de los factores.....	106
Ilustración 6-39: Gráfico de Codo para obtener número de factores, dataset N°2 .....	107
Ilustración 6-40: Dataset N°2, valores de los factores para cada dato campo.....	108
Ilustración 6-41: Dataset N°2, estadística de los factores.....	109
Ilustración 7-1: Diseño N°1 de arquitectura de una red neuronal artificial, una salida.....	113
Ilustración 7-2: Diseño N°2 de arquitectura de una red neuronal artificial, dos salidas .....	114
Ilustración 7-3: Validación de modelos de arquitectura N°23, matriz de confusión .....	126
Ilustración 7-4: Validación de modelos de arquitectura N°27, matriz de confusión .....	127
Ilustración 7-5: Validación de modelos de arquitectura N°36, matriz de confusión .....	128
Ilustración 7-6: Ventana principal del prototipo.....	131
Ilustración 10-1: Porcentajes de estados en los correos enviados Enero 2020.....	144
Ilustración 10-2: Porcentajes de estados en los correos enviados Febrero 2020.....	145
Ilustración 10-3: Porcentajes de estados en los correos enviados Marzo 2020 .....	146
Ilustración 10-4: Porcentajes de estados en los correos enviados Abril 2020 .....	147
Ilustración 10-5: Porcentajes de estados en los correos enviados Mayo 2020 .....	148
Ilustración 10-6: Porcentajes de estados en los correos enviados Junio 2020.....	149
Ilustración 10-7: Porcentajes de estados en los correos enviados Julio 2020.....	150
Ilustración 10-8: Porcentajes de estados en los correos enviados Agosto 2020 .....	151
Ilustración 10-9: Porcentajes de estados en los correos enviados Septiembre 2020.....	152
Ilustración 10-10: Porcentajes de estados en los correos enviados Octubre 2020 .....	153
Ilustración 10-11: Porcentajes de estados en los correos enviados Noviembre 2020.....	154
Ilustración 10-12: Porcentajes de estados en los correos enviados Diciembre 2020 .....	155
Ilustración 10-13: Código de fuente, generación de redes neuronales del diseño N°1 .....	171

Ilustración 10-14: Código de fuente, generación de redes neuronales del diseño N°2 .....	172
Ilustración 10-15: Código para realizar predicción y generar matriz de confusión.....	183
Ilustración 10-16: Ventana principal de interfaz, botón "Empezar" deshabilitado .....	184
Ilustración 10-17: Error de ingreso del modelo RNA, archivo incorrecto .....	184
Ilustración 10-18: Error de ingreso del modelo RNA, Arquitectura de salida no soportada .....	185
Ilustración 10-19: Error de ingreso del dataset, archivo o formato incorrecto.....	185
Ilustración 10-20: Error de ingreso del dataset, dimensiones y campos no soportados .....	185
Ilustración 10-21: Ventana de resultados de predicción del prototipo .....	186

## RESUMEN

El proyecto abordado a lo largo de esta contribución es un modelo que permite la estimación respecto a que día y que horario es el indicado para contactar a una persona, el cuál será desplegado mediante una interfaz gráfica que dará apoyo a la investigación. Los datos de entradas fueron brindados por “*Activa Research*”, dentro de estos contiene datos personales de los sujetos a contactar, información estacional y ratios respecto a su historial de contactabilidad. Estos serán ajustados a una red neuronal, cuyas salidas serán la predicción si las personas contestan las encuestas vía correo electrónico.

Para el desarrollo de este proyecto se emplea una metodología muy usada en el área de ciencia de datos llamada CRISP-DM (Cross Industry Standard Process for Data Mining), la cual consta de 5 etapas, estos son las siguientes: **comprensión de negocio, comprensión y preparación de los datos, modelamiento y evaluación del proyecto.**

### PALABRAS CLAVES:

- Modelo predictivo
- Computer Assisted Web Interviewing (CAWI)
- CRISP-DM
- Redes Neuronales Artificiales (RNA)

## ABSTRACT

The project addressed throughout this contribution is a model that allows the estimation of what day and what time is indicated to contact a person, which will be displayed through a graphical interface that will support the investigation. The input data were provided by "*Activa Research*" and within these it contains personal data of the subjects to contact, seasonal information and ratios regarding their contact history. These will be adjusted to a neural network, whose outputs will be the prediction if people answer the surveys via email.

For the development of this project a methodology widely used in data science called CRISP-DM (Cross Industry Standard Process for Data Mining) is used, which consists of 5 stages, these are the following: **business understanding, understanding and data preparation, modeling, and evaluation of the project.**

### KEYWORDS:

- Predictive model
- Computer Assisted Web Interviewing (CAWI)
- CRISP-DM
- Artificial Neural Networks (ANN)

# Capítulo 1.

## Introducción

Los distintos rubros empresariales, en su mayoría, centran su competencia en lograr la diferenciación o calidad de productos/servicios; es decir, esto recae principalmente en generar valor agregado ya sea en productos o servicios entregados. Asimismo, muchas empresas necesitan adquirir cualquier tipo de información relacionada con los datos y opiniones de las personas, para conocer sus satisfacciones, situaciones y/o recibir una retroalimentación para la toma de decisiones. Por lo tanto, las empresas requieren estar informados para desarrollar una posición, postura respecto de este o anticiparse a un hecho. Algunas principales vías para obtener información de personas a larga distancia son mediante encuestas, CAWI y a través de correos electrónicos.

Los encuestadores comúnmente formulan preguntas dirigidas a sus contactos para obtener una información que aporte a la empresa el dato que se requiere. Sin embargo, por cada encuesta no respondida se comprometen recursos los cuales se convierten pérdida de tiempo y costos. Algunos sitios, como El Mostrador (2017) señalan que las encuestas por vía web presentan un problema, ya que existen casos donde se reciben formularios extensos, causando disgusto por contestar o que son catalogados como SPAM.

Para mitigar el problema de baja tasa de respuestas en las encuestas, es necesario definir horarios de envíos donde haya una mayor certeza de que los contactos respondan. Por ende, esta solución conlleva una mejora en la eficiencia de recursos, ya que a pesar de no haber encuestadores es importante no molestar al cliente con correos excesivos, y darle un mejor uso a la base de datos de ellos.

Por lo tanto, para conocer el momento indicado de enviar formularios, se propone un modelo predictivo de contactabilidad CAWI aplicado con técnicas de Redes Neuronales, cuyas entradas son principalmente datos asociados a los encuestados donde se encuentra: los estados de la encuesta, edad, sexo, hora de recepción, día, mes, segmentaciones, agrupaciones. Por lo que, estos datos son procesados y otorgan la variable objetivo predicha, donde esta última se interpreta, si una encuesta será contestada o no.

Esta propuesta consiste en el desarrollo de modelos predictivos, con relación a la contactabilidad de personas para que respondan una determinada encuesta, cuya técnica es CAWI. Por lo tanto, el propósito de este proyecto es desarrollar un modelo predictivo que atienda de manera efectiva a la comunicación entre usuarios para generar el dato requerido por las empresas. Por lo tanto, es importante considerar las respuestas emitidas por las personas para la efectividad en la comunicación, y adquirir los conocimientos sobre los horarios óptimos que los encuestadores pueden aplicar, para así obtener eficiencia en la recopilación de información sobre sus respectivas experiencias y opiniones, ya que se busca el momento apropiado para desplegar encuestas, lo cual ahorra costos, tiempo y mejora la tasa de contactabilidad (respuestas).

# Capítulo 2.

## Marco Teórico

En este capítulo, consta de la exposición del conjunto de teorías y conceptos en que se basa con el proyecto, los cuales se detallan lo que es la técnica CAWI, y los modelos en general, donde este incluye análisis Factorial y, principalmente, Redes Neuronales Artificiales (RNA). Para este último caso, se mencionan a grandes rasgos los parámetros, funciones y optimizadores conocidos en el área de inteligencia artificial.

### 2.1. CAWI (Computer Assisted Web Interviewing)

La encuesta CAWI o por su sigla en inglés “*Computer Assisted Web Interviewing*”, según Question Pro (2018), es un cuestionario digital distribuido a través de medios online, correos electrónicos principalmente, que permite recabar información desde las respuestas que una persona responde. Asimismo, esta técnica es más usada debido al bajo costo, capacidad de enviar una encuesta a varias personas, su volumen de respuestas que una empresa reciba y la autonomía que beneficia aquellas organizaciones que requiera obtener información de las personas encuestadas. Cabe destacar que las personas suelen pasar gran parte de su tiempo conectadas a internet, por lo que ejecutar exitosamente la mayor parte de operaciones de encuestadores, permitirá obtener una mejor tasa de respuesta para la empresa.

CAWI está estrechamente relacionada con el avance tecnológico de la época, estas encuestas se terminan generando al usar una combinación perfecta entre una encuesta e internet. Esta técnica se caracteriza por encuestas o cuestionarios digitales, que generan un bajo costo, son bastante rápidas y se logra obtener información con gran velocidad. Son eficientes al momento de conseguir un buen número de respuestas, pero a diferencia de las telefónicas, aquí puede existir la mala interpretación de conceptos y alguno que otro error gramatical, de cohesión o coherencia que no permite entender lo que realmente se responde.

## 2.2. Modelos

En este proyecto se construyen modelos predictivos que, según Agencia B12 (Agencia B12, 2020) y Arimetrics (Senra, 2020), son modelos matemáticos que contiene un conjunto de procesos los cuales son ejercidos mediante técnicas de análisis de datos, permitiendo inferir las probabilidades que ocurrán determinadas situaciones previas a su consecución. Estos modelos tienen como objetivo principal, tratar los datos recopilados para convertirlos en información que entregue valor para la organización.

Dentro del modelo predictivo, se utilizan estadísticas para predecir los resultados que, por lo general, pueden aplicarse a cualquier tipo de evento desconocido, independientemente de cuándo ocurrió. En la realidad, el análisis predictivo cumple un papel importante dentro de una empresa, debido que los modelos matemáticos se utilizan para adecuar las decisiones de negocio lo cual buscan mitigar los riesgos, mejorar la recopilación de información del cliente, aumentar la capacidad para predecir cualquier comportamiento, reducir costos, incrementar beneficios, y por último simplificar las reglas del negocio aumentando su efectividad.

### **2.2.1. Análisis Factorial:**

El Análisis Factorial, según Claudio Collao (Collao, 2020), Santiago Fernández (Fernández, 2011) y Irini Mavrou (2015), es una técnica multivariada de reducción de la dimensionalidad de los datos, que se encarga de analizar la varianza común a todas las variables, permitiendo simplificar el tamaño de un problema sin demasiada pérdida de información. Este método agrupa una serie de procedimientos de análisis multivariable, que se encarga encontrar la relación mutua entre variables, permitiendo estudiar la interdependencia entre un conjunto de estos.

La idea fundamental en el análisis factorial es, como dice su nombre, analizar la correlación existente entre una serie de variables, con el propósito de descubrir alguna estructura latente (no directamente observable). Se busca la reducción de la información proporcionada por variables observadas, con la menor pérdida posible de información, en un número inferior de variables no observadas. Además, la reducción o agrupación de variables en factores o componentes principales se caracteriza por:

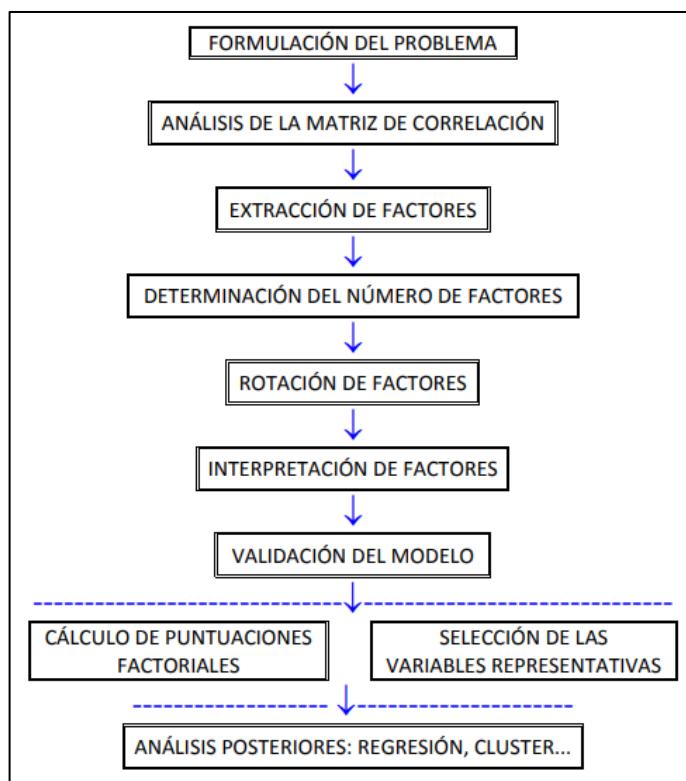
- 1) Aglutinar bajo cada factor o componente variables que estén muy correlacionadas entre ellas.
- 2) Garantizar que las variables agrupadas en distintos factores o componentes están poco correlacionadas.

De hecho, al considerar la relación entre factores o componentes, existen casos donde la correlación sea igual a cero. Esta característica indica que cada factor o componente mide o representa una dimensión distinta en los datos. Dentro del análisis factorial, es importante considerar los siguientes procesos:

- Estudiar la matriz de correlaciones
- Extraer los factores identificados
- Realizar rotación para facilitar la interpretación
- Reflejar mediante representaciones gráficas.

A partir de la ilustración [2-1], se visualiza el esquema sobre la forma de cómo trabaja el análisis factorial.

*Ilustración 2-1: Esquema de un Análisis Factorial*



Fuente: Fernández, 2011

Por lo tanto, el análisis factorial permite la búsqueda de pesos para localizar medidas distintas a partir de las variables originales, y de manera que, a poder ser, entre todas las nuevas medidas agoten o expliquen toda la varianza presente en las variables originales.

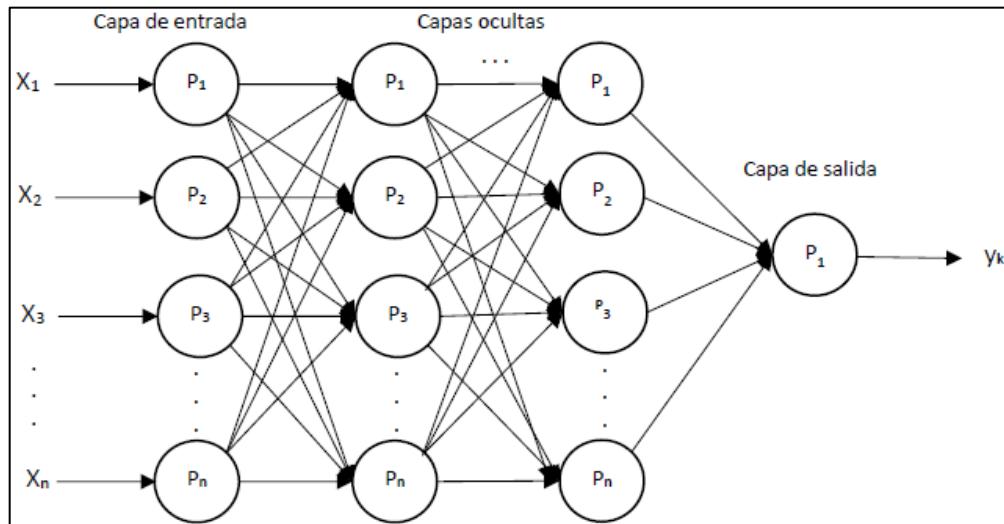
### **2.2.2. Redes Neuronales Artificiales (RNA):**

Las redes neuronales artificiales (RNA), en inglés “*Artificial Neural Networks*” (ANN), según Carlos Rebato (2020), International Business Machines Corporation (s. f.) y Eiber Galindo (2020), son un conjunto de neuronas conectadas entre sí y que trabajan en conjunto, sin que haya una tarea concreta para cada una. La principal idea de estas redes es la idea de imitar el funcionamiento de las redes neuronales de los organismos vivos, ya que a medida que se va adquiriendo la experiencia, las neuronas van creando y reforzando ciertas conexiones para **aprender** algo que se queda fijo. Cabe destacar que este es un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo, por lo que una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones. Este modelo, a partir de la ilustración [2-2], consta de tres capas distintas en el procesamiento, de los cuales son las siguientes:

- **Capas de entrada:** Representan campos entrantes de los datos a utilizar.
- **Capas ocultas:** Representan procesamiento de datos con sus respectivos cálculos y/o estimaciones, donde el modelo aprende mediante entrenamiento. Cabe destacar que la arquitectura puede tener 0, una o más capas.
- **Capas de salida:** Representan los campos de salidas como resultados calculados.

Dentro de ello, como afirma Rebato (2020), las Redes Neuronales Artificiales (RNA) utilizan grafos y funciones, conformadas por elementos de proceso (nodos) y conexiones (enlaces). Procesan entradas y generan salidas que ayudan a resolver problemas. En algunos modelos se utiliza memoria local en los nodos o elementos de proceso. Los nodos y conexiones de la red neuronal se organizan en capas.

Ilustración 2-2: Estructura general de una red neuronal



Fuente: Fernández, 2011

Las redes neuronales, como menciona Royo (2021), a diferencia de otros modelos como análisis de regresión, árboles de decisión y análisis bayesianos, contiene algoritmos de inteligencia artificial con la capacidad de aprender automáticamente, una rama conocida como “*Machine Learning*”. Estos son muy usados en la realidad, ya que lo utilizan para:

- Crear sistemas inteligentes para la toma de decisiones.
- Modelar una predicción.
- Desarrollar reconocimiento de tendencias, patrones y gestión de riesgo.
- Crear artefactos inteligentes con capacidad de aprendizaje.
- Desarrollar sistemas de visión computacional y detección.

En base a la documentación de Keras (s.f.) y TensorFlow (s.f.), dentro del modelo de redes neuronales, existen distintos parámetros que afectan los resultados del entrenamiento y de predicción, donde cada uno se configura para la creación y compilación de arquitecturas. A continuación, se mencionan estos parámetros que se debe configurar para el modelamiento:

- **Número de neuronas ( $n$ ):** Corresponde a la cantidad de nodos que tienen cada capa (entrada, oculta y salida). Dado que las redes neuronales presentan una baja interpretabilidad, ya que no es posible conocer con certeza hasta qué punto este parámetro influye en una predicción, por lo que se debe probar el resultado con distintos valores de este.
- **Épocas (epoch):** Corresponde a la cantidad de veces que se ejecutarán los algoritmos de entrenamiento, donde se van a pasando los datos por la red. Asimismo, por cada ciclo (epoch) todos los datos de entrenamiento pasan por la red neuronal para que esta aprenda sobre ellos, cuyo tamaño depende de la cantidad de datos que están analizando, correspondiente al otro parámetro “Batch size”. Mientras mayor cantidad de épocas se asignan, más lento es el proceso, la red neuronal puede aprender con más profundidad en la predicción de variables.
- **Batch Size:** Corresponde a la cantidad de datos que se introducen en la red neuronal, lo cual se utiliza entrenar dentro del algoritmo. Cabe destacar que, si el número que se asigna es pequeño, significa que la red tiene en memoria poca cantidad de datos, y entrena más rápido. Sin embargo, es posible que no aprenda las características y detalles que pueden ser significativos en la predicción del modelo. Por lo tanto, si el conjunto de datos es grande, es necesario asignar un valor más alto, ya que es más probable que tenga en cuenta los casos más importantes a la hora de aprender, entrena más lento.
- **Función de activación (Activation):** Es el encargado de devolver una salida que será generada por la neurona dada por un conjunto de datos de entrada. Cada una de las capas que conforman la red neuronal tienen una función de activación, que permite reconstruir o predecir.

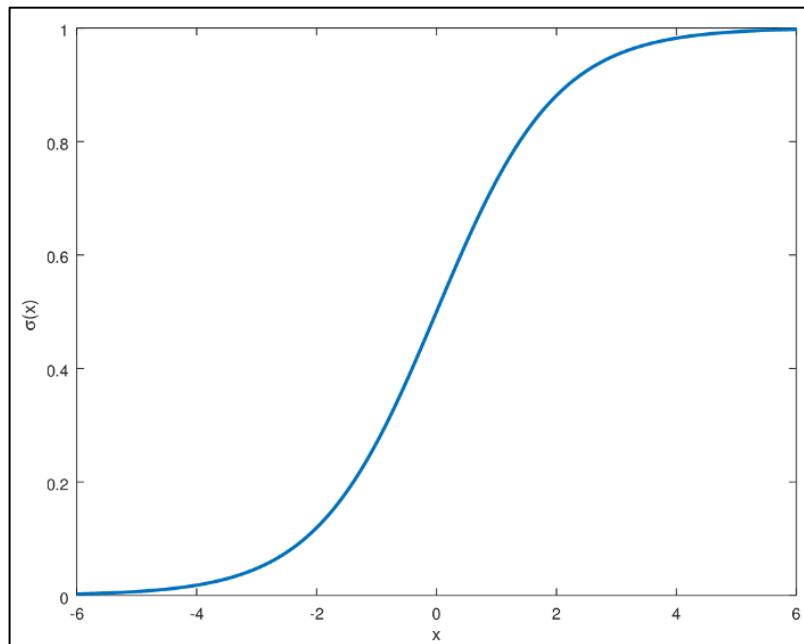
- **Función de pérdida (Loss):** Es el encargado de evaluar la desviación entre las predicciones realizadas por la red neuronal y los valores reales de las observaciones utilizadas durante el aprendizaje. Cuanto menor es el valor de la pérdida, implica que la red neuronal es más eficiente.
- **Optimizador (optimizer):** Es el encargado de minimizar el valor de la pérdida durante el proceso de entrenamiento. Entre ellos, existen diversos algoritmos que permita optimizar en las redes neuronales, los más conocidos se encuentran SGD y ADAM.

Como se ha definido anteriormente, a continuación, se mencionan las funciones de activación, los cuales generalmente harán que los modelos sean no lineales, basado en conceptos dichos Alberto (2020) y Calvo (2018).

1. **Sigmoide (sigmoid):** Es una función de regresión logística que transforma los datos a una escala (0,1), donde los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a 0. Las propiedades matemáticas que caracteriza esta función son derivables, continuas y su rango de valores es entre 0 y 1 ([0,1]). Cabe destacar que esta función de activación es muy usada para problemas de clasificación binaria, cuya formula y gráfico se visualiza a continuación:

$$f(X) = \frac{1}{1 + e^{-X}}$$

Ilustración 2-3: Gráfico de función de activación, Sigmoid

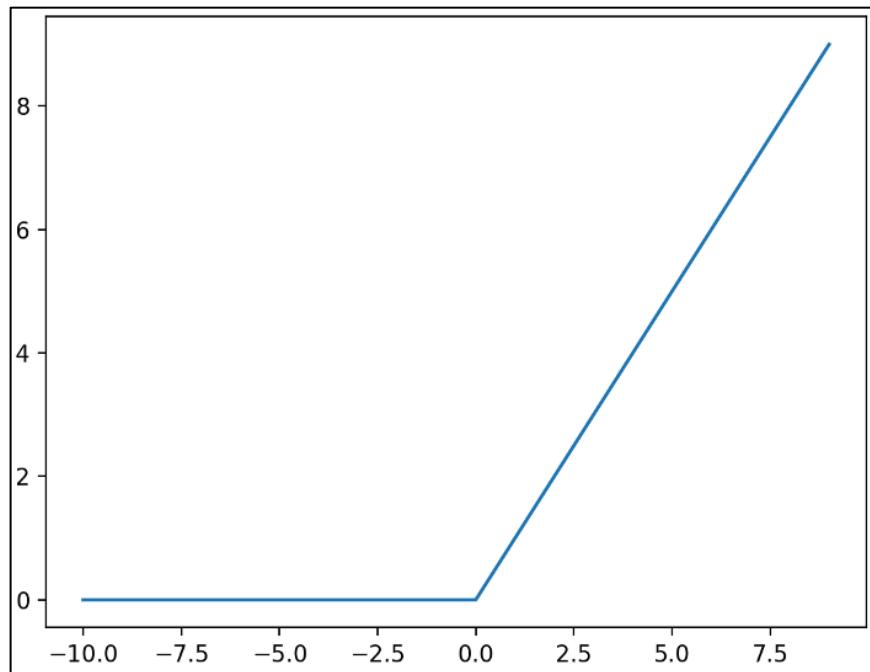


Fuente: Alberto, 2020

2. **Rectificador Lineal Unitario (ReLU):** Es una función que transforma los datos introducidos a la capa, donde anulan los valores negativos y dejan solo otros positivos. Esta función de activación es muy usada para problemas de clasificación de imágenes, cuyas propiedades matemáticas son continuas, no es derivable y su rango de valores es entre 0 (inclusive) al infinito positivo, cuya formula y gráfico se visualiza a continuación:

$$f(X) = \max(0, X)$$

Ilustración 2-4: Gráfico de función de activación, rectificador Lineal Unitario (RELU)

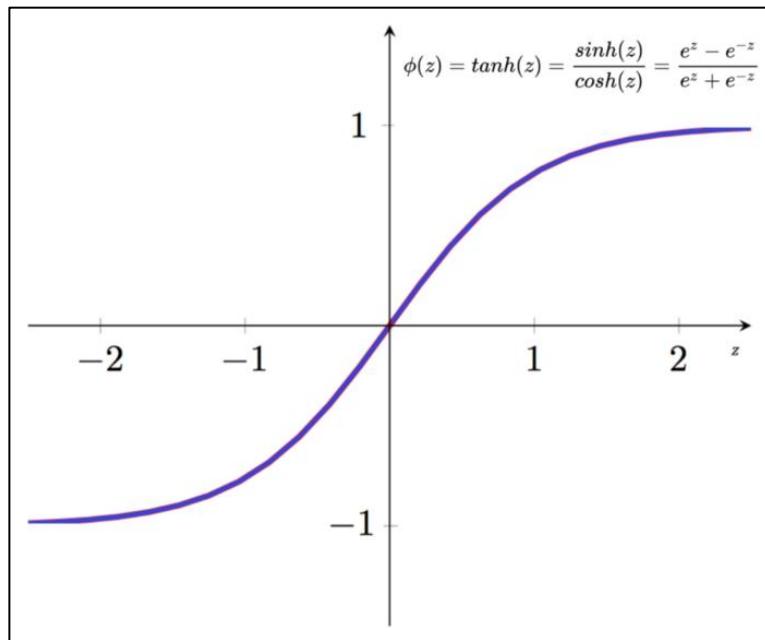


Fuente: Alberto, 2020

3. **Tangente Hiperbólica (Tanh):** Es una función que transforma los datos introducidos a una escala entre -1 y 1 (inclusive), donde los valores altos tienden de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a -1. Las propiedades matemáticas que caracteriza esta función son derivables, continuas y su rango de valores, como se ha mencionado anteriormente, es entre -1 y 1 ( $[-1,1]$ ), cuya formula y gráfico se visualiza a continuación:

$$f(X) = \frac{2}{1 + e^{-2X}} - 1 = \frac{e^X - e^{-X}}{e^X + e^{-X}}$$

Ilustración 2-5: Gráfico de función de activación, Tangente Hiperbólica



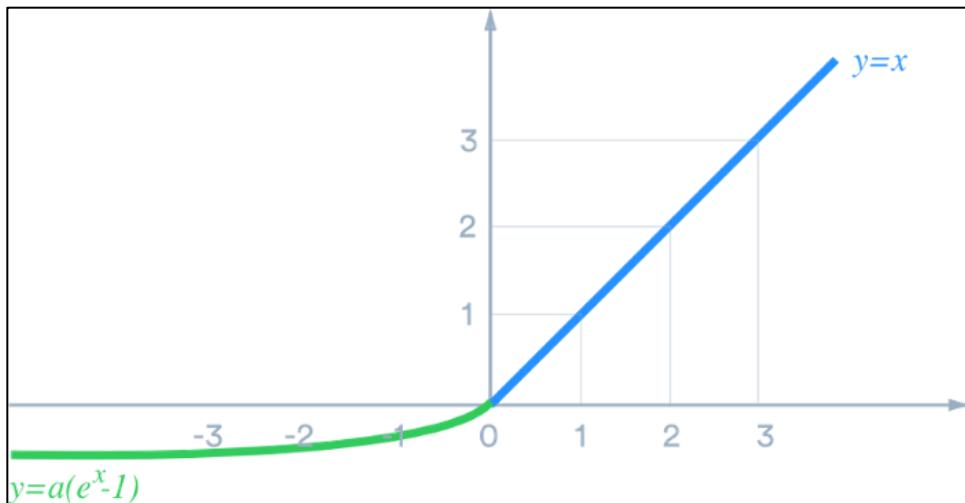
Fuente: Alberto, 2020

- 4. Unidades Lineales Exponenciales (Exponential Linear Units, ELU):** Es una función similar a RELU, transforma los datos introducidos a la capa y dejan solo otros positivos. Sin embargo, para valores que son negativos se reemplaza mediante aplicación de fórmula exponencial, haciendo que estos valores sean más cercanos al 0. Por lo tanto, las propiedades matemáticas que caracteriza esta función son continuas y sus valores corresponden a cualquier valor positivo y negativo, cuya formula y gráfico se visualiza a continuación:

$$f(X) = e^X - 1, \quad Si X < 0$$

$$f(X) = X, \quad Si \geq 0$$

Ilustración 2-6: Gráfico de función de activación, ELU



Fuente: Alberto, 2020

5. **Softmax:** Por último, según Ramírez (2021), esta función de activación es una generalización de la regresión logística que puede ser aplicada a datos continuos. Cabe destacar que este es muy usado para problemas de clasificación multinomial, además que se convierte en el recurso principal utilizado en las capas de salida de un clasificador. Esta función de activación devuelve la distribución de probabilidad de cada una de las clases soportadas en el modelo, dependiendo las dimensiones de salidas. Por lo tanto, las propiedades matemáticas que caracteriza esta función son continuas, su rango de valores es entre 0 y 1 ([0,1]), y la suma de todas las probabilidades, de una fila, será igual a uno. Asimismo, la fórmula que denota esta función se visualiza a continuación, siendo “n” como cantidad de eventos:

$$f(X_i) = \frac{e^{X_i}}{\sum_{j=0}^n e^j}, \text{ donde } i = 0, 1, 2, \dots, n$$

En cuanto a los algoritmos de optimizadores, existen muchos que pueden ser utilizado para mejorar el rendimiento del entrenamiento y/o reducir las pérdidas de valores, entre ellos se encuentran SGD, ADAM, ADAGRAD y ADAMAX. A continuación, se detallan estos algoritmos, los cuales son usados para el modelamiento de redes neuronales.

- ***Gradiente Descendente Estocástico (SGD)***

En este algoritmo, según Rivera (2018) y Ruder (2018), se encuentran los parámetros que disminuyen el error de la función objetivo, lo cual reduce también la esperanza del error sobre la distribución generadora de datos, todo esto a partir del promedio de una muestra. Dentro del algoritmo, la tasa de aprendizaje se mantiene en un punto, pero cambia de dirección controladamente hasta alcanzar un valor máximo. La única gran desventaja de utilizar este algoritmo es en cuanto el efecto de los outliers en el gradiente de una muestra, lo cual puede afectar más fuertemente y desviar al algoritmo de su trayectoria de convergencia. A partir de la ilustración [2-7], se visualiza una formula o ecuación de gradiente descendente estocástico en general, lo cual también incluye su dirección de descenso de gradiente, donde  $\Omega$  es el conjunto de datos y  $\#\Omega$  es la cantidad de elementos en el conjunto:

Ilustración 2-7: *Gradiente descendente estocástico y su dirección de descenso de gradiente*

$\arg \min_x f(x) \stackrel{\text{def}}{=} \frac{1}{\#\Omega} \sum_{i \in \Omega} f_i(x)$
$\begin{aligned} p^t &= -\nabla f^t \\ &= -\frac{1}{\#\Omega} \sum_{i \in \Omega} \nabla f_i^t \end{aligned}$ <p style="text-align: center;"><i>Dirección de descenso de gradiente</i></p>

Fuente: Rivera (2018)

Cabe mencionar que la dirección de convergencia de la tasa de aprendizaje, en un punto determinado, corresponde a la suma del gradiente en el punto actual más la dirección de aprendizaje del punto anterior en un porcentaje cercano a su totalidad, lo cual habitualmente es cercano al 90%. Para el caso del nuevo punto del gradiente, este será la diferencia entre el punto anterior del gradiente y la multiplicación de la dirección de convergencia con la tasa de aprendizaje.

- ***Adaptive Moment Estimation (ADAM):***

El algoritmo Estimación Adaptativa de Momentos (ADAM), según Brutalk (2021) y Programador Clic (2021), es un optimizador estocástico que permite solventar el problema con la fijación de la tasa de aprendizaje, cómo estén distribuidos los parámetros. Asimismo, este optimizador calcula el promedio móvil exponencial del gradiente, y los hiperparámetros beta1 y beta2 controlan la tasa de disminución de estos promedios móviles. El valor inicial de la media móvil y los valores de beta1 y beta2 están cerca de 1 (valor recomendado), por lo que la desviación de la estimación de momento está cerca de 0. La desviación se mejora calculando primero la estimación con desviación y luego calculando la estimación después de la corrección de desviación.

Por lo tanto, la ventaja de utilizar este algoritmo es que incorpora una corrección del sesgo a la hora de estimar los momentum. por lo que, si los parámetros están muy dispersos y existe una alta correlación en los valores (datos), entonces la tasa de aprendizaje aumentará. Asimismo, la tasa de aprendizaje se reserva adaptativamente para cada parámetro, como resultado, el algoritmo tiene un excelente rendimiento en problemas no estacionarios y en línea.

- ***Adaptative Gradient Algorithm (ADAGRAD):***

El Algoritmo de Gradiente Adaptativo, según Duchi (2011), es un optimizador que permite que adaptar la ratio de aprendizaje a los parámetros, lo cual se basa en gradientes. Este algoritmo se caracteriza principalmente por realizar muchas actualizaciones en los parámetros que son pocos frecuentes, y, por otro lado, otras pocas actualizaciones en caso de que sean muy frecuentes. Además, es destacable que cada actualización que se requiera utiliza una tasa de aprendizaje fija.

Sin embargo, la desventaja de utilizar este optimizador es la posibilidad de que la tasa de aprendizaje para una variable decrezca demasiado rápidamente, y esto es debido a la acumulación de altos valores del gradiente al comienzo del entrenamiento. Por lo tanto, si esto ocurriera, puede llevar a que el entrenamiento no sea capaz de aproximarse al mínimo en dicha dimensión.

- ***ADAMAX:***

Por último, el optimizador ADAMAX, según Kingma (2015), es un algoritmo proveniente de una variante de Adam que, en lugar de un momento de segundo orden, se utiliza un momento de orden infinito. Este algoritmo optimizador proporciona un rango más simple para el límite superior de la tasa de aprendizaje. Asimismo, este algoritmo optimizador proviene de la distribución de algoritmo similar a Adam, teniendo como prioridad a la maximización de la regla de actualización. Por lo tanto, esto hace que la función final no sea tan susceptible a la varianza de los datos.

# Capítulo 3.

## Marco Metodológico

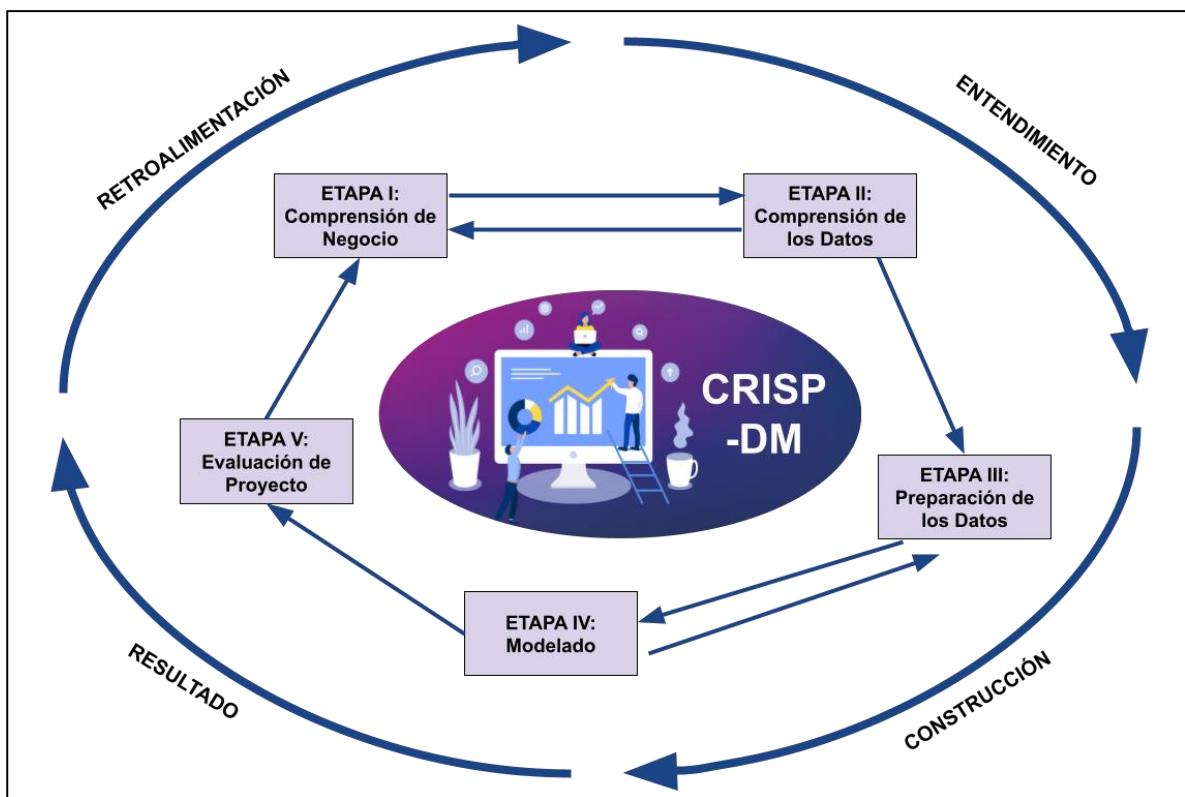
En este capítulo, consiste en los detalles respecto a las metodologías que son utilizadas para el desarrollo completo del proyecto, referente a CRISP-DM. En ella, se detallan las etapas con contiene en ella, donde cada uno incluye fases (actividades) que se debe realizar. Además, se mencionan todas las herramientas que se utiliza para análisis estadístico, manipulación y estructuración de los datos, y creación de modelos de redes neuronales artificiales.

### 3.1. Metodología

El proyecto de investigación aplicada se basa en una metodología que permite crear un modelo de minería de datos que se adapte a sus necesidades concretas, y su enfoque se centra en el área de Ciencia de Datos e Inteligencia Artificial, por lo que, la metodología a trabajar es CRISP-DM con sus siglas “*Cross-Industry Standard Process for Data Mining*”. Para su aplicación adecuada, es importante analizar todos los tipos de datos, teniendo en cuenta su fiabilidad, agrupar la información, identificar el comportamiento de ello y estar alineados al contexto, objetivos y solución propuesta. Esta metodología, a partir de la ilustración [3-1], consta de 5 fases los cuales son comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación de proyecto.

Según Sngular (2019), IBM Knowledge Center (s. f) y Healthdataminer (2019), CRISP-DM es el proceso estándar de la industria para la minería de datos que proporciona una descripción normalizada del ciclo de vida de un proyecto, específicamente el de análisis de datos, lo cual se destaca por su flexibilidad y se pueden personalizar dependiendo de las actividades que se llevarán a cabo. Esta metodología contempla el proceso de análisis de datos como un proyecto profesional, por esta razón se establece que la existencia de un cliente no es parte del equipo, ya que está relacionado con otros proyectos donde es preciso documentarlo de forma exhaustiva. De esta forma, los otros equipos de proyectos utilizarán el conocimiento adquirido y puedan trabajar a partir de este.

*Ilustración 3-1: Ciclo de vida CRISP-DM para el desarrollo del proyecto*

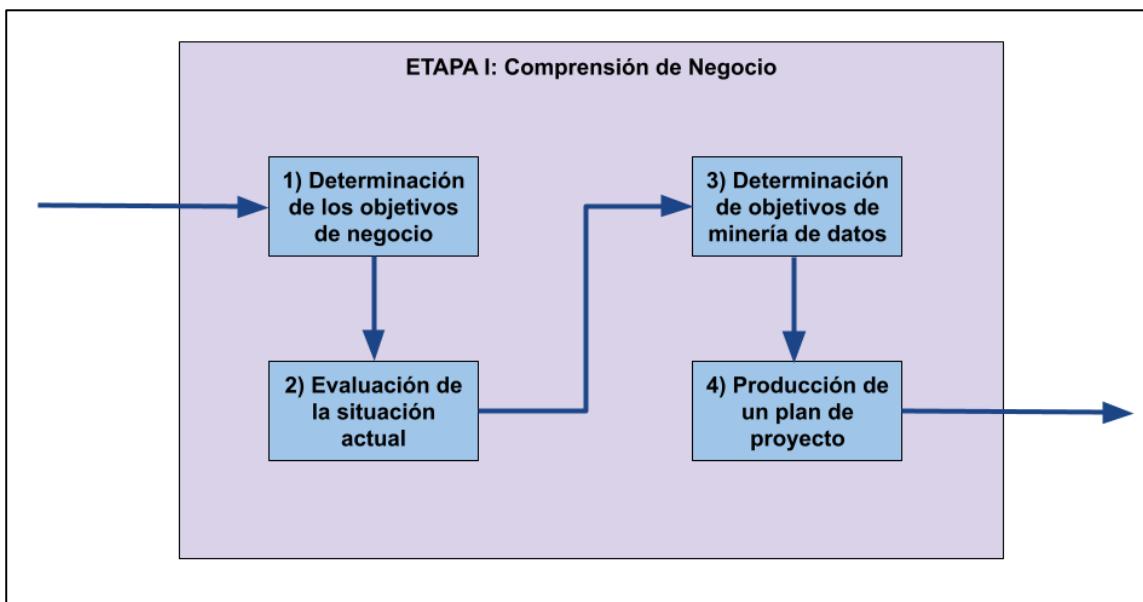


Fuente: Elaboración propia

### 3.1.1. Etapa I, Comprensión del negocio:

En esta primera etapa, se establece los objetivos de negocio a nivel de minería de datos, se evalúa la situación actual para visualizar el problema en relación con la contactabilidad de los clientes, y se define un plan para llevar a cabo el proyecto. A partir de la ilustración [3-2] se detalla las fases que tendrá dentro de esta etapa inicial del trabajo.

Ilustración 3-2: Etapa I de CRISP-DM, Comprensión de negocio



Fuente: Elaboración propia

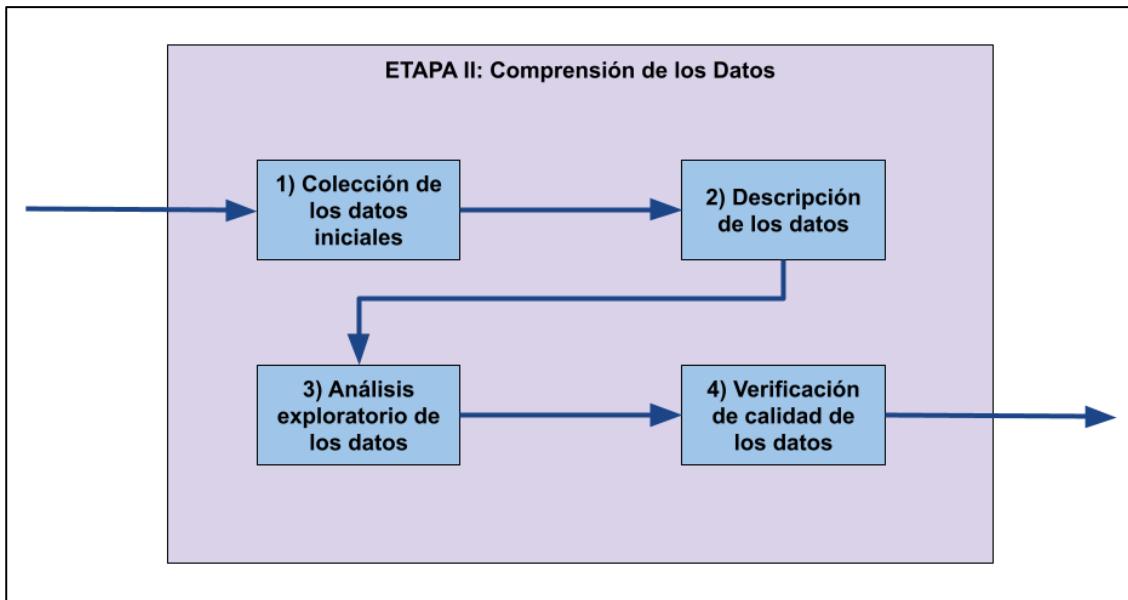
- 1. Determinación de los objetivos:** Esta primera fase, se efectúa una recopilación de los antecedentes respecto al tema del proyecto, donde se definen problema, objetivos y solución del negocio.
- 2. Evaluación de la situación actual:** En esta segunda fase, se realiza la evaluación sobre el estado actual que se encuentra el negocio, respecto a los supuestos, limitaciones y proyecciones en el futuro que el proyecto puede tener.

- 3. Determinación de objetivos de minería de datos:** En esta tercera fase, se definen los objetivos de minería de datos para desarrollar una solución al tema.
- 4. Producción de un plan de proyecto:** En esta última fase, se crea un plan de proyecto, planteando un conjunto de tareas y actividades estimadas para alcanzar los objetivos de minería de datos.

### 3.1.2. Etapa II, Comprensión de los datos:

En esta segunda etapa, se coleccionan los datos iniciales desde registros de encuestas, lo que implica acceder, explorar, analizar, estudiar y verificar la calidad en ellos. A partir de la ilustración [3-3] se detallan las fases que tendrá dentro de esta etapa del proyecto.

Ilustración 3-3: Etapa II de CRISP-DM, Comprensión de los datos



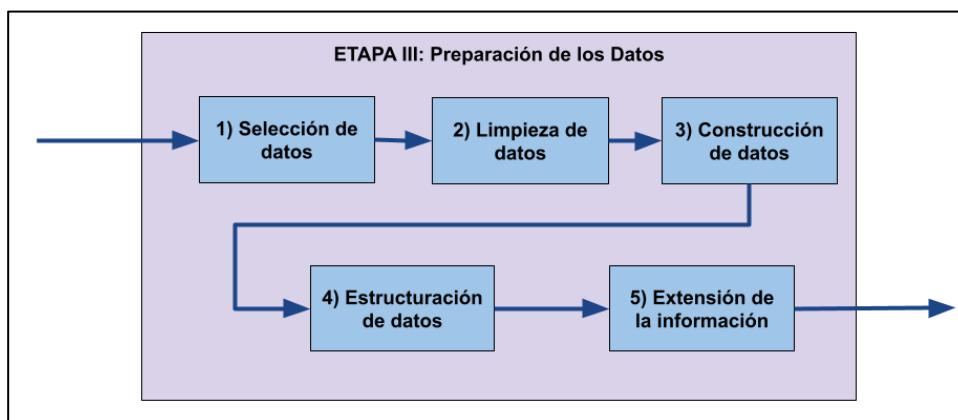
Fuente: Elaboración propia

- 1. Colección de los datos iniciales:** En esta primera fase, se recopilan conjuntos de datos provenientes de la organización.
- 2. Descripción de los datos:** En esta segunda fase, se describen las características que tienen los datos recopilados, respecto a cantidad, tipo de valores y esquemas de codificación que lo contienen.
- 3. Análisis exploratorio de los datos:** En esta tercera fase, se efectúa una exploración de los datos mediante tablas, gráficos y/u otras herramientas de visualización, con el fin de poder analizarlas y obtener alguna hipótesis.
- 4. Verificación de calidad de los datos:** En esta fase final, consta en la búsqueda de problemas en los registros recopilados, los cuales pueden tener:
  - **Datos perdidos:** Son aquellos datos que se encuentran ausentes dentro de un registro.
  - **Errores de datos:** Son aquellos datos que no cumplen el tipo o formato adecuado.
  - **Errores de mediciones en los datos:** Son datos que se escapan de un criterio establecido respecto a unidad de medida.
  - **Outliers:** Son datos que tienen un comportamiento distinto al resto, que por lo general afecta a la distribución del conjunto.
  - **Incoherencia en la codificación:** Son datos que no coincide con alguna definición o categoría definida.
  - **Selección de datos poblados:** Tras descartar datos vacíos, se seleccionan aquellos datos que contribuyen con la información posible y en gran volumen.

### 3.1.3. Etapa III, Preparación de los datos:

En esta tercera etapa, se construyen uno o más conjuntos de datos mediante selección, limpieza y transformación de atributos, tablas y registros. A partir de la ilustración [3-4] se detallan las fases que tendrá la etapa de preparación en los datos del proyecto.

Ilustración 3-4: Etapa III de CRISP-DM, Preparación de los datos



Fuente: Elaboración propia

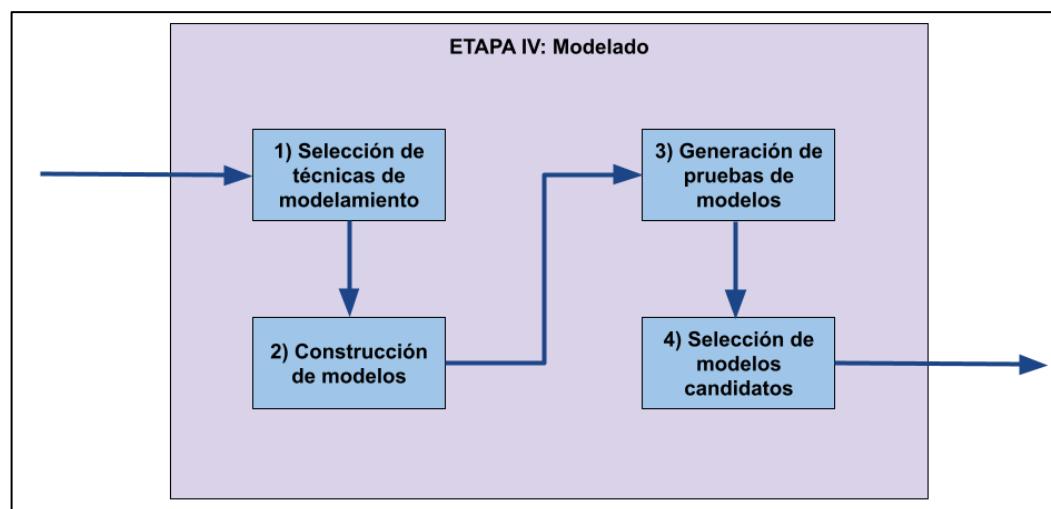
1. **Selección de datos:** Esta primera fase, se seleccionan los elementos, atributos y/o características relevantes para el cumplimiento de los objetivos de minería de datos.
2. **Limpieza de datos:** Esta segunda fase, se procede a solucionar problemas en los registros de datos seleccionados.
3. **Construcción de datos:** Esta tercera fase, a través de derivación de atributos (columnas) se construyen nuevos conjuntos de datos (con ratios calculados).
4. **Estructuración de datos:** Esta última fase, se construyen estructura de los conjuntos de datos generados, los cuales serán utilizados mediante técnicas de modelamiento.

**5. Extensión de la información:** Para complementar información a la ya existente, los datos serán puestos en diversos análisis como: Análisis de reducción de información, Análisis PCA y factorial y Análisis de clustering. La finalidad de esta fase se enfoca en reducir la redundancia existente en la información y lograr una optimización de los datos de entrada.

### 3.1.4. Etapa IV, Modelado:

En esta cuarta etapa, con los conjuntos de datos ya estructurados, se aplican técnicas de modelamiento adecuadas para encontrar la solución al contexto, calibrando parámetros y valores óptimos durante la construcción de modelos. A partir de la ilustración [3-5] se detallan las fases que tendrá dentro de esta etapa.

Ilustración 3-5: Etapa IV de CRISP-DM, Modelado



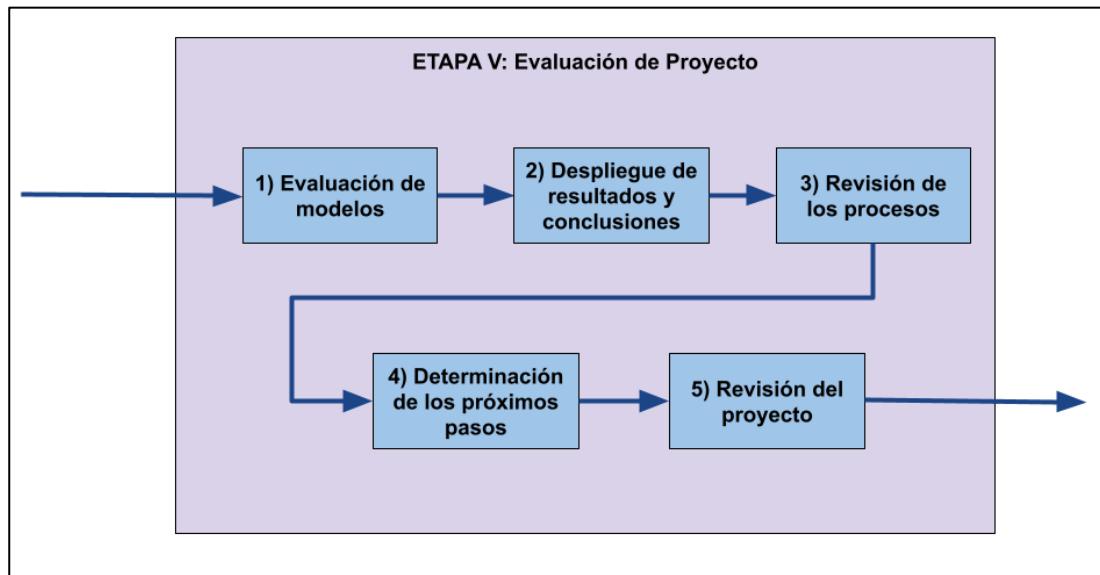
Fuente: Elaboración propia

- 1. Selección de técnicas de modelamiento:** Esta primera fase consta de selección de técnicas adecuadas de modelado, dependiendo de las estructuras de datos generados.
- 2. Construcción de modelos:** En esta segunda fase, se aplican técnicas de modelado y se realiza una configuración de parámetros en el algoritmo.
- 3. Generación de pruebas de modelos:** En esta tercera fase, se ejecutan una serie de pruebas en los modelos construidos, esto incluye revisión de los parámetros y sus despliegues de resultados durante su ejecución.
- 4. Selección de modelos candidatos:** En esta última fase, se verifican los resultados generados de cada modelo, identificando si existe algún problema, inconveniente o incongruencia, y se selecciona algunos de estos para ser evaluados.

### **3.1.5. Etapa V, Evaluación de proyecto:**

En esta última etapa, se realiza la evaluación de varios modelos candidatos generados, a través de análisis, comparaciones entre ellos y revisión de todos los procesos realizados en el modelamiento. Una vez terminado, se procede a determinar los próximos pasos para mejorar el proyecto de investigación aplicada y se efectúa su revisión completa con el fin de documentarlas. A partir de la ilustración [3-6] se detallan las fases que tendrá dentro de esta etapa.

Ilustración 3-6: Etapa V de CRISP-DM, Evaluación de Proyecto



Fuente: Elaboración propia

1. **Evaluación de modelos:** En esta primera fase, se efectúa un análisis y comparación de resultados obtenidos durante la evaluación de los modelos generados, con el fin de que estos se adapten de acuerdo con los objetivos a alcanzar. Cabe destacar, que para lograr a cabalidad esta fase, se hará uso de la validación cruzada.
2. **Despliegue de resultados y conclusiones:** En esta segunda fase, se despliega los resultados del algoritmo, donde esto se visualiza mediante ejecución de códigos y generación de una interfaz gráfica, que aporte a la investigación. Además, se sintetiza los puntos relevantes y conocimientos explorados a lo largo del proyecto.
3. **Revisión de los procesos:** En esta tercera fase, se procede a una revisión de todos los procesos de minería de datos realizados en las etapas anteriores, con el fin de analizarlos y generar sugerencias de mejoras.

**4. Determinación de los próximos pasos:** En esta penúltima fase, se crea un listado de posibles acciones que se tomarán de acuerdo con la revisión anterior, con el fin de mejorar la realización de actividades en el próximo ciclo.

**5. Revisión del proyecto:** Por último, en esta fase se ejecuta una revisión completa del proyecto, lo cual se concluye con el ciclo de esta investigación, identificando resultados que son relevantes, las lecciones y experiencias que se tuvieron durante su desarrollo completo.

En la metodología CRISP-DM, como se visualiza su ciclo de vida en la ilustración [3-1], contiene flechas que indican las dependencias más importantes y frecuentes, cuya secuencia, en la mayoría de los proyectos, avanzan y retroceden entre estas etapas dependiendo de su necesidad y resultado. En la segunda etapa de la metodología, comprensión de los datos, es importante verificar si los conjuntos de datos recopilados no contengan problemas de calidad irreparables o de gran magnitud, y por otro lado que no estén muy alejados a los objetivos del negocio, ya que puede afectar considerablemente el desarrollo de la investigación del proyecto en cuanto su efectividad. Para este caso, si en la fase 4 de esta etapa, los datos presentan problemas graves de calidad y coherencia que puede afectar en el desarrollo de la solución, se tomará una decisión de si se debe retroceder a la etapa anterior (compresión de negocio), implicando que se debe realizar nuevamente todas las fases correspondientes.

En cuanto a la cuarta etapa de la metodología (modelado), es importante que haya más de un modelo candidato para ser evaluado, los cuales no contengan errores de codificación del algoritmo y en las obtenciones de resultados durante la fase de generación de pruebas de modelos. Para este caso, si se ha seleccionado solo un modelo o no existe ningún candidato, se tomará la decisión de retroceder a la tercera etapa, que es preparación de los datos, implicando así repetir todas las fases.

Para la última etapa de metodología (evaluación de proyecto), es importante realizar evaluaciones de manera profunda, incluyendo la revisión de los pasos que se han ejecutado para construcción de los modelos generados y comparación entre ellos. Además, se construye un prototipo con el fin de obtener resultados adicionales, y que aporte a la investigación. Para este caso, si las evaluaciones no fueran positivas y/o los modelos seleccionados no se acercan a los objetivos planteados, se determinará una lista de acciones para corregir errores y detalles que surgieron en el proyecto. Una vez concluido el ciclo, se vuelve a la etapa inicial (comprensión de negocio), con la gran diferencia que todos los procesos serán mucho más rápidos por la experiencia adquirida, ya que se afinarán detalles, resultados y conclusiones sobre esta investigación.

### **3.2. Herramientas**

Las herramientas que se mencionarán a continuación se pueden trabajar con cualquier sistema operativo, que tenga compatibilidad con los recursos para así trabajar en el desarrollo del proyecto. A continuación, se detallarán lenguajes de programación, paquetes, herramientas y frameworks que se utilizan durante el transcurso del proyecto:

- **Python:** Es un lenguaje de programación multiparadigma y con amplias librerías, de escritura rápida, escalable, robusta y de código abierto, que permite construir aplicaciones web, analizar datos (caso principal de este proyecto), automatizar operaciones y crear aplicaciones empresariales fiables y escalables. La versión que se debe utilizar en este lenguaje es 3.8.2, ya que presenta mayor estabilidad para soportar las bibliotecas TensorFlow y Keras.
- **Jupyter Notebook:** Es una aplicación que está enfocada al cliente y servidor para computación interactiva con lenguajes de programación, lo cual se caracteriza por la creación y compartición de documentos en formato JSON, los que siguen un esquema versionado y poseen una lista de manera ordenada de entradas y salidas. Generalmente, Jupyter Notebook es usado como un “*colador de datos*”, el cual distingue los datos que son o no importantes, también para modelamiento estadístico con el fin de estimar la probabilidad estadística, visualización de datos, entre otros.
- **Google Colaboratory:** Como alternativa de Jupyter Notebook, es un entorno de ejecución y trabajo interactivo que permite ejecutar y programar códigos de lenguaje Python de manera online, entregando acceso gratuito a GPUs y que se pueda compartir contenidos de forma sencilla. Para este caso, es esencial que el computador que se utilizará tenga acceso a internet.
- **Pandas:** Es un paquete de Python que permite la creación y manipulación de Dataframe. Estos últimos corresponde a una clase especial de objetos, que representa conjuntos de datos organizados en filas y columnas, con la capacidad de almacenar distintos tipos de datos, para ser analizados estadísticamente.

- **Sklearn:** Es un paquete que tiene la capacidad de estructurar los sistemas de análisis datos y modelado estadístico, ya que cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. Cabe destacar que este paquete se integra con otras bibliotecas de Python (matplotlib, plotly, scipy, entre otros).
- **TensorFlow:** Es un paquete para computación numérica, que utiliza gráficos de flujo de datos. Lo más destacable de esta biblioteca de software libre es que proporciona una infraestructura completa para trabajar con el aprendizaje profundo, ya que así se detectan patrones complejos en los datos sobre los sistemas para tomar buenas decisiones. Además, de ser la librería base de RNA
- **Keras:** Es una biblioteca de Python que ayuda a la creación de redes neuronales. En comparación a otros frameworks, este software es de código abierto y no se utiliza para operaciones sencillas de bajo nivel, ya que ocupan las bibliotecas de frameworks de aprendizaje automático. Las capas de la red neuronal que generalmente requieren configurarse se relacionan entre sí en base al principio modular, sin que el usuario de Keras tenga que comprender o controlar directamente el Backend del framework escogido.
- **Microsoft Excel 365:** Es una herramienta de hojas de cálculo, desarrollada por Microsoft, que permite realizar operaciones matemáticas mediante funciones, aplicación de fórmulas, generación de gráficos y transformación de datos.
- **QT Designer:** Por último, esta herramienta permite crear interfaz gráfica de usuario con widgets del marco GUI. Este se caracteriza por su uso multiplataforma y utiliza los recursos del sistema para dibujar ventanas, controles, botones, entre otros. Por lo que, su aplicación tendrá una apariencia nativa.

# Capítulo 4.

# Comprensión de Negocio

En el presente capítulo, se describe todo relacionado con la comprensión de negocio respecto al tema del proyecto, donde se detallan el contexto de la problemática, con relación a la baja tasa de contactabilidad, su solución propuesta, objetivos, alcances y limitaciones del proyecto, y por último un planteamiento de una hipótesis.

## 4.1. Definición del Problema

Las empresas que realizan encuestas para obtener informaciones relevantes de los contactos, utilizando la técnica de CAWI, busca que las personas respondan la mayor cantidad de información sin necesidad de insistir. La falta de respuestas de las personas hace que sea un problema permanente, donde se tiene una baja tasa de contactabilidad. De acuerdo con la empresa “Activa Research”, las principales razones de que **no respondan** las encuestas son las siguientes:

- 1) Las personas han recibido las encuestas, pero no se encuentran disponibles en el horario en que se las enviaron.
- 2) Las personas están conscientes de haber recibido la encuesta, sin embargo, están desmotivadas, no desean responderla o se les han olvidado al postergarlo.

También puede pasar que los contactados respondan a las encuestas, pero sin la suficiente honestidad o calidad de información, y esto ocurre por la desmotivación que se ha mencionado en el segundo punto. En virtud, a esta información se consideran que las posibles consecuencias que trae este problema son las siguientes:

- **Costo adicional:** Por cada operación que los encuestadores realicen (encuestas) hacia los contactos, implica un costo, por lo que, si un contacto no responde a ellas, seguirá recibiendo encuestas con el fin de insistir hasta ser respondida. Estos pueden provocar una ineficiencia en el área de recursos humanos y falta de comunicación en los contactos.
- **Pérdida de tiempo:** Al no lograr que los contactos respondan encuestas solicitadas, los encuestadores tienen el deber de seguir enviando estos, con las mismas preguntas, hasta ser respondidas. Por lo tanto, este problema genera una pérdida de tiempo por operaciones ejecutadas sin éxito.
- **Falta de retroalimentación:** Lo ideal en la operación es que las encuestas logren ser enviadas y receptadas a la mayor parte del universo de los contactos, teniendo así una mayor eficiencia en la comunicación. Sin embargo, hay casos que las encuestas mediante vía correos electrónicos fueran recibidas como SPAM, ya que los sistemas lo clasifican como llamadas y correos no deseados, donde lo identifican con fines comerciales y asociados a esta reiteración en intentos de contactos. Por este motivo, las personas no se encuentran al tanto de haber recibido las encuestas por medio de llamadas y correos, considerando que una empresa necesita recopilar toda información que contenga una variedad de opiniones de sus contactos.

- **Molestias de los contactos:** Al enviar las encuestas en reiteradas ocasiones, esto puede ser tomado como un **acto insistente** lo que provoca incomodidad, desagrado y enojo en los contactados, por lo que es menos probable que estas personas las respondan. Debido a no tener un horario definido, los encuestadores proceden a realizar encuestas en horarios donde los encuestados no se encuentran disponibles.

Para el caso de la encuesta aplicada con la técnica de CAWI, sucede que, al recibir muchos formularios de preguntas, las personas suelen molestar al recibir esa gran cantidad, provocando catalogar estos mensajes como SPAM (Correo no deseado) o que el sistema de correos electrónicos lo haga por el usuario, por esta razón, difícilmente las personas podrían estar al tanto de la situación.

## 4.2. Solución Propuesta

La solución de la propuesta presentada está orientada a mejorar la eficiencia de operaciones en los encuestadores, lo cual implica aumentar la tasa de respuestas y disminuir la cantidad de intentos de contactos. En consecuencia, esto evitaría la pérdida de tiempo, costo adicional, y actos de recurrencia de llamadas a los encuestados; que puede afectar en el estado de ánimo del contacto, además de repercutir en los sistemas telefónicos y correos electrónicos que detecten y clasifiquen las encuestas como SPAM. Asimismo, se espera adquirir un mayor conocimiento respecto a los momentos que son más probables que las personas respondan las encuestas, por lo que se propone ajustar un modelo matemático que permita estimar la probabilidad de que una persona responda una encuesta, mediante correos electrónicos, en un determinado horario.

Para el modelamiento de la solución, se utilizarán algoritmos mediante Redes Neuronales, que genere modelos predictivos con capacidad de predecir probabilidades (véase a ilustración [2-2]). Cabe destacar que los datos que se utilizan son registros sobre envíos de encuestas de CAWI del año 2020, proporcionados por la empresa “*Activa Research*”, una vez analizado estos archivos, cada uno contiene columnas y atributos (filas), donde posteriormente se realiza una limpieza y se arma uno o más conjuntos de datos muestrales.

Por último, se genera varios modelos utilizando las distintas arquitecturas de redes neuronales con los parámetros configurados, para después evaluar los resultados y elegir candidatos que ajustan mejor a los objetivos considerando su precisión. Al identificar los candidatos en los modelos de redes neuronales, estos se llevan a cabo mediante validaciones, donde consiste comparar las cantidades reales de encuestas que son respondidas y otros que son predichas por una red neuronal artificial, todos estos a partir de una interfaz gráfica y matriz de confusión.

### **4.3. Objetivos**

En la presente sesión se aborda el objetivo general, que hace referencia a la idea principal del proyecto y su finalidad, junto a los objetivos específicos, los cuales indican en detalle, los pasos a seguir para que este objetivo principal se cumpla.

#### **4.3.1. Objetivo general:**

Desarrollar modelos predictivos de contactabilidad, con capacidad de predicción de horarios efectivos para responder encuestas, a través de la técnica de CAWI, con el objeto de mejorar la eficiencia de operaciones mediante el aumento de la tasa de contactabilidad.

#### **4.3.2. Objetivos específicos:**

- Analizar la problemática y situación que se encuentra el contexto del negocio respecto las encuestas con técnicas de CAWI.
- Analizar los datos recopilados a partir de los registros sobre envíos de las encuestas mediante técnicas de CAWI.
- Construir conjuntos de datos que serán utilizados para el modelamiento del proyecto.
- Construir modelos de predicción aplicando las técnicas de redes neuronales artificiales.
- Evaluar los modelos predictivos, verificando sus precisiones, errores y valores perdidos.
- Desplegar resultados del modelo matemático mediante interfaz gráfica, obteniendo así la conclusión del proyecto.

#### **4.4. Alcances y Limitaciones**

En esta sesión se definen los alcances del proyecto, con la finalidad de analizar la llegada de trabajo, esto requiere todo lo necesario para cumplir con los objetivos planteados anteriormente, y sus respectivas limitaciones, indicando aspectos que quedan fuera de la cobertura de dicho proyecto.

##### **4.4.1. Alcances:**

- Los modelos predictivos tendrán capacidad de predecir si las personas contestarán las encuestas o no, en un determinado horario y mediante técnicas de CAWI.

- Los datasets que se utilizarán son recopiladas a partir de los datos cuyo origen proviene de registros de encuestas de los clientes, independiente si son contestadas o no, conseguidas por la empresa “*Activa Research*”.
- Este proyecto puede ser un complemento para otros proyectos, ya que éste tiene un fin de entregar conocimientos respecto a los horarios óptimos donde las personas si están disponibles y contesten las encuestas, mejorando así la eficiencia.

#### **4.4.2. Limitaciones:**

- El proyecto no abarca escenarios sobre otras técnicas que se usan en las encuestas, como es el caso de CAPI (Computer-Assisted Personal Interviewing), CATI (Computer-Assisted Telephone Interviewing) e IVR (Interactive Voice Response).
- Los datos no son de uso público, por lo que estos son anonimizados debido a temas de confidencialidad en la empresa.
- Los modelos generados estarán ajustados para un determinado contexto y público objetivo (cliente), por lo que estos no serán generalizados para otros.
- No es alcance del proyecto de investigación implementar el modelo predictivo en la organización.

#### **4.5. Hipótesis**

*“Existe un modelo que permite predecir si una persona responderá una encuesta o no, con un 90% de precisión y en un determinado horario, mediante técnica de CAWI.”*

# Capítulo 5.

## Comprensión de los Datos

En este capítulo, se detallan todo relacionado con los datasets iniciales del proyecto, donde se realiza una previa revisión de los datos correspondientes al historial de envíos de Sendinblue, base de datos de clientes y de respuesta. Además, se seleccionan los campos que contiene cada de estos universos, los cuales son relevantes para ser llevado a cabo en preparación de los datos, y se detallan lo que consiste en cada uno de ellos.

### 5.1. Descripción del Dataset

Para abordar esta problemática, en relación con las encuestas CAWI, se tienen registros de estas obtenidas por la empresa "Activa Research". Dentro de estos registros cuenta una base de datos del cliente, donde en ella se registran las definiciones de variables, datasets (Archivos .CSV) construidas mediante "Sendinblue" y otros almacenadas por la misma empresa como "*Universo de cliente*". Además, se cuenta con un conjunto de datos almacenados por SPSS Statistics (Archivo .SAV), donde en ella se registran historial de respuestas de los clientes que hayan contestado algunas encuestas.

En universo de cliente, se almacena el universo de datos personales de los clientes de la empresa y sus datasets están divididas por trimestre. En cuanto a Sendinblue, se encuentran datos respecto a los correos destinatarios, fechas, horas, etiquetas y, la más importante, estados de cada envío que se encuentran. Asimismo, la complejidad de trabajar con estos datasets es debido a su gran cantidad de registros y campos. Por otra parte, existen diversos CSV por cada mes con versiones distintas, donde se hace necesario compactarlos en uno solo.

### **5.1.1. Sendinblue:**

Sentinblue, como afirma Ordoñez (2021), es una plataforma de marketing enfocada a correos electrónicos, con el fin de mejorar las relaciones entre clientes y empresa. Esta plataforma tiene la función de enviar correos electrónicos, mensajes de texto, generar respuestas automáticas, limpiar la base de datos de los clientes entre otras más. Para este caso, Sendinblue trabaja con los envíos que se dirigen directamente a un destinatario en concreto y que se activan mediante la realización de sus acciones, donde en el apartado “Log” se visualizan y recaban datos respecto los estados de estos envíos. Por lo tanto, como detalla en el sitio web de Sendinblue (s. f.), cada correo se compone de distintos estados característicos dependiendo de la situación individual de cada uno, estos son los estados que proporciona la plataforma:

- 1. Enviado:** Indica que el email ha sido enviado al destinatario.
- 2. Entregado:** Indica que el e-mail ha sido recibido.
- 3. Primera apertura:** Indica que el destinatario ha abierto el e-mail por primera vez.
- 4. Abierto:** Indica que el destinatario ha abierto el e-mail.

- 5. Rebote suave:** Indica que el e-mail no ha sido entregado, en este caso cuando el servidor del destinatario no está disponible o su buzón de entradas está lleno.
- 6. Aplazado:** Indica que se abrió el e-mail, pero no se respondió la encuesta, provocando que la respuesta de la encuesta se aplace.
- 7. Bloqueado:** Indica que la dirección de correo electrónico del cliente está incluida en una lista negra, no se ha podido enviar el e-mail a su destinatario.
- 8. Rebote duro:** Indica que el e-mail no ha sido entregado, en este caso debido a que el correo electrónico no existe o está bloqueado.
- 9. E-mails no válidos:** Indica que el e-mail no existe o está mal escrito
- 10. Queja:** Indica que el destinatario ha marcado el e-mail como correo no deseado o SPAM.
- 11. Cliqueado:** Indica que el número de clics que se ha hecho en alguno de los e-mails.
- 12. Suscripción cancelada:** Indica que la suscripción del cliente se encuentra cancelada.

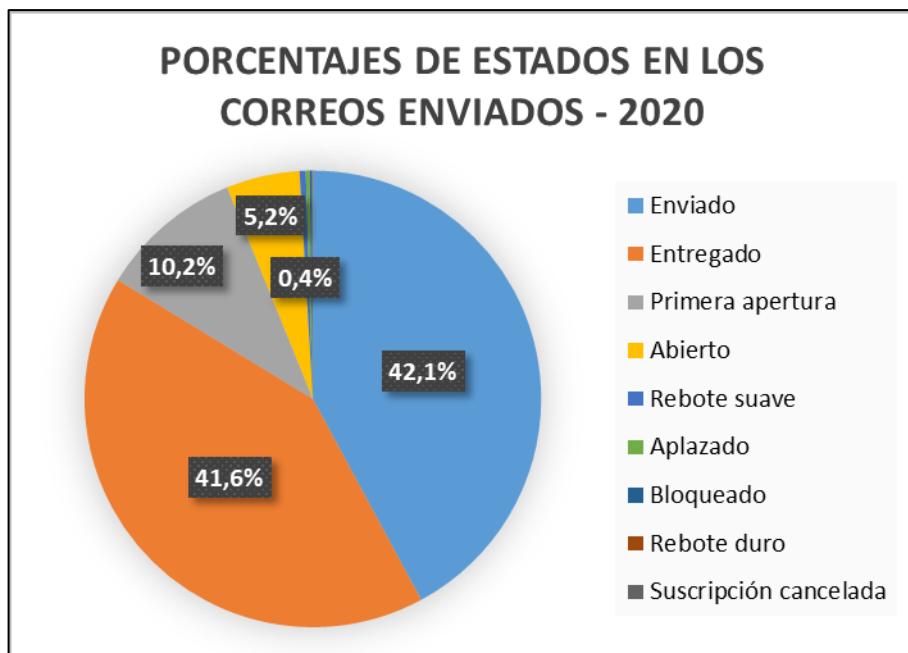
En los datasets correspondiente al SendinBlue, para el año 2020, se registraron una cantidad total de 25.865.946 estados a sus respectivos destinatarios, donde los estados “Enviado” y “Entregado” llevan porcentajes más alto que los otros. A partir de la tabla [5-1] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados, y en la ilustración [5-1] se logra visualizar un gráfico de torta donde indica los porcentajes que contiene en los estados. Además, para mayores detalles respecto estas cantidades de envíos y los estados que se encontraban, véase al anexo 10.1.

Tabla 5-1: Cantidad de encuestas enviadas con su respectivo estado 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
<b>Enviado</b>	10.898.670	42,14%
<b>Entregado</b>	10.755.598	41,58%
<b>Primera apertura</b>	2.626.666	10,15%
<b>Abierto</b>	1.344.721	5,20%
<b>Rebote suave</b>	112.812	0,44%
<b>Aplazado</b>	68.946	0,27%
<b>Bloqueado</b>	41.313	0,16%
<b>Rebote duro</b>	10.146	0,04%
<b>Suscripción cancelada</b>	5.096	0,02%
<b>E-mails no válidos</b>	1.037	< 0,1%
<b>Queja</b>	931	< 0,1%
<b>Clicado</b>	10	< 0,1%
<b>TOTAL</b>	25.865.946	

Fuente: Elaboración propia

Ilustración 5-1: Porcentajes de estados en los correos enviados 2020



Fuente: Elaboración propia

### **5.1.2. Base de datos de respuestas:**

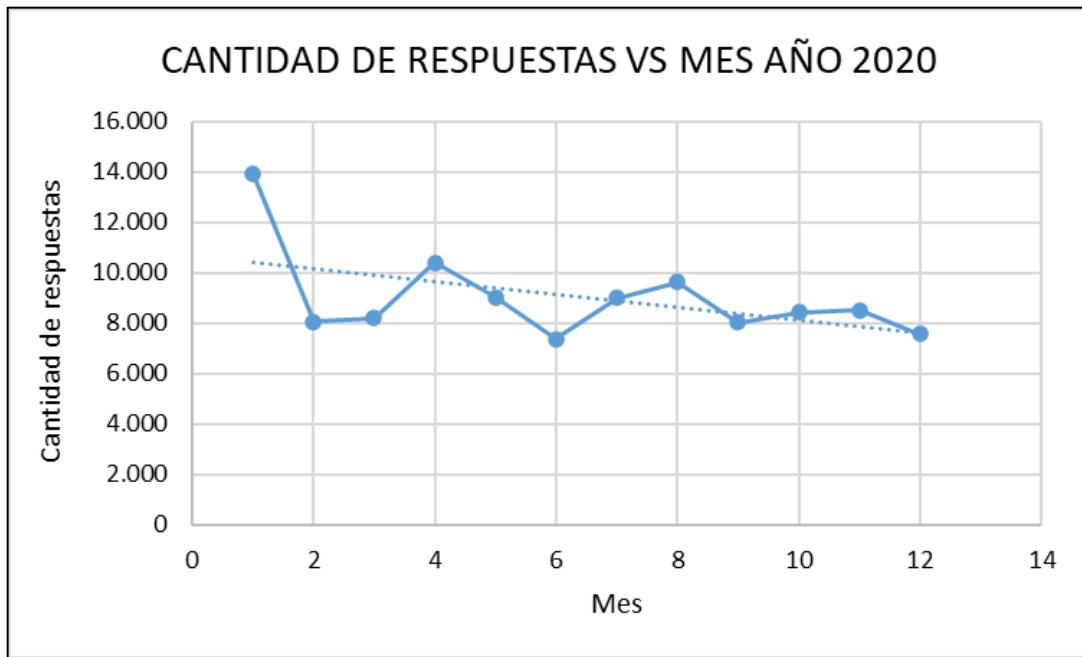
La empresa "Activa Research" ha proporcionado un historial de respuestas, almacenadas en una base de datos de la organización, donde cuenta con un total de 108.149 respuestas en el año 2020. Cabe mencionar que el mes de Enero ha recibido un porcentaje más alto de respuestas del año, con 12,9%, mientras que el otro mes de Junio tiene una cantidad más baja de dicho periodo, con otro porcentaje de 6,8%. A partir de la tabla [5-2] se observa la cantidad y frecuencia de encuestas que son respondidas, y en la ilustración [5-2] se logra visualizar un gráfico de cantidad de respuestas vs mes, donde indica un decrecimiento con una diferencia de 6.346 respuestas, aproximadamente 5,9% respecto al total, entre enero y diciembre.

*Tabla 5-2: Cantidad de encuestas respondidas desde la BD de respuesta 2020*

MES	CANTIDAD (RESPUESTAS)	FRECUENCIA (%)
1	13.910	12,9%
2	8.055	7,4%
3	8.186	7,6%
4	10.404	9,6%
5	9.040	8,4%
6	7.379	6,8%
7	8.998	8,3%
8	9.637	8,9%
9	8.027	7,4%
10	8.438	7,8%
11	8.511	7,9%
12	7.564	7,0%
<b>TOTAL</b>	<b>108.149</b>	

*Fuente: Elaboración propia*

*Ilustración 5-2: Cantidad de respuestas vs mes, encuestas del año 2020*



*Fuente: Elaboración propia*

## 5.2. Descripción de Variables

Los archivos de tipo CSV, sin duda, contiene un gran número de campos o atributos (columnas) que son valores únicos lo cual proporcionan una estructura a los registros (filas) de cada datasets. Además, los datos de cada campo pueden ser de diferentes tipos, donde pueden ser numéricos, códigos con significado, decimales, lógicos (verdadero “1” o falso “0”), textos, etc. Para este caso no se consideran informaciones personales de los clientes, ya que la construcción de datasets de pruebas incluirán datos de forma anónima, como es el nombre completo, RUT, área donde trabaja, entre otros, y también todas las segmentaciones (categorías) se reemplaza por letras y números.

Sin embargo, es necesario utilizarlos, principalmente correo electrónico y teléfonos, para generar datos numéricos, edad, género y segmentación, lo cual se incluirá en dicha construcción del conjunto de pruebas, esto permitirá identificar el sexo, edad y segmentaciones en los registros. Para cada universo, en los siguientes apartados, se mencionan campos que son utilizados para la generación de datos y construcción de datasets de pruebas para CAWI.

❖ **Universo Cliente**

1. **E-mail:** Corresponde a correos electrónicos de los encuestados
2. **Edad:** Como menciona su nombre, corresponde a la edad de una persona lo cual implica el tiempo que ha vivido contando desde su nacimiento.
3. **Segmentos:** Corresponde al nombre del segmento que una persona pertenece, donde representa como categoría y existe un total de 5, los cuales son A, B, C, D y E.
4. **Subsegmentos:** Corresponde al nombre del subsegmento que una persona pertenece dentro de un segmento, donde representa como subcategoría y existe un total de 31, los cuales son E1A, E1B, E2A, E2B, E3, E4, E5, C1, C2, C3, B1, B2, B3A, B3B, B4A, B4B, B5, B6, B7, D1A, D1B, D2A, D2B, D2C, D3, A1, A2A, A2B, A3, A4 y A5. Cabe destacar que, en cada subsegmento, la primera letra o carácter representa el segmento que se encuentra adentro.
5. **Segmentos agrupados:** Corresponde al nombre de agrupación con varios segmentos entre personas, donde existe un total de 5 los cuales son E1S, E2S, D1S, A15 y C1S.

**6. Carterizado:** Este campo, según Bernués (2021), corresponde a una clasificación de los clientes con relación a sus volúmenes de facturación, rentabilidad, potencial de crecimiento y/o nivel de vinculación. Por lo que, si una persona se encuentra registrado con una cantidad mayor de ventas en el negocio y se encuentra sobre del promedio, esta persona es denominado como cliente carterizado.

❖ **Sendinblue**

- 1. Estado:** Como se menciona su nombre, corresponde al estado del envío de la encuesta que se encontraba durante su periodo, los cuales son mencionadas anteriormente en la tabla [5-1].
- 2. Fecha:** Corresponde a un tiempo determinado en que una encuesta se ha enviado, lo cual tiene formato de “DIA-MES-AÑO HH:MM:SS AM/PM”.
- 3. Para (destinatario):** Corresponde al correo del contacto quien responde las encuestas.
- 4. Id del mensaje (Id de envío):** Corresponde al código de identificación de los envíos
- 5. Etiquetas:** Como apoyo en el momento de procesar datos, corresponde a una información lo cual orienta el sexo y agrupación que identifica al destinatario.

❖ **Base de datos de respuestas**

1. **E-mail:** Como se menciona su nombre, corresponde al correo de una persona quien respondió una encuesta.
2. **Fecha de respuesta (FECHAFIN):** Corresponde a la fecha que un cliente ha respondido una encuesta, donde se toma en cuenta solamente el periodo del año 2020.
3. **Fuente:** Corresponde a la técnica de encuesta que se ha utilizado para recabar información, donde se considera solamente CAWI.

Tras una revisión de los datos mediante “*Jupyter Notebook*”, en la fuente de datos proveniente de “*Sendinblue*” y la base de datos de respuesta, no se presentan datos vacíos, valores ni formato erróneos en “Estado”, “Fecha”, “Destinatarios (para)” y “Etiquetas”. Sin embargo, los textos de correos electrónicos presentan diferencias debido el uso de mayúsculas y minúsculas, y también la presencia de los espacios vacíos en ambos extremos, por lo que es recomendable normalizar los caracteres y eliminar dichos espacios. Por otro lado, existe la opción de realizar limpieza y transformaciones de datos solamente en “Etiquetas”, ya que en los textos de cada fila contiene datos correspondientes al sexo y agrupación de segmentos. Pero, en algunos casos, no indica estos dichos datos, por lo que es más conveniente extraer desde el universo de cliente.

Por otro lado, al referirse al universo de clientes, en los campos edad, sexo, segmentos, subsegmentos y segmentados no agrupados no se encontraron errores de valores y formatos en los datasets. Sin embargo, en algunos registros de correos desde dicho universo presentan datos vacíos, nulos e indeterminados, por lo que puede existir dificultades en la búsqueda, identificación y extracción de otros campos requeridos. Para el caso de los datos outliers, pueden ocurrir que exista una mayor diferencia entre distribución de las edades y que, en algunos casos, presentan un valor atípico, por lo que afectará a la precisión y coherencia de los resultados. En cuanto a las variables categóricas, como son las segmentaciones, sexo, carterizado, agrupaciones y estados, puede ocurrir que exista una gran diferencia entre frecuencias lo cual tiene un gran impacto a los resultados.

Por otra parte, es importante destacar que la empresa “*Activa Research*” envían encuestas por medio de correos electrónicos con un horario entre 8:00 a 21:00 horas. Asimismo, si esto presenta un valor fuera del rango implica la existencia de un delay (retraso) y/o la mala configuración de horario en los sistemas. Por último, al momento de combinar dos archivos de distintas fuentes, entre Sendinblue y Universo del cliente, se prioriza aquel conjunto que tenga mayor población de datos, respecto a los meses registrados.

# Capítulo 6.

# Preparación de los Datos

En este capítulo, se describen los procesos de construcción y estructuración del conjunto de datos, utilizando los archivos iniciales con relación al historial de envíos de encuestas, universo de cliente y de respuestas. Además, se detalla todo relacionado con análisis descriptivo mediante estadísticas, detección y tratamiento de outliers, análisis factorial, y definición de variables independientes y objetivos del conjunto de datos construido.

## 6.1. Construcción y estructuración del conjunto de datos de prueba

Para la construcción del dataset, se manipulan los datos mediante Jupyter Notebook, cuyos datos de entrada provienen de archivos de base de datos de clientes, Sendinblue y base de datos de respuesta. Por lo que, como se visualiza en la ilustración [6-1], en Sendinblue se procesan datos correspondientes a las columnas “E-mail” (Para), “Fecha”, “Estado” y “Id de envío”. En cuanto a Universo de cliente, se utilizarán otros datos de columnas “E-mail”, “Edad”, “Segmento”, “Subsegmento”, “Agrupación de segmentos” y “Carterizado”. Por último, para la base de datos de respuesta, solo se utilizarán “Fecha”, “E-mail” y “Fuente”. Todos estos datos serán utilizados en el proceso de estructuración de datos, entregando como resultado un conjunto de datos que contiene los siguientes campos:

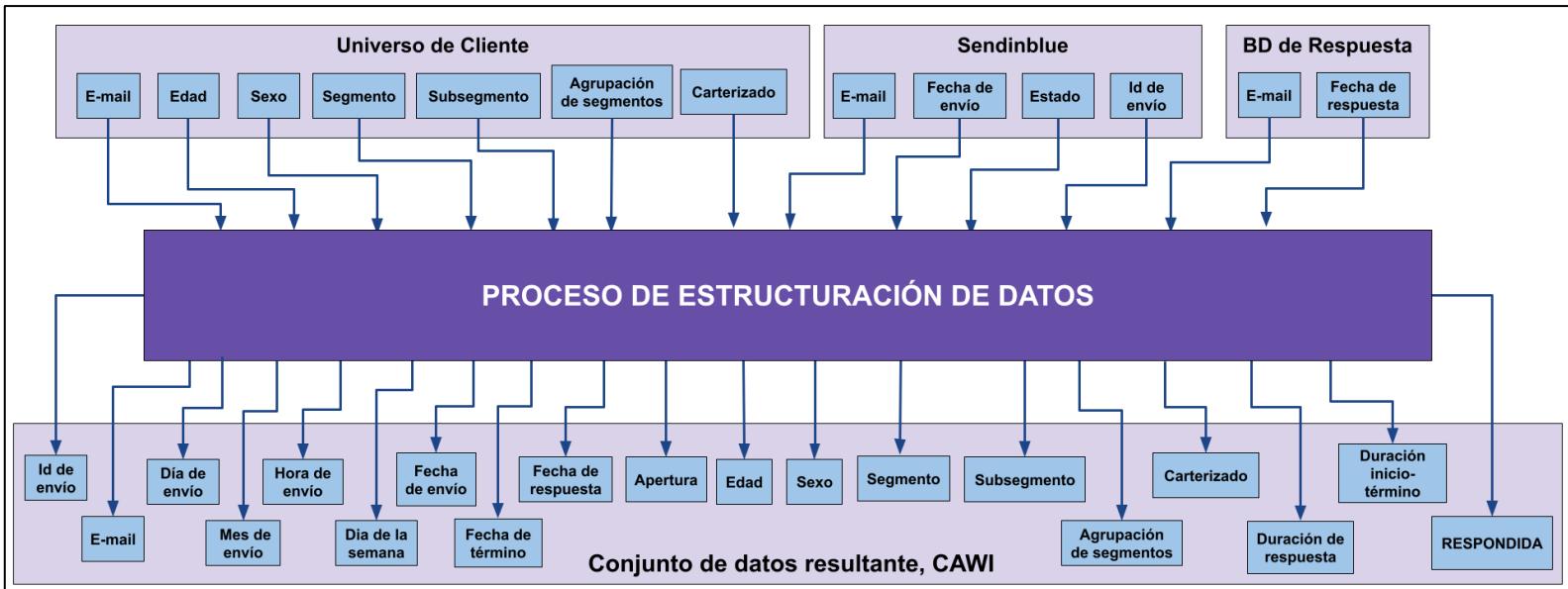
- **Código de envío (Id\_Envio):** Corresponde al código de identificación único de cada envío realizado.

- **E-mail:** Corresponde al correo electrónico del receptor, cuyo valor resultante será anonimizado por números.
- **Día de Envío:** Corresponde el día en que se ha efectuado el envío de la encuesta y que el destinatario logre recibirla.
- **Mes de Envío:** Corresponde al mes en que se ha efectuado el envío de la encuesta.
- **Hora de Envío:** Es un campo relacionado con la hora en que el correo destinatario logre recibir un envío de forma exitosa.
- **Dia de la semana:** Corresponde al día de la semana (lunes a domingo) donde se ha efectuado el envío.
- **Fecha de envío (Fecha):** Es un campo relacionado con la fecha que se ha enviado una encuesta, concretando con otros campos como son la hora, día y mes de envío. Esto se utiliza para comparación, aplicación de cálculos y verificación de lógica con otras fechas, y en su posterior se elimina antes del inicio del modelamiento de redes neuronales.
- **Fecha de Término:** Este campo corresponde a la fecha límite que la encuesta pueda ser respondida, donde este representa el próximo envío o 01/01/2021 en caso de no existir más registro. Esta columna, al igual que la fecha de envío, se elimina una vez que se concluya el proceso de estructuración de datos
- **Fecha de Respuesta (Fecha\_R):** Este campo corresponde a la fecha que se haya recibido una encuesta respondida. En caso de no existir alguna fecha se determina como S.F. (Sin Fecha). Además, como sucede con otras dos fechas mencionadas anteriormente, esta columna será eliminado una vez terminado dicho proceso.

- **Apertura:** Corresponde a la cantidad de veces que una persona ha abierto el correo, cuyos estados de Sendinblue están incluidos “Primera Apertura” y “Abierto”.
- **Edad:** Como menciona su nombre, es un campo relacionado a la edad del destinatario.
- **Sexo:** Corresponde al género lo cual distingue si la persona es hombre (H), mujer (M) o sin especificar (S).
- **Segmento:** Corresponde al nombre del segmento que un destinatario pertenece.
- **Subsegmento:** Corresponde al nombre del subsegmento que un destinatario pertenece.
- **Agrupación de Segmentos (Segto\_Agrup):** Corresponde al nombre de agrupación con varios segmentos entre destinatarios.
- **Carterizado:** Este campo, como menciona su nombre, corresponde si una persona pertenece al conjunto de carterizado o no.
- **Duración de respuesta (Duración):** Este campo corresponde a la cantidad de días que se ha demora en que una persona responda una encuesta. En caso de que la respuesta no se encuentra en el rango entre fecha de envío y de término, o nunca se ha respondido, por defecto se obtiene días transcurrido entre fecha de envío y de término. Este campo será utilizado solamente para generar estadísticas y para aquellos envíos que tiene respuesta.

- **Duración inicio-término (Duracion\_i\_f):** Este campo corresponde a la cantidad de días que se ha demora en terminar el plazo para verificar si la encuesta ha respondida o no, entre fecha de envío y otra fecha de próximo envío (o 01-01-2021 en caso no existir este último).
- **RESPONDIDA:** Por último, este campo determina si el destinatario ha respondido la encuesta o no, donde esto es obtenido mediante estimación de contactabilidad.

Ilustración 6-1: Flujos de datos y su procesamiento para construcción de datos CAWI



Fuente: Elaboración propia

### **6.1.1. Extracción de muestra y estructuración de Sendinblue:**

Dentro del proceso de estructuración de datos, como se visualiza en la ilustración [6-2], se da inicio al leer y extraer los datos de universo de cliente, y realizar la misma acción con los archivos de Sendinblue por cada mes. Por lo que, para cualquier acción en el proceso, es recomendable que todos los resultados de procesamiento de datos sean almacenados en un archivo .CSV, ya que al manipular demasiados datos requiere mayor tiempo en su ejecución. Además, se debe normalizar los textos de los correos electrónicos en todos los archivos, transformando todos los caracteres en minúsculas, y eliminar espacios que se encuentra en ambos extremos de ellos, ya que una mínima diferencia entre ellos surge problemas de igualdad en sus datos para unir otros campos.

En los datasets de Sendinblue, antes de generar estadísticas de estados mostradas en la tabla e ilustración [5-1], se procedió a unir todas las versiones de registros de envíos por cada mes y a su vez, eliminar los que se encuentran duplicados, con una condición que exista coincidencia en los campos de id de envíos, correos electrónicos del cliente, estado en que se encuentra el envío y la fecha que se registró en dicho estado. Luego, en los datasets solo se consideraron 3 estados, los cuales son: “*Entregados*”, “*Primera Apertura*” y “*Abierto*”.

En cuanto a otros estados, estos no fueron considerados debido a que proporcionan una mínima cantidad de datos, pueden presentar dificultades al momento de construir el dataset de prueba y en que el algoritmo de aprendizaje entregue resultados. Debido a esto, se procedió a eliminar las filas que contengan estados equivalentes a “*Bloqueado*”, “*Rebote duro*”, “*Rebote suave*”, “*Aplazado*”, “*Suscripción cancelada*”, “*E-mails no válidos*”, “*Queja*”, “*Cliqueado*” y “*Suscripción cancelada*”. A su vez, la diferencia del horario operación entre estados enviados y entregados en los registros son

diminutos, con un intervalo de entre 3 a 37 segundos, por lo que se procedió eliminar las filas que contengan estado “*Enviado*”, para asegurar de que se esté tomando en cuenta aquellas encuestas que fueron recepcionadas.

Luego, se utilizan los correos electrónicos de los contactos para buscar en los archivos del universo de cliente según el mes correspondiente desde Sendinblue, obteniendo así los campos de carterizado, segmento, edad, sexo, subsegmento y segmentos agrupados. Para este caso, se considera solamente archivos que tengan mayor población de datos y sus coincidencias, con el fin de extraer solo muestras asegurando todos los campos. Una vez hecho, se logran combinar en un conjunto de datos, y se eliminan columnas que no se utilizarán, los cuales son extraídos desde el Universo de cliente, y se obtiene datos muestrales de cada mes.

Tras realizar este proceso, se debe repetir los mismos pasos para otros meses correspondientes al año 2020. Cabe mencionar que es importante que todas las columnas tengan los mismos nombres de encabezado, ya que, si se presentan diferencias, en el momento de combinar todos los meses puede generar nuevos campos innecesarios. Una vez que se haya recorrido todos los meses y que todos tengan campos con los mismos nombres, se procede a unir y, en su posterior, eliminar el resto de las filas duplicadas y anonimizar los datos. Posteriormente a esto, se logra extraer datos muestrales de todos, cuya cantidad de estados de “*Entregado*”, “*Primera apertura*” y “*Abierto*” se visualiza en la tabla [6-1]. Cabe mencionar que cada cantidad representa una fila del conjunto de datos, por lo que se debe realizar cálculos y extraer las fechas claves por cada envío, lo cual se reduce el volumen del dataset en construcción.

Tabla 6-1: Extracción de muestra entre Sendinblue y BD de cliente

ESTADO	DATOS TOTALES SENDINBLUE	DATOS DE MUESTRA EXTRAIDO	PORCENTAJE DE DATOS EXTRAIDOS
Entregado	10.755.598	4.621.984	43,0%
Primera apertura	2.626.666	1.237.546	47,1%
Abierto	1.344.721	665.869	49,5%

Fuente: Elaboración propia

Obteniendo el conjunto de datos muestrales, se procedió con los cálculos de estados por envíos que se encuentran almacenados, por lo que se tomó en cuenta “Entregado”, “Primera apertura” y “Abierto”. Cabe destacar que los envíos contienen ID únicos, por lo que con esto se logra identificar las actividades que están relacionadas en cada uno de los envíos. Para este paso, se ha revisado todas las filas con relación a la cantidad de estados que contiene primera apertura y abiertos, para así generar otra columna “Primera apertura” y otra “Cantidad de abiertos”. En cuanto a la actividad de “Entregado”, donde todos los envíos deben tener uno, se registra su fecha y hora de envío respectivamente.

Continuando con lo anterior mencionado, se debe comprobar que todos los envíos tengan una actividad “Entregado”, en caso de que uno no lo contenga, se debe eliminar este envío por falta de datos de fecha y hora. Por otro lado, los envíos pueden tener o no una actividad “Primera apertura”, por lo que, si uno de ellos registra más de una de dicha actividad, entonces esa cantidad se reemplaza por 1 ya que esta actividad solo indica si ha abierto por primera vez el envío, formando así un valor booleano. Por último, si en algunos envíos contienen uno o más de una actividad correspondiente a “Abierto”, debe contener otra “Primera apertura”, en caso de no contarlo, se reemplaza por valor 1 para mantener la coherencia y preservar los datos.

Después, para el caso del campo fecha desde los datasets de Sendinblue, se procede a dividir el formato “*DIA-MES-AÑO HH:MM:SS AM/PM*” mediante Jupyter Notebook, esto genera nuevos campos de tipo entero y pero se tomarán en cuenta solamente el mes, día y horario (HH:MM:SS AM/PM) que se operó la encuesta, y en su posterior, se introducirá el conjunto de datos (dataset) de CAWI. En este caso, se excluye el año debido a que el proyecto solamente está considerando el periodo 2020, en cuanto a los minutos y segundos, estos pueden afectar la precisión y resultados de la predicción por ser muy específico o caso muy puntual.

Luego de separar dichos campos, se procedió con otra transformación de datos en la columna “*Horario*” para el formato de 24 horas (HH:MM:SS), lo cual en su posterior se extrae únicamente la hora que se ha recepcionado correctamente la encuesta. Una vez hecho lo anterior, se filtra las horas de envío que se encuentren fuera de rango del horario 8:00 a 21:00, donde esto se procede a cambiar valores según la cercanía en dichas horas, es decir, se reemplaza por 21 aquellos valores que son 22, 23, 0, 1 y 2 horas, en cuanto 3, 4, 5, 6 y 7 son sustituidos por 8 horas. Cabe mencionar que al reemplazar por 21 para el caso de 0, 1 y 2 afecta el valor de día, por lo que se debe restar por uno para estos casos específicos. Por último, se realizó una transformación de datos con relación a las fechas de envío por el día de la semana, donde esto puede indicar si se ha enviado el lunes, martes, miércoles, jueves, viernes, sábado o domingo.

Una vez procesado las horas correspondientes, se debe cambiar al formato de fecha a “AÑO-MES-DÍA”, ya que se utiliza la librería Datetime para realizar comparaciones de distintas fechas y realizar cálculos de duraciones. Por lo que, al cambiar el tipo de dato (String como origen) y no contar dicho formato, ocurre problemas de consistencia en el día y mes en la fecha. Por último, se ordena el conjunto de datos en desarrollo por E-mails y fechas de envío, de acuerdo con el orden alfabético y acontecimientos viejo a más reciente.

Continuando lo anterior, se construye un programa donde permite determinar la fecha de término de plazo de cada envío en un determinado correo, identificando el próximo envío, es decir, desde el actual envío se consulta su próxima fecha al que se ha realizado otro. Por un lado, en caso de no existir un próximo envío durante el año 2020, este se asigna para la fecha 2021-01-01, concluyendo así el ciclo de un correo y empezar con el siguiente restante. Por lo tanto, una vez que todos los correos completen sus ciclos de envíos en el año, se procede a calcular duración en días entre la fecha de envío y fecha de término (*Duracion\_i\_f*). Posteriormente a esto, se eliminan aquellas filas cuyo valor es 0, debido que se ha presentado dos envíos en un día y solamente se considera uno de ellos.

Después, para saber si la persona ha contestado la encuesta, es necesario considerar los registros que se encuentran en la base de datos de respuesta para el año 2020. Para este caso, se requiere solamente campos de fuente, E-mail y fecha que se ha emitido la encuesta, lo cual es importante considerar que los datos estén normalizados y con formato correcto. Por lo tanto, en la base de datos de respuesta, se filtra todas las filas que pertenezca a “CAWI” en el campo fuente, se ha modificado el formato de fecha de respuesta por “AÑO-MES-DIA” y se ha normalizado los correos electrónicos la misma forma que se han hecho a los otros.

Para el paso final de proceso, se genera un campo importante, una variable de salida, donde este recibe el nombre de “RESPONDIDA”, cuyo valor es “Sí” (1), si en caso de que el contacto respondió, y “No” (0), que es el caso contrario del anterior. Por lo que, para estimar la contactabilidad, primero se realizó una unión de todos los envíos por coincidencias de E-mails, y luego se realizó comparaciones entre fechas de envío, de respuesta y de término para determinar su valor, definiendo así las siguientes lógicas:

- 1) Si la fecha de respuesta se encuentra entre la fecha de envío y de término, y, a la vez, es diferente a la fecha de término, entonces este envío si ha respondido teniendo un valor igual a “Sí”.
- 2) Si la fecha de respuesta coincide con la de envío, entonces la persona si ha respondido esta encuesta, ya que menciona que ha respondido al mismo día que se efectuó el envío y recepción en su correo electrónico.
- 3) Si la fecha de respuesta coincide con la de envío, entonces la persona si ha respondido esta encuesta, ya que menciona que se ha efectuado la respuesta al mismo día que se realizó el envío y recepción en su correo electrónico.

- 4) Si la fecha de respuesta coincide con la fecha de término, implica que terminó el plazo para determinar si ha respondido una encuesta y se tomara en consideración para el siguiente envío, cumpliendo en el tercer punto anteriormente mencionado. Por lo tanto, para este envío se asigna con un valor “No” a la columna “RESPONDIDA”.
- 5) Por último, en caso de que la fecha de respuesta se encuentra fuera del rango entre la fecha de envío y de término, o que el correo electrónico no coincide con el registro de base de datos de respuesta 2020, generando fecha de respuesta sin S. F., implica que definitivamente no hay respuesta. Por lo tanto, se asigna un valor de “No” para estos casos mencionados.
- 6) Resumiendo todo lo anterior con expresiones matemáticas, la lógica que se aplica si hay respuesta en los envíos son las siguientes:

*“Si (Fecha de envío) ≥ (Fecha de respuesta) < (Fecha de término), entonces RESPONDIDA de dicha fila es igual 1, en caso contrario, se asigna un valor igual a 0.”*

Tras al realizar una unión y haber aplicado la lógica, para filas que si tienen respuestas (“RESPONDIDA” = “Sí”), se procede a calcular los días que se han transcurrido respecto a cuanto se demoró una persona en responder la encuesta. En caso de que se encuentre fuera del rango o no se haya registrado la fecha de respuesta, por defecto se considera solamente la diferencia entre fecha de envío y de término (mismos valores de “Duración\_i\_f”) para no generar valores nulos.

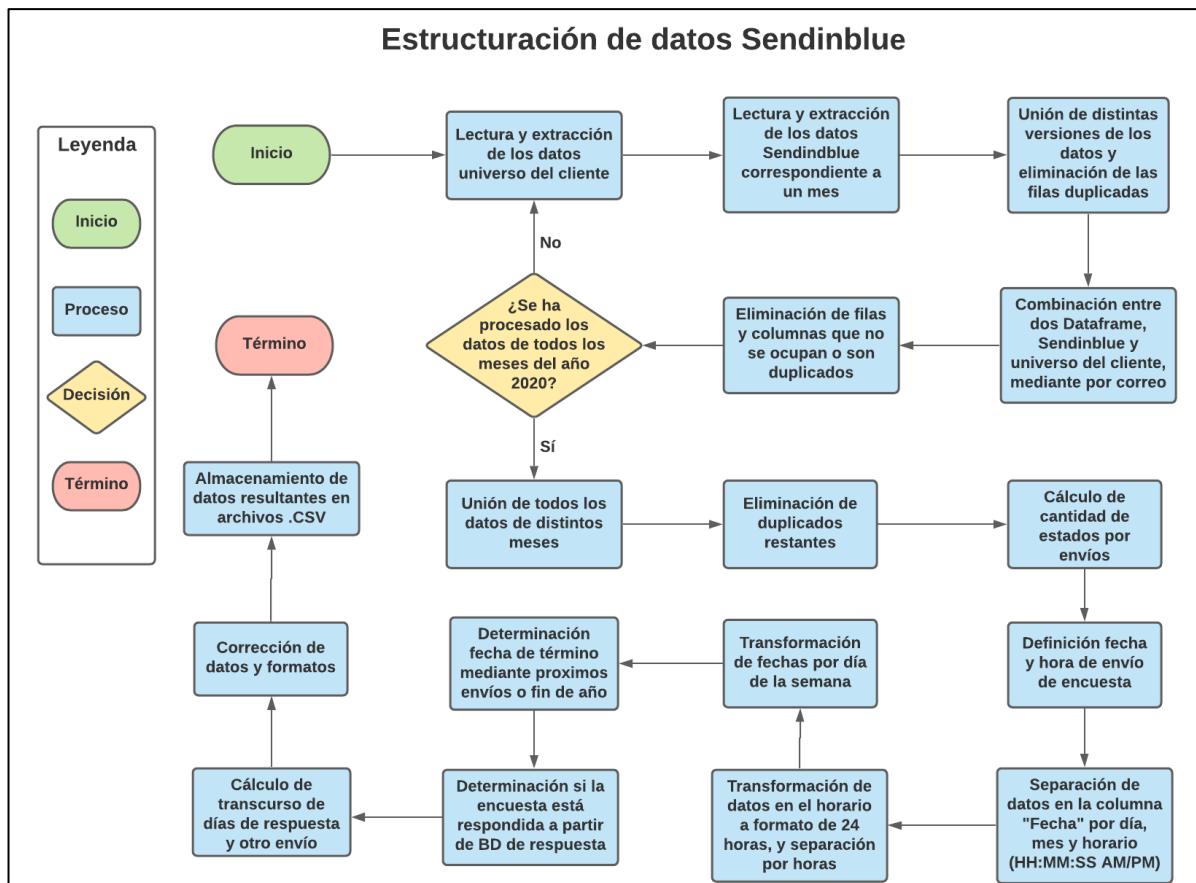
Cabe mencionar que hay personas que han respondido más de una vez para una misma encuesta, como efecto de esto, se genera filas nuevas y casi duplicadas, con fechas y duraciones de respuesta distintas unidas en caso de existir más de un valor. Por lo que, se debe agrupar y fusionar los envíos casi duplicados por fecha, duración de respuesta, y duración entre fecha de envío y de termino, donde cada valor distinto de dichas columnas se une junto con separador “;”. Una vez hecho esto, de acuerdo con la agrupación en columna RESPONDIDA, se debe corregir las consistencias y formato, manteniendo los envíos con respuestas registradas, lo cual se menciona a partir de los siguientes criterios:

- 1) Se mantendrá todos los valores aquellas filas que solo contenga un valor en las columnas de fecha de respuesta, “*Duración\_i\_f*” y “*Duracion*”.
- 2) En caso de que algunas filas se registran dos respuestas, y ambas indica que si ha respondido el envío de acuerdo a la columna “RESPONDIDA” (“Sí;Sí”), entonces esos valores se conservan para ser separado.
- 3) En caso de que algunas filas se registran dos respuestas, y ambas indica que no ha respondido el envío de acuerdo a la columna “RESPONDIDA” (“No;No”), entonces se debe eliminar cualquiera de los dos valores en base a una posición (0 o 1) para columnas “RESPONDIDA”, fecha de respuesta”, “Duracion”, “*Duración\_i\_f*”.
- 4) En caso de que algunas filas se registran dos respuestas, y ambas presentan resultados distintos de acuerdo a la columna “RESPONDIDA” (“Sí;No” o “No;Sí”), entonces se debe eliminar valores de columnas “RESPONDIDA”, fecha de respuesta, “Duración” y “*Duración\_i\_f*”, de acuerdo a la posición que se encuentra el valor “No”. Como resultado, solo tendrá valores que corresponde aquellos que, si contiene respuesta.

- 5) En caso de que algunas filas se registran dos respuestas, y ambas presentan resultados distintos de acuerdo a la columna “RESPONDIDA” (“Sí;No” o “No;Sí”), entonces se debe eliminar valores de columnas “RESPONDIDA”, fecha de respuesta, “Duración” y “Duración\_i\_f”, de acuerdo a la posición que se encuentra el valor “No”. Como resultado, solo tendrá valores que corresponde aquellos que, si contiene respuesta.
- 6) En caso de que algunas filas se registran tres respuestas, y todas indica que si ha respondido el envío de acuerdo a la columna “RESPONDIDA” (“Sí;Sí;Sí”), entonces esos valores se conservan para ser separado.
- 7) En caso de que algunas filas se registran tres respuestas, y todas indica que no ha respondido el envío de acuerdo a la columna “RESPONDIDA” (“No;No;No”), entonces se debe eliminar dos de los tres valores en base a sus posiciones para columnas “RESPONDIDA”, fecha de respuesta”, “Duracion”, “Duración\_i\_f”, conservando datos de uno de ellos.
- 8) Por último, para en caso de que algunas filas se registran tres respuestas, y uno o dos de ellos indica que si ha respondido el envío de acuerdo a la columna “RESPONDIDA” (“Sí;No;Sí”, “No;No;Sí”, “Sí;No;No”, “No;Si;No”, “Si;Si;No”, “No;Sí;Sí”), entonces se debe eliminar todos valores de columnas “RESPONDIDA”, fecha de respuesta, “Duración” y “Duración\_i\_f”, de acuerdo a las posiciones que se encuentra el valor “No”. A consecuencia de esto, solo tendrá valores que corresponde aquellos que, si contiene respuesta.

Una vez terminado de aplicar criterios anteriormente expuestas, se procede a separar todas aquellas filas que contenga más de dos respuestas, los cuales si cumple con el rango de fecha entre el de envío y de término. Después, se eliminan aquellos campos (columnas) que ya no son necesarios o fueron utilizados durante el proceso de estructuración de Sendinblue, y se almacena quedando la misma forma como se muestra en la anterior ilustración [6-1]. Para terminar, se almacena el dataset resultante en un archivo .CSV, para realizar estadísticas descriptivas acciones para los datos outliers, y generar dataset de prueba para redes neuronales. En relación con el código de fuente utilizado, véase a la carpeta “Códigos de fuentes”, donde se adjuntan archivos y un documento.

Ilustración 6-2: Proceso de estructuración de datos



Fuente: Elaboración propia

## 6.2. Análisis Descriptivo de los Datos

Al terminar el proceso de construcción de dataset, como resultado, se obtiene un total de 4.593.606 registros (filas) y 14 campos variables (columnas), sin contar con código de envío, E-mail, fecha de envío, de término y de respuesta. En los envíos, la cantidad de días transcurridos hasta el próximo envío son aproximadamente 54 días (54,45 días), cuya desviación estándar es 55,8. Esto indica mayor dispersión en los valores, por lo que no es posible detectar un patrón sobre programación envíos. Cabe destacar que la cantidad de personas de la muestra, mediante correos únicos del conjunto de datos, son 911.371. Asimismo, mientras que el 74.829 de ellas han respondido la encuesta por lo menos una vez, otras 836.542 personas nunca lo han hecho durante en el periodo de 2020. Por lo tanto, esto indica que hay más personas desinteresadas en contestar o no están atentos al recibir formularios a través de correos electrónicos.

Por otro lado, al comparar con el conjunto inicial de los datos muestrales (véase a tabla [6-1]), se tiene una pérdida total de 6,54% con respecto a los estados. Cabe destacar que apertura es la suma entre la cantidad de estados “*Primera apertura*” y “*Abierto*” por envío, por lo que, al aplicar el criterio de que cada envío debe tener 0 o 1 “*Primera apertura*” (dejando con valores booleanos), se ha restado valores aquellos que tienen más de una. En cuanto a cantidad de envíos, lo cual representa cada estado “*Entregado*” de Sendinblue, para pocos casos se ha detectado dos envíos en un día, por lo que se eliminan aquellos que tengan 0 días transcurridos para que haya próximo envío. Todo lo anterior se visualiza en la tabla [6-2].

*Tabla 6-2: Pérdida de datos en estructuración de Sendinblue*

DESCRIPCIÓN	CANTIDAD MUESTRAL (EXTRAIDA)	CANTIDAD RESULTANTE (PROCESADA)	DIFERENCIA ENTRE REAL Y ESTIMADA	PÉRDIDA DE DATOS (%)
Envíos	4.621.984	4.593.606	28.378	0,61%
Apertura	1.903.415	1.790.573	112.842	5,93%

*Fuente: Elaboración propia*

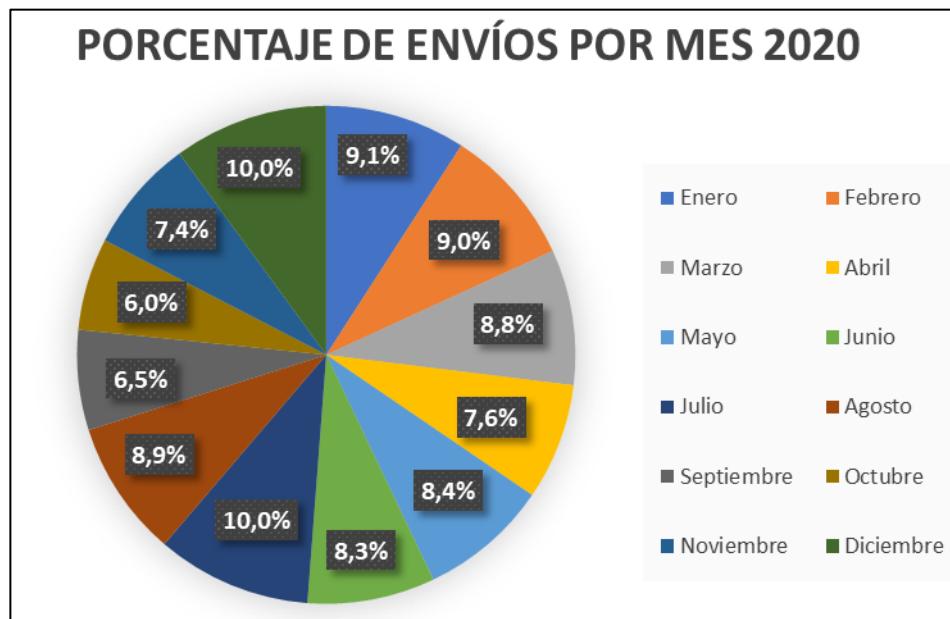
En cuanto las cantidades de envíos por meses, como se visualiza en la tabla e ilustración [6-3], la mayor frecuencia de envíos está más concentrado en los meses Julio y Diciembre, mientras que Octubre se ha registrado menor frecuencia para otros envíos.

*Tabla 6-3: Cantidad de envíos por mes*

MES DE ENVÍO	CANTIDAD	FRECUENCIA (%)
Enero	418.771	9,1%
Febrero	415.060	9,0%
Marzo	404.268	8,8%
Abril	347.791	7,6%
Mayo	385.665	8,4%
Junio	380.770	8,3%
Julio	461.385	10,0%
Agosto	406.959	8,9%
Septiembre	298.601	6,5%
Octubre	275.028	6,0%
Noviembre	338.851	7,4%
Diciembre	460.457	10,0%
<b>TOTAL</b>	<b>4.593.606</b>	

*Fuente: Elaboración propia*

Ilustración 6-3: Porcentaje de envíos por mes 2020



Fuente: Elaboración propia

#### 6.2.1. Estadística con relación al horario de envíos:

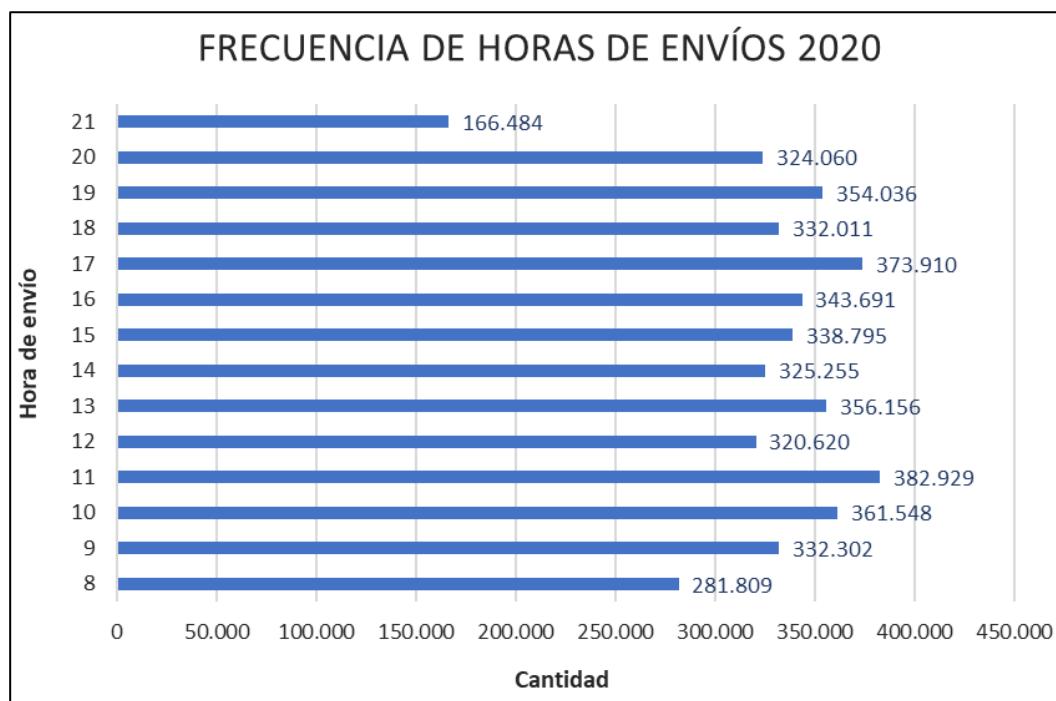
Para el caso del horario de envío, es importante saber la hora que se debe enviar un formulario a los encuestados, y en qué día de la semana tiene mayor respuesta. Para el caso de las frecuencias en las horas de envío (aproximadamente), como se visualiza en la tabla e ilustración [6-4], se tiene una mayor frecuencia de envíos a las 11:00 horas con 8,3% y seguida a las 17:00 horas con 8,1%. En cuanto los otros envíos realizados a las 21:00 horas, dado que es el horario de cierre del trabajo en la organización, se ha registrado un porcentaje de frecuencia menor con 3,6%.

Tabla 6-4: Frecuencia con relación a horas de envíos

HORA DE ENVÍO	CANTIDAD	FRECUENCIA (%)
8	281.809	6,1%
9	332.302	7,2%
10	361.548	7,9%
11	382.929	8,3%
12	320.620	7,0%
13	356.156	7,8%
14	325.255	7,1%
15	338.795	7,4%
16	343.691	7,5%
17	373.910	8,1%
18	332.011	7,2%
19	354.036	7,7%
20	324.060	7,1%
21	166.484	3,6%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

Ilustración 6-4: Gráfico de barras sobre frecuencia de horas de envíos 2020



Fuente: Elaboración propia

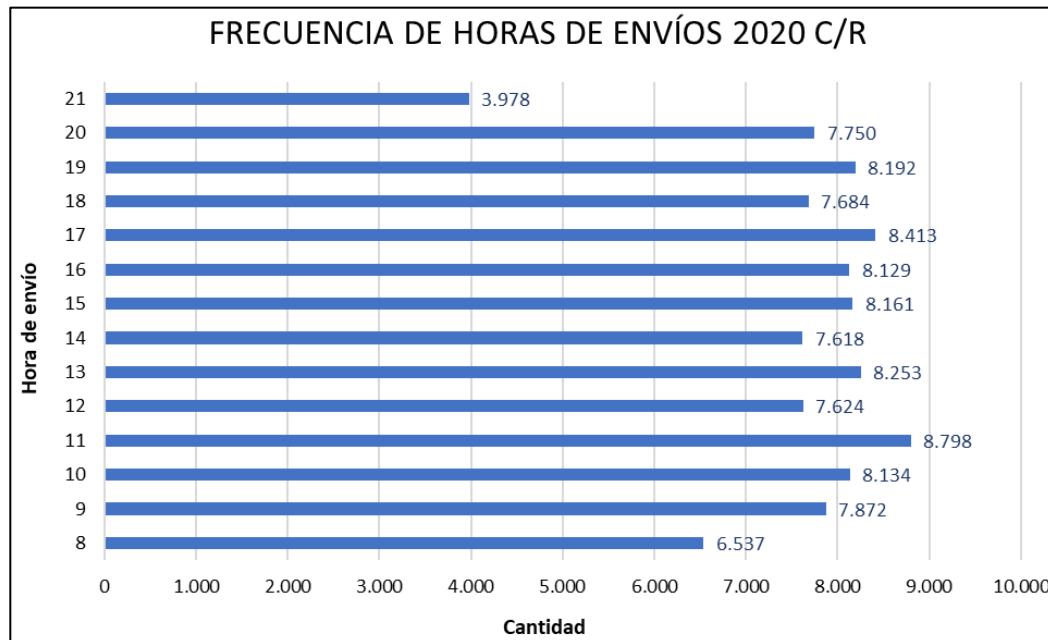
Asimismo, si se compara entre las horas de envíos que tienen respuestas y otras no, como se visualiza en las ilustraciones [6-5], [6-6] y la tabla [6-5], no se presentan diferencias significativas, ya que los de frecuencias entre ellos son similares. Por lo que, al tener datos muy similares a la tabla anterior [6-4], indica que a las 11:00 y 17:00 horas lleva un mayor porcentaje, mientras que a las 21:00 horas es el que lleva menor frecuencia. Por lo tanto, no se detecta algún patrón ni diferencias importantes entre las horas de envíos.

*Tabla 6-5: Comparativa frecuencia con relación a horas de envíos con y sin respuesta*

HORA DE ENVÍO	ENVÍOS SIN RESPUESTA		ENVÍOS CON RESPUESTA	
	CANTIDAD	FRECUENCIA (%)	CANTIDAD	FRECUENCIA (%)
8	275.272	6,1%	6.537	6,1%
9	324.430	7,2%	7.872	7,3%
10	353.414	7,9%	8.134	7,6%
<b>11</b>	<b>374.131</b>	<b>8,3%</b>	<b>8.798</b>	<b>8,2%</b>
12	312.996	7,0%	7.624	7,1%
13	347.903	7,8%	8.253	7,7%
14	317.637	7,1%	7.618	7,1%
15	330.634	7,4%	8.161	7,6%
16	335.562	7,5%	8.129	7,6%
<b>17</b>	<b>365.497</b>	<b>8,1%</b>	<b>8.413</b>	<b>7,9%</b>
18	324.327	7,2%	7.684	7,2%
19	345.844	7,7%	8.192	7,6%
20	316.310	7,1%	7.750	7,2%
21	162.506	3,6%	3.978	3,7%
<b>TOTAL</b>	<b>4.486.463</b>		<b>107.143</b>	

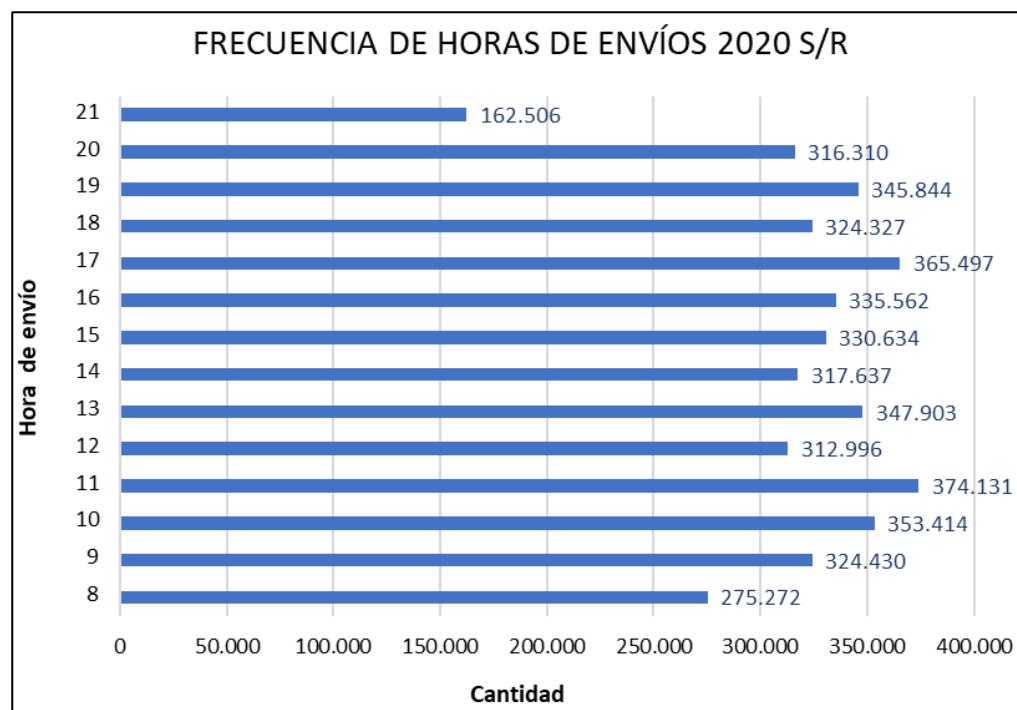
*Fuente: Elaboración propia*

Ilustración 6-5: Gráfico de barras sobre frecuencia de horas de envíos con respuesta 2020



Fuente: Elaboración propia

Ilustración 6-6: Gráfico de barras sobre frecuencia de horas de envíos sin respuesta 2020



Fuente: Elaboración propia

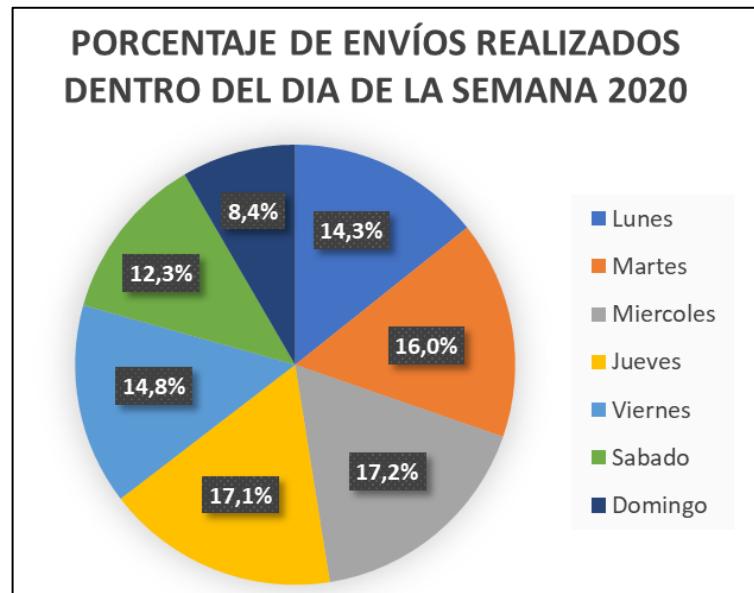
En cuanto a los días de semana, referente a lunes, martes, miércoles, jueves, viernes, sábado y domingo, como se visualiza en la tabla [6-6] e ilustración [6-7], existe mayor concentración de envíos los miércoles, con 17,2%, y jueves, con 17,1%. Mientras que los días domingo presenta un porcentaje menor de frecuencia, de 8,4%.

*Tabla 6-6: Frecuencia de envíos con respecto a los días de semana*

DÍA DE LA SEMANA	CANTIDAD	FRECUENCIA (%)
<b>Lunes</b>	656.526	14,3%
<b>Martes</b>	735.993	16,0%
<b>Miércoles</b>	<b>788.368</b>	<b>17,2%</b>
<b>Jueves</b>	<b>785.624</b>	<b>17,1%</b>
<b>Viernes</b>	679.215	14,8%
<b>Sábado</b>	562.904	12,3%
<b>Domingo</b>	384.976	8,4%
<b>TOTAL</b>	4.593.606	

*Fuente: Elaboración propia*

*Ilustración 6-7: Porcentaje de envíos realizados dentro del día de la semana 2020*



*Fuente: Elaboración propia*

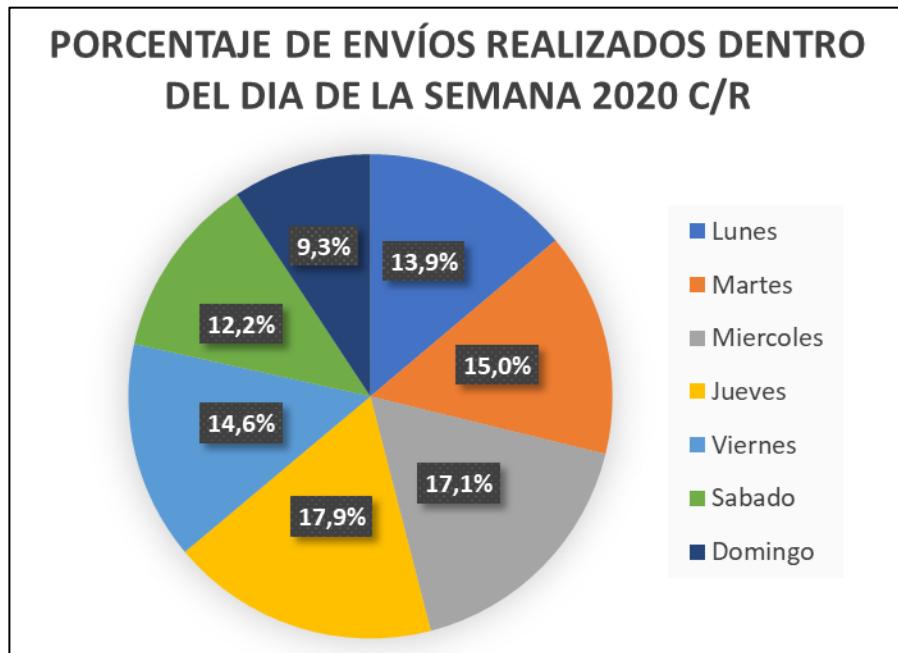
Al comparar con los días de semana entre envíos con y sin respuesta, como se visualiza en la tabla [6-7] e ilustraciones [6-8] y [6-9], existe una leve diferencia entre porcentajes de frecuencias, ya que todos los envíos realizados en los jueves tienen más concentración de respuestas. Mientras que otros envíos sin respuesta, en su mayoría, ocurre al mandar en los miércoles de la semana. Por otro lado, para mayores detalles respecto el conjunto de datos resultante, véase a los reportes generado por “*Profiling pandas*” en la carpeta “/Reportes - Profiling pandas/ Estadísticas Descriptiva”.

*Tabla 6-7: Comparativa sobre los envíos en los días de semana, con y sin respuesta*

DÍA DE LA SEMANA	ENVÍOS SIN RESPUESTA		ENVÍOS CON RESPUESTA	
	CANTIDAD	FRECUENCIA (%)	CANTIDAD	FRECUENCIA (%)
<b>Lunes</b>	641.672	14,3%	14.854	13,9%
<b>Martes</b>	719.957	16,0%	16.036	15,0%
<b>Miércoles</b>	<b>769.994</b>	<b>17,2%</b>	18.374	17,1%
<b>Jueves</b>	766.428	17,1%	<b>19.196</b>	<b>17,9%</b>
<b>Viernes</b>	663.591	14,8%	15.624	14,6%
<b>Sábado</b>	549.804	12,3%	13.100	12,2%
<b>Domingo</b>	375.017	8,4%	9.959	9,3%
<b>TOTAL</b>	4.486.463		107.143	

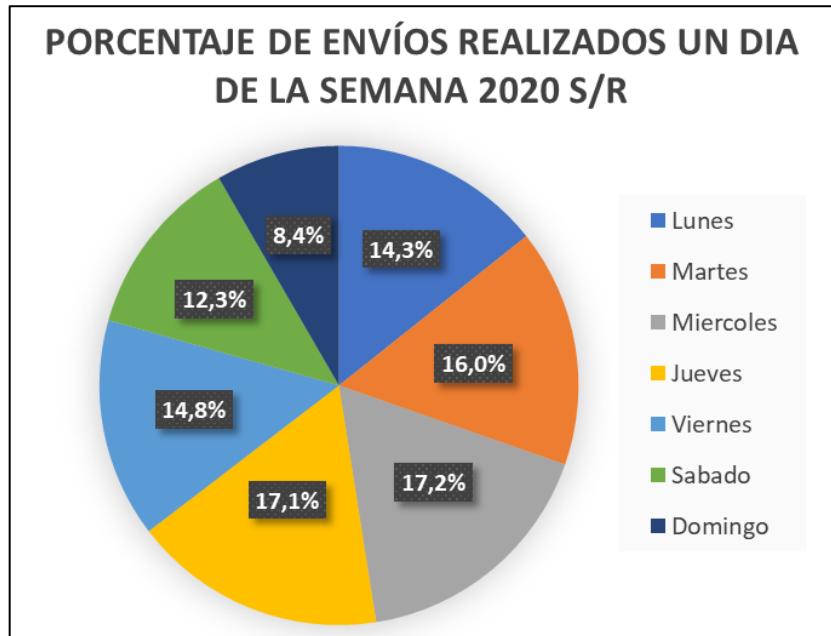
*Fuente: Elaboración propia*

Ilustración 6-8: Porcentaje de envíos realizados dentro del día de la semana 2020 con respuesta



Fuente: Elaboración propia

Ilustración 6-9: Porcentaje de envíos realizados dentro del día de la semana 2020 sin respuesta



Fuente: Elaboración propia

Dado todos los antecedentes anteriores mencionadas, se procede a calcular los porcentajes de frecuencias de que personas hayan contestado una encuesta, en base al horario de envío. Como resultado, se determina que la probabilidad más alta de que una persona responda es el jueves a las 11:00, mientras que el domingo a las 21:00 tiene la probabilidad muy baja de que el encuestado haya contestado a dicho horario. Asimismo, a partir de la tabla [6-8] se visualiza las frecuencias de envíos contestados en horario, destacando que sus cálculos se realizaron utilizando el código de fuente elaborado “*Calculo\_de\_frecuencia\_horaria.ipynb*”.

*Tabla 6-8: Porcentaje de frecuencia de respuestas en un determinado horario*

HORAS DE ENVÍO / DÍA DE SEMANA	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo	GENERAL HORA DE ENVÍO
<b>8:00</b>	0,85%	0,91%	1,05%	1,09%	0,89%	0,75%	0,57%	6,11%
<b>9:00</b>	1,02%	1,10%	1,26%	1,32%	1,07%	0,90%	0,68%	7,35%
<b>10:00</b>	1,05%	1,14%	1,30%	1,36%	1,11%	0,93%	0,71%	7,60%
<b>11:00</b>	1,14%	1,23%	1,41%	<b>1,47%</b>	1,20%	1,00%	0,76%	8,21%
<b>12:00</b>	0,99%	1,07%	1,22%	1,27%	1,04%	0,87%	0,66%	7,12%
<b>13:00</b>	1,07%	1,15%	1,32%	1,38%	1,12%	0,94%	0,72%	7,70%
<b>14:00</b>	0,99%	1,06%	1,22%	1,27%	1,04%	0,87%	0,66%	7,11%
<b>15:00</b>	1,06%	1,14%	1,31%	1,36%	1,11%	0,93%	0,71%	7,62%
<b>16:00</b>	1,05%	1,14%	1,30%	1,36%	1,11%	0,93%	0,71%	7,60%
<b>17:00</b>	1,09%	1,18%	1,35%	1,41%	1,15%	0,96%	0,73%	7,87%
<b>18:00</b>	0,99%	1,07%	1,23%	1,28%	1,05%	0,88%	0,67%	7,17%
<b>19:00</b>	1,06%	1,14%	1,31%	1,37%	1,11%	0,93%	0,71%	7,63%
<b>20:00</b>	1,00%	1,08%	1,24%	1,30%	1,05%	0,88%	0,67%	7,22%
<b>21:00</b>	0,51%	0,56%	0,64%	0,67%	0,54%	0,45%	0,35%	3,72%
<b>GENERAL DÍA DE LA SEMANA</b>	13,87%	14,97%	17,16%	17,91%	14,59%	12,22%	9,31%	100%

Fuente: Elaboración propia

### 6.2.2. Estadísticas con relación a envíos con respuesta:

En cuanto a comparación en relación con cantidad de respuestas registradas en cada envío, a partir de la tabla [6-9] se observa la cantidad y frecuencia de envíos de encuestas, con relación de que ha recibido una respuesta o no, y en la ilustración [6-10] se logra visualizar un gráfico de torta donde indica los porcentajes de ellos. Cabe destacar que hay una enorme diferencia entre sus frecuencias, ya que esto indica que existe una tasa muy baja de respuesta, por lo que en el modelamiento puede presentar alguna dificultad en relación con la precisión de predicción.

Ilustración 6-10: Porcentaje de envíos con y sin respuesta 2020



Fuente: Elaboración propia

Tabla 6-9: Cantidad de envíos con y sin respuesta 2020

¿RESPONDÍO LA ENCUESTA?	CANTIDAD	FRECUENCIA (%)
Sí	107.143	2,3%
No	4.486.463	97,7%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

En cuanto a la diferencia entre cantidad de respuestas reales (véase la tabla [5-2]) y otras estimadas mediante el proceso de estructuración de Sendinblue por meses, como se muestra en la tabla [6-10], existe un mayor porcentaje de error en el mes de enero, esto es debido a que no se ha considerado el historial de envíos para el mes de diciembre del año 2019. Otra razón, en cuanto a todos los meses, al procesar una gran cantidad de datos en la estructuración de dataset se pierden datos, y esto se justifica las diferencias que se presentan al comparar con la cantidad de respuestas que hay en la BD de respuestas (incluyendo enero), destacando que el mes de abril se ha logrado extraer todas las respuestas con 100% de precisión

*Tabla 6-10: Cantidad de respuestas estimadas y comparación con la BD de respuesta*

MES	CANTIDAD (RESPUESTAS)	FRECUENCIA (%)	DIFERENCIA ENTRE REAL Y ESTIMADA	ERROR (%)
1	13.068	12,2%	842	6,05%
2	8.012	7,5%	43	0,53%
3	8.180	7,6%	6	0,07%
4	10.404	9,7%	0	0,00%
5	9.035	8,4%	5	0,06%
6	7.377	6,9%	2	0,03%
7	8.998	8,4%	0	0,00%
8	9.628	9,0%	9	0,09%
9	8.014	7,5%	13	0,16%
10	8.424	7,9%	14	0,17%
11	8.465	7,9%	46	0,54%
12	7.538	7,0%	26	0,34%
<b>TOTAL</b>	<b>107.143</b>	<b>100,0%</b>	<b>1.006</b>	<b>0,93%</b>

*Fuente: Elaboración propia*

Por otro lado, respecto a las frecuencias que cada persona responde la encuesta, como se aprecia en la tabla [6-11], es destacable que 836.542 (91,8%) individuos nunca ha respondido una encuesta, mientras que otras 74.829 (8,2%) si ha contestado por lo menos una vez un formulario por correo electrónico. En cuanto a los días transcurridos hasta recibir respuesta por parte del encuestado, como se visualiza en la tabla [6-12], se destaca que el 30,2% de las personas respondió la encuesta al mismo día que han recibido (0 días), mientras que 41,8% se ha demorado entre 1 y 4 días en responder, y otros 70% restantes se demoró más de 4 días, siendo 54 como el valor máximo.

*Tabla 6-11: Cantidad de veces que se ha respondido una encuesta por personas*

¿Cuántas veces ha respondido la encuesta en el año 2020?	CANTIDAD DE PERSONAS	FRECUENCIA (%)	CANTIDAD DE RESPUESTAS
Nunca (0)	836.542	91,8%	0
Una vez (1)	60.754	6,7%	60.754
Dos veces (2)	5.942	0,7%	11.884
Tres veces (3)	945	0,1%	2.835
Cuatro veces (4)	4.991	0,5%	19.964
Cinco veces (5)	1.640	0,2%	8.200
Seis veces (6)	414	<0,1%	2.484
Siete veces (7)	122	<0,1%	854
Ocho veces (8)	21	<0,1%	168
<b>TOTAL, DE PERSONAS</b>	<b>911.371</b>	<b>TOTAL, DE RESPUESTAS</b>	<b>107.143</b>

*Fuente: Elaboración propia*

Tabla 6-12: Frecuencia sobre días transcurridos hasta recibir una respuesta

DÍAS TRANSCURRIDO PARA RESPONDER ENCUESTA	CANTIDAD	FRECUENCIA (%)
Al mismo día del envío	32.336	30,2%
Entre 1 a 4 días	44.823	41,8%
Entre 5 a 9 días	4.450	4,2%
Entre 10 a 14 días	3.279	3,1%
Entre 15 a 19 días	6.193	5,8%
Entre 20 a 29 días	6.373	5,9%
Entre 30 a 39 días	5.516	5,1%
40 o más días	4.173	3,9%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

Para el caso de aperturas en los envíos, como se visualiza en la tabla [6-13], dada la alta incidencia de encuestas sin respuesta por parte de una persona, se tiene un 74,9% sobre que un encuestado nunca abrió un envío. Mientras que el 25,1% restante, otros individuos han abierto un envío por lo menos una vez.

Tabla 6-13: Frecuencia con relación a N° de aperturas por envío

N° DE APERTURAS	CANTIDAD	FRECUENCIA (%)
0	3.439.712	74,9%
1	763.125	16,6%
2	205.611	4,5%
3	130.068	2,8%
4	53.585	1,2%
5	467	<0,1%
6	388	<0,1%
7 o más	650	<0,1%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

Al comparar con los N° de aperturas entre envíos con y sin respuesta, como se visualiza en la tabla [6-14], es importante destacar que todos los envíos que tengan registrados una cantidad de abiertos mayor que 4, se asegura de que este envío haya recibido respuesta alguna. Para otro caso, si dicha cantidad se da igual a 3, existe la probabilidad de 54,2% que la persona responda la encuesta, mientras que si N° de apertura es igual a 4, tiene otra probabilidad de 67,5% que el envío sea contestado. Por lo tanto, si el N° de apertura es menor que 3, se asegura que no habrá respuesta alguna por parte del encuestado.

*Tabla 6-14: Comparativa N° de aperturas en los envíos con y sin respuesta*

ENVIOS SIN RESPUESTA			ENVIOS CON RESPUESTA		
N° DE APERTURAS	CANTIDAD	FRECUENCIA (%)	N° DE APERTURAS	CANTIDAD	FRECUENCIA (%)
0	3.439.712	76,7%	3	70.451	65,8%
1	763.125	17,0%	4	35.187	32,8%
2	205.611	4,6%	5	467	0,4%
3	59.617	1,3%	6	388	0,4%
4	18.398	0,4%	7 o más	650	0,6%
<b>TOTAL</b>	<b>4.486.463</b>		<b>TOTAL</b>	<b>107.143</b>	

*Fuente: Elaboración propia*

### 6.2.3. Estadística con relación al perfil de encuestados:

Se muestra a continuación un conjunto de estadísticas descriptivas de las personas que recibieron encuestas en el año 2020 (un total aproximado de 4,5 millones de encuestas), según el sexo, la edad, segmento, subsegmento, agrupación de segmentos y carterización. Además, se realiza un análisis estadístico mediante gráficos de torta y tablas de frecuencia, con relación a personas quienes responden encuestas y otras no, para así identificar el perfil de cada uno de ellos.

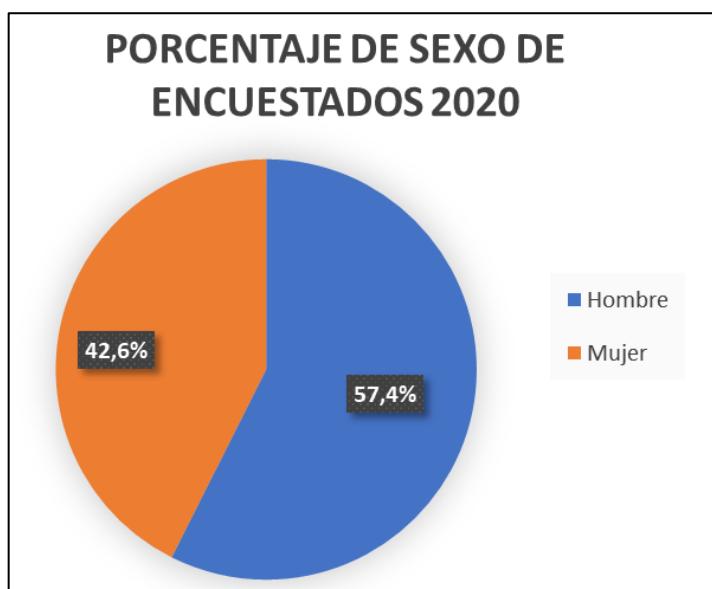
- **Perfil de sexo:** De acuerdo con la tabla [6-15] e ilustración [6-11], se observa una predominancia de los hombres encuestados con un 57,7% de las encuestas recibidas, contra el 42,6% de las encuestas las cuales son recibidas por mujeres.
  - **Encuestas respondidas y no respondidas:** A partir de los datos presentes en las tablas [6-16] y [6-17], se calcula que solo el 2,59% de los hombres respondió la encuesta (el 97,41% no la contestó) frente al 2,00% de las mujeres (98,00% no contestó la encuesta).

*Tabla 6-15: Frecuencia con relación a sexo de los encuestados 2020*

SEXO	CANTIDAD	FRECUENCIA (%)
Hombre	2.637.448	57,40%
Mujer	1.956.158	42,60%
<b>TOTAL</b>	<b>4.593.606</b>	

*Fuente: Elaboración propia*

*Ilustración 6-11: Porcentaje de sexo de encuestados 2020*



*Fuente: Elaboración propia*

Tabla 6-16: Frecuencia con relación a sexo de los encuestados con respuesta 2020

SEXO	CANTIDAD	FRECUENCIA
Hombre	68.192	63,65%
Mujer	38.951	36,35%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

Ilustración 6-12: Porcentaje de sexo de encuestados C/R 2020



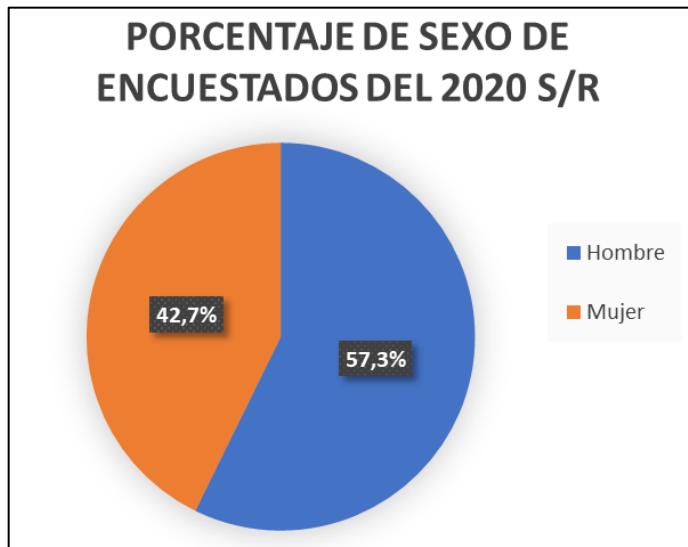
Fuente: Elaboración propia

Tabla 6-17: Frecuencia con relación a sexo de los encuestados S/R 2020

SEXO	CANTIDAD	FRECUENCIA
Hombre	2.569.256	57,27%
Mujer	1.917.207	42,73%
<b>TOTAL</b>	<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-13: Porcentaje de sexo de encuestados S/R 2020



Fuente: Elaboración propia

- **Perfil de edad:** Según la tabla [6-18], todas las personas encuestadas son adultas, donde el 95% de los encuestados tienen entre 18 y 68 años (inclusive).

Tabla 6-18: Frecuencia con relación a la edad de los encuestados 2020

ESTADÍSTICAS DE CUANTILES	VALOR
MINIMO	18
5to PERCENTIL	23
1ER CUARTIL	31
MEDIANA	39
3ER CUARTIL	50
95to PERCENTIL	68
MAXIMO	121
DESVIACION ESTANDAR	13,7513681
VARIANZA	189,1001246
PROMEDIO	41,62141551

Fuente: Elaboración propia

Tabla 6-19: Frecuencia con relación a la edad de los encuestados C/R 2020

ESTADÍSTICAS DE CUANTILES	VALOR
MINIMO	18
5to PERCENTIL	25
1ER CUARTIL	33
MEDIANA	42
3ER CUARTIL	54
95to PERCENTIL	71
MAXIMO	119
DESVIACION ESTANDAR	14,24417692
VARIANZA	202,8965762
PROMEDIO	44,63249116

Fuente: Elaboración propia

Tabla 6-20: Frecuencia con relación a la edad de los encuestados S/R 2020

ESTADÍSTICAS DE CUANTILES	VALOR
MINIMO	18
5to PERCENTIL	23
1ER CUARTIL	31
MEDIANA	39
3ER CUARTIL	50
95to PERCENTIL	68
MAXIMO	121
DESVIACION ESTANDAR	13,7313145
VARIANZA	188,5489978
PROMEDIO	41,54950682

Fuente: Elaboración propia

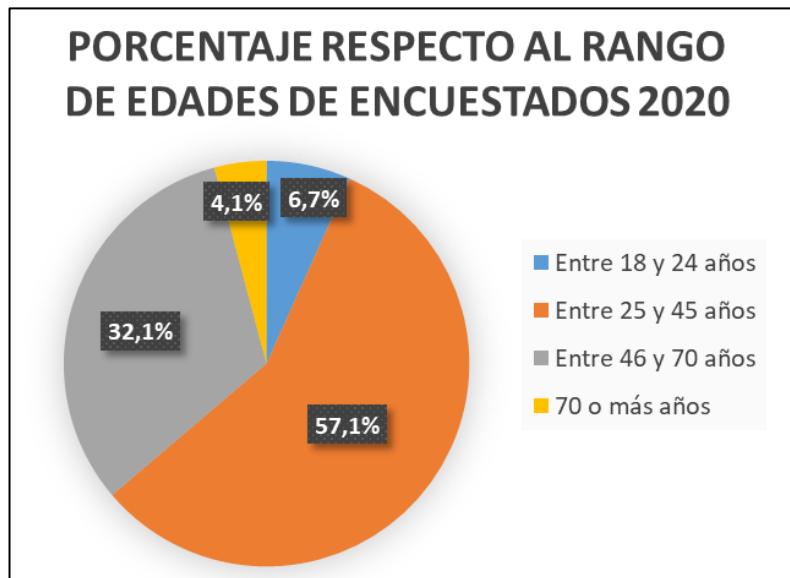
Al determinar los rangos de edades, como se puede apreciar en la tabla [6-21] y las siguientes ilustraciones, se destaca que la mayoría de los encuestados, tanto quienes respondieron envíos y otras no, tienen entre 25 y 45 años.

Tabla 6-21: Frecuencia con relación al rango de edades 2020

RANGO DE EDADES	GLOBAL		ENVÍOS CON RESPUESTA		ENVÍOS SIN RESPUESTA	
	CANTIDAD	FRECUENCIA (%)	CANTIDAD	FRECUENCIA (%)	CANTIDAD	FRECUENCIA (%)
Entre 18 y 24 años	309.102	6,7%	4.155	3,9%	304.947	6,8%
Entre 25 y 45 años	2.620.671	57,1%	55.235	51,6%	2.565.436	57,2%
Entre 46 y 70 años	1.473.473	32,1%	40.887	38,2%	1.432.586	31,9%
70 o más años	190.360	4,1%	6.866	6,4%	183.494	4,1%
<b>TOTAL</b>	<b>4.593.606</b>		<b>107.143</b>		<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-14: Porcentaje respecto al rango de edades de encuestados 2020



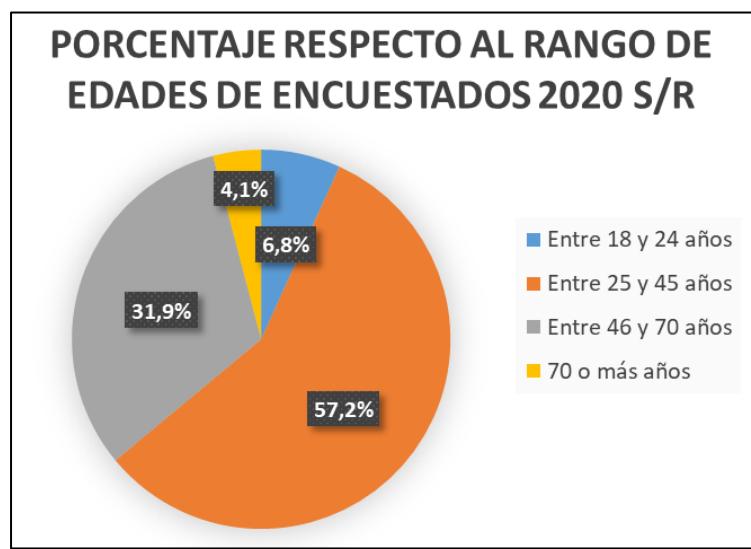
Fuente: Elaboración propia

Ilustración 6-15: Porcentaje respecto al rango de edades de encuestados 2020 C/R



Fuente: Elaboración propia

Ilustración 6-16: Porcentaje respecto al rango de edades de encuestados 2020 S/R



Fuente: Elaboración propia

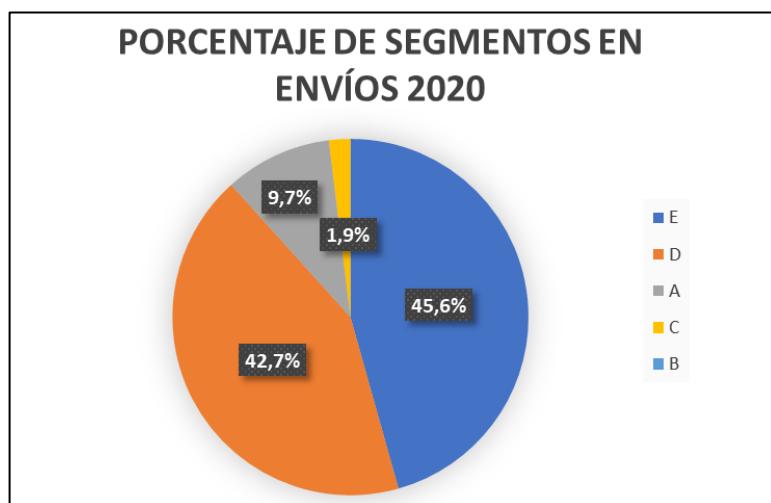
- **Perfil de segmento:** De acuerdo con la tabla [6-22] e ilustración [6-17], los segmentos E y D recibieron la mayor cantidad de encuestas, correspondientes al 45,60% y 42,70% del total, respectivamente.
  - **Encuestas respondidas y no respondidas:** A partir de los datos presentes en las tablas e ilustraciones, se calcula que solo el 2,08% del segmento E respondió la encuesta frente al 2,57% del segmento D.

*Tabla 6-22: Frecuencia con relación al segmento de los encuestados 2020*

SEGMENTO	CANTIDAD	FRECUENCIA (%)
E	2.096.962	45,60%
D	1.959.776	42,70%
A	447.601	9,70%
C	89.246	1,90%
B	21	<0,1%
<b>TOTAL</b>	<b>4.593.606</b>	

*Fuente: Elaboración propia*

*Ilustración 6-17: Porcentaje de segmentos de encuestados 2020*



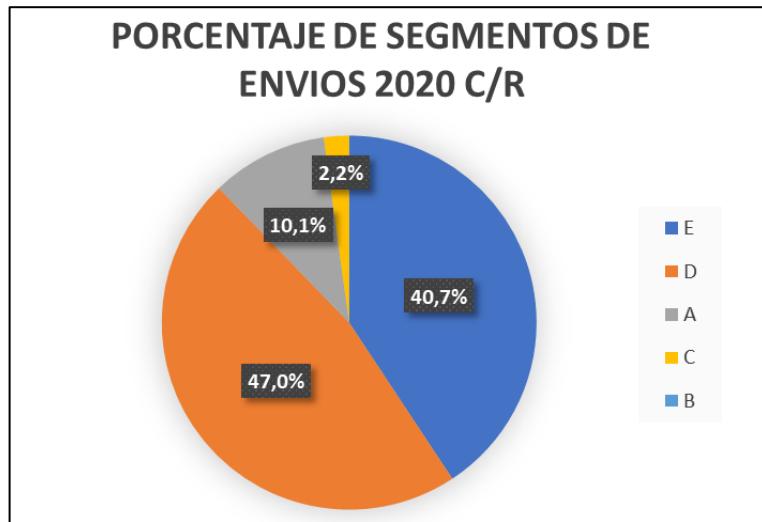
*Fuente: Elaboración propia*

Tabla 6-23: Frecuencia con relación al segmento de los encuestados C/R 2020

SEGMENTO	CANTIDAD	FRECUENCIA (%)
E	43.631	40,72%
D	50.410	47,05%
A	10.784	10,07%
C	2.316	2,16%
B	2	<0,1%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

Ilustración 6-18: Porcentaje de segmentos de encuestados C/R 2020



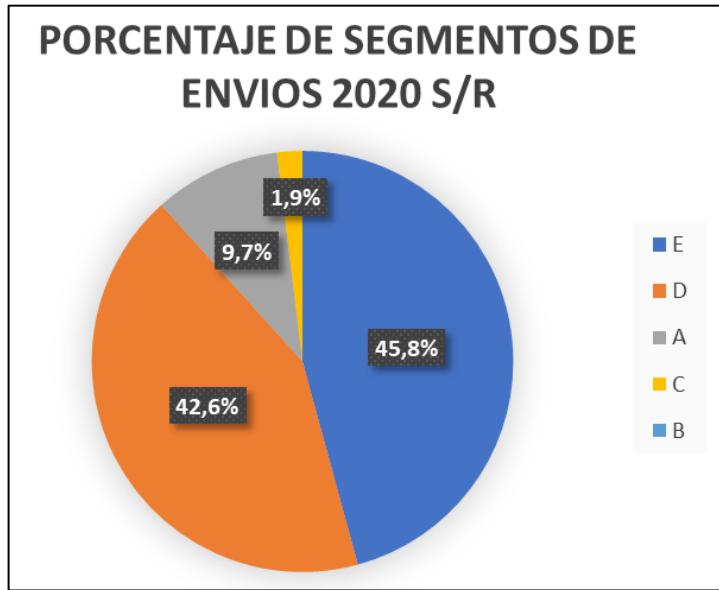
Fuente: Elaboración propia

Tabla 6-24: Frecuencia con relación al segmento de los encuestados S/R 2020

SEGMENTO	CANTIDAD	FRECUENCIA (%)
E	2.053.331	45,77%
D	1.909.366	42,56%
A	436.817	9,74%
C	86.930	1,94%
B	19	0%
<b>TOTAL</b>	<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-19: Porcentaje de segmentos de encuestados S/R 2020



Fuente: Elaboración propia

- **Perfil de subsegmento:** De acuerdo con la tabla [6-25] e ilustración [6-20], los subsegmentos D2A y E1B recibieron un 27,9% y un 14,0% de las encuestas, respectivamente. Se observa junto a las estadísticas del perfil de segmento, una distribución más homogénea de encuestados en el segmento E por sobre el segmento D.

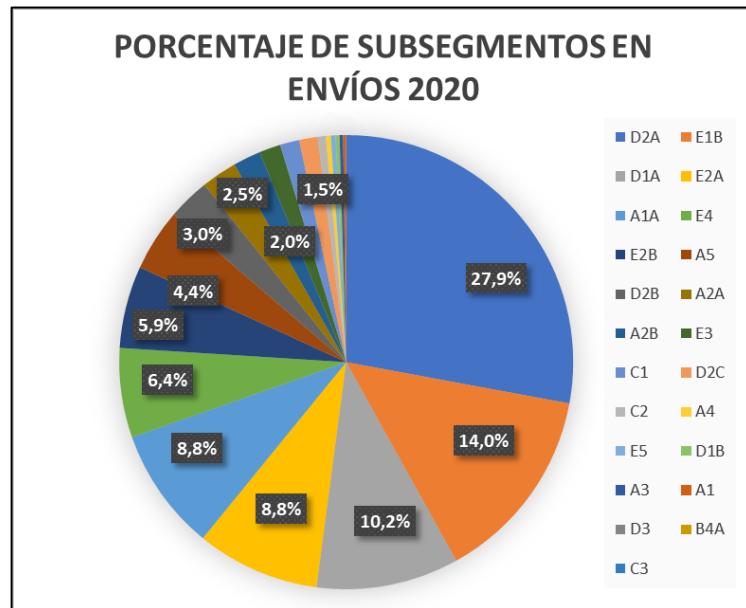
Tabla 6-25: Frecuencia con relación al subsegmento de los encuestados 2020

SUBSEGMENTO	CANTIDAD	FRECUENCIA (%)
D2A	1.282.302	27,9%
E1B	644.420	14,0%
D1A	466.358	10,2%
E2A	403.669	8,8%
A1A	402.022	8,8%
E4	292.390	6,4%
E2B	268.913	5,9%
A5	203.522	4,4%
D2B	137.124	3,0%

A2A	116.617	2,5%
A2B	90.089	2,0%
E3	70.431	1,5%
C1	62.973	1,4%
D2C	60.506	1,3%
C2	26.272	0,6%
A4	16.402	0,4%
E5	15.117	0,3%
D1B	12.988	0,3%
A3	11.924	0,3%
A1	9.047	0,2%
D3	498	<0,1%
B4A	21	<0,1%
C3	1	<0,1%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

Ilustración 6-20: Porcentaje de subsegmentos de encuestados 2020



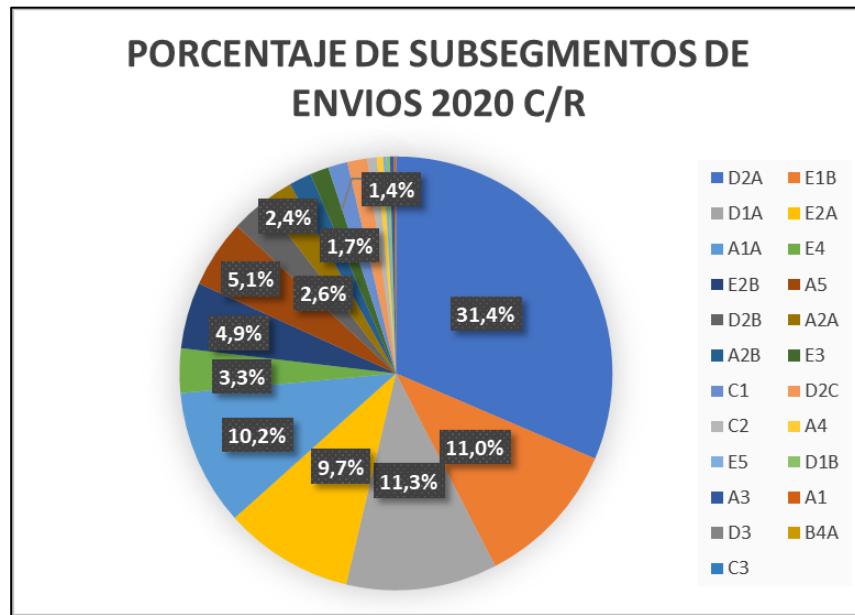
Fuente: Elaboración propia

Tabla 6-26: Frecuencia con relación al subsegmento de los encuestados C/R 2020

SUBSEGMENTO	CANTIDAD	FRECUENCIA (%)
D2A	33.670	31,4%
E1B	11.772	11,0%
D1A	12.100	11,3%
E2A	10.375	9,7%
A1A	10.882	10,2%
E4	3.551	3,3%
E2B	5.294	4,9%
A5	5.445	5,1%
D2B	2.759	2,6%
A2A	2.563	2,4%
A2B	1.806	1,7%
E3	1.495	1,4%
C1	1.562	1,5%
D2C	1.600	1,5%
C2	754	0,7%
A4	498	0,5%
E5	262	0,2%
D1B	274	0,3%
A3	317	0,3%
A1	155	0,1%
D3	7	<0,1%
B4A	2	<0,1%
C3	0	<0,1%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

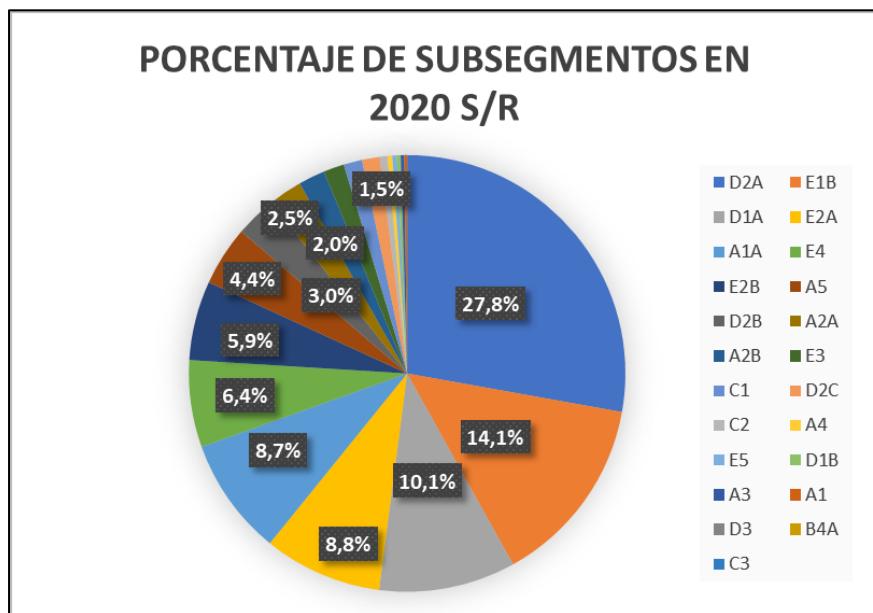
Ilustración 6-21: Porcentaje de subsegmentos de encuestados C/R 2020



C2	25.518	0,6%
A4	15.904	0,4%
E5	14.855	0,3%
D1B	12.714	0,3%
A3	11.607	0,3%
A1	8.892	0,2%
D3	491	<0,1%
B4A	19	<0,1%
C3	1	<0,1%
<b>TOTAL</b>	<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-22: Porcentaje de subsegmentos de encuestados S/R 2020



Fuente: Elaboración propia

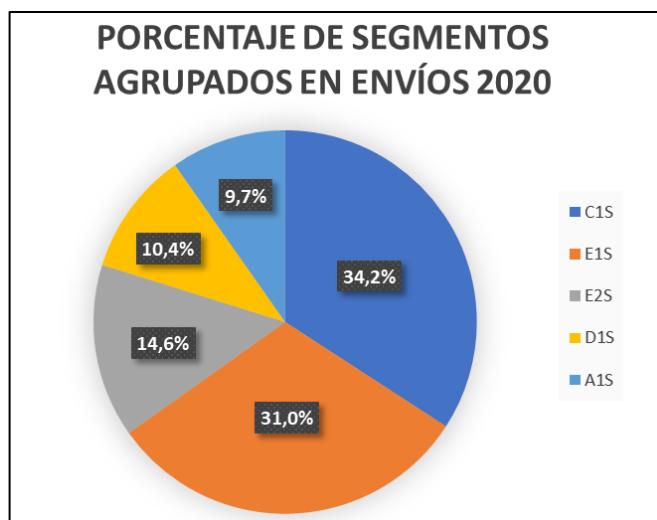
- **Perfil de agrupación de segmentos:** De acuerdo con la tabla [6-28] e ilustración [6-23], los grupos de segmentos C1S y E1S predominan la recepción de encuestas con un 34,20% y un 31,00%, respectivamente.

Tabla 6-28: Frecuencia con relación al segmento agrupado de los encuestados 2020

SEGTO AGRUPADO	CANTIDAD	FRECUENCIA (%)
C1S	1.569.178	34,20%
E1S	1.424.401	31,00%
E2S	672.582	14,60%
D1S	479.844	10,40%
A1S	447.601	9,70%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

Ilustración 6-23: Porcentaje de segmentos agrupados de encuestados 2020



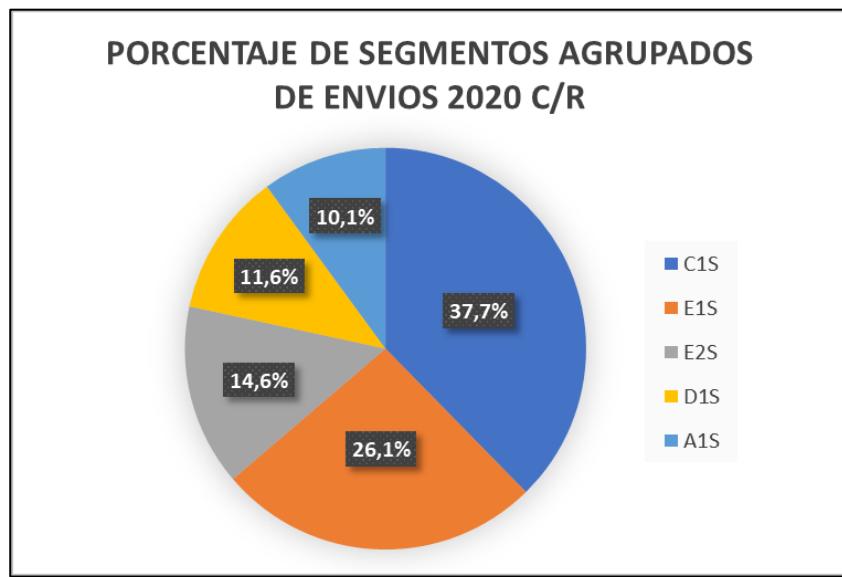
Fuente: Elaboración propia

Tabla 6-29: Frecuencia con relación al segmento agrupado de los encuestados C/R 2020

SEGTO AGRUPADO	CANTIDAD	FRECUENCIA (%)
C1S	40.345	37,66%
E1S	27.964	26,10%
E2S	15.669	14,62%
D1S	12.381	11,56%
A1S	10.784	10,07%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

Ilustración 6-24: Porcentaje de segmentos agrupados de encuestados C/R 2020



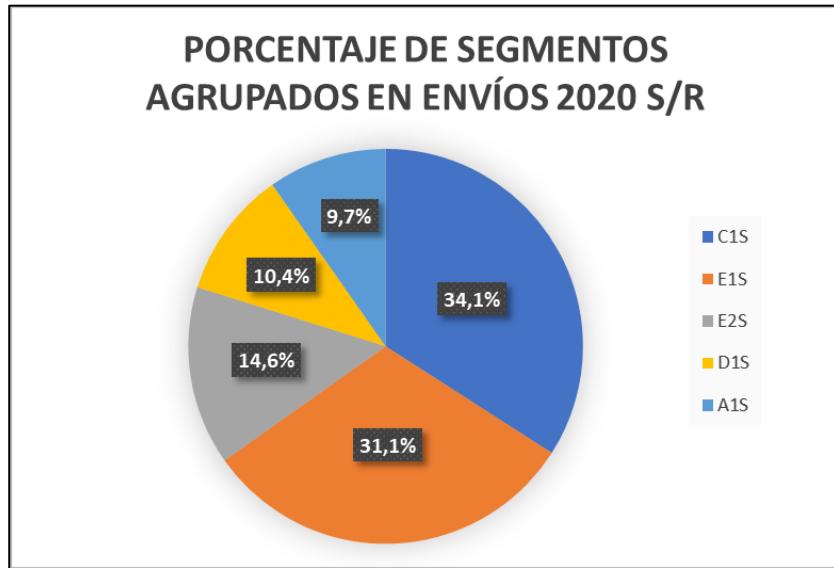
Fuente: Elaboración propia

Tabla 6-30: Frecuencia con relación al segmento agrupado de los encuestados S/R 2020

SEGMENTO AGRUPADO	CANTIDAD	FRECUENCIA (%)
C1S	1.528.833	34,08%
E1S	1.396.437	31,13%
E2S	656.913	14,64%
D1S	467.463	10,42%
A1S	436.817	9,74%
<b>TOTAL</b>	<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-25: Porcentaje de segmentos agrupados de encuestados S/R 2020



Fuente: Elaboración propia

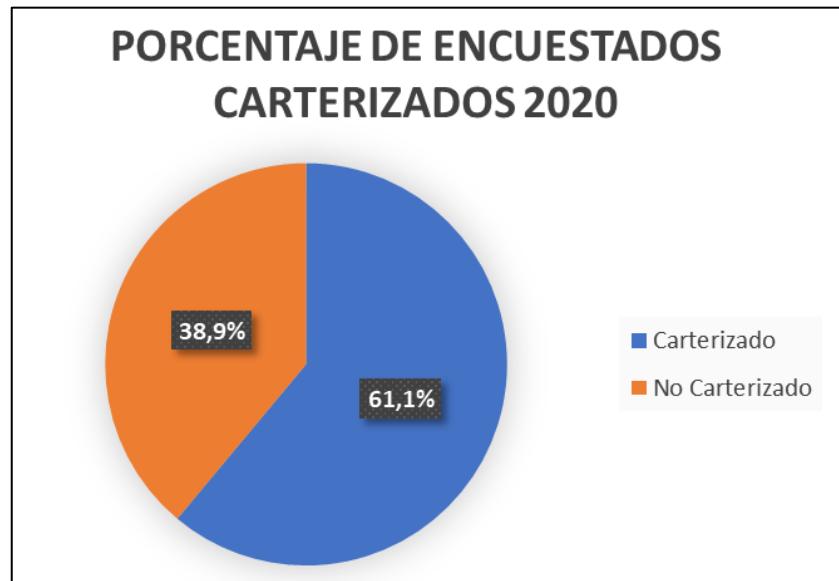
- **Perfil de carterización:** De acuerdo con la tabla [6-31] e ilustración [6-26], la cantidad de encuestados carterizados es mayor con un 61,10% del total.

Tabla 6-31: Frecuencia con relación a la carterización de los encuestados 2020

¿PERTENECE A CARTERIZADO?	CANTIDAD	FRECUENCIA (%)
Carterizado	2.805.557	61,10%
No Carterizado	1.788.049	38,90%
<b>TOTAL</b>	<b>4.593.606</b>	

Fuente: Elaboración propia

Ilustración 6-26: Porcentaje de encuestados carterizados 2020



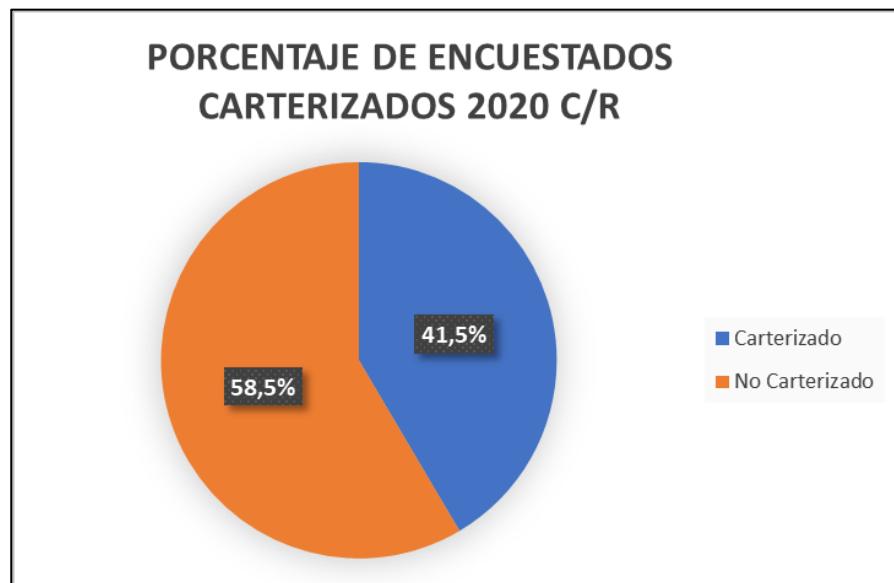
Fuente: Elaboración propia

Tabla 6-32: Frecuencia con relación a la carterización de los encuestados C/R 2020

¿PERTENECE A CARTERIZADO?	CANTIDAD	FRECUENCIA (%)
Si	44.469	41,50%
No	62.674	58,50%
<b>TOTAL</b>	<b>107.143</b>	

Fuente: Elaboración propia

Ilustración 6-27: Porcentaje de encuestados carterizados C/R 2020



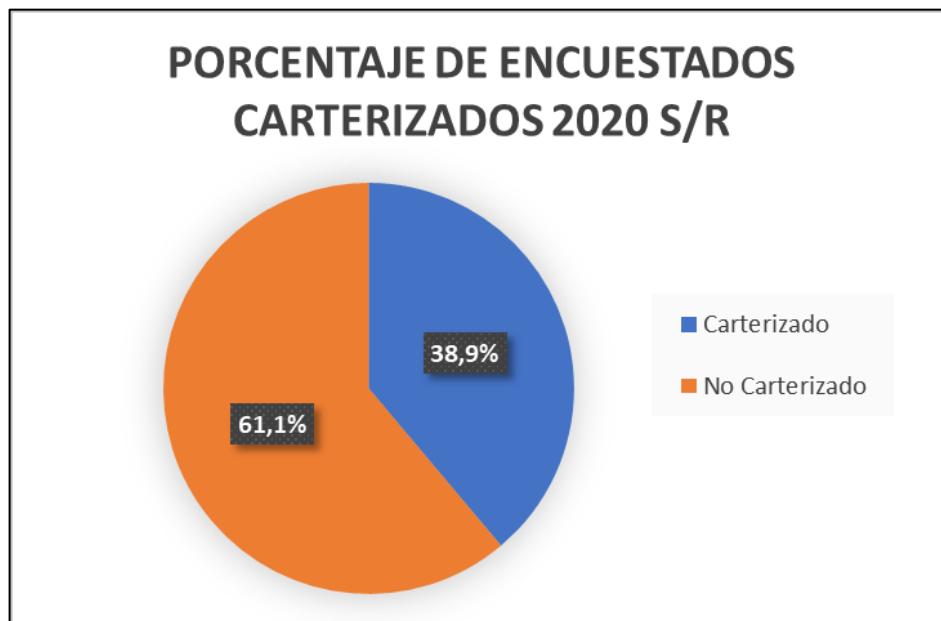
Fuente: Elaboración propia

Tabla 6-33: Frecuencia con relación a la carterización de los encuestados S/R 2020

¿PERTENECE A CARTERIZADO?	CANTIDAD	FRECUENCIA (%)
Carterizado	1.743.580	38,86%
No Carterizado	2.742.883	61,14%
<b>TOTAL</b>	<b>4.486.463</b>	

Fuente: Elaboración propia

Ilustración 6-28: Porcentaje de encuestados carterizados S/R 2020



Fuente: Elaboración propia

Para mayores detalles respecto a las estadísticas con relación al perfil de encuestados, por meses, véase los reportes generados (HTML) en la carpeta “Reportes - Profiling pandas”. Además, para visualizar mejor las frecuencias y porcentajes de los campos, véase a la carpeta “Hojas de cálculo/Estadística descriptiva”.

### **6.3. Detección y tratamiento de outliers:**

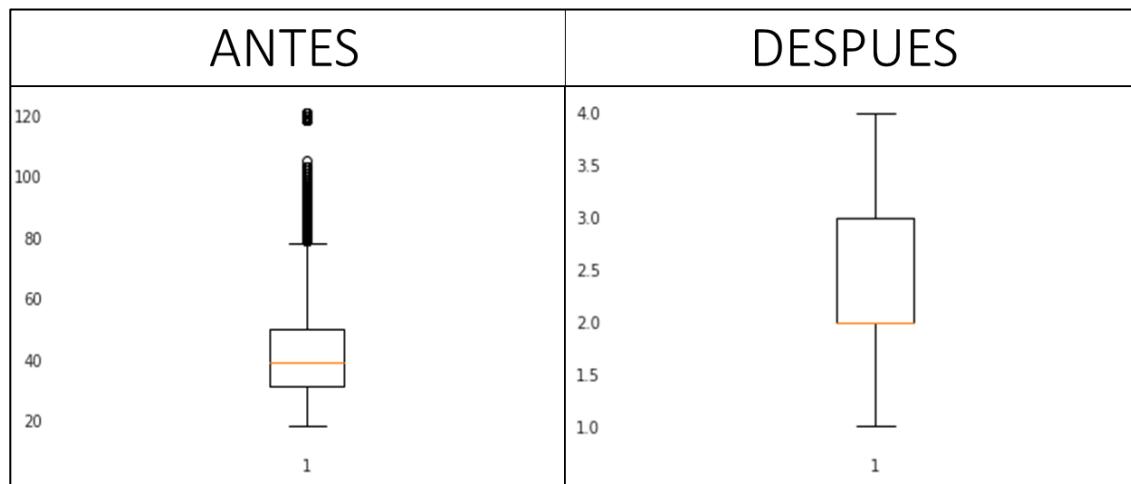
Los datos atípicos (outliers) pueden ser un problema para precisar mejor los resultados estadísticos y de predicción, y como se busca predecir la contactabilidad de los clientes. Para este caso, es importante tener definido una estrategia para reducir su impacto, ya que, si se considera un dato con valor muy extremo, tiene mayores consecuencias en la estimación de respuestas en los envíos. En el ámbito de inferencia, las pruebas de hipótesis son sensibles al incumplimiento de supuestos en los modelos y a la presencia de outliers, pero no significa que todos estos datos sean erróneos. Por lo tanto, cuando los datos no cumplen con estos supuestos disminuye la capacidad de detectar efectos reales, por lo que cualquier interpretación de los datos pueden ser erróneas.

En el datasets, se ha detectado outliers en los campos de edad, segmentos, subsegmentos, agrupación de segmentos, aperturas y, principalmente, “RESPONDIDA”. Para el caso de edad, dado a la gran cantidad de outliers detectados se procede a categorizar los valores por intervalos de la siguiente manera:

- Todas las personas que tienen entre 18 y 24 años, es categorizada como “*Edad agrupada 1*”.
- Todas las personas que tienen entre 25 y 45 años, es categorizada como “*Edad agrupada 2*”.
- Todas las personas que tienen entre 46 y 70 años, es categorizada como “*Edad agrupada 3*”.
- Todas las personas que tienen 70 o más años, es categorizada como “*Edad agrupada 4*”.

Como resultado, visualizado en la ilustración [6-29], se puede apreciar la desaparición de outliers (círculos fuera de la caja) en el campo edad del conjunto de datos.

*Ilustración 6-29: Comparativa de BoxPlot, resultado de mitigación de outliers en campo edad*



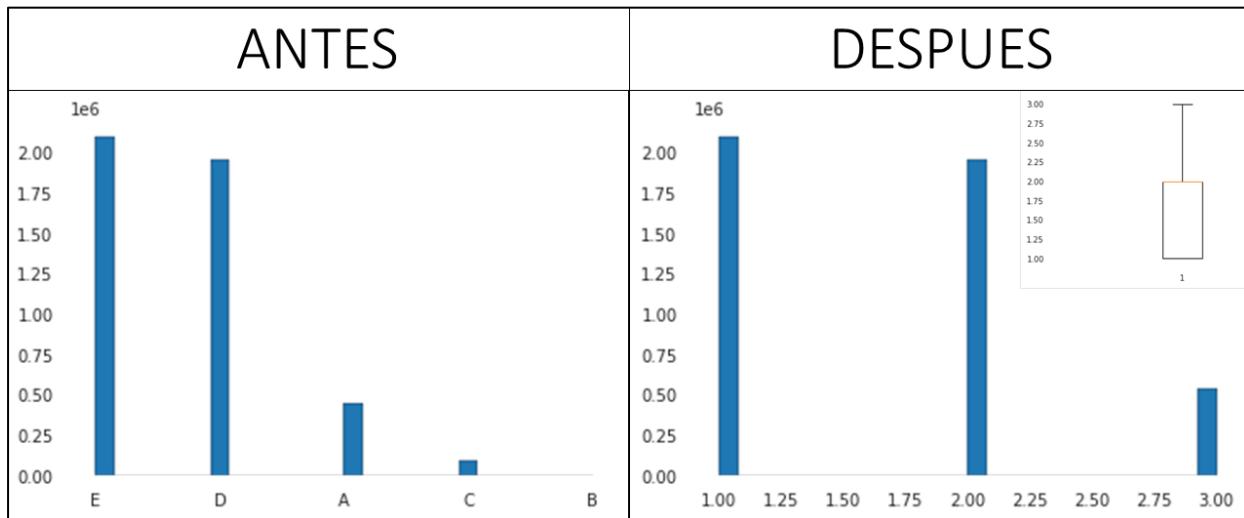
Fuente: Elaboración propia

Para el campo “Segmento”, lo cual se ha detectado por su alta correlación en los datos categóricos, se ha agrupado de la siguiente manera:

- Para el segmento “E”, dado que tiene alto porcentaje de frecuencia, es agrupada como “Segmento 1”.
- Para el segmento “D”, al igual que el punto anterior con relación a su frecuencia, es agrupada como “Segmento 2”.
- Para los segmentos “A”, “B” y “C”, son agrupadas como “Segmento 3”.

Como resultado, visualizado en la ilustración [6-30], se puede apreciar la mitigación de alta correlación del campo segmento en el conjunto de datos.

Ilustración 6-30: Comparativa de histograma, mitigación de outliers en Segmento



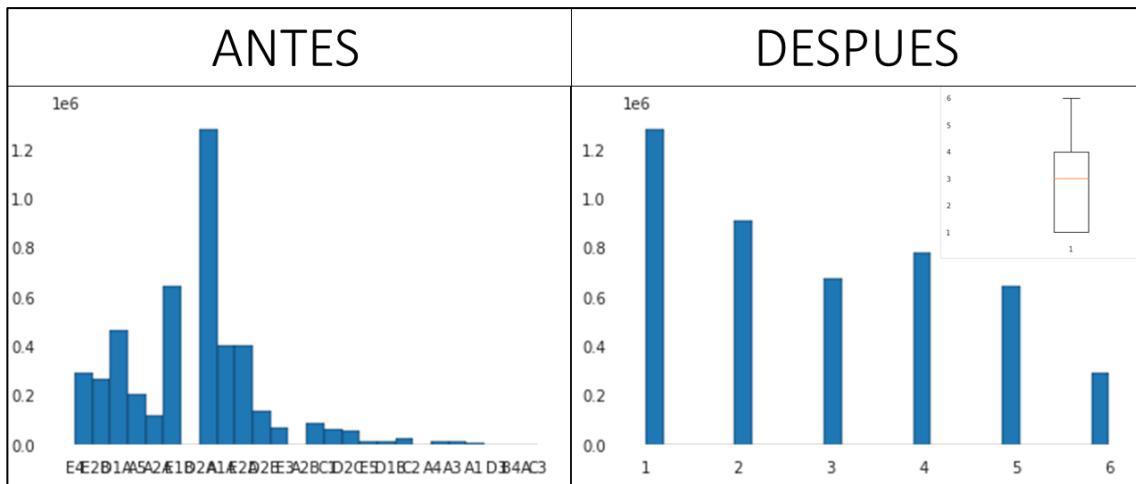
Fuente: Elaboración propia

Para otro campo “Subsegmento”, al igual que el campo anterior, se ha detectado por la alta correlación en los datos categóricos, por lo que se ha agrupado de la siguiente manera:

- Para el subsegmento “D2A”, dado que tiene alto porcentaje de frecuencia, es agrupada como “Subsegmento 1”.
- Para los subsegmentos “E1B” y “E2B”, al igual que el punto anterior con relación a su frecuencia, son agrupados como “Subsegmento 2”.
- Para los subsegmentos “D1A”, “D3”, “D1B”, “D2B” y “D2C”, son agrupadas como “Subsegmento 3”.
- Para los subsegmentos “E2A”, “E5”, “E3” y “E4”, son agrupadas como “Subsegmento 4”.
- Para los subsegmentos “A1”, “A1A”, “A2A”, “A2B”, “A3” y “A4”, son agrupadas como “Subsegmento 5”.
- Para los subsegmentos “C1”, “C2”, “C3”, “B4A” y “A5”, son agrupadas como “Subsegmento 6”.

Como resultado, visualizado en la ilustración [6-31], se puede apreciar la mitigación de alta correlación del campo subsegmento en el conjunto de datos.

*Ilustración 6-31: Comparativa de histograma, mitigación de outliers en Subsegmento*



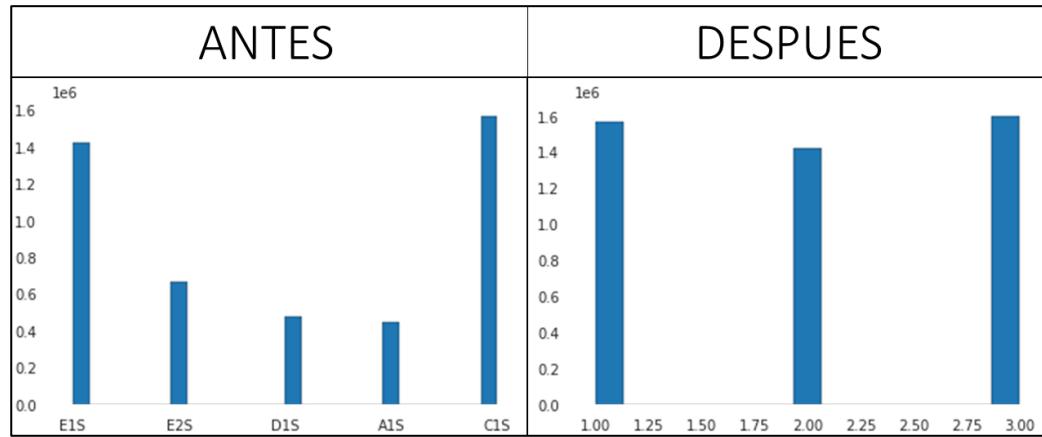
Fuente: Elaboración propia

Para otro campo “Agrupación de segmento”, al igual que los dos puntos anteriores, se ha detectado por la alta correlación en los datos, por lo que se ha agrupado de la siguiente manera:

- Para la agrupación de segmentos “C1S”, dado que tiene alto porcentaje de frecuencia, es agrupada como “Agrupación segmento 1”.
- Para la agrupación de segmentos “E1S”, al igual que el punto anterior con relación a su frecuencia, es agrupada como “Agrupación segmento 2”.
- Para la agrupación de segmentos “E2S”, “D1S” y “A1S”, son agrupadas como “Agrupación segmento 3”.

Como resultado, visualizado en la ilustración [6-32], se puede apreciar la mitigación de alta correlación del campo subsegmento en el conjunto de datos.

Ilustración 6-32: Comparativa de histograma, mitigación de outliers en Agrupación de Segmentos



Fuente: Elaboración propia

En cuanto al campo de aperturas, se decide sustituir las cantidades que supere 5 por 6, lo cual permite reducir la dispersión de valores para todos los envíos que contengan respuestas. Como resultado, visualizado en la ilustración [6-33], se puede apreciar la mitigación de outliers de este campo, donde se observa que los otros datos anormales, como son los valores 4, 5 y 6, son necesarios para la predicción del modelo.

Ilustración 6-33: Comparativa de BoxPlot, resultado de mitigación de outliers en campo apertura



Fuente: Elaboración propia

Por último, para el campo de “RESPONDIDA”, dado la gran cantidad de envíos sin respuestas es necesario tomar decisiones, y es que hay existencia de envíos que, sin considerar Id de envío, E-mails, día de envío y otros campos con relación a las fechas (fecha de envío, de término y de respuesta), coinciden con los datos. Por lo que, se tiene opción de eliminar los registros que son muy parecidos, con el fin de reducir las incidencias de encuestas que no tiene respuesta. Para este caso, es importante mantener los registros de envíos con respuestas, y existen dos ramas donde cada uno genera un dataset de prueba distinto.

1. Antes de aplicar mitigación de outliers para los campos de edad, segmentos, subsegmentos, agrupación de segmentos y aperturas, se eliminan registros que contiene valores muy parecidos. Por lo que, reduce el volumen de envíos sin perder mucha información, eliminando 13,4% de los registros. Como resultado, se tiene un conjunto de datos con un total de 3.991.687 envíos, donde su porcentaje de distribución de frecuencia se aprecian a partir de la tabla [6-34] e ilustración [6-34].

*Tabla 6-34: Dataset N°1, frecuencia de envíos con y sin respuesta*

DATASET DE PRUEBA 1		
SITUACIÓN DEL ENVÍO	CANTIDAD	FRECUENCIA (%)
Sin respuesta	3.884.544	97,3%
Con respuesta	107.143	2,7%
<b>TOTAL</b>		3.991.687

*Fuente: Elaboración propia*

Ilustración 6-34: Dataset N°1, porcentaje de envíos con y sin respuesta



Fuente: Elaboración propia

2. Después de aplicar mitigación de outliers de los campos anteriormente mencionados en el punto anterior, se eliminan registros que contiene valores muy parecido, lo cual reduce el volumen de envíos drásticamente. Sin embargo, se pierden datos que posiblemente aporta a los resultados, ya que elimina 46,7% de los registros. Como resultado, se tiene un conjunto de datos con un total de 2.499.069 envíos, donde su porcentaje de distribución de frecuencia se aprecian a partir de la tabla [6-35] e ilustración [6-35].

Tabla 6-35: Dataset N°2, frecuencia de envíos con y sin respuesta

DATASET DE PRUEBA 2		
SITUACIÓN DEL ENVÍO	CANTIDAD	FRECUENCIA (%)
Sin respuesta	2.391.926	95,7%
Con respuesta	107.143	4,3%
<b>TOTAL</b>	<b>2.499.069</b>	

Fuente: Elaboración propia

Ilustración 6-35: Dataset N°2, porcentaje de envíos con y sin respuesta



Fuente: Elaboración propia

## 6.4. Transformación de Valores en los Datos

Al realizar alguna predicción usando técnicas de redes neuronales, es importante que los tipos de datos en los valores generados tenga compatibilidad en los algoritmos que los paquetes de Python contengan, por lo que es muy recomendable trabajar todos los valores flotantes, del rango entre 0 y 1. Para los campos de “sexo”, ‘día de semana’ “segmento”, “subsegmento”, “segmentos agrupados” y “carterizado”, durante el proceso de estructuración de Sendinblue, se codifica por valores enteros, donde estos son utilizados para aplicación de técnicas de redes neuronales con mayor compatibilidad de datos. A partir de la tabla [6-36], se aprecia los significados de cada codificación en los campos de los datasets.

Tabla 6-36: Codificación de datos del dataset estructurado

CAMPOS	VALOR ORIGEN	VALOR CODIFICADO	CAMPOS	VALOR ORIGEN	VALOR CODIFICADO
Sexo	H	0	Segmentos agrupados	Agrupación segmento 1	1
	M	1		Agrupación segmento 2	2
Edades agrupadas	Edad agrupada 1	1		Agrupación segmento 3	3
	Edad agrupada 2	2	Carterizado	Carterizado	1
	Edad agrupada 3	3		No Carterizado	0
	Edad agrupada 4	4	Dia de la semana	Lunes	1
Segmento	Segmento 1 (E)	1		Martes	2
	Segmento 2 (D)	2		Miércoles	3
	Segmento 3	3		Jueves	4
Subsegmento	Subsegmento 1	1		Viernes	5
	Subsegmento 2	2		Sábado	6
	Subsegmento 3	3		Domingo	7
	Subsegmento 4	4	RESPONDIDA	Si (con respuesta)	1
	Subsegmento 5	5		No (sin respuesta)	0
	Subsegmento 6	6			

Fuente: Elaboración propia

Una vez codificado los valores de tipo String por enteros, se procede a normalizar todos los valores para transformarlo en tipo flotante (decimales). Para este caso, es importante que todos los valores se encuentren en un rango entre 0 y 1, ya que esto permite que las funciones de activación puedan ejecutarse correctamente. Por lo que, se aplica la siguiente fórmula para cada celda de los datasets, es importante que todos los campos sean enteros y que se haya codificado correctamente, como menciona en la tabla [6-34] anterior:

$$\text{Valor Celda Normalizada} = \frac{\text{Valor Celda actual} - \text{Valor mínimo columna}}{\text{Valor máximo columna} - \text{Valor mínimo columna}}$$

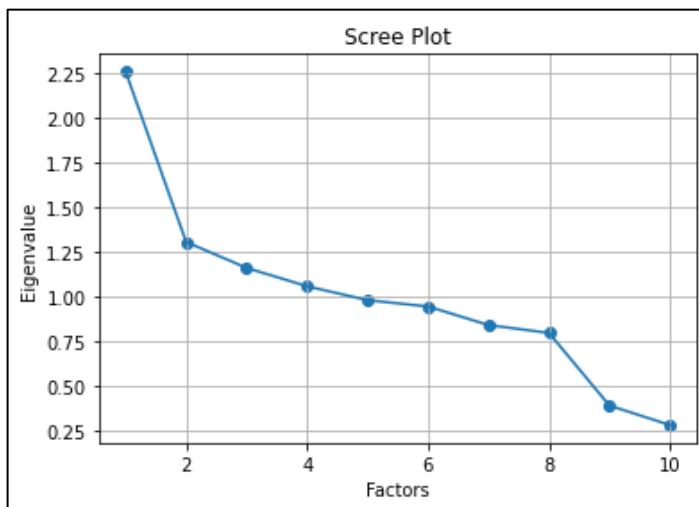
El código utilizado para mitigar a los outliers y transformar datos, se encuentra en el archivo “Analisis de datos y generacion de dataset de pruebas.ipynb”.

## 6.5. Análisis Factorial

Después de haber culminado la etapa anterior, se hace necesario analizar estos factores, para ello cobra relevancia el módulo factor analyzer. El cual, en primera instancia crea el factor de análisis y su desempeño para posteriormente graficarlo e interpretar con método del codo la forma idónea de reducir las componentes. Para recordar, este método se basa en encontrar el punto de inflexión más notorio entre la función, cuyo valor puede identificarse en cuatro. A continuación, se crea el análisis y el desempeño usando esta vez el valor encontrado del método del codo, disponiendo de tal manera, como cada componente incide en cada factor. Finalmente, se otorga una gráfica de estos factores reducidos ilustrando en primer lugar la suma de cargas al cuadrado, enseguida la proporción de la varianza, y en último lugar la varianza acumulada.

Para el primer dataset de prueba, el punto de inflexión cuando se comparan los “Eigenvalue” respecto a los “Factors” con dicho método se encuentra en valor 4, lo cual se visualiza en la ilustración [6-36]. Lo que significa que, al reducir las dimensiones con este valor, no se perdería información y las variables tendrán distinta incidencia por cada “nueva dimensión”. Cabe destacar que la primera ruptura de la curva ocurre cuando hay presencia de dos factores, mientras que otro quiebre ocurre cuando se trata de que hay ocho factores, lo cual los valores de “Eigenvalue” desciende hasta llegar acercar a un valor igual a 0.

Ilustración 6-36: Gráfico de Codo para obtener número de factores, dataset N°1



Fuente: Elaboración propia

Entonces, las variables independientes por estos “nuevos factores” su incidencia se encuentra en la ilustración [6-37], donde se destaca: La variable “Sexo” solo guarda incidencia en el Factor 1, Factor 2 y Factor 4, destacando una mayor contribución en el primero. En cuanto a la variable “Edad” solo incide en el Factor 3. Para “Segmento”, tiene cerca de un 64% de aporte en el Factor 3, un 6.74% en el Factor 1 y casi un 0% en el Factor 4. Al analizar “Subsegmento” tiene un aporte considerable en el “Factor 1” llegando al 81.36%, para el “Factor 2” alcanza un 29%, en cambio en el “Factor 3” acumula tan solo un 13.27% y en el “Factor 4” llega a un 8.49%.

Al analizar la variable “Segto\_Agrup” su aporte se encuentra en los Factores 1, 2 y 4, acaparando su gran aporte en el primero, para después disminuir considerablemente en los restantes. “Carterizado” solo incide en los Factores 3 y 4, con una diferencia de casi el 50%. Por otra parte, “Apertura” y “Hora\_envio” solo guarda relación y en baja cantidad en el Factor 3. La variable “Dia\_semana” concentra un 2.25% en el Factor 1 y un 97% de incidencia para el Factor 2. Finalmente, “Mes\_envio” su aporte no supera un 1% en los Factores 1 y 3, pero en el Factor 4 llega a un 31.30%.

Ilustración 6-37: Dataset N°1, valores de los factores para cada dato campo

	Factor1	Factor2	Factor3	Factor4
Sexo	0.058532	0.021602	-0.098809	0.087891
Edad	-0.081964	-0.012208	0.371726	-0.022947
Segmento	0.067475	-0.006275	0.640750	0.006059
SubSegmento	0.813666	0.029062	0.132779	0.084936
Segto_Agrup	0.722723	0.020299	-0.254624	0.016543
Carterizado	-0.675611	-0.003963	0.638810	0.063168
Apertura	-0.002545	-0.041616	0.019455	-0.257473
Hora_envio	-0.004138	-0.103253	0.004107	-0.034840
Dia_semana	0.022553	0.970949	-0.029986	-0.225818
Mes_envio	0.008432	-0.041040	0.010126	0.313059

Fuente: Elaboración propia

En la ilustración [6-38], se almacena la suma de las cargas al cuadrado, puesto que se utiliza una matriz de correlación, donde la suma de todos los factores equivale al número de variables utilizadas. Destacando, que el Factor 4 tiene una baja tasa respecto a las demás. Siguiendo con “Proportion Var”, lo cual indica que parte de la varianza a nivel general, explica el factor al tomar en cuenta todas las variables. Se debe destacar, que el Factor 1 y Factor 3 superan el 10%. Finalmente, la suma acumulada de la “Proportion Var” indica que el Factor 3 y el Factor 4 se acumulan más de un 36%.

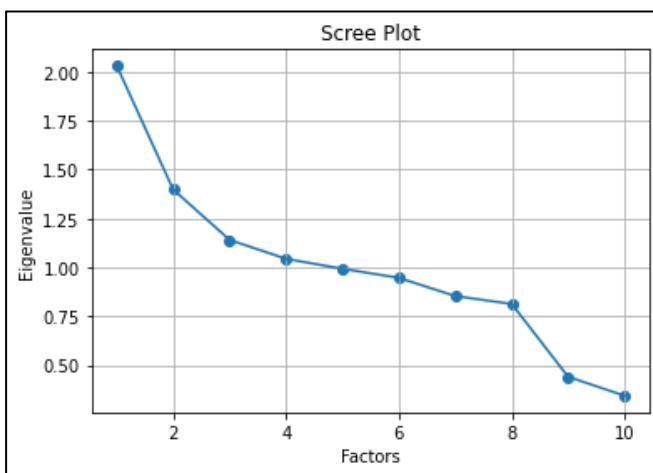
Ilustración 6-38: Dataset N°1, estadística de los factores

	Factor1	Factor2	Factor3	Factor4
SS Loadings	1.656131	0.958747	1.050442	0.236272
Proportion Var	0.165613	0.095875	0.105044	0.023627
Cumulative Var	0.165613	0.261488	0.366532	0.390159

Fuente: Elaboración propia

Para otro caso del conjunto de datos, referente al dataset de prueba N°2, a grandes rasgos persigue el mismo comportamiento del apartado anterior. Al utilizar el método del codo, como se visualiza en la ilustración [6-39], también en valor arroja cuatro factores, es decir, que en este número de componentes (factores) se pueden reducir las variables sin perder información. También, al igual que la anterior, la primera ruptura de la curva ocurre cuando hay presencia de dos factores, mientras que otro quiebre ocurre cuando se trata de que hay ocho factores, lo cual los valores de “Eigenvalue” desciende hasta llegar acercar a un valor igual a 0.

*Ilustración 6-39: Gráfico de Codo para obtener número de factores, dataset N°2*



*Fuente: Elaboración propia*

En la siguiente matriz, ilustración [6-40] como referencia, se puede ver la incidencia de las variables respecto a la reducción de Factores, caso muy similar sobre análisis factorial del dataset de prueba N°1. Donde se puede destacar que la variable “Sexo” está fuertemente ligada al Factor 4 y su aporte se reduce considerablemente en el Factor 2. Al llegar a la variable “Edad” solo está relacionada al Factor 3 en un 34.89%. Casi distinta para “Segmento” cuyo aparte ocurre en el Factor 1 y Factor 3, siendo significante en este último en un 61.16%.

Cabe destacar que, en el campo “Subsegmento”, es la única variable independiente que está relacionada a todos los Factores, teniendo un 77.42% de aporte en el Factor 1. Siguiendo adelante con las variables, hasta llegar a “Segto\_Agrup”, indica que su aporte alcanza un 67.36% en el Factor 1, un 3.89% para el Factor 2 y un 4.49 para el Factor 4. En cuanto al campo “Carterizado”, tiene un gran aporte en el Factor 3 acumulando un 61.15% y reduciendo a un 2.31 % para el Factor 4. En el campo “Apertura” no supera el 1% de impacto en los Factores 2 y 3, mientras que otra variable “Hora\_envio” tiene relación indirecta con los Factores. En cuanto a la variable de “Dia\_semana” su aporte es significativo para el Factor 2 e insignificante respecto al Factor 3. Finalmente, “Mes\_envio” supera el 50% tanto en el Factor 3 y 4.

Ilustración 6-40: Dataset N°2, valores de los factores para cada dato campo

	Factor1	Factor2	Factor3	Factor4
Sexo	-0.017706	0.006353	-0.077524	0.353410
Edad	-0.076473	-0.019652	0.348960	-0.074616
Segmento	0.086128	-0.025235	0.611610	-0.029805
SubSegmento	0.774242	0.019142	0.288261	0.202175
Segto_Agrup	0.673684	0.038902	-0.172115	0.044966
Carterizado	-0.684850	-0.062773	0.611551	0.023146
Apertura	-0.061446	0.009940	0.001513	-0.132145
Hora_envio	-0.000950	-0.085844	-0.002187	-0.009125
Dia_semana	-0.013000	0.997391	0.012390	-0.009035
Mes_envio	-0.030826	-0.112649	0.058073	0.062817

Fuente: Elaboración propia

La suma de las cargas al cuadrado, como se observa en la ilustración [6-41], arroja un aporte significativo solo en el Factor 1, 2 y 3. Al analizar “Proportion Var” tiene una baja tasa en comparación al resto. Y al llegar al “Cumulative Var” se acumula más de 35% en el Factor 3 y 4. De este análisis, se puede concluir que al escoger entre cualquiera de los dos dataset el comportamiento es similar, teniendo un margen menor de variación de resultados. Esto concuerda en cuanto a la reducción de dimensiones, los aportes de las variables en dichos Factores, como en la estadística inmersa por cada factor. En cuanto al código de fuente utilizado, véase al archivo “*Analisis Factorial.ipynb*”

*Ilustración 6-41: Dataset N°2, estadística de los factores*

	Factor1	Factor2	Factor3	Factor4
SS Loadings	1.540794	1.021831	0.992096	0.196360
Proportion Var	0.154079	0.102183	0.099210	0.019636
Cumulative Var	0.154079	0.256262	0.355472	0.375108

*Fuente: Elaboración propia*

## 6.6. Definición de Variables Input y Target

Una vez generado los conjuntos de datos de pruebas, haber realizado sus respectivos análisis estadísticos y factorial, se definen las variables independientes y dependientes para ser utilizado en el modelamiento de redes neuronales. Asimismo, es importante destacar que todos los valores en las variables son de tipo flotante (decimales) y se encuentra en un rango entre 0 y 1. A continuación, se menciona las siguientes variables independientes (Input) que se utilizan, cuyo total es 10 dimensiones:

1. Sexo.
2. Edad.
3. Segmento.
4. Subsegmento.
5. Agrupación de segmento.
6. Carterizado.
7. Aperturas.
8. Hora de envío.
9. Mes de envío.
10. Día de la semana.

Cabe mencionar que, como se ha mencionado en muchas oportunidades, la variable objetivo (Target) es el campo “RESPONDIDA”, cuyo valor es binario. Para el campo día de envío, no se incluye debido a que este presenta diferencias con otros meses (ejemplo, febrero tiene 29 y diciembre cuenta 31 días). Por otro lado, los campos “Duración” y “Duración\_i\_f” son valores (días) aproximados, donde estos solo se usan para generar estadísticas con relación a los días transcurridos para que el envío sea respondido y, por otro lado, que ocurra otro envío. En cuanto a las fechas, estos se utilizan únicamente para extraer y calcular campos. Por último, no se utiliza E-mail ya que solo representa una persona y son valores únicos, y esto incluye Id de envío con mención a este último punto.

# Capítulo 7.

## Modelado y Evaluación

En este capítulo, se describe la segmentación de los conjuntos de datos de prueba para entrenamiento y evaluación (prueba), los diseños y parámetros que se asigna al modelo de redes neuronales. Además, se detalla los resultados de evaluaciones de distintos modelos generados, donde en este apartado se selecciona aquellos que muestra una mejor precisión y de baja pérdida de evaluación, en ellas se realiza predicción para validar las cantidades de envíos con respuesta. Por último, se muestra la interfaz gráfica confeccionada (prototipo), donde en ella se detalla el uso del programa y la forma de como despliegan los resultados, con relación a comparación entre cantidad de envíos con respuestas reales y predicha.

### 7.1. Segmentación Data en Grupo de Entrenamiento y Evaluación

Antes de empezar con modelamiento de redes neuronales, es importante dividir los conjuntos de datos, donde uno se utiliza para enviar los datos a las redes neuronales para entrenar su capacidad de predecir, y por otro lado se usa para realizar pruebas y evaluar las predicciones de contactabilidad, donde se compara las cantidades de respuesta real y otras predichas. Para este caso, dado la alta incidencia de envíos sin respuesta, no es recomendable particionar los datasets de prueba mediante distribución de porcentajes, tanto para entrenamiento (train) como de evaluación (test), ya que existe un gran riesgo de que los envíos con respuesta estén uno de ellos y no ambos.

Como consecuencia, las redes neuronales pueden reconocer todos los envíos que tienen respuesta, pero en el dataset de prueba no existen otros que estén contestadas (o hay muy pocos), peor aún si es un caso inverso. Por lo tanto, se decide dividir utilizando el campo de mes de envío como de referencia. Para este caso, los registros de datos para el entrenamiento se consideran los primeros seis meses del año ( $\text{Mes\_envio} < 0,5$ ), es decir, enero, febrero, marzo, abril, mayo y junio. Mientras que, para el resto de los otros 6 meses restantes, se consideran para realizar pruebas de predicción y obtener así las precisiones y porcentaje de pérdida de valores.

Tras particionar los dos conjuntos de datos, como se visualiza en la tabla [7-1], se logra observar las cantidades de envíos que se encuentran distribuidos, donde el 54,85% del dataset de prueba N°1 se utiliza en el proceso del entrenamiento, mientras que el 45,15% restante se usa para evaluar los modelos generados. En cuanto al dataset de prueba N°2, el 60,86% del dicho conjunto de datos son destinados para procesar en el entrenamiento del modelo, mientras que el resto de 39,14% se utiliza para la evaluación de la red neuronal.

*Tabla 7-1: División de los datasets para entrenamiento y de prueba*

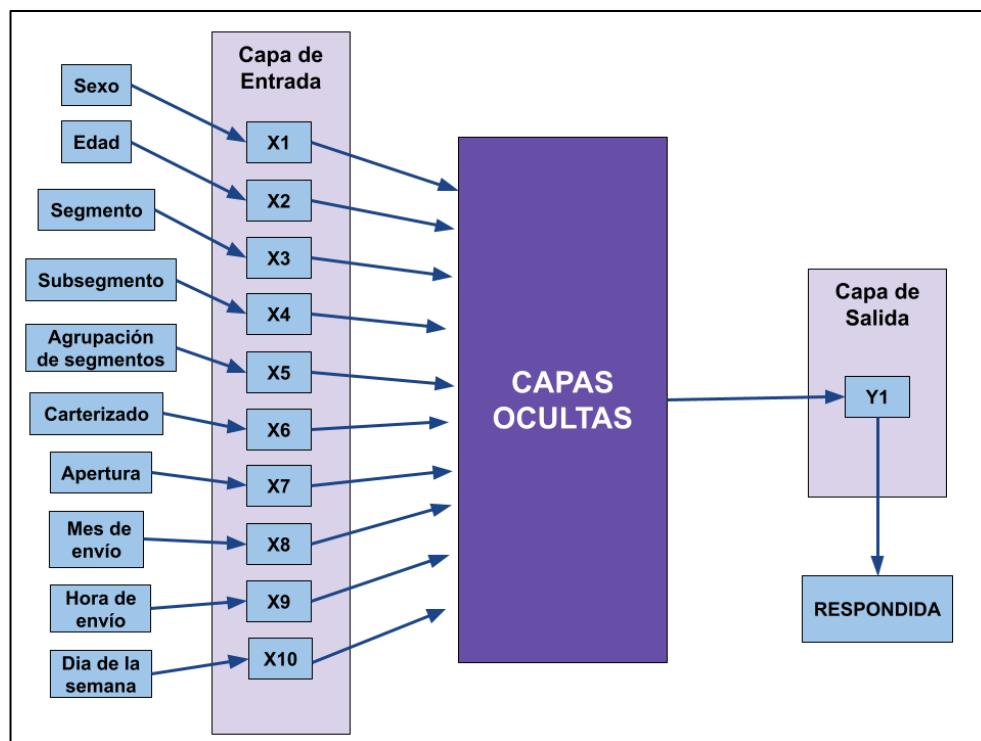
<b>DESCRIPCIÓN</b>	<b>ENTRENAMIENTO</b>		<b>PRUEBA</b>	
	<b>DATASET N°1</b>	<b>DATASET N°2</b>	<b>DATASET N°1</b>	<b>DATASET N°2</b>
Cantidad de envíos	2.189.349	1.520.840	1.802.338	978.229
Porcentaje de partición	54,85%	60,86%	45,15%	39,14%
Cantidad de envíos con respuesta	56.944	56.944	50.199	50.199
Cantidad de envíos sin respuesta	2.132.405	1.463.896	1.752.139	978.229

*Fuente: Elaboración propia*

## 7.2. Diseño de Arquitectura y Definición de Parámetros

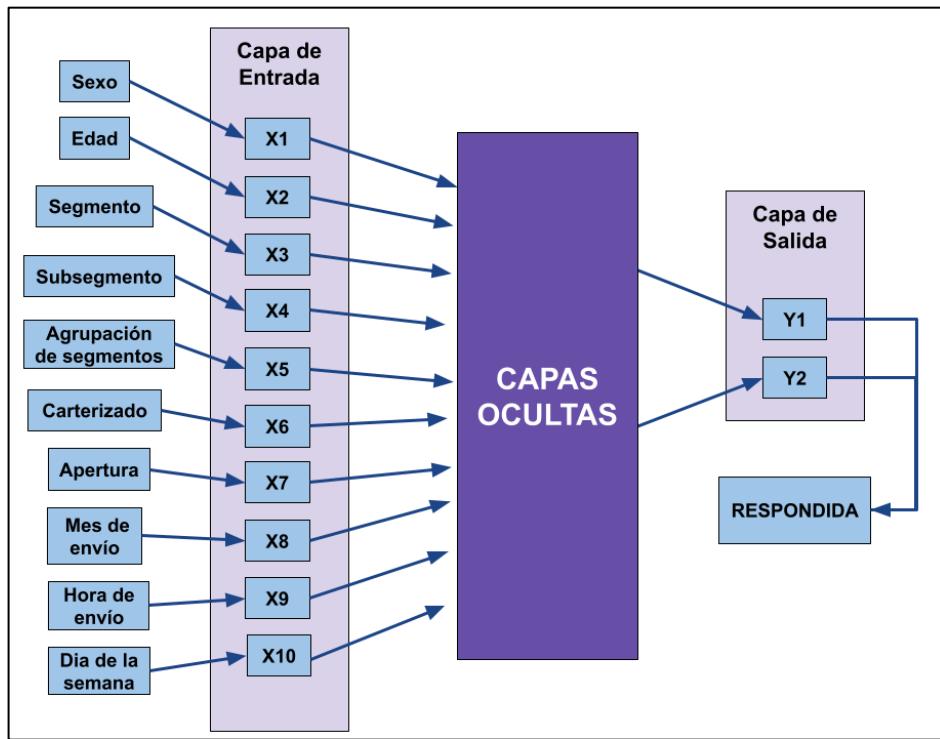
En el ámbito de inteligencia artificial, existen muchas formas de diseñar arquitecturas de redes neuronales, donde este incluye las cantidades de capas ocultas y sus respectivos parámetros. Como se ha mencionado en Marco Teórico (2.2.2), los parámetros que se configuran para la elaboración de arquitectura son N° de neuronas, funciones de activación, de pérdida y optimizadores. También se encuentra el inicializador de Kernel, lo cual se considera como el filtro que se aplica a un conjunto de datos, con el fin de extraer ciertas características importantes o patrones de esta. Para empezar, se propone dos diseños generales para formar arquitecturas, donde uno contiene un nodo de salida y otra contiene dos de ellas. Estos se visualizan a partir de las siguientes ilustraciones.

Ilustración 7-1: Diseño N°1 de arquitectura de una red neuronal artificial, una salida



Fuente: Elaboración propia

Ilustración 7-2: Diseño N°2 de arquitectura de una red neuronal artificial, dos salidas



Fuente: Elaboración propia

Para ambos diseños, se plantea un estándar con relación a los parámetros de capas de entrada, de salida, inicializador de Kernel, épocas, Batch Size y función de pérdida. Para el caso de inicializador de Kernel, en ambos diseños, se utiliza la distribución “Normal” para aplicar filtros al conjunto de datos de entrada. En cuanto a la función de pérdida, se utiliza la entropía cruzada, lo cual es una medida de la distancia entre distribuciones de probabilidad. Por el lado de la capa de entrada, como son 10 variables independientes, se asigna un total de 10 cantidades de nodos (neuronas) para ambos diseños. Además, para dicha capa se utiliza la función de activación “RELU”, ya que este tiene la ventaja de que reciben tal cual todos los datos mayor o igual a 0, y, como el rango de todos los campos es entre 0 y 1, no existe ninguna pérdida de valores alguna.

En cuanto a la capa de salida, para el primer diseño (ilustración [7-1]) se asigna un nodo y una función de activación “RELU”, con el fin de ver el comportamiento que tiene al desplegar los resultados de precisión de entrenamiento, predicción y probabilidad de que la encuesta sea respondida. Mientras que, para el segundo diseño (ilustración [7-2]), se asigna dos nodos y una función de activación “Softmax”, ya que se tienen dos clases de variables de salida. Para este caso, es necesario aplicar la función “*np\_utils.to\_categorical(y)*”, ya que este genera dos dimensiones en la variable objetivo, donde el valor [1,0] corresponde a envíos que no tienen respuestas, y el otro valor [0,1] indica que otros envíos si han respondido. Por lo tanto, los resultados de salida tienen relación la probabilidad de que el envío se haya respondido o no, quedando de la siguiente forma:

- $P_1(X)$ : Probabilidad de que la encuesta no sea respondida.
- $P_2(X)$ : Probabilidad de que la encuesta sea respondida

$$Y = "[P_1(X), P_2(X)]"$$

Por otro lado, los parámetros de épocas y Batch size (tamaño de datos) se asignan distintos valores de acuerdo con el tamaño de los conjuntos de datos a entrenar, y la cantidad de veces que se ejecuten los algoritmos, donde en este caso se tienen dos distintas. Para el caso del dataset N°1, se asigna 1000 épocas y 2200 de tamaño, mientras que, para el dataset N°2, se asignan también 1000 épocas, pero con un tamaño de 1550 de datos. Cabe destacar que, mientras mayor sea el valor de la época, ya que cada vez mejora la precisión, y que en la ejecución estén involucrado muchos datos, mejor será la predicción del modelo de red neuronal artificial.

Sin embargo, dado lo anterior, implica un consumo de recursos computacionales, por lo que se necesita más tiempo hasta terminar con el entrenamiento. Para ahorrar tiempo y recursos, se implementa una función “Early Stopping” (parada temprana), que, según Peltarion (2021), permite detener automáticamente el entrenamiento cuando una métrica elegida ha dejado de mejorar, evitando así un sobreajuste. Para este caso, se tienen 3 parámetros que influye la función, donde estos se mencionan a continuación:

- “***monitor***”: Corresponde a la métrica que se analiza durante el entrenamiento del modelo, lo cual este se asigna como “loss” para monitorear las pérdidas de valores.
- “***min\_delta***”: Corresponde a la condición de que tan bajo puede ser la diferencia del monitor, lo cual este se asigna un valor “0”, implicando que entre épocas debe tener los mismos valores de pérdida.
- “***patience***”: Corresponde la paciencia que desea que se detenga temprano para usar. Por lo tanto, el valor que se asigna son números de épocas que se cumple la condición de que monitor sea “*min\_delta*”, antes de parar. Para este caso, se asigna valor 5, lo cual significa que, si se cumple 5 veces dicha condición, se detiene el entrenamiento del modelo, para así proceder con su evaluación.

Una vez definido el estándar del diseño, se debe asignar al azar los parámetros de optimizador y ciertas cantidades de capas ocultas ( $N$ ), donde cada uno contiene número de nodos y función de activación. Asimismo, los optimizadores que se utilizan para entrenar son SGD, ADAM, Adamax y Adagrad. En cuanto a las funciones de activación, se usan RELU, Sigmoid, Tanh y Softmax. Por otro lado, para asignar la cantidad nodos en cada una de las capas ocultas ( $CO$ ), es importante que se cumplan los siguientes criterios:

1. En la primera capa oculta ( $CO_1$ ), los números de nodos debe ser siempre mayor que 10. En otras palabras, debe superar la cantidad de nodos asignados en la capa de entrada.
2. En la capa oculta intermedia o más profunda ( $CO_{N/2}$ ), debe estar asignado una cantidad de nodos más alto que el resto.
3. Para las primeras capas oculta, hasta llegar a una intermedia ( $CO_1, CO_2, \dots, CO_{N/2}$ ), se debe asignar número de nodos de forma creciente.
4. Para las ultimas capas ocultas, después de llegar a la capa intermedia (desde  $CO_{N/2}$  a  $CO_N$ ), se debe asignar número de nodos de forma decreciente.
5. En la última capa oculta ( $CO_N$ ), los números de nodos debe ser siempre mayor que otras cantidades de nodos correspondiente a la capa de salida. Es decir, para el diseño N°1 debe asignar más de un nodo, y en cuanto al diseño N°2 debe tener más de dos nodos.

Por lo tanto, si se cumplen todos estos criterios, puede existir una proporcionalidad en los nodos de cada capa, incluyendo las de entrada y salida. Para terminar, como se debe enviar al entrenamiento dos datasets de pruebas distintas, se debe generar dos resultados del modelo, precisión y valores perdidos en cada arquitectura y por separado. Además, se deben proceder el entrenamiento 5 veces (iteraciones), con el fin de analizar el comportamiento y estabilidad en los resultados de entrenamiento y evaluación de distintos modelos. En este apartado, se genera 40 arquitecturas de redes neuronales en total, donde cada 20 de ellas corresponde a un diseño distinto, 20 del diseño N°1 y otros 20 del diseño N°2. Todos estos diseños y parámetros se pueden ver en el anexo 10.2.

### 7.3. Resultados de Evaluación de Modelos

Tras ejecutar los códigos de fuentes elaborados en distintos entornos de desarrollo y ejecución, donde este se menciona en el anexo 10.3, se tienen resultados distintos con relación a la evaluación de modelos, obteniendo los porcentajes de precisión, error y valores de pérdida. Para el caso del diseño N°1 propuesto (véase a ilustración [7-1]), todos los modelos generados, arquitecturas de RNA N°1 a 20 como referencia, presentan los mismos resultados de evaluación (precisión, error y pérdida de valores), donde estos se visualizan en la tabla [10-53] proveniente desde el anexo 10.4. Además, al momento de aplicar la predicción de algunos modelos este diseño, mediante la función “*predict(X)*” de Keras, la mayoría los valores de salida son iguales o se encuentran cerca del 0.

Sin embargo, ningún valor de probabilidad está cerca del 0,5, dado que no es posible determinar la probabilidad de que envíos sean respondidas con mayor certeza. Por lo tanto, el diseño N°1 de arquitectura, junto con su estándar y parámetros que se asignan, fracasaron, y esto implica por dos principales motivos.

- Las configuraciones de los parámetros no son adecuadas al diseño, principalmente el estándar que se planteó (funciones de activación de salida, de pérdida y Kernel).
- Por la alta incidencia de envíos que no tienen respuesta en ambos datasets.

En cuanto al diseño N°2 (véase a ilustración [7-2]), a diferencia de la anterior, los resultados de evaluación si han mejorado, implicando que los ajustes de parámetros si son adecuados. Para este caso, existen diversos resultados en los modelos, por lo que se eligen las 3 arquitecturas (6 modelos) que muestra un mejor desempeño en la evaluación, para utilizarlos en la validación de resultados mediante matriz de confusión, y la predicción de contactabilidad mediante interfaz gráfica. Los criterios para seleccionar las arquitecturas que presentan un mejor desempeño, se mencionan a continuación:

1. Todas aquellas arquitecturas que presentan un promedio de precisión alta, bajo porcentaje de error y con desviación estándar muy baja, donde el valor 0 implica que no hay diferencia en las 5 iteraciones, en ambos modelos (uno entrenado con dataset N°1 y otra dataset N°2), son buenos candidatos.

2. Todas las arquitecturas que presentan porcentajes de pérdidas de valores más bajas son buenos candidatos, es decir, si dos modelos tienen promedios de precisiones similares, se elige uno que ha perdido menos de estos valores.
3. Se eligen aquellas arquitecturas que presentan precisiones muy similares entre dos modelos, donde uno es entrenado usando el dataset N°1 y N°2 en cuanto a otro. Esto indica que el modelo presenta una estabilidad en el entrenamiento.

Al recopilar todos los datos con relación a la evaluación de modelos, se seleccionan tres arquitecturas de redes neuronales artificiales, que demuestren su mejor desempeño y que cumpla los criterios de selección. Como resultado, las arquitecturas de Redes Neuronales Artificiales N°23 (tabla [10-23]), 27 (tabla [10-39]) y 36 (tabla [10-48]), son los que se destacan como modelos mejor evaluados, siendo que este último (arquitectura N°36) presenta un mejor rendimiento.

Como mención honorifica, la arquitectura N°28 (tabla [10-40]) se ha presentado con un porcentaje de precisión más alto y estable que otros modelos, lo cual se refleja en la tabla [10-59]. Sin embargo, este se muestra un porcentaje de pérdida de valores más alto que las tres arquitecturas, y por esta razón no se ha seleccionado como el mejor. A partir de las siguientes tablas, se visualizan los resultados de precisión, error, el promedio de precisión y la desviación estándar por cada modelo, según el dataset que se ha utilizado para el entrenamiento. Para ver los resultados de evaluación en las otras arquitecturas, véase a anexo 10.4.

Tabla 7-2: Resultados de evaluación arquitectura de RNA N°23

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,15%	5,08%	1,85%	98,53%	0,002
2	98,77%	2,73%	1,23%		
3	98,50%	3,68%	1,50%		
4	98,64%	2,66%	1,36%		
5	98,58%	2,68%	1,42%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,33%	2,74%	1,67%	98,45%	0,003
2	98,38%	2,72%	1,62%		
3	97,98%	2,84%	2,02%		
4	98,76%	2,49%	1,24%		
5	98,77%	2,52%	1,23%		

Fuente: Elaboración propia

Tabla 7-3: Resultados de evaluación arquitectura de RNA N°27

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,12%	2,57%	1,88%	98,26%	0,001
2	98,31%	2,77%	1,69%		
3	98,30%	2,77%	1,70%		
4	98,27%	2,77%	1,73%		
5	98,30%	2,77%	1,70%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,14%	2,78%	1,86%	98,04%	0,003
2	97,44%	2,77%	2,56%		
3	98,23%	2,73%	1,77%		
4	98,21%	2,78%	1,79%		
5	98,21%	2,79%	1,79%		

Fuente: Elaboración propia

Tabla 7-4: Resultados de evaluación arquitectura de RNA N°36

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,62%	4,06%	1,38%	98,66%	0,000
2	98,62%	3,92%	1,38%		
3	98,67%	2,59%	1,33%		
4	98,70%	2,51%	1,30%		
5	98,69%	3,17%	1,31%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,74%	2,55%	1,26%	98,67%	0,001
2	98,71%	2,53%	1,29%		
3	98,76%	3,26%	1,24%		
4	98,69%	2,57%	1,31%		
5	98,44%	4,59%	1,56%		

Fuente: Elaboración propia

### 7.3.1. Análisis y observaciones en los resultados:

Tras conseguir todos los resultados de evaluaciones, lo cual se encuentra en el anexo 10.4, y haber identificado las tres mejores arquitecturas en cuanto a su desempeño, se logra adquirir las siguientes observaciones al respecto:

- El algoritmo ADAM ha demostrado ser el optimizador indicado para la generación de modelos, ya que demuestra tener mayor estabilidad en los resultados de precisiones, errores y bajas pérdidas de valores. Además, las tres mejores arquitecturas mencionadas anteriormente tienen asignado este optimizador. Por lo tanto, para el caso de los dos dataset, se logra aprovechar la ventaja de que su tasa de aprendizaje se adapte en el entrenamiento de modelos, a medida que se vaya ejecutando las épocas.

- En las capas ocultas, al asignar una función de activación Softmax, se presenta inestabilidad en los resultados de precisiones, y aumenta el porcentaje de valores perdidos. Por lo tanto, para futuras configuraciones, no es recomendable utilizar esta función y solo considerarlo como salida.
- Al asignar reiteradas veces la función de activación RELU en las capas ocultas, si bien, los resultados de evaluación y predicción pueden ser buenas al existir una fuerte relación directa entre variables, y que ningún valor se encuentre debajo de 0. Sin embargo, como el dataset solo presenta relación directa entre campos “Apertura” y “RESPONDIDA”, y el resto demuestra tener leve relación (o ninguna), por lo que no mejora la precisión del modelo.
- En cuanto a los números de neuronas que se asignan en las capas ocultas, todas las arquitecturas que tienen precisiones altas y con mayor estabilidad en las iteraciones, esto ocurre debido a que existe una buena distribución entre cantidades de nodos, similar a lo que se refleja en la ilustración [2-3] del marco teórico. Por otro lado, no se presenta cambios al asignar números de nodos altos, y sucede lo mismo al sumar cantidad de capas ocultas. Por lo tanto, se puede generar un buen modelo de red neuronal si existe una clara proporcionalidad sobre los números de nodos en todas las capas (entrada, oculta y salida).

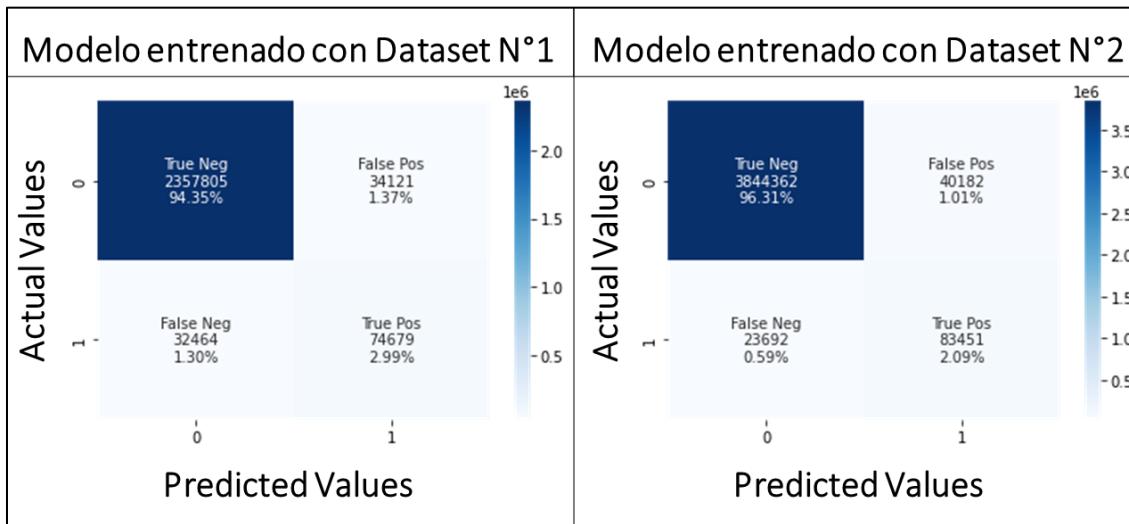
- Al usar algoritmo de optimización Adamax, demuestra tener una mayor inestabilidad en las precisiones que otros optimizadores. Además, este optimizador presenta mucha pérdida de valores mientras que la red neuronal se encuentre en entrenamiento. Por lo que, para problemas de alta incidencia de un evento (envíos sin respuesta) y baja en el otro evento (envíos con respuesta), no es recomendable usarlo.
- El algoritmo SGD ha demostrado ser eficiente, en cuanto a la mitigación sobre pérdidas de valores. Sin embargo, las precisiones y errores que se generan no son mejores si se compara con el otro optimizador ADAM. Por lo tanto, si se busca una configuración los parámetros adecuados en las capas ocultas, sin duda, los resultados de evaluación pueden mejorar. En caso de que este no mejore, es debido a que se enfrenta el problema de alta incidencia de uno de los dos eventos.
- Para el caso del algoritmo Adagrad, si se compara con los resultados entre arquitectura N°24 (tabla [10-56]) y 31 (tabla [10-62]), no demuestra ser un optimizador adecuado para aquellos modelos que tenga más de 3 capas ocultas, ya que presenta inestabilidad en las precisiones, errores y con mayor pérdida de valores. Además, si se observa en las asignaciones de funciones de activación, en la arquitectura N°31 (tabla [10-43]) se encuentra presenta la función sigmoide, mientras que la otra arquitectura N°24 (tabla [10-36]) no cuenta con dicha función. Por lo tanto, es posible que esta función afecta de forma negativa los resultados de evaluación para aquellos modelos que utilice Adagrad.

## 7.4. Validación de Modelos, Matriz de Confusión

Tras identificar las tres arquitecturas mejor evaluadas, se procede a validar los resultados de la predicción, mediante la matriz de confusión. Para este caso, se utiliza los datasets completos (sin particionar) para cada modelo, donde se extraen las variables independientes para predecir. En cuanto a la variable objetivo (RESPONDIDA), se usa en el momento de generar matriz de confusión, para comparar los valores de salidas obtenidos mediante predicción de modelos. Por lo tanto, para proceder con la ejecución de matriz de confusión, se utiliza el dataset N°1 para todos los modelos que haya sido entrenados mediante dataset N°2, y viceversa en cuanto a otro modelo. A continuación, se detallan los resultados de validación de estos modelos por cada arquitectura, donde se mencionan los porcentajes de aciertos, indicando las coincidencias en los valores de salidas de predicción con otros reales. Dada la alta incidencia de los envíos sin contestar (0), se analiza únicamente los que si se han respondido (1)

- En las matrices de confusiones para la arquitectura N°23, como se visualiza en la ilustración [7-3], se tiene 108.800 envíos con respuestas predichas para el modelo entrenado con dataset N°1, donde el 68,64% coinciden con los valores objetivos reales. Mientras que con los otros 31,36% restantes, se produjeron errores, en los cuales se detecta que los envíos supuestamente habían sido respondidos, lo cual no fue así. En cuanto al otro modelo entrenado con dataset N°2, se obtiene 123.633 envíos contestados, donde en ella indica que hay 67,50% de aciertos. Sin embargo, los otros 32,50% de estos señalan desaciertos en los valores objetivos.

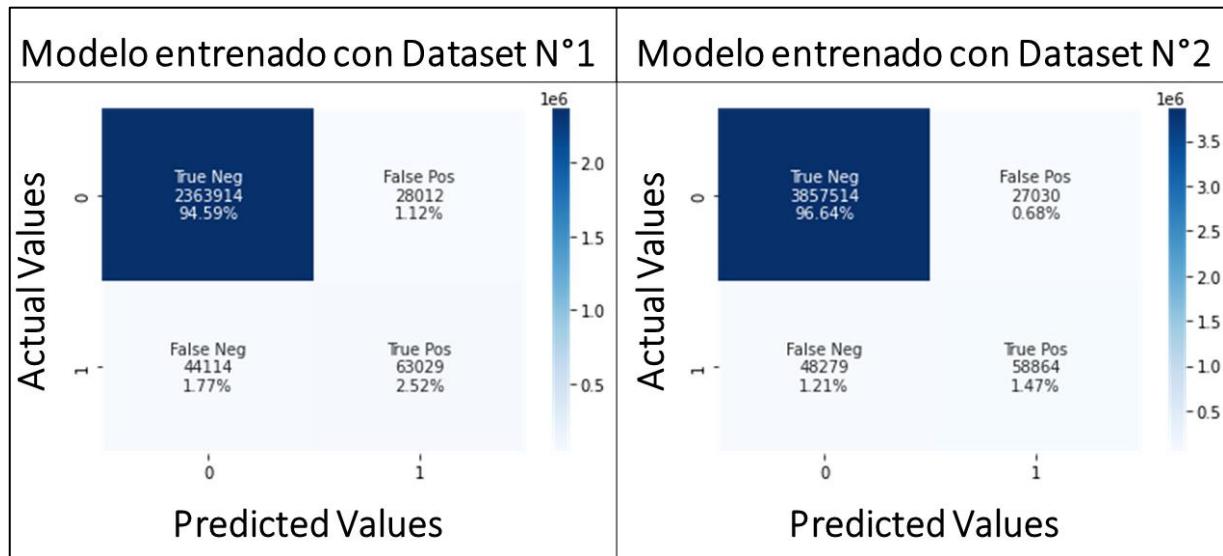
Ilustración 7-3: Validación de modelos de arquitectura N°23, matriz de confusión



Fuente: Elaboración propia

- En las matrices de confusiones para la arquitectura N°27, como se visualiza en la ilustración [7-4], se tiene 91.041 envíos con respuestas predichas para el modelo entrenado con dataset N°1, donde el 69,23% coinciden con los valores objetivos reales. Mientras que con los otros 30,77% restantes, se produjeron errores, en los cuales se detecta que los envíos supuestamente habían sido respondidos, lo cual no fue así. En cuanto al otro modelo entrenado con dataset N°2, se obtiene 85.894 envíos contestados, donde en ella indica que hay 68,53% de aciertos. Sin embargo, los otros 31,47% de estos señalan desaciertos en los valores objetivos.

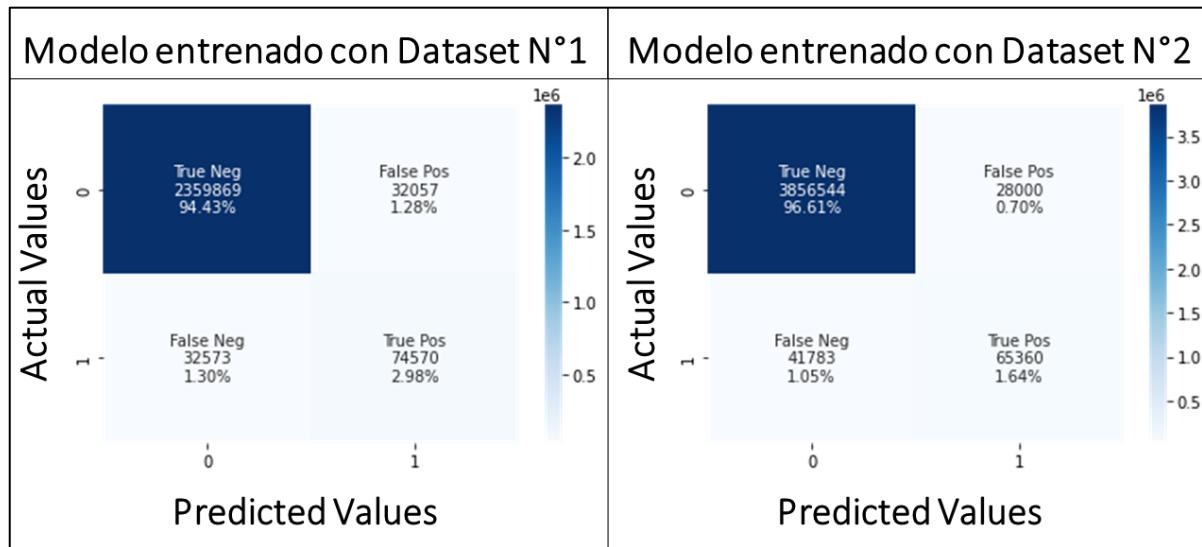
Ilustración 7-4: Validación de modelos de arquitectura N°27, matriz de confusión



Fuente: Elaboración propia

- En las matrices de confusiones para la arquitectura N°36, como se visualiza en la ilustración [7-5], se tiene 106.627 envíos con respuestas predichas para el modelo entrenado con dataset N°1, donde el 69,94% coinciden con los valores objetivos reales. Mientras que con los otros 30,06% restantes, se produjeron errores, en los cuales se detecta que los envíos supuestamente habían sido respondidos, lo cual no fue así. En cuanto al otro modelo entrenado con dataset N°2, se obtiene 93.360 envíos contestados, donde en ella indica que hay 70,01% de aciertos. Sin embargo, los otros 29,99% de estos señalan desaciertos en los valores objetivos.

Ilustración 7-5: Validación de modelos de arquitectura N°36, matriz de confusión



Fuente: Elaboración propia

Por lo tanto, tras analizar todas las validaciones en las predicciones de la variable objetivo, la arquitectura N°36 ha tenido un porcentaje de aciertos más altos que otras. Además, en otro punto de vista, esta arquitectura se ha presentado muy bajos desaciertos que el resto, por lo que los modelos se han demostrado estables con la detección de envíos que tienen respuestas o no. Asimismo, si se compara con dos modelos de esta arquitectura, se ha demostrado que el segundo (entrenado con dataset N°2) ha tenido un mejor desempeño, y esto es producto sobre la reducción drástica sobre registros de envíos sin responder. En el anexo 10.5, se encuentran los detalles con relación al código utilizado para generar matriz de confusión.

## 7.5. Prototipo Elaborado

Tras evaluar todos los modelos de redes neuronales, identificar los tres modelos con mejor desempeño, y analizar las matrices de confusión para validar los valores generados en la predicción, se construye una interfaz gráfica con el fin de seguir realizando experimentos. Dado el fracaso del primer diseño de arquitectura de RNA, este prototipo soporta solamente modelos correspondientes al diseño N°2 (de dos salidas). Este tiene una funcionalidad principal de generar resultados de la predicción del modelo, en base al dataset normalizado que se ingresa, donde este es generado a partir del capítulo 6 sobre tratamiento de outliers (6.3) y transformación de los datos (6.4). Para este caso, se utiliza los mismos dos archivos CSV de pruebas, referenciando al dataset N°1 y 2. Asimismo, dado la alta incidencia de envíos sin respuesta, se prioriza analizar solamente los resultados de predicción de envíos respondidas. Cabe destacar que, en este prototipo, se encuentran implementado las validaciones de archivos entrantes, donde estos se detallan en el anexo 10.6.

Para empezar, la interfaz gráfica del prototipo se encuentra ubicada en la carpeta “/prototipo”. Una vez ubicado, es necesario instalar todas las librerías que son necesarias para su ejecución, cuyas indicaciones se encuentran detallada en “*Readme.txt*” de la misma carpeta. Una vez ejecutado el programa, en caso de no existir inconveniente, se genera una ventana principal, lo cual se visualiza en la ilustración [7-6]. Para este caso, se debe buscar los dos archivos adecuados, pulsando el botón “Buscar”, donde uno corresponde al modelo de red neuronal entrenado (h5 de extensión) y otro al dataset que se utiliza para predecir (CSV de extensión, con “;” como separador). Tras ingresar estos dos archivos requeridos, se debe pulsar el botón “*Empezar*”, para proceder con la estimación de respuesta horaria y la predicción.

Después, se debe esperar entre 1 a 5 minutos, mientras que el programa esté en proceso. Cuando termine, se despliega una ventana de resultados, donde muestra las cantidades totales de respuestas reales, precedida, y como estos se encuentran distribuidos en el horario. Cabe destacar que, en la vista previa de esta ventana, se visualiza en la ilustración [10-21] proveniente del anexo 10.6. Para proceder con el experimento, al igual que en la sección de validación de modelos, se realiza predicciones en las tres arquitecturas mejores evaluadas, donde consta de 6 modelos dado que cada uno son entrenado con dos datasets distintos. Por lo que, se utiliza el dataset N°1 y 2 para la predicción de cada modelo, y así comparar los resultados.

Cuando esta ejecución termine, se despliega una ventana donde se tiene resultados de comparación entre cantidades reales de envíos respondidas y otras predichas. Además, se visualiza la distribución de estas respuestas en un determinado horario, cuyas proporciones de frecuencias son basadas en la tabla [6-8]. Como resultado, en cuanto a los datos reales, se tiene un horario de lunes a domingo, entre a las 8:00 a 21:00, donde cada celda representa cantidad de envíos enviados y que se haya contestado. Este horario de contactabilidad se visualiza en la tabla [7-5], donde el jueves a las 11:00 horas se tiene mayor retención de respuestas.

Por otra parte, las bajas retenciones de respuestas se encuentran concentradas los fines de semana, es decir, sábado y domingo. En cuanto a las horas, si se envían encuestas a las 8:00 y 21:00, también se detectan baja tasa de respuesta, por lo que es recomendable enviarlas entre a las 10:00 a 17:00 horas para los miércoles y jueves. Este resultado se repite si se utiliza los mismos datasets de pruebas, y también sirve como de referencia al comparar los resultados de predicción, para visualizar que tan cerca se encuentra con la cantidad de respuestas reales.

Ilustración 7-6: Ventana principal del prototipo



Fuente: Elaboración propia

Tabla 7-5: Resultados estimados de las respuestas reales dentro de un horario determinado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	911	975	1125	1168	954	804	611
9	1093	1179	1350	1414	1146	964	729
10	1125	1221	1393	1457	1189	996	761
11	1221	1318	1511	1575	1286	1071	814
12	1061	1146	1307	1361	1114	932	707
13	1146	1232	1414	1479	1200	1007	771
14	1061	1136	1307	1361	1114	932	707
15	1136	1221	1404	1457	1189	996	761
16	1125	1221	1393	1457	1189	996	761
17	1168	1264	1446	1511	1232	1029	782
18	1061	1146	1318	1371	1125	943	718
19	1136	1221	1404	1468	1189	996	761
20	1071	1157	1329	1393	1125	943	718
21	546	600	686	718	579	482	375

Fuente: Elaboración propia

En cuanto a los resultados de predicción de contactabilidad, sabiendo que la cantidad real de envíos respondidas son 107.143, como se visualiza en la tabla [7-6], se muestra una comparación de resultados sobre respuestas predichas, porcentajes precisión y error predicción. Tras analizar esta comparativa, se destaca que, ya demostrado en validaciones de modelos, que la arquitectura N°36 ha estado muy cerca del resultado real, con 99,04% de precisión, cuyo modelo que ha sido entrenado con dataset N°1. Sin embargo, como ya se ha demostrado su matriz de confusión (ilustración [7-5]), existen registros que detectan mal en el momento de predecir, es decir, algunas cantidades de envíos contestados son falsos. Por lo tanto, a pesar de la precisión, se requiere seguir ajustando parámetros al modelo para mejorar la detección de envíos con y sin respuesta, lo cual implica generar nuevos modelos a través del entrenamiento, pero con arquitectura similar.

*Tabla 7-6: Resultados estimados general de las respuestas predichas*

Nº	NOMBRE DEL MODELO (.h5)	DATASET N° 1			DATASET N°2		
		RESPUESTAS PREDICHAS	PRECISIÓN DE PREDICCIÓN (%)	ERROR DE PREDICCIÓN (%)	RESPUESTAS PREDICHAS	PRECISIÓN DE PREDICCIÓN (%)	ERROR DE PREDICCIÓN (%)
1	Red23_D1S2	110.620	96,75%	3,25%	108.800	98,45%	1,55%
2	Red23_D2S2	123.633	84,61%	15,39%	121.771	86,35%	13,65%
3	Red27_D1S2	92.344	86,19%	13,81%	91.041	84,97%	15,03%
4	Red27_D2S2	85.894	80,17%	19,83%	84.748	79,10%	20,90%
5	Red36_D1S2	108.173	99,04%	0,96%	106.627	99,52%	0,48%
6	Red36_D2S2	93.360	87,14%	12,86%	92.538	86,37%	13,63%

Fuente: Elaboración propia

En cuanto a los resultados de respuesta predicha dentro de un horario determinado, las observaciones son las mismas con los datos reales de acuerdo con la concentración de tasa de respuesta, tanto en las horas de envíos como los días de semana. Para el caso del modelo que ha sido entrenado con dataset N°1, correspondiente a la arquitectura N°36, los resultados de predicción horaria se visualizan en las siguientes tablas, donde primero se utilizó el dataset N°1 para predecir y la segunda con dataset N°2. En cuanto a los otros resultados de otros 8 modelos, se encuentran en el anexo 10.11.

*Tabla 7-7: Resultado de predicción horaria del Modelo Red36\_D1S2, dataset N°1 usado*

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	919	984	1.136	1.179	963	811	617
9	1.103	1.190	1.363	1.428	1.157	974	736
10	1.136	1.233	1.406	1.471	1.201	1.006	768
11	1.233	1.331	1.525	1.590	1.298	1.082	822
12	1.071	1.157	1.320	1.374	1.125	941	714
13	1.157	1.244	1.428	1.493	1.212	1.017	779
14	1.071	1.147	1.320	1.374	1.125	941	714
15	1.147	1.233	1.417	1.471	1.201	1.006	768
16	1.136	1.233	1.406	1.471	1.201	1.006	768
17	1.179	1.276	1.460	1.525	1.244	1.038	790
18	1.071	1.157	1.331	1.385	1.136	952	725
19	1.147	1.233	1.417	1.482	1.201	1.006	768
20	1.082	1.168	1.341	1.406	1.136	952	725
21	552	606	692	725	584	487	379

Fuente: Elaboración propia

*Tabla 7-8: Resultado de predicción horaria del Modelo Red36\_D1S2, dataset N°2 usado*

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
<b>8</b>	906	970	1.120	1.162	949	800	608
<b>9</b>	1.088	1.173	1.344	1.407	1.141	960	725
<b>10</b>	1.120	1.216	1.386	1.450	1.184	992	757
<b>11</b>	1.216	1.312	1.503	1.567	1.280	1.066	810
<b>12</b>	1.056	1.141	1.301	1.354	1.109	928	704
<b>13</b>	1.141	1.226	1.407	1.471	1.194	1.002	768
<b>14</b>	1.056	1.130	1.301	1.354	1.109	928	704
<b>15</b>	1.130	1.216	1.397	1.450	1.184	992	757
<b>16</b>	1.120	1.216	1.386	1.450	1.184	992	757
<b>17</b>	1.162	1.258	1.439	1.503	1.226	1.024	778
<b>18</b>	1.056	1.141	1.312	1.365	1.120	938	714
<b>19</b>	1.130	1.216	1.397	1.461	1.184	992	757
<b>20</b>	1.066	1.152	1.322	1.386	1.120	938	714
<b>21</b>	544	597	682	714	576	480	373

Fuente: Elaboración propia

Para terminar, este prototipo se caracteriza por contar una interfaz gráfica simple y amigable, todo gracias a las ventajas de diseño que proporciona la herramienta Qt Designer, haciendo que su desarrollo sea más intuitivo y rápido. Además, la principal característica de este prototipo es su extensibilidad, ya que cualquier persona, quienes desean continuar con este proyecto, pueda añadir funcionalidades adicionales, sumando así más valor agregado al proyecto. Por otro lado, como Qt Designer, lenguaje Python y sus librerías se pueden utilizar y ejecutar en cualquier sistema operativo, Windows, Linux y Mac como referencia, entonces el prototipo también cuenta con la misma característica, correspondiente a la adaptabilidad.

# Capítulo 8.

## Conclusiones

Tras completar todas las etapas de CRISP-DM, se logra concluir, mediante análisis de datos y despliegue de resultados (sobre estimación de las respuestas reales dentro de un horario determinado), que hay mayor retención de respuestas al realizar un envío para el jueves a las 11:00 horas. Si bien, los días miércoles y jueves han recibido mayor cantidad de respuestas, por ende, es factible enviarlos entre las 10:00 a 17:00, ya que los demás horarios demuestran una baja contestabilidad en las personas encuestadas. Además, no es factible enviar las encuestas durante los fines de semana (sábado y domingo), dado que, posiblemente, los encuestados tienden a desconectarse de los dispositivos electrónicos (computador, celulares y tablets). Por lo tanto, dado todos los antecedentes y conocimientos generados en el proyecto, la empresa “*Activa Research*” debe tomar medidas para mejorar la tasa de respuesta de sus clientes.

Por otro lado, tras generar modelos de redes neuronales con ajustes de parámetros diferentes, se logra seleccionar tres arquitecturas (6 modelos) con relación al segundo diseño propuesto (ilustración [7-2]). El principal motivo de que estos fueron seleccionados, es que presentan precisiones altas y estables, con un bajo porcentaje de pérdida de valores. Tras proceder con la validación de estos modelos, se comprueba que la arquitectura N°36 ha tenido mayor porcentaje de aciertos respecto a la detección de envíos que tienen o no respuestas, siendo que, el modelo, que ha sido entrenado usando el dataset N°2, tuvo un mejor desempeño en la predicción.

Por último, se elaboró un prototipo para agregar valor y realizar experimentos a esta investigación, donde se logra comparar entre datos de la variable objetivo real y predecido, y cómo estas se distribuyen en un determinado horario. Tras utilizar los dos datasets de pruebas y cargar los 6 modelos con mejor desempeño en la evaluación, se destaca que la arquitectura N°36 ha estado muy aproximada del resultado real (cantidad de respuestas reales), siendo que el modelo (que ha sido entrenado usando el dataset N°1), tuvo una mayor precisión en cuanto a la aproximación del resultado real. Sin embargo, con los resultados de validación de modelos obtenidos mediante la matriz de confusión, esta arquitectura alcanza hasta un 70% de coincidencias con los datos reales, por lo que aún se requiere una mejora en la generación de estos modelos entrenados.

Por lo tanto, dada la hipótesis propuesta en el capítulo 4, no existe aún un modelo que permite predecir si una persona responderá una encuesta (con un 90% de precisión y en un determinado horario). Para que esta hipótesis se cumpla, es necesario contar con frecuencias más equilibradas entre cantidad de envíos contestados y respuestas. Además, se debe mejorar en el modelamiento de redes neuronales artificiales, donde esto implica ajustar nuevos parámetros, modificar la cantidad de capas ocultas, utilizar otro estándar (capas de entrada, salida, función de pérdida y Kernel).

Para los futuros trabajos de este proyecto, quienes desean continuar con su desarrollo, pueden utilizar la misma metodología CRISP-DM, para así actualizar al nuevo contexto de la problemática, mejorar las soluciones dadas, recopilar y analizar nuevos datos sobre historial de Sendinblue y características de las personas. Además, puede tener distintas formas de preparar un nuevo conjunto de datos, donde se espera que se alcance un aumento de incidencias de envíos contestados y que exista relación entre las variables independientes y objetivo.

Para el caso del modelamiento de redes neuronales artificiales, existen muchas formas de crear arquitecturas y ajustar los parámetros de entrenamiento del modelo. Cabe señalar que existen más optimizadores a probar, entre ellos se encuentran Adadelta, FTRL, Nadam y RMSprop. Por otro lado, además de la distribución “Normal”, existen otros filtros (inicializadores de Kernels) que se aplican a un conjunto de datos, lo cual se puede ver un listado de ellos en proyecto de grado, desarrollado por Arreaza (2008). Entre ellos, el otro filtro más conocido es la distribución “Uniforme”.

En cuanto al prototipo elaborado con interfaz gráfica, es mejorable, por lo que, dado que este cuenta con la característica de extensibilidad, se propone añadir funcionalidades como las predicciones unitarias de acuerdo con los datos de entrada que ingresa el usuario. Para este caso, se requiere implementar al prototipo una lógica que implique la transformación de datos entrantes, obligando a que todos los valores sean de tipo flotante y que se encuentren dentro del rango [0,1]. Otras de las funcionalidades que se puede agregar, es que reciban modelos de distintos diseños de redes neuronales. También, como se está obligando a que los datasets contengan su variable objetivo (dado por la funcionalidad de comparar resultados de predicción), es bueno recibir una gran cantidad de registro que solo predijera cuántos envíos responden y cuantos no, generando así nuevas estadísticas de predicción.

# Capítulo 9.

## Referencias Bibliográficas

- El Mostrador (2017, junio). “*Eduardo Engel critica tasas de respuesta de las encuestas: Son un motivo para no creerles a los márgenes de errores*”. Recuperado de 9 de abril de 2021, de <https://www.elmostrador.cl/noticias/pais/2017/06/04/eduardo-engel-critica-tasas-de-respuesta-de-las-encuestas-son-un-motivo-para-no-creerles-a-los-margenes-de-errores/>
- Question Pro (2018, diciembre). “*Encuestas CAWI: ¿qué son y cómo usarlas?*”. Recuperado de 9 de abril de 2021, de <https://www.questionpro.com/blog/es/encuestas-cawi/>
- Sngular. (2019, agosto). “*CRISP-DM: La metodología para poner orden en los proyectos*”. Recuperado de 27 de marzo de 2021, de <https://www.sngular.com/es/data-science-crisp-dm-metodologia/>
- IBM Knowledge Center (s. f.). “*Conceptos básicos de ayuda de CRISP-DM*”. Recuperado de 27 de marzo de 2021, de [https://www.ibm.com/support/knowledgecenter/es/SS3RA7\\_sub/modeler\\_crispdm\\_ddita/clementine/crisp\\_help/crisp\\_overview.html](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html)
- Healthdataminer (2019, 4 noviembre). “*CRISP-DM: Una metodología para minería de datos en salud*”. Recuperado de 27 de marzo del 2021, de <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>

- Agencia B12 (2020, febrero). “Qué es un modelo predictivo y cómo se aplica al negocio”. Recuperado de 04 de abril del 2021, de <https://agenciab12.com/noticia/que-es-modelo-predictivo-como-aplica-negocio#:~:text=Un%20modelo%20predictivo%20es%20un,vez%2C%20detectar%20oportunidades%20de%20negocio.>
- I. Senra (2020, junio). “¿Qué es modelo predictivo? Definición, significado y ejemplos”. Recuperado de 02 de Mayo del 2021, de. <https://www.arimetrics.com/glosario-digital/modelo-predictivo>
- C. Collao (2020). “Análisis de Componentes Principales y de Factorial”. Class notes INFB8104, Departamento de informática, Universidad Tecnológica Metropolitana.
- S. Fernández (2011). “Análisis factorial”. Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid. Recuperado de 09 de Mayo del 2021, de <https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIAL/analisis-factorial.pdf>
- P. Royo (2021, marzo). “Qué son las redes neuronales y cuál es su aplicación en el marketing”. Recuperado de 09 de Mayo del 2021, de. <https://artyco.com/que-son-las-redes-neuronales-y-cual-es-su-lingüística-en-el-marketing/>
- I. Mavrou (2015). “Análisis factorial exploratorio”. Universidad Antonio de Nebrija. Recuperado de 09 de Mayo del 2021, de <https://www.nebrija.com/revista-lingüística/analisis-factorial-exploratorio.html>
- C. Rebato (2020, abril). “¿Qué son las redes neuronales artificiales y cómo funcionan?”. Recuperado de 09 de Mayo del 2021, de. <https://empresas.blogthinkbig.com/redes-neuronales-artificiales/>

- IBM Corporation. (s. f.). “*El modelo de redes neuronales*”. Recuperado 08 de mayo de 2021, de <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model>
- E. Galindo., J. Perdomo & J. Figueroa (2020, febrero). “*Estudio comparativo entre máquinas de soporte vectorial multiclasificación, redes neuronales artificiales y sistema de inferencia neuro-difuso auto organizado para problemas de clasificación*”. Información Tecnológica, Vol. 31. Recuperado de 08 de mayo de 2021, de [https://scielo.conicyt.cl/scielo.php?script=sci\\_arttext&pid=S0718-07642020000100273](https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-07642020000100273)
- Team Keras (s. f.). “*Keras documentation: Keras API reference*”. Recuperado de 24 de Noviembre del 2021, de. <https://keras.io/api/>
- TensorFlow (s. f.). “*TensorFlow Documentation: tf.keras.Sequential*”. Recuperado de 24 de Noviembre del 2021, de. [https://www.tensorflow.org/api\\_docs/python/tf/keras/Sequential](https://www.tensorflow.org/api_docs/python/tf/keras/Sequential)
- R. Alberto (2020, octubre). “*Explicación de las Funciones de activación en Redes Neuronales y práctica con Python*”. Recuperado de 24 de Noviembre del 2021, de. <https://rubialesalberto.medium.com/explicaci%C3%B3n-funciones-de-activaci%C3%B3n-y-pr%C3%A1ctica-con-python-5807085c6ed3>
- D. Calvo (2018, diciembre). “*Función de activación – Redes neuronales*”. Recuperado de 24 de Noviembre del 2021, de. <https://www.diegocalvo.es/funcion-de-activacion-redes-neuronales/>
- F. Ramírez (2021, junio). “*Las matemáticas del Machine Learning: Funciones de activación*”. Recuperado de 24 de Noviembre del 2021, de. <https://empresas.blogthinkbig.com/las-matematicas-del-machine-learning-funciones-de-activacion/>

- M. Rivera (2018, agosto). “*Descenso de Gradiente Estocástico (SGD)*”. Recuperado de 25 de Noviembre del 2021, de. [http://personal.cimat.mx:8181/%7Emrivera/cursos/optimizacion/descenso\\_grad\\_estocastico/descenso\\_grad\\_estocastico.html#descenso-de-gradiente-estoc%C3%A1stico-sgd](http://personal.cimat.mx:8181/%7Emrivera/cursos/optimizacion/descenso_grad_estocastico/descenso_grad_estocastico.html#descenso-de-gradiente-estoc%C3%A1stico-sgd)
- Brutalk (2021). “*Introducción suave al algoritmo de optimización de Adam para el aprendizaje profundo*”. Recuperado de 25 de Noviembre del 2021, de. <https://www.brutalk.com/en/news/brutalk-blog/view/introduccion-suave-al-algoritmo-de-optimizacion-de-adam-para-el-aprendizaje-profundo-60471b4df29fa>
- Programador Clic (2021). “*Algoritmo de optimizador común (Adam SGD)*”. Recuperado de 25 de Noviembre del 2021, de. <https://programmerclick.com/article/1375926681/>
- J. Duchi (2011). “*Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*”. Journal of Machine Learning Research. Recuperado de 27 de Noviembre del 2021, de. <https://www.jmlr.org/papers/v12/duchi11a.html>
- D. Kingma & J. Lei Ba (2015, julio). “*ADAM: A Method for Stochastic Optimization*”. Paper, Vol. 8. Recuperado de 28 de Noviembre del 2021, de. <https://arxiv.org/pdf/1412.6980v8.pdf>
- J. Ordóñez (2021, abril). “*Sendinblue*”. Recuperado de 22 de mayo de 2021, de <https://jordiob.com/herramientas-ecommerce/sendinblue/>
- Sendinblue (s. f.). “*Entender los informes estadísticos de sus e-mails*”. Recuperado de 22 de mayo de 2021, de <https://help.sendinblue.com/hc/es/articles/208858829-Entender-los-informes-estad%C3%Adsticos-de-sus-e-mails>

- S. Bernués (2021, octubre). *La carterización de Clientes*. Recuperado de 03 de octubre de 2021, de <https://www.marketingdepymes.com/comercial/gestion-de-clientes/la-carterizacion-de-clientes/>
- Peltarion (2021). “*Early stopping of deep learning experiments*”. Recuperado de 3 de Diciembre del 2021, de. <https://peltarion.com/knowledge-center/documentation/modeling-view/run-a-model/early-stopping>
- E. Arreaza (2008, octubre). “*Impacto de distintos Kernels en el tamaño de muestra para generar variables aleatorias con Bootstrap*”. Universidad de los Andes, Venezuela. Recuperado de 18 de Diciembre del 2021, de [http://bdigital.ula.ve/storage/pdftesis/pregrado/tde\\_arquivos/8/TDE-2009-09-29T16:37:16Z-634/Publico/Arreaza%20Eden.pdf](http://bdigital.ula.ve/storage/pdftesis/pregrado/tde_arquivos/8/TDE-2009-09-29T16:37:16Z-634/Publico/Arreaza%20Eden.pdf)

# Capítulo 10.

## Anexos

### 10.1. Estadísticas CAWI, Estados de Encuestas Enviadas:

#### 10.1.1. Enero del año 2020

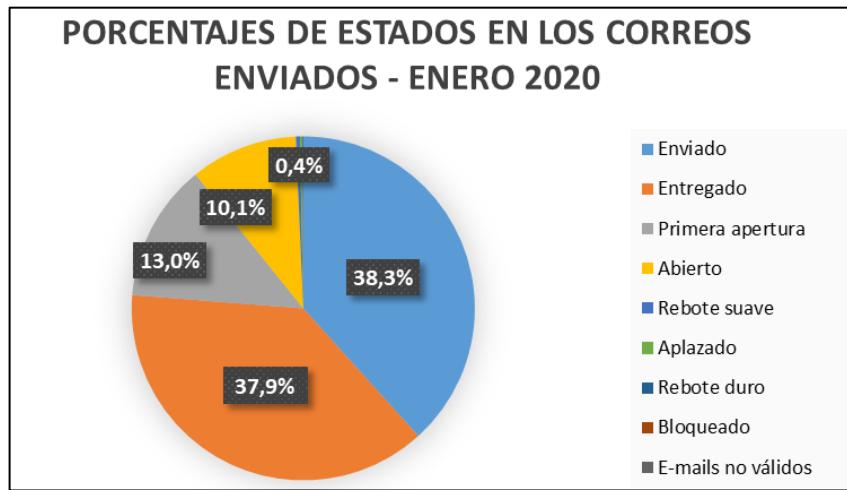
A partir de la tabla [10-1] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de enero del 2020, junto con su gráfico circular donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-1].

Tabla 10-1: Cantidad de encuestas enviadas con su respectivo estado Enero del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	794.465	38,3%
Entregado	786.903	37,9%
Primera apertura	268.688	13,0%
Abierto	209.776	10,1%
Rebote suave	8.033	0,4%
Aplazado	4.235	0,2%
Rebote duro	1.495	0,1%
Bloqueado	262	< 0,1%
E-mails no válidos	103	< 0,1%
Queja	85	< 0,1%
Suscripción cancelada	5	< 0,1%
Clicado	1	< 0,1%
<b>TOTAL</b>	<b>2.074.051</b>	

Fuente: Elaboración propia

Ilustración 10-1: Porcentajes de estados en los correos enviados Enero 2020



Fuente: Elaboración propia

### 10.1.2. Febrero del año 2020

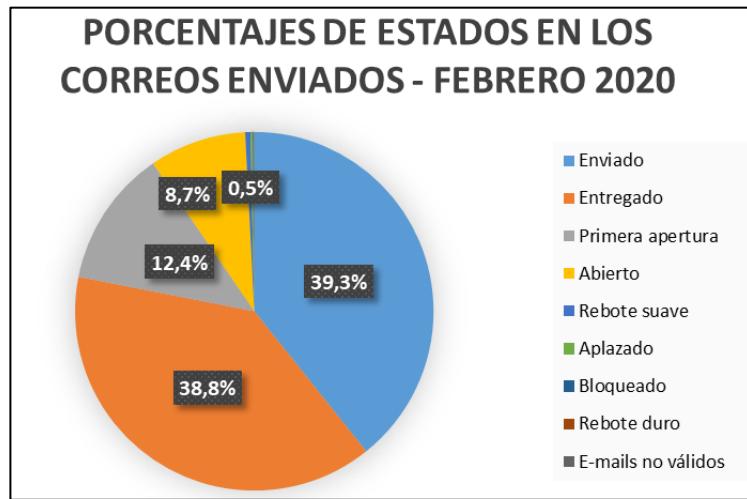
A partir de la tabla [10-2] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de febrero del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-2].

Tabla 10-2: Cantidad de encuestas enviadas con su respectivo estado Febrero del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	678.358	39,3%
Entregado	670.374	38,8%
Primera apertura	213.598	12,4%
Abierto	150.946	8,7%
Rebote suave	8.205	0,5%
Aplazado	3.733	0,2%
Bloqueado	1.529	0,1%
Rebote duro	461	< 0,1%
E-mails no válidos	101	< 0,1%
Queja	90	< 0,1%
Clicado	3	< 0,1%
Suscripción cancelada	2	< 0,1%
<b>TOTAL</b>	<b>1.727.400</b>	

Fuente: Elaboración propia

Ilustración 10-2: Porcentajes de estados en los correos enviados Febrero 2020



Fuente: Elaboración propia

### 10.1.3. Marzo del año 2020

A partir de la tabla [10-3] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de marzo del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-3].

Tabla 10-3: Cantidad de encuestas enviadas con su respectivo estado Marzo del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	949.210	39,7%
Entregado	928.509	38,8%
Primera apertura	294.986	12,3%
Abierto	196.081	8,2%
Rebote suave	11.175	0,5%
Aplazado	6.570	0,3%
Bloqueado	3.567	0,1%
Rebote duro	609	< 0,1%
Queja	238	< 0,1%
E-mails no válidos	118	< 0,1%
Clicado	5	< 0,1%
Suscripción cancelada	4	< 0,1%
<b>TOTAL</b>	<b>2.391.072</b>	

Fuente: Elaboración propia

Ilustración 10-3: Porcentajes de estados en los correos enviados Marzo 2020



Fuente: Elaboración propia

#### 10.1.4. Abril del año 2020

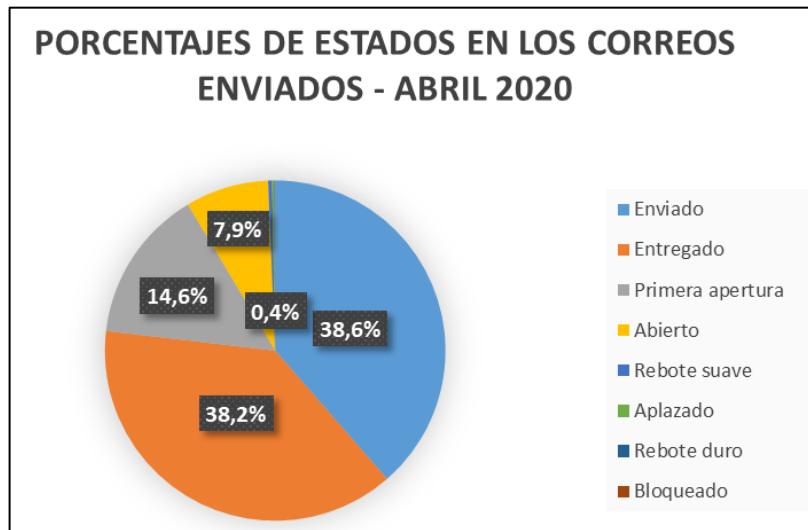
A partir de la tabla [10-4] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de abril del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-4].

Tabla 10-4: Cantidad de encuestas enviadas con su respectivo estado Abril del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	405.421	38,6%
Entregado	401.761	38,2%
Primera apertura	153.207	14,6%
Abierto	83.009	7,9%
Rebote suave	3.917	0,4%
Aplazado	2.187	0,2%
Rebote duro	600	0,1%
Bloqueado	180	< 0,1%
E-mails no válidos	55	< 0,1%
Queja	38	< 0,1%
Suscripción cancelada	1	< 0,1%
Clicado	1	< 0,1%
<b>TOTAL</b>	<b>1.050.377</b>	

Fuente: Elaboración propia

Ilustración 10-4: Porcentajes de estados en los correos enviados Abril 2020



Fuente: Elaboración propia

### 10.1.5. Mayo del año 2020

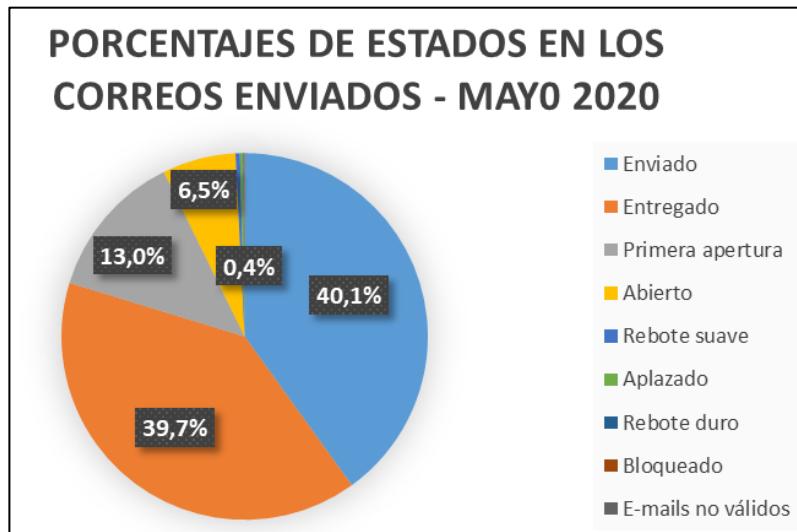
A partir de la tabla [10-5] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de mayo del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-5].

Tabla 10-5: Cantidad de encuestas enviadas con su respectivo estado Mayo del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	623.924	40,1%
Entregado	617.837	39,7%
Primera apertura	202.490	13,0%
Abierto	100.781	6,5%
Rebote suave	6.176	0,4%
Aplazado	4.263	0,3%
Rebote duro	1.022	0,1%
Bloqueado	940	0,1%
E-mails no válidos	55	< 0,1%
Queja	20	< 0,1%
<b>TOTAL</b>	<b>1.557.508</b>	

Fuente: Elaboración propia

Ilustración 10-5: Porcentajes de estados en los correos enviados Mayo 2020



Fuente: Elaboración propia

### 10.1.6. Junio del año 2020

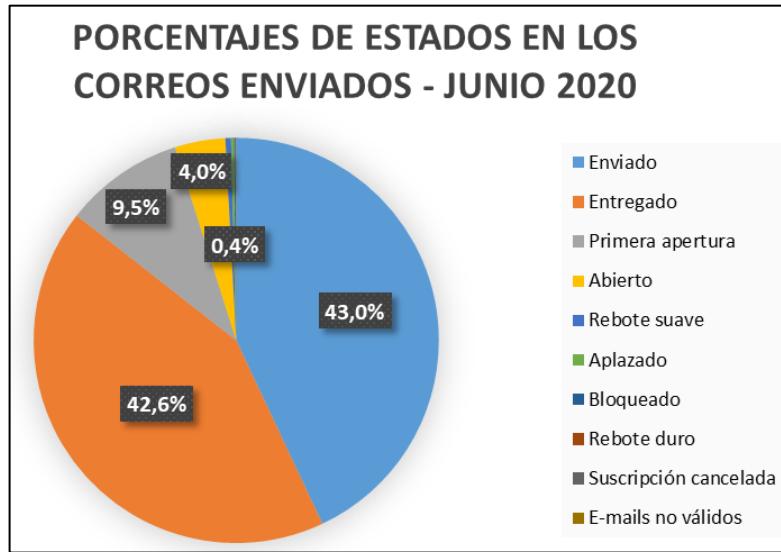
A partir de la tabla [10-6] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de junio del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-6].

Tabla 10-6: Cantidad de encuestas enviadas con su respectivo estado Junio del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	934.210	43,0%
Entregado	924.396	42,6%
Primera apertura	207.413	9,5%
Abierto	87.314	4,0%
Rebote suave	9.712	0,4%
Aplazado	5.102	0,2%
Bloqueado	2.469	0,1%
Rebote duro	817	< 0,1%
Suscripción cancelada	344	< 0,1%
E-mails no válidos	92	< 0,1%
Queja	71	< 0,1%
<b>TOTAL</b>	<b>2.171.940</b>	

Fuente: Elaboración propia

Ilustración 10-6: Porcentajes de estados en los correos enviados Junio 2020



Fuente: Elaboración propia

#### 10.1.7. Julio del año 2020

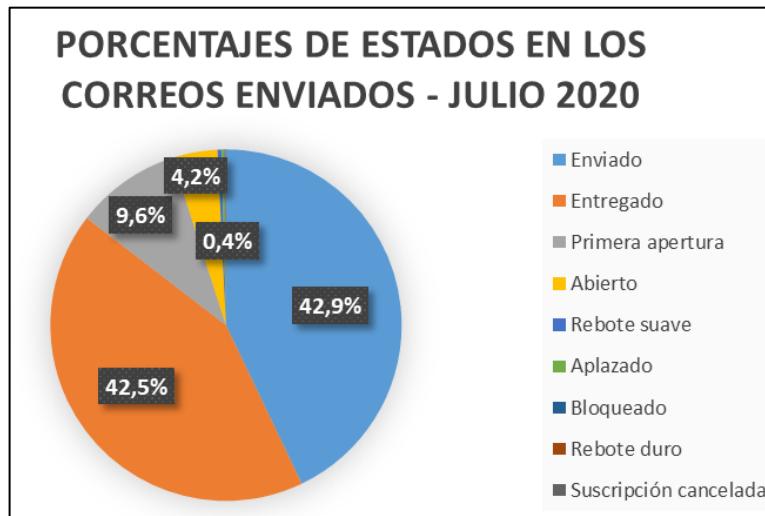
A partir de la tabla [10-7] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de julio del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-7].

Tabla 10-7: Cantidad de encuestas enviadas con su respectivo estado Julio del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	767.242	42,9%
Entregado	760.577	42,5%
Primera apertura	171.784	9,6%
Abierto	74.857	4,2%
Rebote suave	6.823	0,4%
Aplazado	4.296	0,2%
Bloqueado	976	0,1%
Rebote duro	923	0,1%
Suscripción cancelada	626	< 0,1%
E-mails no válidos	70	< 0,1%
Queja	69	< 0,1%
<b>TOTAL</b>	<b>1.788.243</b>	

Fuente: Elaboración propia

Ilustración 10-7: Porcentajes de estados en los correos enviados Julio 2020



Fuente: Elaboración propia

#### 10.1.8. Agosto del año 2020

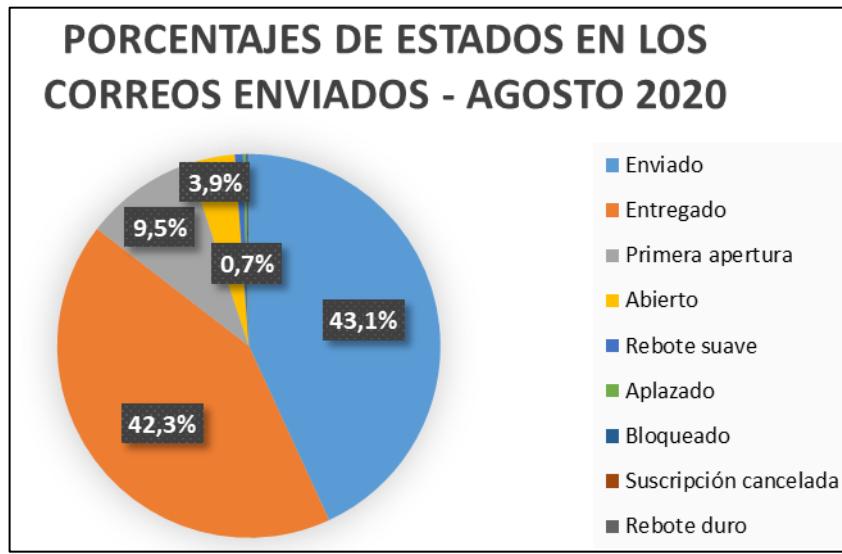
A partir de la tabla [10-8] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de agosto del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-8].

Tabla 10-8: Cantidad de encuestas enviadas con su respectivo estado Agosto del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	1.017.992	43,1%
Entregado	999.641	42,3%
Primera apertura	223.469	9,5%
Abierto	92.376	3,9%
Rebote suave	16.491	0,7%
Aplazado	5.766	0,2%
Bloqueado	3.859	0,2%
Suscripción cancelada	802	< 0,1%
Rebote duro	660	< 0,1%
E-mails no válidos	86	< 0,1%
Queja	71	< 0,1%
<b>TOTAL</b>	<b>2.361.213</b>	

Fuente: Elaboración propia

Ilustración 10-8: Porcentajes de estados en los correos enviados Agosto 2020



Fuente: Elaboración propia

#### 10.1.9. Septiembre del año 2020

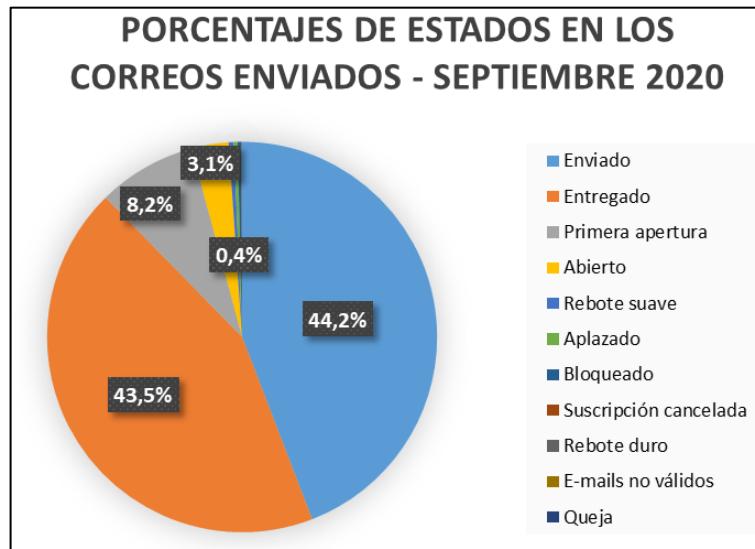
A partir de la tabla [10-9] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de septiembre del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-9].

Tabla 10-9: Cantidad de encuestas enviadas con su respectivo estado Septiembre del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	1.379.574	44,2%
Entregado	1.359.567	43,5%
Primera apertura	255.081	8,2%
Abierto	95.937	3,1%
Rebote suave	13.132	0,4%
Aplazado	9.867	0,3%
Bloqueado	9.357	0,3%
Suscripción cancelada	972	< 0,1%
Rebote duro	789	< 0,1%
E-mails no válidos	117	< 0,1%
Queja	96	< 0,1%
<b>TOTAL</b>	<b>3.124.489</b>	

Fuente: Elaboración propia

Ilustración 10-9: Porcentajes de estados en los correos enviados Septiembre 2020



Fuente: Elaboración propia

#### 10.1.10. Octubre del año 2020

A partir de la tabla [10-10] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de octubre del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-10].

Tabla 10-10: Cantidad de encuestas enviadas con su respectivo estado Octubre del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	1.616.494	44,2%
Entregado	1.591.992	43,5%
Primera apertura	296.162	8,1%
Abierto	111.708	3,1%
Rebote suave	14.848	0,4%
Bloqueado	13.546	0,4%
Aplazado	10.440	0,3%
Suscripción cancelada	1.112	< 0,1%
Rebote duro	441	< 0,1%
E-mails no válidos	122	< 0,1%
Queja	93	< 0,1%
<b>TOTAL</b>	<b>3.656.958</b>	

Fuente: Elaboración propia

Ilustración 10-10: Porcentajes de estados en los correos enviados Octubre 2020



Fuente: Elaboración propia

#### 10.1.11. Noviembre del año 2020

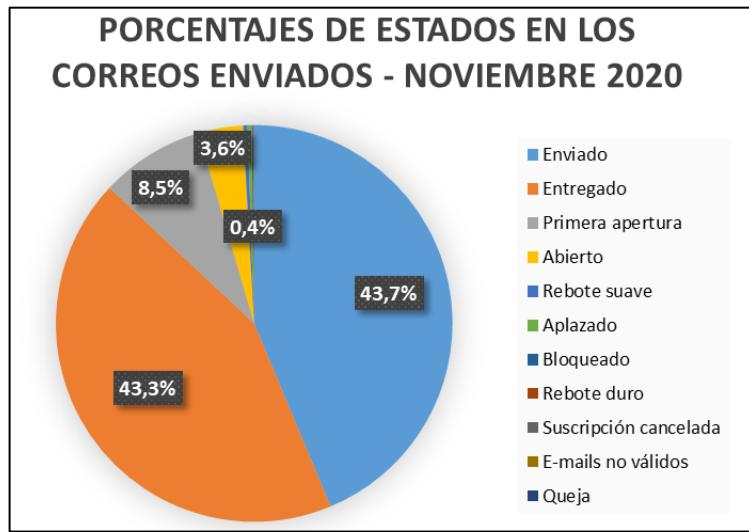
A partir de la tabla [10-11] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de noviembre del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-11].

Tabla 10-11: Cantidad de encuestas enviadas con su respectivo estado Noviembre del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	901.230	43,7%
Entregado	892.215	43,3%
Primera apertura	174.981	8,5%
Abierto	75.004	3,6%
Rebote suave	7.513	0,4%
Aplazado	6.915	0,3%
Bloqueado	2.116	0,1%
Rebote duro	1.138	0,1%
Suscripción cancelada	659	< 0,1%
E-mails no válidos	62	< 0,1%
Queja	45	< 0,1%
<b>TOTAL</b>	<b>2.061.878</b>	

Fuente: Elaboración propia

Ilustración 10-11: Porcentajes de estados en los correos enviados Noviembre 2020



Fuente: Elaboración propia

### 10.1.12. Diciembre del año 2020

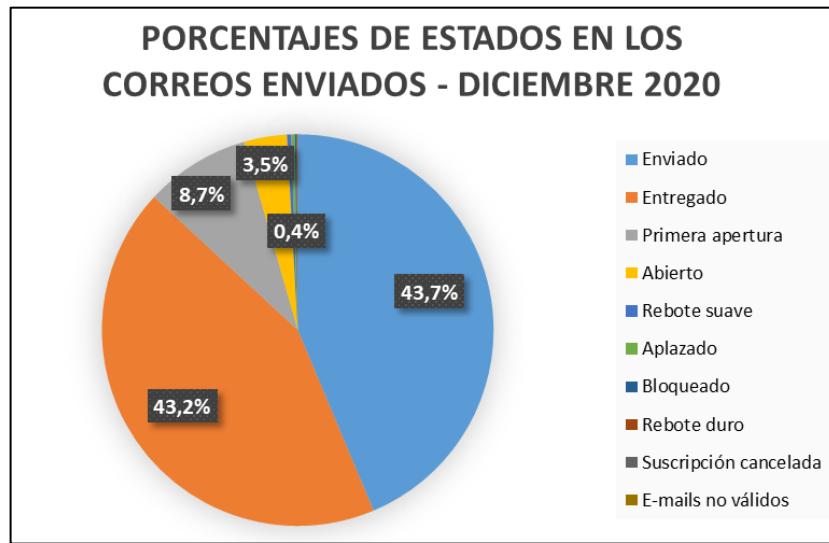
A partir de la tabla [10-12] se observa la cantidad y frecuencia de envíos de encuestas con sus respectivos estados para el mes de diciembre del 2020, junto con su gráfico de torta donde indica los porcentajes que contiene en los estados, lo cual se visualiza en la ilustración [10-12].

Tabla 10-12: Cantidad de encuestas enviadas con su respectivo estado Diciembre del 2020

ESTADOS	CANTIDAD	FRECUENCIA (%)
Enviado	830.550	43,7%
Entregado	821.826	43,2%
Primera apertura	164.807	8,7%
Abierto	66.932	3,5%
Rebote suave	6.787	0,4%
Aplazado	5.572	0,3%
Bloqueado	2.512	0,1%
Rebote duro	1.191	0,1%
Suscripción cancelada	569	< 0,1%
E-mails no válidos	56	< 0,1%
Queja	15	< 0,1%
<b>TOTAL</b>	<b>1.900.817</b>	

Fuente: Elaboración propia

Ilustración 10-12: Porcentajes de estados en los correos enviados Diciembre 2020



Fuente: Elaboración propia

## 10.2. Listado de Diseño de Arquitecturas de Redes Neuronales:

En este apartado, se tiene un listado de arquitecturas para el modelamiento de redes neuronales, donde se tienen asignados cantidades variadas de capas ocultas y con distintos parámetros, como son las funciones de activación, número de neuronas y optimizadores. Además, las arquitecturas pueden ser desplegadas a partir de la función “Summary()”, para este caso es necesario cargar el modelo que se ha generado. Estos están presentados en tablas, los cuales se mencionan a continuación, donde las 20 primeras arquitecturas corresponden al diseño N°1 propuesto en el capítulo 7, mientras que los otros 20 restantes pertenecen al diseño N°2.

Tabla 10-13: Diseño y parámetros de arquitectura RNA N°1

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	40	Relu
	Capa oculta 3	100	Sigmoid
	Capa oculta 4	50	Sigmoid
	Capa oculta 5	25	Sigmoid
	Capa oculta 6	10	Relu
PARÁMETROS	Salida	1	Relu
	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-14: Diseño y parámetros de arquitectura RNA N°2

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	15	Tanh
	Capa oculta 2	30	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-15: Diseño y parámetros de arquitectura RNA N°3

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Sigmoid
	Capa oculta 2	10	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-16: Diseño y parámetros de arquitectura RNA N°4

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Tanh
	Capa oculta 2	40	Elu
	Capa oculta 3	10	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	Adagrad	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-17: Diseño y parámetros de arquitectura RNA N°5

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	40	Tanh
	Capa oculta 2	80	Sigmoid
	Capa oculta 3	35	Sigmoid
	Capa oculta 4	10	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-18: Diseño y parámetros de arquitectura RNA N°6

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Elu
	Capa oculta 2	30	Tanh
	Capa oculta 3	30	Elu
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-19: Diseño y parámetros de arquitectura RNA N°7

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Relu
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	40	Relu
	Capa oculta 4	10	Sigmoid
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-20: Diseño y parámetros de arquitectura RNA N°8

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Relu
	Capa oculta 2	50	Tanh
	Capa oculta 3	40	Sigmoid
	Capa oculta 4	5	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-21: Diseño y parámetros de arquitectura RNA N°9

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Relu
	Capa oculta 2	65	Tanh
	Capa oculta 3	100	Sigmoid
	Capa oculta 4	65	Relu
	Capa oculta 5	30	Tanh
	Capa oculta 6	10	Sigmoid
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-22: Diseño y parámetros de arquitectura RNA N°10

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	60	Tanh
	Capa oculta 3	90	Sigmoid
	Capa oculta 4	60	Relu
	Capa oculta 5	30	Tanh
	Capa oculta 6	15	Sigmoid
PARÁMETROS	Salida	1	Relu
	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-23: Diseño y parámetros de arquitectura RNA N°11

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Sigmoid
	Capa oculta 2	55	Relu
	Capa oculta 3	90	Tanh
	Capa oculta 4	65	Relu
	Capa oculta 5	35	Relu
	Capa oculta 6	10	Sigmoid
PARÁMETROS	Salida	1	Relu
	Optimizador	Adagrad	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-24: Diseño y parámetros de arquitectura RNA N°12

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Relu
	Capa oculta 3	90	Tanh
	Capa oculta 4	120	Relu
	Capa oculta 5	80	Relu
	Capa oculta 6	50	Sigmoid
	Capa oculta 7	20	Sigmoid
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-25: Diseño y parámetros de arquitectura RNA N°13

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Relu
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	25	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-26: Diseño y parámetros de arquitectura RNA N°14

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Elu
	Capa oculta 2	40	Sigmoid
	Capa oculta 3	60	Tanh
	Capa oculta 4	35	Sigmoid
	Capa oculta 5	15	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-27: Diseño y parámetros de arquitectura RNA N°15

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Elu
	Capa oculta 3	70	Relu
	Capa oculta 4	100	Tanh
	Capa oculta 5	65	Relu
	Capa oculta 6	25	Sigmoid
	Capa oculta 7	10	Elu
	Salida	1	Relu
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-28: Diseño y parámetros de arquitectura RNA N°16

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Sigmoid
	Capa oculta 2	55	Elu
	Capa oculta 3	75	Relu
	Capa oculta 4	100	Tanh
	Capa oculta 5	70	Relu
	Capa oculta 6	35	Sigmoid
	Capa oculta 7	15	Elu
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-29: Diseño y parámetros de arquitectura RNA N°17

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Relu
	Capa oculta 3	75	Elu
	Capa oculta 4	100	Sigmoid
	Capa oculta 5	70	Tanh
	Capa oculta 6	35	Relu
	Capa oculta 7	15	Elu
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-30: Diseño y parámetros de arquitectura RNA N°18

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Elu
	Capa oculta 2	50	Relu
	Capa oculta 3	25	Elu
	Capa oculta 4	10	Sigmoid
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-31: Diseño y parámetros de arquitectura RNA N°19

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	45	Softmax
	Capa oculta 3	20	Sigmoid
	Capa oculta 4	5	Sigmoid
	Salida	1	Relu
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-32: Diseño y parámetros de arquitectura RNA N°20

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Softmax
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	25	Tanh
	Capa oculta 4	10	Relu
	Salida	1	Relu
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-33: Diseño y parámetros de arquitectura RNA N°21

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	40	Relu
	Capa oculta 3	100	Sigmoid
	Capa oculta 4	50	Sigmoid
	Capa oculta 5	25	Sigmoid
	Capa oculta 6	10	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-34: Diseño y parámetros de arquitectura RNA N°22

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	15	Tanh
	Capa oculta 2	30	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-35: Diseño y parámetros de arquitectura RNA N°23

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Sigmoid
	Capa oculta 2	10	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-36: Diseño y parámetros de arquitectura RNA N°24

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Tanh
	Capa oculta 2	40	Elu
	Capa oculta 3	10	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	Adagrad	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-37: Diseño y parámetros de arquitectura RNA N°25

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	40	Tanh
	Capa oculta 2	80	Sigmoid
	Capa oculta 3	35	Sigmoid
	Capa oculta 4	10	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-38: Diseño y parámetros de arquitectura RNA N°26

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Elu
	Capa oculta 2	30	Tanh
	Capa oculta 3	30	Elu
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-39: Diseño y parámetros de arquitectura RNA N°27

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Relu
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	40	Relu
	Capa oculta 4	10	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-40: Diseño y parámetros de arquitectura RNA N°28

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Relu
	Capa oculta 2	50	Tanh
	Capa oculta 3	40	Sigmoid
	Capa oculta 4	5	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-41: Diseño y parámetros de arquitectura RNA N°29

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Relu
	Capa oculta 2	65	Tanh
	Capa oculta 3	100	Sigmoid
	Capa oculta 4	65	Relu
	Capa oculta 5	30	Tanh
	Capa oculta 6	10	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-42: Diseño y parámetros de arquitectura RNA N°30

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	60	Tanh
	Capa oculta 3	90	Sigmoid
	Capa oculta 4	60	Relu
	Capa oculta 5	30	Tanh
	Capa oculta 6	15	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-43: Diseño y parámetros de arquitectura RNA N°31

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Sigmoid
	Capa oculta 2	55	Relu
	Capa oculta 3	90	Tanh
	Capa oculta 4	65	Relu
	Capa oculta 5	35	Relu
	Capa oculta 6	10	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	Adagrad	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-44: Diseño y parámetros de arquitectura RNA N°32

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Relu
	Capa oculta 3	90	Tanh
	Capa oculta 4	120	Relu
	Capa oculta 5	80	Relu
	Capa oculta 6	50	Sigmoid
	Capa oculta 7	20	Sigmoid
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-45: Diseño y parámetros de arquitectura RNA N°33

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Relu
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	25	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-46: Diseño y parámetros de arquitectura RNA N°34

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Elu
	Capa oculta 2	40	Sigmoid
	Capa oculta 3	60	Tanh
	Capa oculta 4	35	Sigmoid
	Capa oculta 5	15	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-47: Diseño y parámetros de arquitectura RNA N°35

ITEMS	DESCRIPCIÓN	N° DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Elu
	Capa oculta 3	70	Relu
	Capa oculta 4	100	Tanh
	Capa oculta 5	65	Relu
	Capa oculta 6	25	Sigmoid
	Capa oculta 7	10	Elu
	Salida	2	Softmax
PARÁMETROS	Optimizador	Adamax	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-48: Diseño y parámetros de arquitectura RNA N°36

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	30	Sigmoid
	Capa oculta 2	55	Elu
	Capa oculta 3	75	Relu
	Capa oculta 4	100	Tanh
	Capa oculta 5	70	Relu
	Capa oculta 6	35	Sigmoid
	Capa oculta 7	15	Elu
PARÁMETROS	Salida	2	Softmax
	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-49: Diseño y parámetros de arquitectura RNA N°37

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Sigmoid
	Capa oculta 2	50	Relu
	Capa oculta 3	75	Elu
	Capa oculta 4	100	Sigmoid
	Capa oculta 5	70	Tanh
	Capa oculta 6	35	Relu
	Capa oculta 7	15	Elu
PARÁMETROS	Salida	2	Softmax
	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-50: Diseño y parámetros de arquitectura RNA N°38

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Elu
	Capa oculta 2	50	Relu
	Capa oculta 3	25	Elu
	Capa oculta 4	10	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-51: Diseño y parámetros de arquitectura RNA N°39

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	25	Relu
	Capa oculta 2	45	Softmax
	Capa oculta 3	20	Sigmoid
	Capa oculta 4	5	Sigmoid
	Salida	2	Softmax
PARÁMETROS	Optimizador	ADAM	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

Tabla 10-52: Diseño y parámetros de arquitectura RNA N°40

ITEMS	DESCRIPCIÓN	Nº DE NEURONAS	FUNCIÓN DE ACTIVACIÓN
CAPAS	Entrada	10	Relu
	Capa oculta 1	20	Softmax
	Capa oculta 2	50	Sigmoid
	Capa oculta 3	25	Tanh
	Capa oculta 4	10	Relu
	Salida	2	Softmax
PARÁMETROS	Optimizador	SGD	
	Función de pérdida	Entropía cruzada	

Fuente: Elaboración propia

### 10.3. Código de Fuente Utilizado, Redes Neuronales Artificiales

Para este caso, existen dos funciones declaradas con relación al entrenamiento, evaluación y generación de modelos de redes neuronales, según el diseño de arquitectura planteado en capítulo 7. Cada una de estas funciones reciben como parámetros el dataset de prueba normalizado, optimizador, función de pérdida, un arreglo de capas ocultas, donde cada posición representa N° de neuronas (0) y función de activación (1), cantidad de épocas y Batch size (tamaño de datos). Al finalizar con la ejecución de estas funciones, devuelve tres variables, donde estas son: Precisión, pérdida de valores y modelo generado. A continuación, las funciones elaboradas con lenguaje Python se visualizan en las siguientes ilustraciones.

Ilustración 10-13: Código de fuente, generación de redes neuronales del diseño N°1

```
def red_neuronal_una_salida(df, optimizador, f_perdida, capas, epocas, tamano):
    #Partición de datos de entrenamiento y prueba
    Train = df[(df['Mes_envio']<0.5)]
    X_train = Train.drop('RESPONDIDA', axis=1).to_numpy()
    y_train = Train["RESPONDIDA"].to_numpy()
    Test = df[(df['Mes_envio']>=0.5)]
    X_test = Test.drop('RESPONDIDA', axis=1).to_numpy()
    y_test = Test["RESPONDIDA"].to_numpy()
    #Arquitectura del modelo
    model = Sequential()
    model.add(Dense(10, input_dim = 10, activation='relu'))
    for neuronas, funcion_activacion in capas:
        model.add(Dense(neuronas, activation = funcion_activacion))
    model.add(Dense(1, activation = 'relu', kernel_initializer='normal'))
    model.compile(loss=f_perdida, optimizer=optimizador, metrics=['binary_accuracy'])
    # Ajuste del modelo
    callEar = EarlyStopping(monitor='loss', min_delta=0, patience=5, verbose=1)
    model.fit(X_train, y_train, epochs = epocas, batch_size = tamano, callbacks=[callEar], verbose = 1)
    scores = model.evaluate(X_test, y_test)
    precision = scores[1]
    perdida = scores[0]
    return precision, perdida, model
```

Fuente: Elaboración propia

Ilustración 10-14: Código de fuente, generación de redes neuronales del diseño N°2

```
def red_neuronal_dos_salidas(df, optimizador, f_perdida, capas, epochas, tamano):
    #Particion de datos de entrenamiento y prueba
    Train = df[(df['Mes_envio']<0.5)]
    X_train = Train.drop('RESPONDIDA', axis=1).to_numpy()
    y_train = Train["RESPONDIDA"].to_numpy()
    Test = df[(df['Mes_envio']>=0.5)]
    X_test = Test.drop('RESPONDIDA', axis=1).to_numpy()
    y_test = Test["RESPONDIDA"].to_numpy()
    y_test = np_utils.to_categorical(y_test)
    y_train = np_utils.to_categorical(y_train)
    #Arquitectura del modelo
    model = Sequential()
    model.add(Dense(10, input_dim = 10, activation='relu'))
    for neuronas, funcion_activacion in capas:
        model.add(Dense(neuronas, activation = funcion_activacion))
    model.add(Dense(2, activation = 'softmax', kernel_initializer='normal'))
    model.compile(loss=f_perdida, optimizer=optimizador, metrics=['binary_accuracy'])
    # Ajuste del modelo
    callEar = EarlyStopping(monitor='loss', min_delta=0, patience=5, verbose=1)
    model.fit(X_train, y_train, epochs = epochas, batch_size = tamano, callbacks=[callEar], verbose = 1)
    scores = model.evaluate(X_test, y_test)
    precision = scores[1]
    perdida = scores[0]
    return precision, perdida, model
```

Fuente: Elaboración propia

En cuanto a la automatización de resultados, se puede ver con mayores detalles en los siguientes archivos:

- “generacion\_modelos.py”, cuyo código es elaborado en Visual Studio Code y se usa para ejecutarlo mediante terminal.
- “Redes\_neuronales\_principal.ipynb”, donde es elaborado mediante Colaboratorio de Google, se utiliza para realizar diversas pruebas, evaluaciones, predicciones y validaciones de modelos, lo cual incluye la generación de matriz de confusión.
- “Col\_Oficial\_Redes\_neuronales.ipynb”, donde este es usado en distinto entornos de ejecución mediante Colaboratorio de Google.

## 10.4. Resultados de Evaluación en RNA por Arquitectura

Tras generar los modelos de redes neuronales entrenados, estos son almacenados en archivos “.h5”, y por otra parte se han almacenado los resultados de evaluaciones al terminar con el entrenamiento, lo cual incluye los porcentajes de precisiones y pérdidas de valores. donde se ha realizado 5 iteraciones. A continuación, se muestran estos resultados de acuerdo con las arquitecturas que se han diseñado (véase a anexo 10.2).

Tabla 10-53: Resultados de evaluación arquitectura de RNA N°1 a 20

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,21%	0,00%	2,79%	97,21%	0,000
2	97,21%	0,00%	2,79%		
3	97,21%	0,00%	2,79%		
4	97,21%	0,00%	2,79%		
5	97,21%	0,00%	2,79%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	94,87%	0,00%	5,13%	94,87%	0,000
2	94,87%	0,00%	5,13%		
3	94,87%	0,00%	5,13%		
4	94,87%	0,00%	5,13%		
5	94,87%	0,00%	5,13%		

Fuente: Elaboración propia

Tabla 10-54: Resultados de evaluación arquitectura de RNA N°21

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,72%	3,64%	2,28%	97,90%	0,006
2	98,47%	3,14%	1,53%		
3	98,46%	4,88%	1,54%		
4	97,17%	4,01%	2,83%		
5	97,67%	3,54%	2,33%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,45%	7,71%	1,55%	97,92%	0,007
2	97,50%	5,23%	2,50%		
3	97,37%	5,02%	2,63%		
4	97,40%	7,51%	2,60%		
5	98,88%	5,44%	1,12%		

Fuente: Elaboración propia

Tabla 10-55: Resultados de evaluación arquitectura de RNA N°22

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,31%	6,45%	1,69%	98,34%	0,002
2	98,64%	2,88%	1,36%		
3	98,45%	2,73%	1,55%		
4	98,06%	8,80%	1,94%		
5	98,24%	6,21%	1,76%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,59%	2,85%	1,41%	98,36%	0,004
2	97,65%	3,71%	2,35%		
3	98,57%	2,76%	1,43%		
4	98,76%	2,57%	1,24%		
5	98,21%	3,16%	1,79%		

Fuente: Elaboración propia

Tabla 10-56: Resultados de evaluación arquitectura de RNA N°24

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	96,94%	2,95%	3,06%	97,72%	0,006
2	97,64%	2,94%	2,36%		
3	98,58%	2,83%	1,42%		
4	97,73%	2,95%	2,27%		
5	97,73%	2,95%	2,27%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,31%	2,88%	1,69%	98,52%	0,002
2	98,64%	2,78%	1,36%		
3	98,63%	2,92%	1,37%		
4	98,74%	2,88%	1,26%		
5	98,31%	2,78%	1,69%		

Fuente: Elaboración propia

Tabla 10-57: Resultados de evaluación arquitectura de RNA N°25

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,52%	2,95%	1,48%	98,43%	0,001
2	98,31%	2,94%	1,69%		
3	98,31%	2,83%	1,69%		
4	98,57%	2,95%	1,43%		
5	98,44%	2,95%	1,56%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,64%	2,71%	1,36%	98,62%	0,004
2	99,34%	2,71%	0,66%		
3	98,33%	2,72%	1,67%		
4	98,46%	2,72%	1,54%		
5	98,31%	2,72%	1,69%		

Fuente: Elaboración propia

Tabla 10-58: Resultados de evaluación arquitectura de RNA N°26

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,52%	2,95%	1,48%	98,43%	0,001
2	98,31%	2,94%	1,69%		
3	98,31%	2,83%	1,69%		
4	98,57%	2,95%	1,43%		
5	98,44%	2,95%	1,56%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,64%	2,71%	1,36%	98,62%	0,004
2	99,34%	2,71%	0,66%		
3	98,33%	2,72%	1,67%		
4	98,46%	2,72%	1,54%		
5	98,31%	2,72%	1,69%		

Fuente: Elaboración propia

Tabla 10-59: Resultados de evaluación arquitectura de RNA N°28

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,40%	4,08%	1,60%	98,58%	0,002
2	98,74%	4,05%	1,26%		
3	98,63%	4,06%	1,37%		
4	98,72%	4,08%	1,28%		
5	98,40%	4,06%	1,60%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,40%	4,03%	1,60%	98,42%	0,000
2	98,43%	4,08%	1,57%		
3	98,42%	4,08%	1,58%		
4	98,43%	4,08%	1,57%		
5	98,40%	4,08%	1,60%		

Fuente: Elaboración propia

Tabla 10-60: Resultados de evaluación arquitectura de RNA N°29

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,23%	2,90%	1,77%	98,27%	0,000
2	98,30%	2,77%	1,70%		
3	98,24%	2,77%	1,76%		
4	98,24%	2,77%	1,76%		
5	98,33%	2,77%	1,67%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,22%	3,92%	2,78%	97,33%	0,005
2	97,01%	5,00%	2,99%		
3	97,21%	4,12%	2,79%		
4	98,21%	2,99%	1,79%		
5	97,01%	5,00%	2,99%		

Fuente: Elaboración propia

Tabla 10-61: Resultados de evaluación arquitectura de RNA N°30

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,13%	4,23%	2,87%	97,19%	0,002
2	97,32%	3,23%	2,68%		
3	97,10%	4,00%	2,90%		
4	97,41%	3,10%	2,59%		
5	97,00%	4,23%	3,00%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,71%	4,91%	2,29%	97,57%	0,001
2	97,56%	5,05%	2,44%		
3	97,51%	5,00%	2,49%		
4	97,54%	4,92%	2,46%		
5	97,51%	5,05%	2,49%		

Fuente: Elaboración propia

Tabla 10-62: Resultados de evaluación arquitectura de RNA N°31

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,21%	14,22%	2,79%	97,21%	0,000
2	97,21%	13,51%	2,79%		
3	97,21%	15,05%	2,79%		
4	97,21%	14,00%	2,79%		
5	97,21%	13,37%	2,79%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	94,87%	20,45%	5,13%	94,87%	0,000
2	94,87%	20,38%	5,13%		
3	94,87%	20,42%	5,13%		
4	94,87%	20,28%	5,13%		
5	94,87%	20,42%	5,13%		

Fuente: Elaboración propia

Tabla 10-63: Resultados de evaluación arquitectura de RNA N°32

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,56%	2,91%	2,44%	98,09%	0,003
2	98,20%	2,79%	1,80%		
3	98,32%	2,76%	1,68%		
4	97,99%	2,83%	2,01%		
5	98,35%	2,76%	1,65%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,21%	4,53%	2,79%	97,17%	0,001
2	97,20%	4,21%	2,80%		
3	97,14%	4,81%	2,86%		
4	97,21%	4,98%	2,79%		
5	97,06%	4,99%	2,94%		

Fuente: Elaboración propia

Tabla 10-64: Resultados de evaluación arquitectura de RNA N°33

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,14%	2,79%	1,86%	98,49%	0,003
2	98,31%	2,96%	1,69%		
3	98,72%	3,15%	1,28%		
4	98,65%	6,00%	1,35%		
5	98,64%	2,83%	1,36%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,51%	2,60%	1,49%	98,25%	0,006
2	98,76%	2,64%	1,24%		
3	97,59%	3,67%	2,41%		
4	98,77%	2,58%	1,23%		
5	97,60%	3,65%	2,40%		

Fuente: Elaboración propia

Tabla 10-65: Resultados de evaluación arquitectura de RNA N°34

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,21%	2,91%	2,79%	98,16%	0,009
2	98,79%	2,51%	1,21%		
3	97,22%	2,60%	2,78%		
4	98,81%	2,60%	1,19%		
5	98,75%	2,51%	1,25%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,80%	4,48%	2,20%	97,86%	0,001
2	97,91%	4,48%	2,09%		
3	97,84%	4,47%	2,16%		
4	97,92%	4,57%	2,08%		
5	97,80%	4,48%	2,20%		

Fuente: Elaboración propia

Tabla 10-66: Resultados de evaluación arquitectura de RNA N°35

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,80%	13,94%	2,20%	97,59%	0,004
2	97,92%	12,76%	2,08%		
3	97,08%	12,87%	2,92%		
4	97,92%	12,79%	2,08%		
5	97,21%	12,85%	2,79%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	94,32%	20,60%	5,68%	95,21%	0,016
2	94,00%	20,41%	6,00%		
3	94,81%	20,26%	5,19%		
4	98,02%	20,24%	1,98%		
5	94,87%	20,25%	5,13%		

Fuente: Elaboración propia

Tabla 10-67: Resultados de evaluación arquitectura de RNA N°37

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,92%	2,82%	2,08%	98,13%	0,002
2	98,14%	2,82%	1,86%		
3	98,32%	2,80%	1,68%		
4	97,92%	2,81%	2,08%		
5	98,33%	2,82%	1,67%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,03%	5,02%	2,97%	97,06%	0,001
2	97,00%	5,03%	3,00%		
3	97,21%	5,08%	2,79%		
4	97,03%	5,02%	2,97%		
5	97,00%	5,09%	3,00%		

Fuente: Elaboración propia

Tabla 10-68: Resultados de evaluación arquitectura de RNA N°38

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,64%	3,87%	2,36%	97,81%	0,003
2	98,11%	2,82%	1,89%		
3	98,15%	2,78%	1,85%		
4	97,60%	3,26%	2,40%		
5	97,55%	3,64%	2,45%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,41%	2,80%	1,59%	97,96%	0,005
2	97,60%	12,19%	2,40%		
3	98,14%	12,01%	1,86%		
4	97,32%	4,44%	2,68%		
5	98,33%	6,38%	1,67%		

Fuente: Elaboración propia

Tabla 10-69: Resultados de evaluación arquitectura de RNA N°39

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	97,33%	13,48%	2,67%	97,20%	0,001
2	97,29%	13,42%	2,71%		
3	97,12%	12,25%	2,88%		
4	97,02%	13,21%	2,98%		
5	97,21%	12,72%	2,79%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	96,82%	6,52%	3,18%	96,69%	0,003
2	96,98%	6,42%	3,02%		
3	96,32%	6,43%	3,68%		
4	97,00%	6,53%	3,00%		
5	96,32%	6,46%	3,68%		

Fuente: Elaboración propia

Tabla 10-70: Resultados de evaluación arquitectura de RNA N°40

DATASET N°1					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,15%	2,80%	1,85%	98,15%	0,006
2	98,15%	2,82%	1,85%		
3	97,21%	12,73%	2,79%		
4	98,56%	2,65%	1,44%		
5	98,66%	2,64%	1,34%		
DATASET N°2					
Iteración	Precisión (%)	Pérdida (%)	Error (%)	Promedio Precisión (%)	Desv. Estándar Precisión
1	98,41%	2,75%	1,59%	98,47%	0,001
2	98,46%	2,71%	1,54%		
3	98,50%	2,72%	1,50%		
4	98,61%	2,68%	1,39%		
5	98,39%	2,74%	1,61%		

Fuente: Elaboración propia

## 10.5. Código de Fuente Utilizado, Matriz de Confusión

A continuación, como se visualiza en la ilustración [10-15], se tiene un código de fuente que permite realizar predicción del modelo que ya está cargado, y también la generación de una matriz de confusión. Cabe mencionar que este código se encuentra en el archivo “Redes\_neuronales\_principal.ipynb”, donde se muestran mayores detalles de cómo se manejan los datasets de pruebas y cómo carga los modelos que ya están generados.

Ilustración 10-15: Código para realizar predicción y generar matriz de confusión

```
#importing confusion matrix
from sklearn.metrics import confusion_matrix
new_predictions = mejor_modelo.predict(X)
confusion = confusion_matrix(Y.argmax(axis=1), new_predictions.argmax(axis=1))
#confusion = confusion_matrix(y_test, new_predictions)
print('Confusion Matrix\n')
print(confusion)
group_names = ['True Neg','False Pos','False Neg','True Pos']
group_counts = ["{0:0.{1}f}".format(value) for value in
               confusion.flatten()]
group_percentages = ["{0:.2%}".format(value) for value in
                      confusion.flatten()/np.sum(confusion)]
labels = [f"\n{v1}\n{v2}\n{v3}" for v1, v2, v3 in
          zip(group_names,group_counts,group_percentages)]
labels = np.asarray(labels).reshape(2,2)
sns.heatmap(confusion, annot=labels, fmt='', cmap='Blues')
```

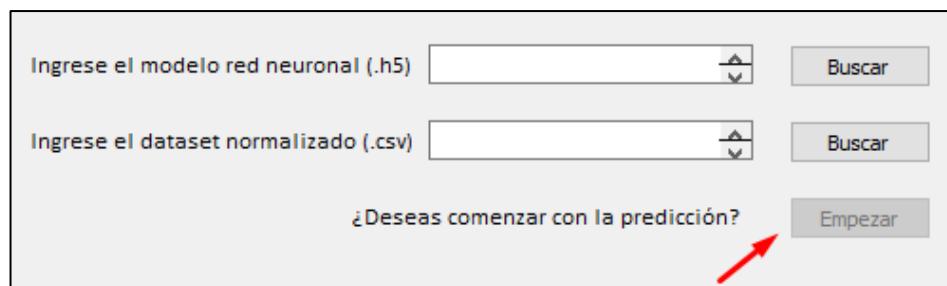
Fuente: Elaboración propia

## 10.6. Validación de Archivos de Entrada en Interfaz Gráfica

Para todas las plataformas, aplicaciones, softwares, entre otros, es importante contar con validación de datos de entrada para el correcto funcionamiento de estos. Como se visualiza en la ilustración [10-16], se observa que el botón “Empezar” se encuentra deshabilitado, y esto se debe a que el usuario deba ingresar el modelo de red neuronal y archivo normalizado CSV. Para habilitar este botón, los archivos deben estar ingresados y, si no ocurriera algún inconveniente, debe estar rellenado con textos (ubicación del archivo) en los cuadros blancos automáticamente. Por otra parte, si el usuario ingresa un tipo de archivo del modelo incorrecto, es decir, que no sea de extensión “.h5”, marcará un mensaje de ventana visualizada en la ilustración [10-17]. Además, si la extensión del archivo es correcta, se verifica que el modelo corresponda a una arquitectura que contiene dos nodos de salidas. De no ser así, se despliega un mensaje de error lo cual se visualiza en la ilustración [10-18].

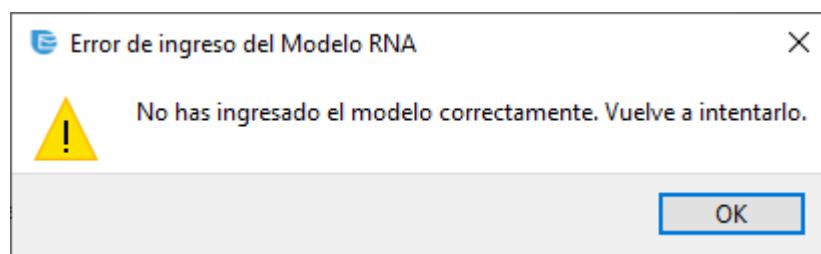
Para otro caso, si el usuario ingresa un tipo de archivo del dataset incorrecto, es decir, que no sea de extensión “.csv”, marcará un mensaje de ventana visualizada en la ilustración [10-19]. También, si la extensión del archivo es correcta, se verifica que la dimensión (campos) del archivo sea de 11, lo cual incluye a la variable objetivo. En caso de que este no cumpliera, se despliega un mensaje de error lo cual se visualiza en la ilustración [10-20]. Por último, si todos los archivos son ingresados correctamente, se procede a ejecutar el programa, y cuando termine, se despliega una ventana de resultados, donde se muestra la cantidad real y predicha de envíos con respuestas en total, y también como estos estarían distribuidos en forma de horario. Esta ventana de resultados de predicción se visualiza en la ilustración [10-21].

Ilustración 10-16: Ventana principal de interfaz, botón "Empezar" deshabilitado



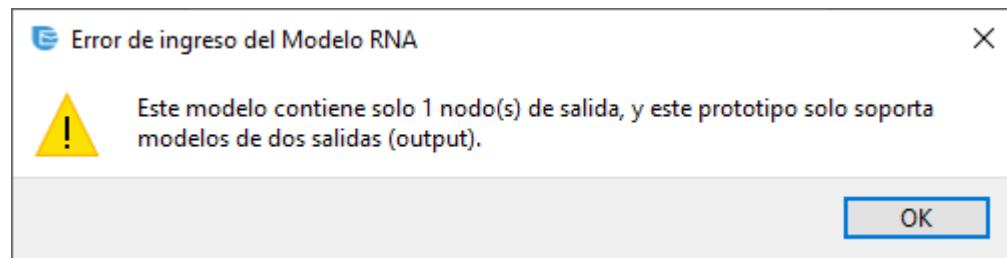
Fuente: Elaboración propia

Ilustración 10-17: Error de ingreso del modelo RNA, archivo incorrecto



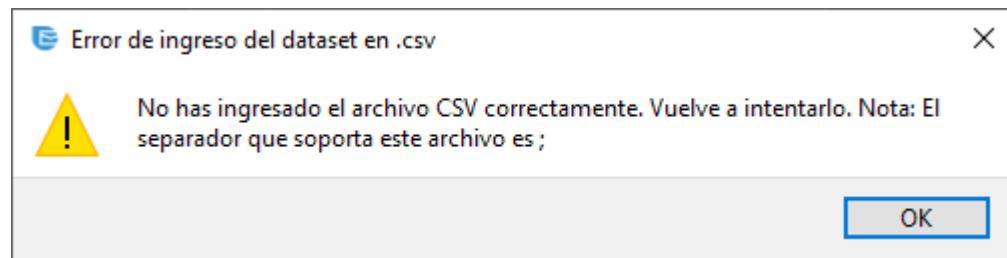
Fuente: Elaboración propia

Ilustración 10-18: Error de ingreso del modelo RNA, Arquitectura de salida no soportada



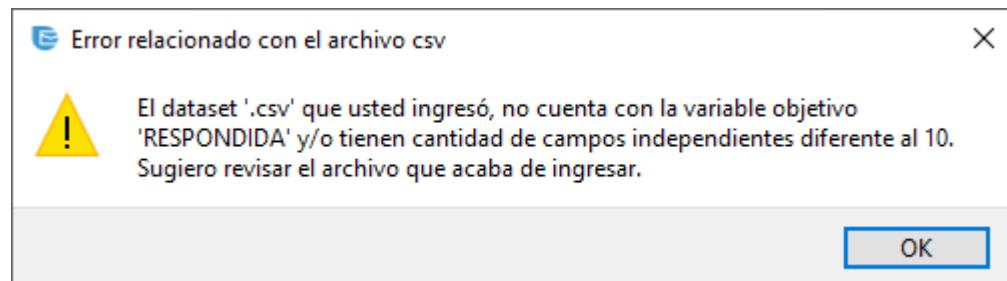
Fuente: Elaboración propia

Ilustración 10-19: Error de ingreso del dataset, archivo o formato incorrecto



Fuente: Elaboración propia

Ilustración 10-20: Error de ingreso del dataset, dimensiones y campos no soportados



Fuente: Elaboración propia

Ilustración 10-21: Ventana de resultados de predicción del prototipo

Resultados de estimación y predicción					
		Datos Reales Cantidad de respuesta estimada: 107143			
	Lunes	Martes	Miercoles	Jueves	Viernes
8	911	975	1125	1168	954
9	1093	1179	1350	1414	1146
10	1125	1221	1393	1457	1189
11	1221	1318	1511	1575	1286
12	1061	1146	1307	1361	1114
13	1146	1232	1414	1479	1200
14	1061	1136	1307	1361	1114

Datos Predichos Cantidad de respuesta estimada: 106627					
	Lunes	Martes	Miercoles	Jueves	Viernes
15	1130	1216	1397	1450	1184
16	1120	1216	1386	1450	1184
17	1162	1258	1439	1503	1226
18	1056	1141	1312	1365	1120
19	1130	1216	1397	1461	1184
20	1066	1152	1322	1386	1120
21	544	597	682	714	576

Fuente: Elaboración propia

## 10.7. Resultados de Predicciones Horaria de Respuestas

A continuación, se detallan los resultados de predicción horaria sobre cantidad de envíos con respuestas, tras utilizar el prototipo elaborado. Cabe destacar que cada predicción se utiliza datasets, entre ellos se encuentra dataset de pruebas N°1 y N°2, y de las tres mejores arquitecturas se tienen modelos distintos, donde uno es entrenado con dataset N°1 (D1) y otro N°2 (D2).

Tabla 10-71: Resultado de predicción horaria del Modelo Red23\_D1S2, dataset N°1 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	940	1.007	1.162	1.206	985	830	631
9	1.128	1.217	1.394	1.460	1.184	996	752
10	1.162	1.261	1.438	1.504	1.228	1.029	785
11	1.261	1.361	1.560	1.626	1.327	1.106	841
12	1.095	1.184	1.350	1.405	1.150	962	730
13	1.184	1.272	1.460	1.527	1.239	1.040	796
14	1.095	1.173	1.350	1.405	1.150	962	730
15	1.173	1.261	1.449	1.504	1.228	1.029	785
16	1.162	1.261	1.438	1.504	1.228	1.029	785
17	1.206	1.305	1.493	1.560	1.272	1.062	808
18	1.095	1.184	1.361	1.416	1.162	973	741
19	1.173	1.261	1.449	1.515	1.228	1.029	785
20	1.106	1.195	1.372	1.438	1.162	973	741
21	564	619	708	741	597	498	387

Fuente: Elaboración propia

Tabla 10-72: Resultado de predicción horaria del Modelo Red23\_D1S2, dataset N°2 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	925	990	1.142	1.186	968	816	620
9	1.110	1.197	1.371	1.436	1.164	979	740
10	1.142	1.240	1.414	1.480	1.208	1.012	772
11	1.240	1.338	1.534	1.599	1.306	1.088	827
12	1.077	1.164	1.327	1.382	1.132	947	718
13	1.164	1.251	1.436	1.501	1.219	1.023	783
14	1.077	1.153	1.327	1.382	1.132	947	718
15	1.153	1.240	1.425	1.480	1.208	1.012	772
16	1.142	1.240	1.414	1.480	1.208	1.012	772
17	1.186	1.284	1.469	1.534	1.251	1.044	794
18	1.077	1.164	1.338	1.393	1.142	957	729
19	1.153	1.240	1.425	1.491	1.208	1.012	772
20	1.088	1.175	1.349	1.414	1.142	957	729
21	555	609	696	729	588	490	381

Fuente: Elaboración propia

Tabla 10-73: Resultado de predicción horaria del Modelo Red23\_D2S2, dataset N°1 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	1.051	1.125	1.298	1.348	1.100	927	705
9	1.261	1.360	1.558	1.632	1.323	1.113	841
10	1.298	1.409	1.607	1.681	1.372	1.150	878
11	1.409	1.521	1.743	1.817	1.484	1.236	940
12	1.224	1.323	1.508	1.570	1.286	1.076	816
13	1.323	1.422	1.632	1.706	1.385	1.162	890
14	1.224	1.311	1.508	1.570	1.286	1.076	816
15	1.311	1.409	1.620	1.681	1.372	1.150	878
16	1.298	1.409	1.607	1.681	1.372	1.150	878
17	1.348	1.459	1.669	1.743	1.422	1.187	903
18	1.224	1.323	1.521	1.583	1.298	1.088	828
19	1.311	1.409	1.620	1.694	1.372	1.150	878
20	1.236	1.335	1.533	1.607	1.298	1.088	828
21	631	692	791	828	668	556	433

Fuente: Elaboración propia

Tabla 10-74: Resultado de predicción horaria del Modelo Red23\_D2S2, dataset N°2 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	1.035	1.108	1.279	1.327	1.084	913	694
9	1.242	1.339	1.534	1.607	1.303	1.096	828
10	1.279	1.388	1.583	1.656	1.352	1.132	865
11	1.388	1.498	1.717	1.790	1.461	1.218	925
12	1.206	1.303	1.486	1.546	1.266	1.059	804
13	1.303	1.400	1.607	1.680	1.364	1.145	877
14	1.206	1.291	1.486	1.546	1.266	1.059	804
15	1.291	1.388	1.595	1.656	1.352	1.132	865
16	1.279	1.388	1.583	1.656	1.352	1.132	865
17	1.327	1.437	1.644	1.717	1.400	1.169	889
18	1.206	1.303	1.498	1.559	1.279	1.072	816
19	1.291	1.388	1.595	1.668	1.352	1.132	865
20	1.218	1.315	1.510	1.583	1.279	1.072	816
21	621	682	779	816	658	548	426

Fuente: Elaboración propia

Tabla 10-75: Resultado de predicción horaria del Modelo Red27\_D1S2, dataset N°1 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	785	840	970	1.007	822	693	526
9	942	1.016	1.164	1.219	988	831	628
10	970	1.053	1.200	1.256	1.025	859	656
11	1.053	1.136	1.302	1.357	1.108	923	702
12	914	988	1.127	1.173	960	803	609
13	988	1.062	1.219	1.274	1.034	868	665
14	914	979	1.127	1.173	960	803	609
15	979	1.053	1.210	1.256	1.025	859	656
16	970	1.053	1.200	1.256	1.025	859	656
17	1.007	1.090	1.247	1.302	1.062	887	674
18	914	988	1.136	1.182	970	813	619
19	979	1.053	1.210	1.265	1.025	859	656
20	923	997	1.145	1.200	970	813	619
21	471	517	591	619	499	416	323

Fuente: Elaboración propia

Tabla 10-76: Resultado de predicción horaria del Modelo Red27\_D1S2, dataset N°2 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	774	828	956	992	810	683	519
9	929	1.001	1.147	1.202	974	819	619
10	956	1.038	1.184	1.238	1.011	847	646
11	1.038	1.120	1.284	1.338	1.092	910	692
12	901	974	1.111	1.156	947	792	601
13	974	1.047	1.202	1.256	1.020	856	655
14	901	965	1.111	1.156	947	792	601
15	965	1.038	1.193	1.238	1.011	847	646
16	956	1.038	1.184	1.238	1.011	847	646
17	992	1.074	1.229	1.284	1.047	874	665
18	901	974	1.120	1.165	956	801	610
19	965	1.038	1.193	1.247	1.011	847	646
20	910	983	1.129	1.184	956	801	610
21	464	510	583	610	492	410	319

Fuente: Elaboración propia

Tabla 10-77: Resultado de predicción horaria del Modelo Red27\_D2S2, dataset N°1 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	730	782	902	936	764	644	490
9	876	945	1.082	1.134	919	773	584
10	902	979	1.117	1.168	953	799	610
11	979	1.056	1.211	1.263	1.031	859	653
12	850	919	1.048	1.091	893	747	567
13	919	988	1.134	1.185	962	807	618
14	850	910	1.048	1.091	893	747	567
15	910	979	1.125	1.168	953	799	610
16	902	979	1.117	1.168	953	799	610
17	936	1.014	1.160	1.211	988	825	627
18	850	919	1.056	1.099	902	756	575
19	910	979	1.125	1.177	953	799	610
20	859	928	1.065	1.117	902	756	575
21	438	481	550	575	464	387	301

Fuente: Elaboración propia

Tabla 10-78: Resultado de predicción horaria del Modelo Red27\_D2S2, dataset N°2 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	720	771	890	924	754	636	483
9	864	932	1.068	1.119	907	763	576
10	890	966	1.102	1.153	941	788	602
11	966	1.042	1.195	1.246	1.017	847	644
12	839	907	1.034	1.076	881	737	559
13	907	975	1.119	1.170	949	797	610
14	839	898	1.034	1.076	881	737	559
15	898	966	1.110	1.153	941	788	602
16	890	966	1.102	1.153	941	788	602
17	924	1.000	1.144	1.195	975	814	619
18	839	907	1.042	1.085	890	746	568
19	898	966	1.110	1.161	941	788	602
20	847	915	1.051	1.102	890	746	568
21	432	475	542	568	458	381	297

Fuente: Elaboración propia

Tabla 10-79: Resultado de predicción horaria del Modelo Red36\_D2S2, dataset N°1 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
8	794	850	980	1.018	831	700	532
9	952	1.027	1.176	1.232	999	840	635
10	980	1.064	1.214	1.270	1.036	868	663
11	1.064	1.148	1.316	1.372	1.120	934	710
12	924	999	1.139	1.186	971	812	616
13	999	1.074	1.232	1.288	1.046	878	672
14	924	990	1.139	1.186	971	812	616
15	990	1.064	1.223	1.270	1.036	868	663
16	980	1.064	1.214	1.270	1.036	868	663
17	1.018	1.102	1.260	1.316	1.074	896	682
18	924	999	1.148	1.195	980	822	626
19	990	1.064	1.223	1.279	1.036	868	663
20	934	1.008	1.158	1.214	980	822	626
21	476	523	598	626	504	420	327

Fuente: Elaboración propia

Tabla 10-80: Resultado de predicción horaria del Modelo Red36\_D2S2, dataset N°2 usado

HORA/DÍA	LUNES	MARTES	MIERCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
<b>8</b>	787	842	972	1.009	824	694	527
<b>9</b>	944	1.018	1.166	1.222	990	833	629
<b>10</b>	972	1.055	1.203	1.259	1.027	861	657
<b>11</b>	1.055	1.138	1.305	1.360	1.110	925	703
<b>12</b>	916	990	1.129	1.175	962	805	611
<b>13</b>	990	1.064	1.222	1.277	1.036	870	666
<b>14</b>	916	981	1.129	1.175	962	805	611
<b>15</b>	981	1.055	1.212	1.259	1.027	861	657
<b>16</b>	972	1.055	1.203	1.259	1.027	861	657
<b>17</b>	1.009	1.092	1.249	1.305	1.064	888	676
<b>18</b>	916	990	1.138	1.184	972	814	620
<b>19</b>	981	1.055	1.212	1.268	1.027	861	657
<b>20</b>	925	999	1.147	1.203	972	814	620
<b>21</b>	472	518	592	620	500	416	324

Fuente: Elaboración propia