PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DESARROLLO DE UN MODELO ALGORÍTMICO BASADO EN ÁRBOLES DE DECISIÓN PARA LA PREDICCIÓN DE LA PERMANENCIA DE UN PACIENTE EN UN PROCESO PSICOTERAPÉUTICO

Tesis para optar el Título de Ingeniero Informático, que presenta el bachiller:

Heli Eliaquin Leon Atiquipa

ASESOR:

Dr. Cesar Beltrán Castañón

Lima, marzo del 2018

DEDICATORIA

A mis padres Victor León y Bertha Atiquipa por brindarme la oportunidad de ser un profesional y darme su apoyo incondicional para el logro de mis metas, a mis hermanos Ruben y Benjamin, quienes durante este camino me brindaron su guía y apoyo, a Mercedes por su apoyo incondicional y constante durante las diferentes etapas de mi vida, incluyendo la universitaria.

RESUMEN

En la actualidad existe una creciente necesidad de atención psicológica en nuestro país, por lo que existen muchas instituciones públicas y privadas que ofrecen esto servicios profesionales. La psicoterapia es parte de estos servicios y quienes lo brindan son profesionales especializados en la materia, los cuales atienden a pacientes de diferentes edades y estratos socioeconómicos. Estos tratamientos suelen durar mucho tiempo, por lo que muchos pacientes, por diferentes circunstancias, abandonan el proceso al poco tiempo de haberlo iniciado.

La institución, el cual es el caso de estudio, maneja ciertos niveles de deserción medibles durante el tiempo. Estos niveles son manejables en el grado en el que se dan, sin embargo, un creciente aumento del mismo podría generar costos para mantener el equilibrio, el cual deberá ser aplicado a los pacientes, los cuales podrían sentir incomodidad y afectar el proceso terapéutico. La necesidad de tener un mayor control sobre los niveles de deserción y reducirlos ayudaría en gran medida a mejorar la calidad de los servicios que se brindan en la institución.

Para la institución, la incertidumbre del abandono en el proceso no permite aplicar medidas correctivas que permitan mejorar los niveles de deserción, sin embargo, la información contenida en la base de datos institucional permite, por cuestiones de investigación, estudiar y analizar los patrones que conllevan al abandono del proceso. Realizar este tipo de análisis sobre una gran cantidad de información implica utilizar métodos computacionales que permitan ayudar a analizar la información de una forma rápida y eficiente. Es por ello, que surge la necesidad de apoyarnos en las ciencias de la computación, específicamente en la minería de datos, para identificar los patrones que permitan predecir y determinar la permanencia de los pacientes durante el proceso.

El presente proyecto de fin de carrera pretende entender las causales de la deserción en un proceso psicoterapéutico con el fin de poder predecir, desde el primer contacto entre el paciente y la institución, la permanencia del paciente. Para esto, se plantea el desarrollo de un prototipo funcional que permita predecir la permanencia de los pacientes haciendo uso de algoritmos de árboles de decisión para la predicción.

Para la elaboración del prototipo funcional y el cumplimiento de los objetivos, se hizo uso de la herramienta Weka, el cual permitió analizar y seleccionar el algoritmo a

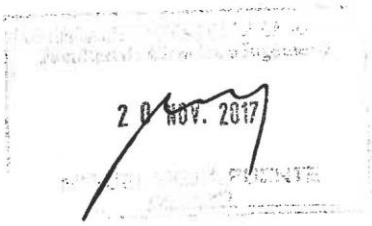
usar para la implementación del prototipo. El desbalanceo de clases dificulto el proceso de análisis algorítmico, por tal motivo, se aplicaron métodos de minería de datos para analizar los conjuntos de datos desbalanceados. El lenguaje de programación usado fue Java y los algoritmos que permitieron la predicción fueron incorporados desde las librerías del API de Weka. Los resultados obtenidos fueron satisfactorios, en base a los datos que fueron extraídos de la base de datos institucional.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

PONTIFICIA
UNIVERSIDAD CATÓLICA
DEL PERÚ

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO:	**Desarrollo de un modelo algorítmico basado en árboles de decisión para la predicción de la permanencia de un paciente en un proceso psicoterapéutico.**
ÁREA:	Ciencia de la Computación.
ASESOR:	Dr. César Armando BELTRÁN CASTAÑÓN
ALUMNO:	Heli Eliaquin LEON ATIQUIPA
CÓDIGO:	20067135
TEMA N°:	#672
FECHA:	San Miguel, 13 de octubre del 2017



DESCRIPCIÓN

La salud mental es una preocupación que todo gobierno debe tener como política de salud prioritaria. Para ello existen instituciones públicas y privadas que brindan servicios psicológicos a diferentes sectores sociales del país, entre los cuales tenemos la psicoterapia, la cual es un proceso que permite a las personas cambiar su conducta y controlarla por ellos mismos mediante el autoconocimiento. Sin embargo, la continuidad del proceso psicoterapéutico muchas veces se ve interrumpida por diferentes factores: económico, de dependencia, situacional, conductual, entre otros. La incertidumbre del abandono del proceso psicoterapéutico no permite aplicar medidas correctivas que permitan la mayor permanencia del paciente en el proceso.

La información contenida en las bases de datos de centros psicoterapéuticos corresponde a la información de los pacientes que se atienden en la misma como la edad, sexo, independencia económica, región geográfica de residencia; así como la del terapeuta asignado y la programación de sus citas. Esta información resulta ser valiosa para el análisis y encontrar uno o más patrones que permita predecir, con cierto grado de precisión aceptable, el tiempo de permanencia de un paciente en el proceso psicoterapéutico. Una característica importante es el tipo de información que se maneja, mayormente cualitativa, lo cual hace inviable que muchos métodos de predicción puedan ser aplicados, especialmente aquellos basados en funciones continuas. En ese sentido, según la literatura, los árboles de decisión se prestan como las técnicas idóneas para trabajar con tipos de datos cualitativos.

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

La propuesta, para el presente proyecto de fin de carrera, es desarrollar una plataforma computacional que contenga tres modelos algorítmicos basados en arboles de decisión los cuales corresponden a cada una de las tres fases del proceso psicoterapéutico (indicación, inicial e intermedia) incorporados para la predicción de la permanencia. Esta plataforma permitiría a la institución aplicar mejoras sobre la atención que se les brinda a los pacientes, así como una mejor asignación de pacientes a los terapeutas del centro.

OBJETIVO GENERAL

Desarrollar un modelo algorítmico basado en árboles de decisión para la predicción de la permanencia en el proceso psicoterapéutico de pacientes en una clínica de psicoterapia.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos son:

OE1. Recolectar, analizar, estructurar y pre-procesar un conjunto de datos a partir de la información de los pacientes y los terapeutas de la institución.

OE2. Aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción del éxito en el proceso psicoterapéutico.

OE3. Aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.

OE4. Desarrollar un prototipo funcional que permita predecir el mejor vinculo terapeuta-paciente en un proceso psicoterapéutico

ALCANCE

El presente proyecto de fin de carrera se desarrollará en el ámbito de las ciencias de la computación y se centrará en predecir la permanencia de un paciente en un proceso psicoterapéutico en una institución donde se brindan estos servicios a pacientes que necesitan ayuda terapéutica. Para esto, se usarán datos, en su mayoría cualitativos, relacionados al terapeuta y paciente de una institución con el fin de determinar patrones que permitan predecir cómo evolucionará el proceso psicoterapéutico. Los

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

datos de los pacientes se encuentran anonimizados así como se cuenta con la autorización del uso de estos por parte de la institución.

Para predecir la permanencia del paciente en el proceso psicoterapéutico, se usará una máquina de aprendizaje la cual será entrenada por medio de información histórica de los pacientes y terapeutas, los cuales serán normalizados, pre-procesados y clasificados. Para la predicción se considerará información referida a aspectos personales, familiares y económicos, así como el horario seleccionado para la primera cita, debido a que el marco del análisis se centra en la predicción de permanencia a partir de los datos obtenidos en el primer contacto entre el paciente y la institución. Como el enfoque del proyecto es desarrollar un prototipo funcional, los resultados serán presentados en una interfaz de fácil acceso y amigable. El resultado del prototipo funcional se limitará a predecir si el paciente permanecerá en el proceso por 16 a más citas efectivas, el cual es equivalente a superar los cuatros meses de permanencia, o en caso contrario, si el paciente superará las cuatro primeras citas efectivas correspondientes a la fase intermedia o si abandonará el proceso en cualquiera de las cuatro primeras citas correspondientes a la fase de indicación e inicial.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

TABLA DE CONTENIDOS

INDICE DE TABLAS

INDICE DE FIGURAS

1. CAPÍTULO 1: PROBLEMÁTICA, OBJETIVOS Y ALCANCE

1.1 Problemática

En la actualidad existen centros terapéuticos los cuales brindan sus servicios al público en general. Por lo general, los pacientes presentan cuadros psicológicos que pueden ser tratados en terapias semanales, las que pueden ser una o dos veces a la semana en una frecuencia baja y de tres a cuatro en una frecuencia alta (BERNARDI, 2014). Existen varios tipos de tratamientos psicoterapéuticos, entre las que se tienen: Psicoterapia psicodinámica, Psicoterapia conductual, Psicoterapia interpersonal, Psicoterapia sistemática, Psicoterapia de apoyo y la Psicoterapia en grupo; que pueden ser muy largos y que un grupo grande abandonan el tratamiento entre la cuarta y decima cita (CASTRO, 2002).

Moreno (MORENO, 2012) menciona que los factores que influyen en el paciente a abandonar su tratamiento terapéutico se encuentran relacionados al terapeuta y otras al paciente. Con respecto al terapeuta, destacan factores relacionados a la experiencia del terapeuta y su edad, con respecto al paciente, destacan los factores económicos, dependencia, sensación de mejora y diagnóstico.

La experiencia del terapeuta es un factor que algunos pacientes podrían percibir ante diferentes circunstancias, por ejemplo, un terapeuta durante el tratamiento podría transmitir sus propias opiniones y consejos, siendo esto inadecuado durante este tipo de tratamiento (BLEGER, 1967). Por otro lado, el desconocimiento del paciente sobre la actitud y posición del terapeuta podría hacer pensar que es normal que el terapeuta le brinde consejos y soluciones, cuando la realidad es otra, pudiendo generar una disconformidad sobre el tratamiento (OROZCO et al, 2014).

Según Benítez (BENÍTEZ et al, 2009), la poca experiencia del terapeuta podría hacer que no transmita y mantenga firme su posición en el tratamiento, deslizando en algunas ocasiones sus propias apreciaciones y manteniendo otro tipo de relaciones con el paciente ya sea del tipo social, afectivo, financiero u otro. También hace mención que la edad del terapeuta puede ser un causal de disconformidad, debido a que el paciente lo relaciona con su experiencia, sin embargo, este solo puede causar impactos negativos en el tratamiento de pacientes en situaciones de gravedad.

Por el lado del paciente, Benítez (BENÍTEZ et al, 2009) menciona lo siguiente:

> Existe una fuerte relación con la deserción en pacientes de edades extremas (vejez y niñez), así como en la adolescencia y soltería. Los niveles de dependencia en estas edades influencian en la continuidad del proceso terapéutico. Los pacientes que tienden a adherirse con mayor fuerza a los tratamientos terapéuticos son los que entienden su necesidad del proceso psicoterapéutico, por lo tanto, asisten a sus sesiones de manera voluntaria y por iniciativa propia. El factor económico suele ser un factor situacional que obliga a los pacientes a abandonar el tratamiento por falta de recursos para costear las sesiones. Existen otros factores situacionales, tales como, descomposición familiar, perdida de un familiar, enfermedad, perdida del trabajo, entre otros, que terminan siendo determinantes en la permanencia del tratamiento terapéutico.

Con respecto al diagnóstico del paciente, algunas características relacionadas al trastorno del paciente pueden determinar la probabilidad de adherencia al tratamiento terapéutico, por ejemplo, los relacionados con la personalidad (BENÍTEZ et al, 2009). Algunas características de los pacientes, por ejemplo, la presencia o no de algún trastorno o discapacidad mental, motora o sensorial, las creencias relacionadas a la salud, la falta de autocuidado, las expectativas negativas frente a tratamientos y enfermedades, el pesimismo entre otros, contribuyen a la escasa adhesión terapéutica (MORENO, 2012).

La adhesión terapéutica se da en algunos casos ante la gravedad sintomática y la necesidad urgente de soluciones; cuando dicha urgencia se atenúa o desaparece, sucede lo mismo con su adhesión (BENÍTEZ et al, 2009). Moreno (MORENO, 2012) menciona lo siguiente con respecto al vínculo paciente-terapeuta:

> Cuando el terapeuta y el paciente tienen un vínculo, la adhesión terapéutica se fortalece por lo que cualquier gesto motivacional a continuar el proceso terapéutico es muy bien recibido. Este vínculo se genera desde el primer encuentro, el cual es esperado en muchas ocasiones por el paciente con muchas ansias, y se da por medio de los intereses y actitudes que presente delante del paciente, así como por su personalidad y rasgos propios de su individualidad.

La institución, el cual será el caso de estudio, utiliza el criterio de asignación por disponibilidad, pero a su vez, añade una serie de reglas que le permite asignar

pacientes de manera equitativa para no abarrotar a un solo terapeuta. Este tipo de asignación resulta eficiente, pero no contribuye en brindar la mejor relación que pueda existir entre terapeuta-paciente.

Los niveles de deserción reportados en la historia de la institución arrojan los siguientes valores que se presentan en la Tabla 1.1:

Tabla 1.1: Porcentaje de deserción mensual en la institución en los doce primeros meses de una muestra de 2256 pacientes [Elaboración propia].

MES	PORCENTAJE DE DESERCIÓN	PORCENTAJE DE DESERCIÓN ACUMULADA
1	30.45%	30.45%
2	25.17%	55.62%
3	12.97%	68.58%
4	7.12%	75.70%
5	3.54%	79.24%
6	2.35%	81.59%
7	2.02%	83.61%
8	1.81%	85.42%
9	1.07%	86.49%
10	0.60%	87.09%
11	0.69%	87.78%
12	0.99%	88.77%
TOTAL	**88.77%**	

Según la Tabla 1.1, una gran cantidad de pacientes abandona el proceso psicoterapéutico en los primeros cuatro meses. El 75.70% es un porcentaje considerable, por lo que se puede decir que este grupo de pacientes no pudo crear un vínculo fuerte terapeuta-paciente. Adicionalmente, se puede observar que los niveles de deserción caen considerablemente entre el cuarto y quinto mes y luego se mantiene en un porcentaje muy reducido. Para el presente proyecto de fin de carrera, se consideró a los pacientes que mantengan el proceso terapéutico más de cuatro meses o dieciséis citas efectivas como procesos exitosos, caso contrario, se tomó como proceso psicoterapéutico abandonado o fallido.

Según lo expuesto anteriormente, es posible reducir estos niveles de deserción si se consideran aspectos relacionados al terapeuta y paciente. Por el lado del terapeuta, su experiencia, edad y género; y por el lado del paciente, los niveles de dependencia, situación económica y la relación vinculante que pueda existir de acuerdo al perfil histórico de pacientes con mejor vinculo terapeuta-paciente.

En el momento en que se realizó el análisis de la base de datos de la institución, el cual es el caso de estudio, se pudo observar que la cantidad de pacientes que por lo menos asistieron a una cita fue de 2621, de los cuales solo se analizó a 2256. Los pacientes que no fueron considerados presentaban información incompleta o incoherente. Según la Tabla 1.2, se puede observar la cantidad de pacientes que abandonaron el proceso durante las primeras fases del proceso psicoterapéutico (1 a 4 citas correspondiente a la fase de indicación e inicial; 5 a 15 citas correspondientes a la fase intermedia) y la cantidad mínima de citas para determinar si el proceso es exitoso o no. El grupo de pacientes que asistieron entre 1-4 citas y posteriormente abandonaron el proceso corresponde a un 53% por lo que es interesante analizar adicionalmente a este grupo.

Tabla 1.2: Cantidad de pacientes por cantidad de citas efectivas durante el proceso [Elaboración propia].

CITAS	1 a 4	5 a 15	16 a más	TOTAL
PACIENTES	1205	699	352	2256
% POR FASE	53%	31%	16%	100%

El interés principal es poder predecir el éxito del proceso psicoterapéutico, o en su defecto, predecir la cantidad de citas efectivas que el paciente tendrá en el proceso psicoterapéutico. Para esto, se tomó en cuenta la información obtenida de anteriores investigaciones, así como la información obtenida de los pacientes en su primer contacto con la institución. Siendo el vínculo terapeuta-paciente muy importante para la permanencia del proceso psicoterapéutico, para el análisis se usó la información de los terapeutas de la institución. El fin de desarrollar este modelo es mejorar la calidad de atención en las citas brindando el mejor servicio a los pacientes y de esta forma mejorar el nivel de permanencia de los pacientes durante el proceso psicoterapéutico.

1.2 Objetivo general

Desarrollar un modelo algorítmico para la predicción de la permanencia en el proceso Psicoterapéutico de pacientes en una Clínica de Psicoterapia basado en árboles de decisión.

1.3 Objetivos específicos

Los objetivos específicos son:

- Objetivo 1: Recolectar, analizar, estructurar y pre-procesar un conjunto de datos a partir de la información de los pacientes y los terapeutas de la institución.
- Objetivo 2: Aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción del éxito o fracaso en el proceso psicoterapéutico.
- Objetivo 3: Aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.
- Objetivo 4: Desarrollar un prototipo funcional que permita predecir el mejor vinculo terapeuta-paciente en un proceso psicoterapéutico usando el método de clasificación seleccionado, así como la cantidad de citas efectivas en caso de fracaso.

1.4 Resultados esperados

Los resultados esperados son:

- Resultado 1 para el objetivo 1: Conjunto de datos seleccionado, pre-procesados y transformados para el proceso de clasificación.
- Resultado 2 para el objetivo 1: Tres conjuntos de datos de entrenamiento estructurados para llevar a cabo el proceso de clasificación.
- Resultado 3 para el objetivo 2: Resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción del éxito o fracaso del proceso psicoterapéutico.
- Resultado 4 para el objetivo 2: Modelo de clasificación optimizado para la predicción del éxito o fracaso del proceso psicoterapéutico.
- Resultado 5 para el objetivo 3: Resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.

- Resultado 6 para el objetivo 3: Modelo de clasificación optimizado para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.
- Resultado 7 para el objetivo 4: Prototipo funcional que permita predecir el éxito o fracaso del proceso psicoterapéutico por medio de la información del paciente.
- Resultado 8 para el objetivo 4: Prototipo funcional que permita predecir la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.

1.5 Herramientas, métodos, metodologías y procedimientos

Para poder alcanzar los resultados esperados identificados es necesario definir qué herramientas se utilizarán como apoyo.

1.5.1 *Herramientas*

A continuación, en la Tabla 1.3, se presentan las herramientas necesarias para el desarrollo de cada una de estas.

Tabla 1.3: Herramientas a usar por cada resultado esperado [Elaboración propia].

Resultados esperados	**Herramientas a usarse**
RE1-OBJ1: Conjunto de datos seleccionado, pre-procesados y transformados para el proceso de clasificación.	**Mysql,** motor de base de datos. **Weka,** herramienta que permite el análisis de minería de datos con diferentes algoritmos de clasificación.
RE2-OBJ2: Tres conjuntos de datos de entrenamiento estructurados para llevar a cabo el proceso de clasificación.	**Weka**
RE3-OBJ2: Resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción del éxito o fracaso del proceso psicoterapéutico.	**Weka** **Spread Subsample** **SMOTE** **Resample** **ClassBalancer**
RE4-OBJ2: Modelo de clasificación optimizado para la predicción del éxito o fracaso del proceso psicoterapéutico.	**Weka**

RE5-OBJ3: Resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.	**Weka** **Spread Subsample** **SMOTE** **Resample** **ClassBalancer**
RE6-OBJ3: Modelo de clasificación optimizado para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.	**Weka**
RE7-OBJ4: Prototipo funcional que me permita predecir el éxito o fracaso del proceso psicoterapéutico por medio de un cuestionario de preguntas.	**Netbeans** IDE como herramienta de desarrollo
RE8-OBJ4: Prototipo funcional que permita predecir la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.	**Netbeans**

1.5.1.1 MYSQL

Mysql es un motor de base de datos de código abierto muy popular, cuyo rendimiento, confiabilidad y facilidad de uso ha sido comprobado[1].

Se usó esta herramienta porque se requiere utilizar un motor de base de datos potente, *open source*, multiplataforma y que sea rápida, razón por la cual se escogió Mysql como motor de base de datos (GUZMÁN, 2006). Su uso se restringirá como fuente para la obtención del conjunto de datos.

1.5.1.2 Weka

Weka es un software libre desarrollado en Java, el cual tiene una serie de paquetes de código con varias técnicas de pre-procesado, clasificación, agrupamiento, asociación y visualización, los cuales permiten el análisis de datos, así como, el rendimiento de los algoritmos (LOPEZ, 2006).

Esta herramienta fue para poder realizar la experimentación del modelo algorítmico a implementar. Al ser posible utilizar diferentes algoritmos de minería de datos, se

[1] MYSQL, 2017. Consulta 10 de marzo del 2017. https://www.oracle.com/lad/mysql/index.html.

podrá seleccionar el mejor método de aprendizaje para el modelo predictivo que se desea implementar.

1.5.1.3 Herramientas para el balanceo de clases

Se hizo uso de estas herramientas para balancear el conjunto de datos de entrenamiento el cual tenía las clases desbalanceadas para el proceso de clasificación.

- **Spread subsample:** Método para la construcción de clasificadores no balanceados, el cual produjo una sub-muestra aleatoria de un conjunto de datos en memoria y eliminó objetos de la clase mayoritaria de manera aleatoria, de tal forma que se ajustó la distribución de clases (Hall et al., 2009).
- **SMOTE:** Método de sobre-muestreo el cual generó objetos sintéticos de la clase minoritaria con el fin de balancear las clases del conjunto de datos por clasificar (Chawla et al., 2002).
- **Resample:** Método hibrido para la construcción de clasificadores con datos no balanceados, el cual eliminó los objetos de la clase mayoritaria y agregó objetos a la clase minoritaria (Hall et al., 2009).
- **Classbalancer:** Método hibrido para la construcción de clasificadores con datos no balanceados, el cual ponderó el peso de las clases y las mantiene equilibradas (Petrosino et al., 2017).

1.5.1.4 Netbeans

Es una IDE de licencia gratuita compatible con diferentes sistemas operativos, el cual permite desarrollar proyectos de software en diferentes lenguajes, entre los cuales están: C/C++, Java, PHP, entre otros[2].

La justificación para el uso de esta herramienta, fue debido a que el producto del presente proyecto de fin de carrera será desarrollado en JAVA. Esta IDE tiene una herramienta gráfica integrada para el fácil desarrollo de la interfaz gráfica.

1.5.2 *Métodos y Procedimientos*

En esta Sección se presenta el método a usar para el proceso de clasificación el cual se encuentra implementado en la herramienta Weka.

[2] NETBEANS, 2017. Consulta 10 de marzo del 2017. http://www.oracle.com/technetwork/developer-tools/netbeans/overview/index.html.

1.5.2.1 Data mining

D*ata mining* o minería de datos se define como “el proceso de descubrimiento de patrones y conocimiento desde grandes cantidades de datos almacenados” (HAN, PEI & KAMBER, 2011). La minería de datos permite obtener este conocimiento desde diferentes fuentes para diferentes sectores, por ejemplo, pueden ser usadas para mejorar la productividad, mejora de productos, incrementar las ventas, etc. (INFANTE, et al., 2010).

Existen dos tipos de modelos de minería de datos: predictivos y descriptivos:

- **Modelos predictivos:** Los modelos predictivos intentan estimar valores futuros basados en variables dependientes con variables independientes, por ejemplo, se puede estimar la demanda de un producto o servicio en función del gasto en publicidad (INFANTE, et al., 2010).
- **Modelos descriptivos:** Los modelos descriptivos identifican patrones con el objetivo de explorar las propiedades de los datos, por ejemplo, en base a la información de ventas en un supermercado, se podría determinar cuál es la tendencia de compras de acuerdo a la edad y sexo de cada persona, de esta manera poner artículos en ubicaciones estratégicas para incrementar las ventas u ofrecerles ofertas especiales (INFANTE, et al., 2010).

1.5.2.2 Aprendizaje de Máquina

Según Xue (XUE, 2009), el aprendizaje de máquina simula el aprendizaje de los seres humanos, con el objetivo de obtener nuevo conocimiento o la habilidad de organizar la estructura del conocimiento, el cual permita mejorar de manera progresiva su propio rendimiento. Además, menciona que es la manera fundamental como la computadora puede tener inteligencia, siendo este el núcleo de la Inteligencia Artificial.

1.5.2.3 Aprendizaje Supervisado

Una de las tareas más comunes en aprendizaje supervisado es la clasificación, el cual se basa en el conjunto de datos de entrenamiento permitiendo supervisar el aprendizaje del modelo de clasificación (HAN, PEI & KAMBER, 2011).

Perversi (PERVERSI, 2007) menciona lo siguiente sobre la clasificación:

Es un método que permite encontrar propiedades comunes de entre el conjunto de datos y los clasifica en clases, de acuerdo al modelo de clasificación. El objetivo de clasificar es analizar los datos de entrenamiento para posteriormente, por el método supervisado, desarrollar un modelo para cada clase utilizando las características disponibles en los datos. Este método es conocido como supervisado, debido a que se conoce la clase de pertenencia y se le indica al modelo si su clasificación es correcta o no.

1.5.2.4 Aprendizaje No Supervisado

El aprendizaje no supervisado, es básicamente el *clustering* o agrupamiento, el cual se basa en no tener conocimiento a priori, por lo que se usa el *clustering* para descubrir las clases dentro de los datos (HAN, PEI & KAMBER, 2011).

El *clustering* consiste en agrupar un conjunto de datos basados en la similitud de sus atributos (PERVERSI, 2007). Según Garre (GARRE, 2007), para medir las similitudes se utilizan diferentes formas, por ejemplo, distancia euclideana, de Manhatan (BUZAI, 2011), de Mahalanobis (PORTILLO, 2015), entre otros. Menciona que una de las desventajas de usar este método es que se pierden detalles de los datos, pero permite su simplificación. También menciona que el método permite muchos tipos de aplicaciones, entre los que se tienen: recuperación de información, exploración de datos científicos, aplicaciones web, marketing, análisis de ADN, entre otras.

1.5.2.5 Aprendizaje Híbrido

Son las que combinan las técnicas de clasificación supervisada y las no-supervisadas en el desarrollo en conjunto de una solución (SEIJAS, 2003). También se les dice híbridas a las técnicas que pueden ser usadas con o sin tener conocimiento a priori de ellas (VERA; BUSTAMANTE, 2007).

1.5.2.6 Reconocimiento de Patrones

Según Ruiz, Guzmán & Martínez (*apud* CARRASCO & MARTINEZ, 2011) definen el reconocimiento de patrones como “la ciencia que se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos o abstractos, con el propósito de extraer información que permita establecer propiedades de o entre conjuntos de dichos objetos los cuales nos permiten interpretar el mundo que nos rodea”.

Existen varios enfoques que permiten resolver los problemas de reconocimiento de patrones: Reconocimiento estadístico de patrones, Reconocimiento sintáctico de patrones, Redes neuronales y el Reconocimiento lógico combinatorio (CARRASCO & MARTINEZ, 2011).

Según Carrasco & Martínez (CARRASCO & MARTINEZ, 2011), existen algunos problemas relacionados con el reconocimiento de patrones, siendo el más importante la selección de atributos de relevantes dentro de un conjunto total de atributos que permitan medir a los objetos de estudio, el cual tiene como objetivo mejorar los procesos y resultados de clasificación y agrupamiento. Adicionalmente, menciona que es necesario encontrar atributos que brinden información relacionada al objeto de estudio, por ejemplo, si el objeto de estudio es identificar la orientación vocacional de una persona, no usaremos información relacionada a sus gestos o rostro, sino se usará información relacionada a sus gustos y aficiones.

Se debe tener en cuenta que no todos los atributos que se tienen pueden ser relevantes para el objeto de estudio. Algunos de estos atributos podrían causar ruido, ser casos atípicos, redundantes o no contribuir en la calidad de los resultados (CARRASCO & MARTINEZ, 2011).

1.5.3 *Metodologías*

A continuación, se muestran las metodologías a aplicar en el presente proyecto de fin de carrera.

1.5.3.1 Metodología para obtención de datos de los pacientes

En la Figura 1.1, se muestra el proceso de obtención de datos de los pacientes desde el primer contacto institucional.

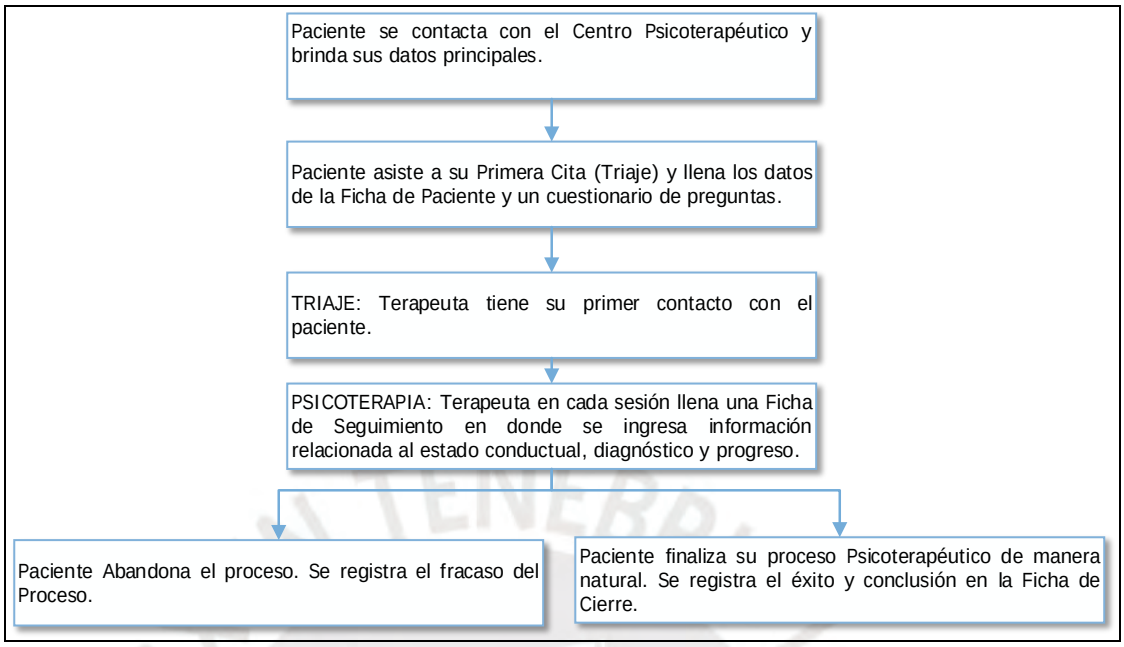


Figura 1.1: Proceso de obtención de datos del paciente [Elaboración propia].

1.5.3.2 Metodología para el procesamiento de datos

Para procesar los datos recolectados, se usó la metodología **SEMMA** de *data mining*, el cual permitió usar buenas prácticas para obtener mejores resultados (CAMARGO, 2011).

Esta metodología contribuye en el proceso de selección, exploración y modelado de grandes cantidades de datos con el fin de obtener patrones. La metodología cuenta con cinco fases: Muestreo, Exploración, Manipulación, Modelado y Valoración que se muestran a continuación (CAMARGO, 2011):

- **Muestreo:** El objetivo de esta fase es poder seleccionar una muestra representativa del problema en estudio. Esta representatividad es importante porque podría invalidar todo el modelo y los resultados. La forma común usada para la selección de la muestra es al azar, pero para el presente proyecto de fin de carrera se seleccionó los datos que nos brinden un alto grado de confianza.
- **Exploración:** En esta fase se explora la información buscando tendencias o anomalías los cuales permitan comprender los datos. Para esto, se apoya de herramientas de visualización o de técnicas estadísticas, como el análisis factorial de correspondencias y agrupaciones, que permitan visualizar las

relaciones entre variables. Se usó la herramienta Weka, el cual nos permitió visualizar estas relaciones de manera visual.

- **Manipulación:** En esta fase se manipulan los datos que fueron explorados en la fase anterior mediante la creación, selección y transformación de variables para definir el formato de los datos para ser introducido al modelo. En esta fase se simplificó y agrupó datos para reducir las variables del modelo.
- **Modelado:** En esta fase se aplican técnicas de modelado en minería de datos tales como: redes neuronales y árboles de decisión. También se pueden utilizar métodos estadísticos tradicionales, tales como: análisis de discriminante y análisis de regresión.
- **Valoración:** En la última fase se valoran los resultados mediante pruebas al modelo. Para esto se puede probar el modelo o los modelos con valores los cuales se tienen resultados conocidos. En esta fase se pueden utilizar métodos comparativos estadísticos para evaluar los resultados.

1.5.3.3 Metodología para el modelado

Para el modelado de datos, se hizo uso de los algoritmos de árboles de decisión. Se tomó esta decisión debido a que la mayoría de datos a analizar eran de orden cualitativo, por lo que, según la literatura, los algoritmos de árboles resultan ser adecuados para el análisis de estos tipos de datos (BARRENO-VEREAU, 2012). Algunas investigaciones realizadas usando arboles de decisión obtuvieron resultados satisfactorios haciendo uso de árboles de decisión sobre datos cualitativos. Algunas de estas investigaciones se muestran en el Estado del Arte.

1.6 Alcance

El presente proyecto de fin de carrera se desarrolla en el ámbito de las ciencias de la computación y se centra en predecir la permanencia de un paciente en un proceso psicoterapéutico en una institución donde se brindan estos servicios a pacientes que necesitan ayuda terapéutica. Para esto, se usan datos relacionados al terapeuta y paciente de una institución con el fin de determinar patrones que permitan predecir cómo evolucionará el proceso psicoterapéutico.

Para predecir la permanencia del paciente en el proceso psicoterapéutico, se utiliza un modelo predictivo el cual fue entrenado por medio de información histórica de los pacientes y terapeutas, los cuales fueron normalizados, pre-procesados y clasificados. Para la predicción se consideran aspectos personales, familiares y

económicos, así como el horario seleccionado para la primera cita, debido a que el marco del análisis se centra en la predicción de permanencia a partir de los datos obtenidos en el primer contacto entre el paciente y la institución. Como el enfoque del proyecto fue desarrollar un prototipo funcional, los resultados fueron presentados en una interfaz de fácil acceso y amigable. El resultado del prototipo funcional se limitará a predecir si el paciente permanecerá en el proceso por 16 a más citas efectivas, el cual es equivalente a superar los cuatro meses de permanencia, o en caso contrario, si el paciente superará las cuatro primeras citas efectivas correspondientes a la fase intermedia o si abandonará el proceso en cualquiera de las cuatro primeras citas correspondientes a la fase de indicación e inicial.

1.7 **Riesgos**

Para el presente proyecto, se han identificado los siguientes riesgos que se muestran en la Tabla 1.4.

Tabla 1.4: Riesgos identificados y medidas correctivas para mitigar.

Descripción	**Síntomas**	**Probabilidad**	**Impacto**	**Severidad**	**Mitigación**	**Contingencia**
Dificultad para acceder a la data real.	Falta de autorización para acceder a la base de datos institucional.	Alto	Alto	Alto	Solicitar de manera anticipada y formal los datos a usar al responsable de la base de datos.	Buscar en internet una base de datos de una institución clínica que brinda servicios de Psicoterapia.
Dificultad para obtener datos que permitan el análisis de predicción.	Conjunto de datos con bajo nivel de precisión.	Alto	Alto	Alto	Realizar muchas pruebas sobre distintos conjuntos de datos desde la misma base de datos.	Buscar otro conjunto de datos en internet.

Deficiente planificación del Proyecto.	Atraso en el Proyecto según la planificación.	Alto	Alto	Alto	Evaluar constantemente el estado actual del proyecto respecto a la planificación.	Ejecutar medidas correctivas para reducir los tiempos de las tareas restantes.
Curva de aprendizaje de las herramientas.	Dificultad al usar las herramientas.	Medio	Medio	Medio	Aprender a usar las herramientas antes del inicio del Proyecto.	Solicitar ayuda a una persona con experiencia en el uso de estas herramientas.

1.8 **Justificación**

La razón por la cual se desarrolla este Proyecto, es por la necesidad de conocer, mejorar y establecer una relación terapeuta-paciente en base al conocimiento adquirido de experiencias con pacientes antiguos con el fin de reducir los niveles de deserción. Según la investigación previa realizada, el establecimiento de esta relación puede ser determinante para el éxito del proceso psicoterapéutico, por lo que, predecir el éxito o fracaso del proceso, así como, en caso de fracaso, determinar en qué fase del proceso el paciente desertara, permitirá al terapeuta tomar medidas que permitan mejorar el tratamiento y por ende mejorar la relación terapeuta-paciente (BENITEZ et al, 2009).

Por lo tanto, según lo mencionado, se pretende desarrollar un prototipo funcional que permita predecir el éxito o fracaso de un proceso terapéutico con el fin de mejorar el tratamiento y la relación terapeuta-paciente. La aplicabilidad de este Proyecto se puede dar en otros ámbitos, por ejemplo, en la predicción de una decisión corporativa basada en el perfil del consumidor. Al finalizar el presente proyecto, este prototipo funcional podrá ser implementado en una herramienta la cual podrá ser utilizada en otras instituciones que pertenezcan al rubro de atención psicoterapéutica.

Finalmente, la investigación y desarrollo de este proyecto permitirá a futuros proyectos de software mejorar el prototipo o utilizar las investigaciones, herramientas y bases teóricas para el desarrollo de otras herramientas basadas en el aprendizaje de máquina. De esta manera, el presente proyecto de fin de carrera aportará conocimientos para nuevas investigaciones relacionadas a la minería de datos aplicada a procesos psicoterapéuticos y desarrollos de futuros proyectos de software afines.

2. CAPÍTULO 2: MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Marco conceptual

A continuación, se describen algunos conceptos y teorías que conforman el escenario del proceso psicoterapéutico, los cuales se deben tomar en cuenta durante el desarrollo del modelo predictivo.

2.1.1 *Marco conceptual relacionado con el proceso terapéutico*

A continuación, se analizan los conceptos relacionados al proceso terapéutico.

2.1.1.1 Psicoterapia

Según Mahoney: "Psicoterapia es la aplicación de procedimientos científicamente evaluados que capacitan a las personas para cambiar sus comportamientos, emociones o conductas mal adaptativas, por ellos mismos." (apud LIRIA, 2001).

Esta definición da a entender que la psicoterapia es un método terapéutico en donde intervienen dos o más personas (terapeuta y paciente) con el fin de capacitar al paciente mediante el autoconocimiento de sí mismo para poder cambiar su conducta.

2.1.1.2 Proceso terapéutico

El proceso terapéutico está dividido en 4 fases: fase de indicación, fases iniciales, fases intermedias y la fase de terminación que se presentan a continuación (LIRIA, 2001):

- **Fase de Indicación:** Esta fase tiene como objetivo determinar si es necesario iniciar un proceso terapéutico con el paciente. Para esto se analiza si este proceso puede ser útil para el paciente mediante una entrevista presencial. Esta fase también es conocida como Triaje.
- **Fases Iniciales:** En esta fase empieza el proceso psicoterapéutico. Durante este proceso se tocan temas como la "Alianza terapéutica" y "el contrato terapéutico". Estos temas permiten al paciente conocer las reglas a usar durante todo el proceso psicoterapéutico, de esta forma, el paciente entiende cual es el trabajo terapéutico y cuáles son las funciones de cada uno de los autores. Durante este proceso se empieza a crear el vínculo terapeuta-paciente, el cual termina siendo determinante para el éxito del proceso.
- **Fases Intermedias:** Durante esta fase, el terapeuta "construye las pautas del problema y el cambio". Para esto, el terapeuta permite al paciente abordar

diferentes etapas de su vida, su conducta, pensamientos, entre otros; con el fin de conocer a fondo el problema que tiene el paciente, para posteriormente guiarlo a cambiar en esos aspectos mediante el autoconocimiento.

- **Fases de Terminación:** Según De Rivera (DE RIVERA, 1990), constituye la "graduación" del paciente como capacitado en el funcionamiento de su propia mente como individuo maduro, autónomo e independiente. El tiempo que puede tomar a llegar a esta fase es variable y dependerá enteramente del paciente.

2.1.1.3 Deserción terapéutica

Según Benitez (BENÍTEZ, 2009), el abandono es la terminación prematura o temprana, ante la inasistencia de sus citas antes de terminar el proceso terapéutico, la cual se puede dar en cualquier fase del proceso, posterior al primer contacto con el terapeuta. El autor menciona muchas razones por las cuales este proceso puede terminar siendo interrumpido, entre las que están, factores económicos, dependencia, obligación a seguir el tratamiento terapéutico, entre otros.

2.1.2 *Marco conceptual relacionado a los arboles de decisión*

A continuación, se analizan los conceptos relacionados a los arboles de decisión.

2.1.2.1 Arboles de Decisión

Según Gervilla (GERVILLA, 2009), los arboles de decisión permiten representar una serie de reglas de forma gráfica con el fin de tomar una decisión sobre la asignación de un valor de salida. Se encuentran compuestos por nodos, ramas y hojas o nodos hoja, cuyas características son:

- Los nodos son los datos de entrada
- Las ramas son grupos de registros en las variables de entrada.
- Los nodos hojas o las hojas son los valores de la variable de salida.

Este tipo de clasificador construye un árbol de decisión de múltiples caminos en el que para cada nodo se busca el atributo que provea mayor ganancia de información para la clase (DUBIAU; ALE, 2013). La forma en que se clasifican los datos es por medio de una o más reglas asociadas que satisfacen ciertas características y asigna un sentido basándose en predicciones. (VÁZQUEZ, 2009).

Por medio de un algoritmo de aprendizaje supervisado se realizan sucesivas divisiones para maximizar la distancia entre grupos por cada división; termina de

formarse cuando todos los registros de una rama tienen como variable de salida un nodo hoja puro (GERVILLA, 2009).

En la Figura 2.1, se puede ver un ejemplo básico de un árbol de decisión el cual por medio de cinco variables se puede determinar el éxito o fracaso de un proceso. Los nodos son representados por rectángulos y las hojas por círculos. Se puede apreciar que dos o más ramas pueden salir de un nodo.

Según Rokach (ROKACH, 2014), cada nodo contiene cierta característica y cada rama un rango de valores. Las hojas terminan siendo el resultado final de las sucesivas divisiones en los nodos. La complejidad de los arboles tienen un efecto sobre la precisión de los resultados. Para medir la complejidad se debe tener en cuenta las siguientes métricas: número de nodos, número de hojas, profundidad del árbol y el número de atributos usados. Los arboles de decisión suelen ser intuitivos y fácil de entender.

Existen varios algoritmos para construir árboles de decisión, pero para el análisis se usó los siguientes: J48, REPtree, Random Tree, LMT y Random Forest. Estos algoritmos se encuentran implementados en la herramienta WEKA, el cual se usó para la determinación del modelo algorítmico que debía usarse para el problema planteado en el presente proyecto de fin de carrera. A continuación, se describen los algoritmos:

- **J48:** Según Martinez et al. (MARTINEZ et al., 2009), es un algoritmo el cual construye iterativamente un árbol, el cual va agregando nodos o ramas que reduzcan la diferencia entre los datos. Tiene la capacidad de utilizar atributos numéricos y vacíos para crear las reglas del árbol. Este algoritmo, con el fin de clasificar una nueva instancia, prueba cada uno de los valores del atributo en base a su estructura hasta que encuentra una hoja, el cual contiene los valores de la clase para cada instancia.
- **REPTree:** Según Kalmegh (KALMEGH, 2015), este algoritmo construye un árbol de decisión con la lógica de un árbol de regresión (JUÁREZ, O., & CASTELLS, 2010) y mediante esta crea múltiples árboles en diferentes iteraciones. Luego selecciona el mejor árbol el cual termina siendo el representativo. Este árbol de decisión usa la ganancia de información como el *splitting criterion* y para reducir el error utiliza el *prunning* o poda.

- **Random Tree:** Según Kalmegh (KALMEGH, 2015), es un clasificador supervisado, cuyo algoritmo genera muchos aprendices individuales a partir de datos nominales y numericos. Utiliza un conjunto de árboles predictores el cual es llamado bosque. La clasificación se realiza tomando las características del vector de entrada y lo clasifica con todos los árboles del bosque, el resultado es la etiqueta de la clase que recibió la mayoría de los votos. Para el caso de regresión, es el promedio de las respuestas de todos los árboles del bosque.
- **Logistic Modelo Tree (LMT):** Según Landwehr et al. (LANDWEHR et al., 2003), es una estructura de árbol de decisión de serie con funciones de regresión logística en las hojas. Para el caso en que un nodo tenga "k" nodos hijos, por tanto "k" ramas, este es clasificado en cada una de sus ramas dependiendo del valor de su atributo. En caso los atributos sean numéricos, el nodo solo tiene dos hijos, para esto se compara el valor del atributo con un umbral. Por ejemplo, se puede tomar un valor intermedio y colocar los datos menores a la izquierda y los mayores iguales a la derecha.
- **Random Forest:** Según Breiman (BREIMAN, 2001), es un algoritmo que utiliza una combinación de árboles predictores de tal manera que cada árbol depende de los valores de un vector aleatorio muestreado de manera independiente con una misma distribución para todos los arboles del bosque. El error de generalización de un bosque de clasificadores converge a un límite mientras va creciendo. Este error depende de la significancia de los arboles individuales en el bosque y la correlación entre ellos.

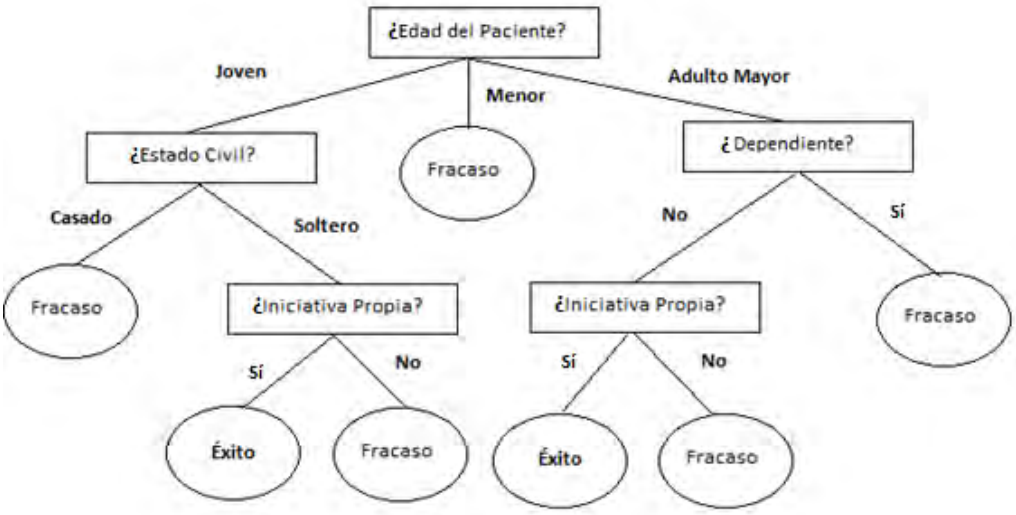


Figura 2.1: Representación gráfica de un árbol de decisión de ejemplo para determinar el éxito de un proceso en base a variables básicas. [Elaboración propia].

Los arboles de decisión, en comparación con otros algoritmos de predicción, suelen obtener muy buenos resultados. Durante el proceso de investigación, se encontró varias investigaciones relacionadas al análisis comparativo entre algoritmos de predicción. En una de las investigaciones de la Universidad Politécnica de Madrid (MARCANO et al, 2011) se investiga sobre el análisis comparativo y se obtienen buenos resultados con los algoritmos de árboles de decisión.

2.2 Estado del arte

La minería de datos es un proceso que permite analizar datos, cuya aplicación se puede dar en la medicina, economía, negocios, educación, entre otros. En esta sección se menciona algunas investigaciones y herramientas existentes que abarcan la minería de datos aplicada a la predicción del éxito en un proceso psicoterapéutico.

2.2.1 *Objetivos de la revisión del estado del arte*

El objetivo de revisar el estado del arte es poder recuperar el conocimiento acumulado actual acerca de la predicción del éxito en un proceso psicoterapéutico. Para esto, se ha realizado una investigación sobre los estudios e investigaciones relacionadas existentes, para poder analizarlos, observar sus características y hacer comparaciones para que en base a los métodos utilizados y soluciones podamos establecer un modelo para predecir la permanencia del paciente durante el proceso psicoterapéutico.

2.2.2 *Método usado en la revisión del estado del arte*

El método utilizado para hacer la revisión del estado del arte es la Revisión Tradicional, el cual permitió obtener información transparente, auditable y confiable. Para hacer uso de este método, se utilizó las siguientes bases de datos: IEEE xplorer, ACM, APA Psycnet y Wiley.

2.2.3 *Formulación de la pregunta*

Para la Revisión Tradicional se formuló la siguiente pregunta: ¿Qué investigaciones se han realizado con respecto al éxito del Proceso Psicoterapéutico usando algunos métodos computacionales?

Para responder estas preguntas se hicieron uso de las siguientes palabras claves: *"success", "process", "psychotherapy", "data mining", "prediction", "computational".*

2.2.4 *Selección de las fuentes*

Para realizar las búsquedas en las bases de datos, se agruparon las palabras claves con los operadores lógicos "*AND*" y "*OR*".

Cadena general de búsqueda: (success "*OR*" process) "*AND*" psychotherapy "*AND*" (data mining "*OR*" classification "*OR*" computational "*OR*" prediction).

En la revisión tradicional, la cadena de búsqueda fue usada en algunos casos de manera disgregada al operador "*OR*" creando diferentes cadenas de búsqueda. El operador "*AND*" fue usado en todas las búsquedas de acuerdo a la cadena general de búsqueda.

Se encontraron algunas investigaciones relacionadas a la permanencia en el proceso terapéutico a nivel clínico, los cuales fueron tomados en consideración como base para identificar los datos que se deben analizar. Adicionalmente, se encontraron investigaciones relacionadas a la minería de datos orientada a la predicción del éxito del proceso psicoterapéutico.

2.2.5 *Investigaciones existentes relacionadas a los árboles de decisión*

A continuación, se presentan las investigaciones más resaltantes en el uso de árboles de decisión sobre datos cualitativos.

2.2.5.1 Análisis comparativo de la deserción universitaria utilizando árboles de decisión y regresión logística

Para la investigación de Barreno (BARRENO-VEREAU, 2012) relacionada a la deserción universitaria, se hizo un análisis comparativo entre un modelo de regresión logística y un árbol de decisión. Los atributos, los cuales fueron seleccionados por el método inductivo, fueron de orden cualitativo y cuantitativo. Para la regresión logística, se tuvieron que transformar las variables cualitativas de dos categorías de forma binaria, para la transformación de variables cualitativas de tres a más, se tuvo que hacer uso de las variables *dummy*, los cuales son un tipo de representación binaria. Por el lado del árbol de decisión, no fue necesario ninguna transformación. Según las conclusiones de esta investigación, ambos modelos presentaron resultados muy similares con un alto nivel de predicción. El modelo de regresión logística presento un 94.27% de precisión, mientras el árbol de decisión un 95.08%, siendo este último ligeramente más preciso.

2.2.5.2 Aplicación de la metodología de Arboles de Decisión para la determinación de la mortalidad de un infarto agudo de miocardio

Según la investigación de Trujillano (TRUJILLANO et al., 2008), al comparar los modelos de árbol de decisión, regresión logística y red neuronal, los resultados de precisión resultaron ser muy cercanos. Por otro lado, se menciona que a pesar de tener resultados muy similares y por tanto no poder decidir que metodología es la mejor para este caso en específico, es útil analizar las características de cada uno de los modelos. Partiendo de ello, los autores de la investigación concluyen que el método basado en un árbol de decisión ofrece como ventaja su simplicidad en su utilización e interpretación debido a que sus procesos de interpretación no necesitan procesos matemáticos.

2.2.5.3 Metodología para la predicción del grado de Riesgo Psicosocial en docentes aplicando Arboles de decisión

En la investigación de Mosquera (MOSQUERA et al., 2016), se utilizaron 114 variables cuantitativas y cualitativas, por lo que se determinó usar los algoritmos Naive Bayes y J48. Los datos tuvieron que pasar un proceso de pre-procesamiento, limpieza y reducción de atributos para que los autores puedan realizar el análisis algorítmico. Luego de la comparación algorítmica usando la matriz de confusión, se obtuvieron resultados muy cercanos, sin embargo, el algoritmo J48 presento una precisión mayor en comparación de Naive Bayes. Entre las conclusiones de los autores, destaca que la metodología seleccionada (J48) presenta ventajas

importantes sobre otras metodologías debido a su: mayor precisión, permisión para clasificar diversas características de la población y capacidad de identificar diversidad de fuentes que pueden ser tratados.

2.2.6 *Investigaciones existentes relacionadas a la permanencia del Proceso Psicoterapéutico*

A continuación, se presentan las investigaciones más resaltantes.

2.2.6.1 Análisis del tiempo del proceso Psicoterapéutico mediante una red neuronal

La investigación de Villmann (VILLMANN, 1999), tiene como objetivo predecir el éxito de un proceso terapéutico por medio de un conjunto de datos de entrenamiento. Los datos de entrenamiento se encuentran basadas en pruebas empíricas con los datos de los pacientes. El análisis de la información de esta investigación se basa en el estudio empírico de los datos de los pacientes referentes a su personalidad, para esto se usan 26 valores obtenidos por los terapeutas mediante un cuestionario, los cuales permiten tener un amplio espectro de la personalidad de los pacientes. Esta investigación menciona que para tener un resultado más a fondo, se puede incorporar la fase del proceso en la cual se encuentra el paciente.

La investigación concluye que los resultados podrían ser mejorados y tener una mayor precisión si se usa una mayor cantidad de datos de entrenamiento (VILLMANN,1999).

2.2.6.2 Predicción diferencial con modelos de tratamiento - respuesta

Esta investigación se basa en el monitoreo de resultados de los pacientes y su progreso de acuerdo a la respuesta del paciente al tratamiento. El progreso del paciente se toma como experiencia previa para mejorar en la selección de la medicación y tratamiento. Esta investigación se basa en el método *Expected Treatment Response (ETR),* en el cual se propone una mejora para ampliar el método, debido a que los predictores trabajaban solo para subconjuntos de pacientes específicos (KRAUSE, HOWARD, y LUTZ, 1998). Para mejorar la precisión de los resultados, la investigación presenta como estrategia desagregar a los pacientes en subgrupos homogéneos, mediante técnicas de vecinos más cercanos, con el fin de generar perfiles de tratamiento más óptimos, los cuales son usados para predecir y realizar un seguimiento del progreso en las diferentes modalidades del tratamiento (LUTZ y SAUNDERS et al., 2006).

2.2.6.3 Predicción basada en el vecino más cercano

La investigación de Lutz et al. (LUTZ et al., 2006), se basa en la premisa que los métodos empíricos tienen mejores resultados que el juicio clínico. Para esto se propone una herramienta que permita realizar predicciones basadas en el método de "vecinos más cercanos" con el fin de modelar un curso de tratamiento esperado para cada paciente.

Lutz et al. (LUTZ et al., 2006), menciona que la forma en cómo se recolecta la información es por medio de un cuestionario de 45 preguntas, los cuales son obtenidos al inicio de cada sesión. Las preguntas abarcan aspectos fundamentales del progreso de los pacientes: malestar subjetivo, las relaciones interpersonales y el desempeño de los roles sociales. Según los resultados de esta herramienta, los métodos estadísticos predictivos resultaron tener mayor precisión que los métodos predictivos clínicos.

2.2.7 *Herramientas existentes*

A continuación, se presentan productos softwares existentes, los cuales brindan herramientas de minería de datos que pueden servir de apoyo para el presente tema de proyecto de fin de carrera.

2.2.7.1 RapidMiner

Según Han, Rodriguez & Beheshti (HAN, RODRIGUEZ, BEHESHTI, 2008), RapidMiner es una herramienta de minería de datos que permite crear modelos de análisis de datos. Cuenta con una función de arrastrar y soltar para crear un flujo de trabajo. Cada bloque del flujo, contiene operaciones de análisis de datos. Es posible implementar operaciones de análisis de datos propias. Esta herramienta de aprendizaje automático, desarrollada en Java, permite realizar tareas de análisis de datos de manera fácil y con una interfaz amigable. Es usado para analizar datos científicos para predicciones. Cuenta con soporte para Windows, Linux, Mac OS.

2.2.7.2 KEEL

Según Alcala-Fdez (ALCALA-FDEZ, 2008), KEEL es una herramienta que permite evaluar el comportamiento de algoritmos evolutivos para los diferentes métodos de minería de datos, tales como: regresión, clasificación, agrupación, entre otros. Incluye una gran librería de algoritmos, por lo que simplifica la labor de análisis, reduciendo de manera considerable los niveles de conocimiento de programación. Simplifica la integración con las diferentes técnicas de procesamiento previo. Esta

herramienta ha sido desarrollada en Java, por lo que puede ser instalada en diferentes sistemas operativos, tales como: Windows, Linux, Mac OS.

En la Tabla 2.1, se resume las herramientas existentes que pueden permitir el desarrollo del análisis de datos para la predicción. En el mismo, se hace mención a Weka, el cual es una herramienta de código abierto, que incluye librerías muy útiles para ser usado en caso de clases desbalanceadas, cuyo problema fue necesario solucionar para el análisis de datos.

Tabla 2.1: Resumen de las herramientas existentes [Elaboración propia].

PRODUCTO	**DESCRIPCIÓN**
WEKA	• Implementado en Java. • OPEN SOURCE • Disponible en Windows, Linux, Mac OS. • Cuenta con una gran cantidad de algoritmos de minería de datos. • Permite la experimentación de análisis de datos, utilizando principalmente las técnicas de aprendizaje automático.
RAPIDMINER	• Implementado en Java. • Disponible en Windows, Linux, Mac OS. • Entorno grafico amigable para poder crear de manera gráfica operaciones de flujos de datos. • Permite implementar operaciones de análisis de datos propia.
KEEL	• Implementado en Java. • OPEN SOURCE • Disponible en Windows, Linux, Mac OS. • Permite evaluar el comportamiento de algoritmos evolutivos para los diferentes métodos de minería de datos. • Incluye una gran cantidad de librerías de algoritmos. • No es necesario tener altos niveles de programación • Simplifica la integración con las diferentes técnicas de procesamiento previo.

2.2.8 *Conclusiones sobre el estado del arte*

Luego de haber realizado la revisión del estado del arte por medio de la Revisión Tradicional, se puede concluir que existen investigaciones relacionadas al presente tema de proyecto de fin de carrera, los cuales nos permiten tener una visión más clara sobre el problema, los cuales podrían aportar de manera significativa para la creación del modelo algorítmico que permite predecir la permanencia del paciente en el proceso psicoterapéutico.

Las herramientas existentes pueden brindarnos una gran ayuda al momento de crear el modelo algorítmico a usar en el prototipo funcional. Tanto las investigaciones previas y herramientas existentes, nos permiten tener una mayor proximidad a la solución del problema planteado.

3. CAPÍTULO 3: SELECCIÓN DE LOS CONJUNTOS DE DATOS

En este Capítulo se desarrolla el primer objetivo específico el cual consiste en recolectar, analizar, estructurar y pre-procesar un conjunto de datos a partir de la información de los pacientes y los terapeutas de la institución. Para desarrollar el objetivo, se cuenta con un conjunto de datos, el cual, en base a la investigación realizada, fue acotada dentro del alcance del proyecto. Como resultado, se obtuvo dos resultados esperados: un conjunto de datos seleccionado, pre-procesados y transformados para el proceso de clasificación y tres conjuntos de datos de entrenamiento estructurados para llevar a cabo el proceso de clasificación.

3.1 Selección, pre-procesado y transformación de datos

En esta Sección se desarrolla el Resultado Esperado 1, para esto, se han dividido en sub-secciones para hacerlo más ordenado y entendible.

3.1.1 *Conjunto de Datos*

El presente Proyecto de fin de carrera tiene como datos de entrenamiento un conjunto de datos obtenidos de una base de datos, el cual pertenece a una clínica que brinda servicios de Psicoterapia. La información contenida en esta base de datos contiene información del paciente, terapeuta, citas, pagos, entre otras. Para el análisis se utilizó solo las tablas que tengan información relacionada a los pacientes y los terapeutas.

En la Figura 3.1 se puede observar el modelo de la base de datos de donde se extrae la información para los datos de entrenamiento. Las tablas contienen información histórica de la institución, desde el año 2014, las cuales en algunos casos se tiene información incompleta, no estandarizada, cualitativa no parametrizada incluyendo errores de digitación. Para el proceso de selección de datos se analizó cada uno de los atributos de las tablas, así como sus relaciones de tal forma que sea convertido en datos de entrenamiento.

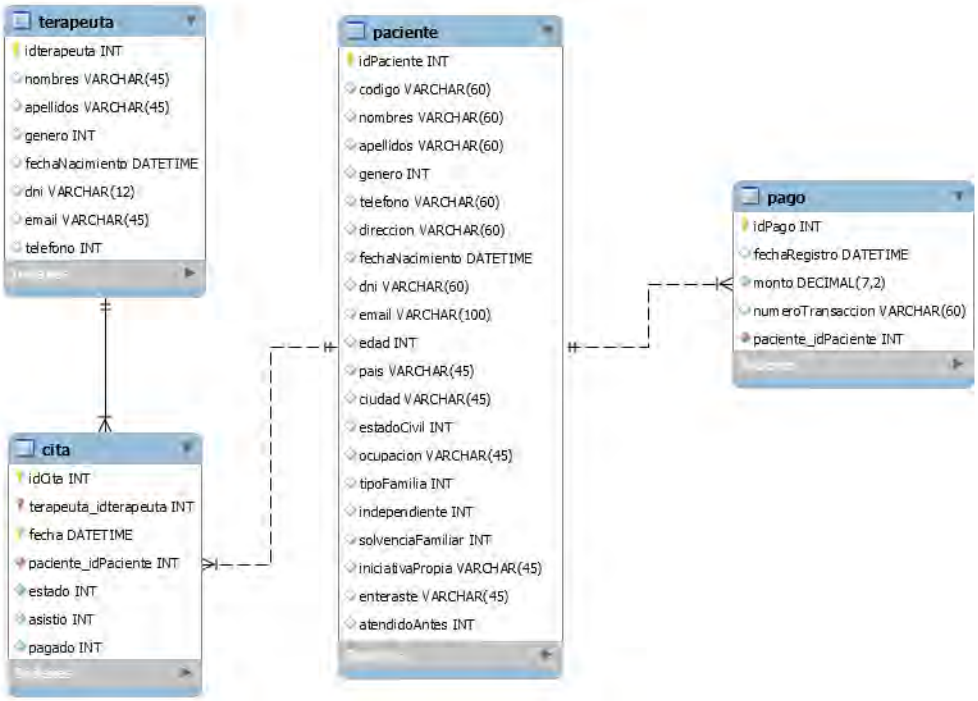


Figura 3.1: Modelo acotado de la Base de Datos con los principales atributos de las tablas [Elaboración propia].

3.1.2 *Selección de los datos de entrenamiento*

Para la selección de datos de entrenamiento se analizó cada uno de los atributos de las tablas Paciente, Terapeuta, Cita y Pago.

Tabla 3.1: Análisis de las tablas Paciente, Terapeuta, Cita y Pago [Elaboración propia].

TABLAS	CANT. DE ATRIBUTOS	CANT. DE REGISTROS
Paciente	22	4400
Terapeuta	8	24
Cita	7	50087
Pago	5	18441

De la Tabla 3.1, se puede observar que la tabla Cita contiene una gran cantidad de datos relacionado al proceso terapéutico de los pacientes, por lo que analizar los atributos de la misma fue importante para obtener patrones relacionada a sus citas. Con respecto a la tabla Paciente, la cantidad de registros es importante, por lo que es posible encontrar patrones relacionados a su información personal, familiar y económico del paciente. La tabla Terapeuta contiene los registros de los terapeutas de los pacientes, los cuales terminan siendo un atributo de los pacientes. Finalmente,

la tabla Pago permitirá determinar el comportamiento relacionado a su compromiso económico con la institución, sin embargo, luego del análisis este atributo no fue considerado.

De la Tabla 3.2, 3.3 y 3.4, se tienen los atributos de las tablas Cita, Terapeuta y Paciente correspondientemente. En cada una se muestran los atributos, tipos de datos, descripción y los posibles valores que tienen. Los atributos que se muestran son los que se usaran en el análisis.

Tabla 3.2: Atributos de la tabla Cita a considerar en el análisis [Elaboración propia].

ATRIBUTOS	TIPO DE DATO	DESCRIPCIÓN	VALORES
terapeuta_idterapeuta	int	Relación entre la tabla Terapeuta y Cita [Llave Primaria]	Entero
fecha	datetime	Fecha de la cita [Llave Primaria]	Fecha en formato Datetime
paciente_idPaciente	int	Relación entre la tabla Paciente y Cita [Llave Foránea]	Entero
estado	int	La cita ¿está activa?	0: Anulado, 1: Activo
asistio	int	El paciente ¿asistió a la cita?	0: No asistió, 1: Asistió
pagado	int	La cita ¿fue pagada?	0: No pagado, 1: Pagado

Tabla 3.3: Atributos de la tabla Terapeuta a considerar en el análisis [Elaboración propia].

ATRIBUTOS	TIPO DE DATO	DESCRIPCIÓN	VALORES
idterapeuta	int	Identificador del Terapeuta [Llave Primaria]	Entero
genero	int	Género del Terapeuta	1: Femenino, 2: Masculino
fechaNacimiento	datetime	Fecha de nacimiento del Terapeuta	Fecha en formato Datetime

Tabla 3.4: Atributos de la tabla Paciente a considerar en el análisis [Elaboración propia].

ATRIBUTOS	TIPO DE DATO	DESCRIPCIÓN	VALORES
genero	int	Género del Paciente	1: Femenino, 2: Masculino
edad	int	Edad del Paciente al ser registrado en la institución	Entero
estadoCivil	int	Estado Civil del Paciente	1: Soltero, 2: Casado, 3: Conviviente, 4: Divorciado, 5: Viudo
ocupacion	varchar	Ocupación del Paciente	Cadena de caracteres
tipoFamilia	int	Tipo de familia de donde proviene el Paciente	1: Nuclear, 2: Extendida, 3: Monoparental, 4: Ensamblada, 5: Conviviente, 6: Solo

independiente	int	El paciente ¿es Independiente?	0: No, 1: Si
solvenciaFamiliar	int	Nivel de socioeconómico del Paciente	1: A, 2: B, 3: C , 4:D, 5:E
iniciativaPropia	varchar	El paciente ¿decidió por sí mismo iniciar el proceso terapéutico?	Caracteres
enteraste	varchar	¿Cómo el Paciente se enteró de la institución?	Caracteres
atendidoAntes	int	¿Fue atendido antes por un terapeuta en otra institución o por consulta privada?	0: No, 1: Si

3.1.3 *Pre-tratamiento de datos*

Después de analizar los atributos, se identificó que algunos de estos contienen datos de poca relevancia, por ejemplo, con respecto al paciente y terapeuta: dirección, correo electrónico, DNI, teléfono y sus respectivos identificadores. Estos datos fueron excluidos del análisis. Luego de seleccionar los atributos, es necesario estandarizar y normalizar los datos para que puedan ser analizados con la herramienta Weka. La Tabla 3.5 resume el proceso de transformación de los atributos que no han sido normalizados, estandarizados y que presentan problemas en digitación.

Para la transformación de datos, se utilizaron *querys*, los cuales en una primera etapa permitió identificar *tokens*. Posteriormente se categorizaron en grupos normalizando los datos.

Tabla 3.5: Transformación de datos de los atributos seleccionados no estandarizados [Elaboración propia].

ATRIBUTO	TIPO DE DATO	VALORES INICIALES	TRANSFORMACIÓN
ocupación_num	int	Cadena de caracteres con errores de digitación no parametrizado. Los valores contenían nombres de profesiones sin categorización.	1: Ingeniero, 2: Licenciados, 3: Técnico, 4: Empleado, 5: Oficio, 6: Ama de casa, 7: Estudiante, 8: Independiente, 9: Desempleado
iniciativaPropia_num	int	Cadena de caracteres no estandarizado. Ejemplo: S, N, SI, si, NO, no	1: Si, 2: No
enteraste_num	int	Cadena de caracteres no estandarizado. Los valores incluían frases como "por recomendación", "sugerido por", "por internet", "fb", "web", etc.	1: Amigo, 2: Familiar, 3: Radio, 4: Internet, 5: Facebook

3.1.4 *Pre-procesamiento de los datos de entrenamiento*

En esta Sección se explica los criterios utilizados para el filtrado sobre los datos para obtener un conjunto en donde se pueda aplicar la minería de datos. Para esto se analizó el conjunto de datos y se eliminó los que generen ruido, sean inconsistentes y tengan valores incompletos.

Como primer filtro, se procedió a eliminar todos los registros que se encuentren incompletos. Se debe tener en cuenta que muchos registros que se encuentran en la tabla “Paciente”, corresponden al registro de personas que se han comunicado con

la institución por medio de sus canales de atención y han sacado una cita, por lo que no se tiene la información completa de los mismos. Siendo la cantidad de citas efectivas importante para el modelo a desarrollar, es necesario retirar los registros que no aporten al estudio. Para filtrar estos elementos, se usó el atributo “código”, el cual es esencial para determinar si es un paciente que por lo menos ha asistido a una cita o no. En caso el paciente haya asistido a una cita, este debe tener un código de paciente, caso contrario, tiene el valor de “NULL”. En la Tabla 3.6 se resume el análisis realizado sobre la tabla “Paciente” con la cantidad de registros válidos para los siguientes análisis.

Tabla 3.6: Análisis de los registros incompletos a eliminar de la tabla Paciente. [Elaboración propia].

TABLAS	CANT. DE REGISTROS INCOMPLETOS	CANT. DE REGISTROS VALIDOS
Paciente	1779	2621

Para los resultados de predicción de los modelos generados, se planteó que esta se encuentre vinculada a la cantidad de citas efectivas del paciente. Según la investigación realizada sobre la institución, el cual es estudio, se tuvo que las probabilidades de deserción posterior a los 4 meses o 16 citas se reducen significativamente, por lo que es muy probable que el paciente haya creado un vínculo con el terapeuta y continúe el proceso terapéutico. Bajo esto, para evitar tener pacientes que no alcanzan los 4 meses o 16 citas debido a la antigüedad del terapeuta en la institución, se eliminaron del análisis todos los terapeutas que tenían menos de 6 meses laborando en la institución. En la Tabla 3.7 se observa la cantidad de registros luego de quitar los registros que causan ruido al análisis.

Tabla 3.7: Análisis de las tablas Paciente y Terapeuta después de la eliminación de terapeutas nuevos [Elaboración propia].

TABLAS	CANT. DE ATRIBUTOS	CANT. DE REGISTROS
Paciente	22	2407
Terapeuta	8	18

Luego de este análisis, se evaluó si los datos presentaban alguna inconsistencia. En la Figura 3.2 se observa que al realizar el análisis gráfico de los datos entre los

atributos “EstadoCivil” y “EdadP”, presentan incoherencia entre ellas. Un menor de edad solo puede tener como Estado Civil: Soltero o valor=1, sin embargo, algunos pacientes presentan un estado civil diferente a soltero siendo menores de edad.

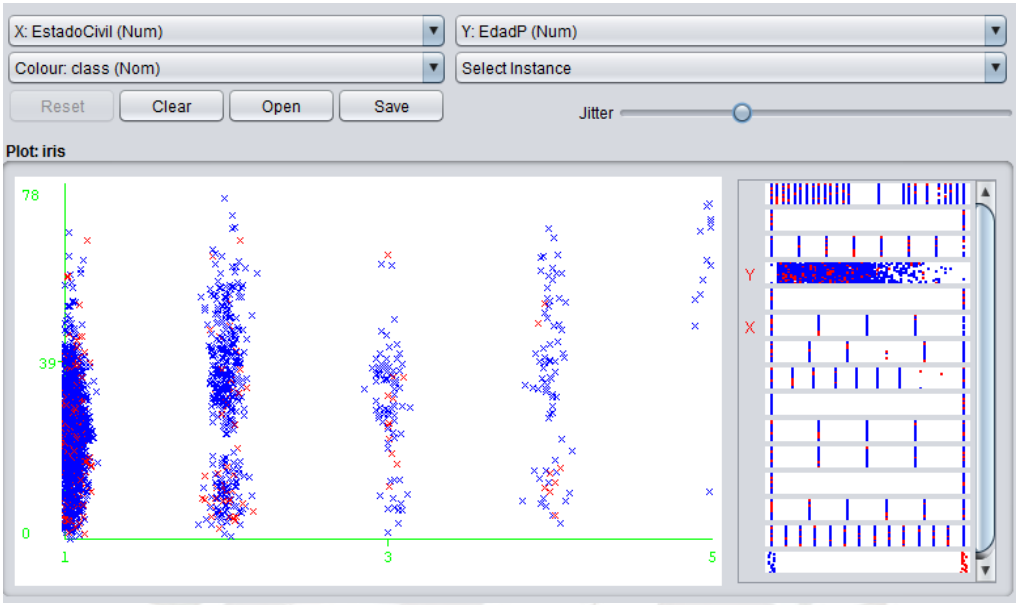


Figura 3.2: Análisis de datos entre “Estado Civil” y “Edad Paciente” [Elaboración propia].

En la Tabla 3.8, se puede observar la cantidad de datos inconsistentes en la relación “EstadoCivil” y “EdadP”. Luego de esto, se procedió a retirar estos datos del análisis.

Tabla 3.8: Incoherencia entre la relación “EstadoCivil” y “EdadP” [Elaboración propia].

VALOR	**CANTIDAD DE DATOS INCOHERENTES**
2= Casado	116
3= Conviviente	15
4= Divorciado	19
5= Viudo	1
TOTAL	**151**
TABLA PACIENTE DESPUES DE FILTRAR DATOS	**2256**

Finalmente, luego de retirar los registros que pueden causar ruido en nuestro análisis, obtenemos un conjunto de datos sobre el cual se podrá trabajar el modelo de predicción para el estudio.

3.1.5 *Transformación de datos*

Para el proceso de transformación de datos fue necesario definir los atributos necesarios a usar en los datos de entrenamiento. Para esto se evaluó los atributos que se tienen en las tablas de la base de datos y las necesidades en datos que se desea obtener.

Para el caso, los terapeutas que atienden en la institución eran recientemente egresados de la universidad, por lo que se asume que no tenían experiencia previa. Por lo tanto, para el estudio se tomó a todos desde su ingreso a la institución sin experiencia.

Según la investigación realizada, la experiencia del terapeuta es un atributo que puede determinar el proceso terapéutico, por lo que es necesario medirlo a partir de los datos que se tienen. Para esto, se tomó al terapeuta con la mayor cantidad de horas efectivas y la cantidad de horas de cada terapeuta según atendía a un paciente nuevo. El valor obtenido de esta relación se convirtió en un valor entero, el cual para el análisis será el atributo "Experiencia". Este valor cambia conforme el terapeuta atienda nuevos pacientes, por lo que este valor se incrementa progresivamente en la historia de sus pacientes.

$$Experiencia = \frac{Horas\ efectivas\ del\ terapeuta\ X\ al\ atender\ un\ paciente\ nuevo}{Máxima\ cantidad\ de\ horas\ efectivas\ de\ un\ terapeuta\ Y}$$

Los niveles de deserción de cada terapeuta ocupan un lugar importante en el análisis. Para esto se calculó el valor correspondiente a los niveles de deserción por cada mes. Según la investigación realizada, los niveles de deserción se reducen de manera significativa después del cuarto mes, por lo que los niveles de deserción por cada terapeuta fueron calculados para los meses 1, 2, 3 y 4. Para esto, se tomó las fechas de inicio y abandono de cada uno de los pacientes para determinar en qué mes del proceso optó por abandonar el proceso. El valor de los mismos correspondió al porcentaje de deserción por mes por terapeuta el cual fue re-calculado por cada paciente nuevo que el terapeuta atendió.

Para determinar el éxito del proceso psicoterapéutico, se estableció un periodo de cuatro meses o dieciséis citas efectivas como meta para determinar que el paciente pudo crear un vínculo con el terapeuta, por ende, existen altas probabilidades que el paciente logré terminar con éxito el proceso psicoterapéutico. El cálculo fue realizado

a partir de la cantidad de citas efectivas de cada paciente y se diferenció el éxito con valor 1 y el fracaso con valor 0.

3.2 Resultado Esperado 2: Conjunto de datos de entrenamiento

En esta Sección se desarrolla el Resultado Esperado 2, el cual contiene el análisis para la división de los conjuntos de datos de entrenamiento para los modelos de predicción.

3.2.1 *Clases y Atributos*

Para el presente Proyecto, se desarrollaron 3 sub-conjuntos de datos para el entrenamiento y validación de los modelos de predicción, por lo cual, se separó 3 conjuntos de datos de entrenamiento para su posterior análisis algorítmico.

El primer modelo a desarrollar permitió predecir el éxito o fracaso del proceso psicoterapéutico. Para realizar esta predicción se tomó como referencia la estadística histórica de los niveles de deserción de la institución durante los primeros cuatro meses o 16 citas de permanencia del paciente en el proceso. El primer conjunto de datos dividió en dos el conjunto de datos: pacientes con menos de 16 citas y pacientes con 16 a más citas. Para justificar esta división, se desarrolló una prueba de hipótesis sobre la media de una distribución continua.

La Figura 3.3 muestra los niveles de deserción por meses entre los periodos Noviembre-2015 y Agosto-2016, periodo en el cual se empezó a tomar muestras para medir los niveles de deserción. Los meses comprendidos entre Setiembre-2014 y Noviembre-2015 se encontraron dentro del análisis pero no son tomados en cuenta por la poca cantidad de pacientes y por nuevos ingresos de terapeutas a la institución.

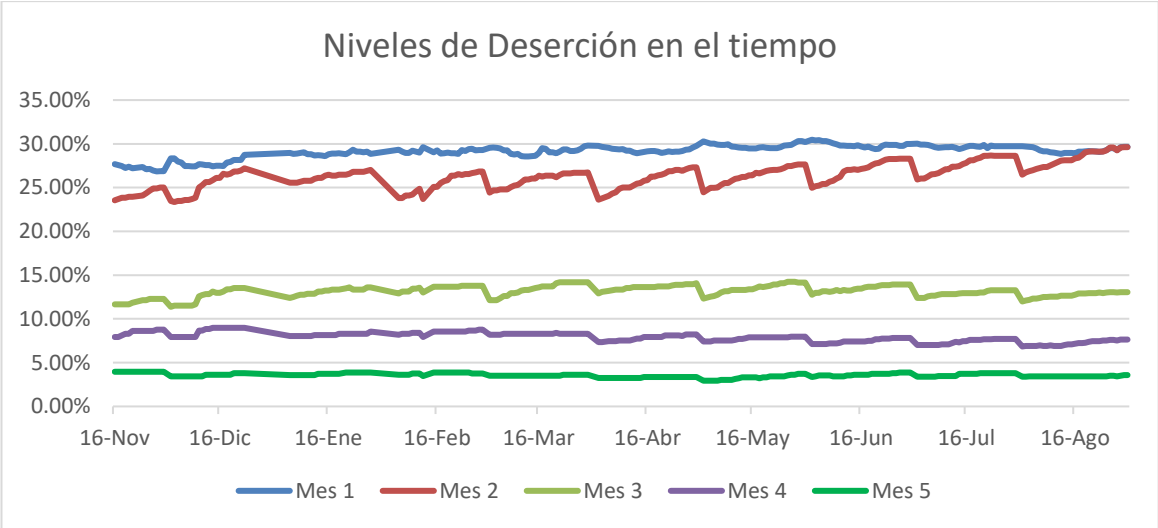


Figura 3.3: Niveles de deserción por meses de permanencia en la institución [Elaboración propia].

La prueba de hipótesis realizada se hizo sobre el histórico de los niveles de deserción del 5to mes, para esto, se definió la siguiente hipótesis:

$$\mu \leq 3.56\% : H\emptyset \; Hipotesis \; nula$$
$$\mu > 3.56\% : H1 \; Hipotesis \; alternativa$$
$$\bar{x} = 3.57\%, \qquad \sigma = 0.24\%, \qquad n = 264$$

La hipótesis nula se definió como el nivel de deserción del 5to mes era menor o igual a 3.56%. Se escogió este último valor debido a que era el inmediato inferior a la media de la deserción en el 5to mes, de esta forma se garantiza la baja probabilidad de abandono al proceso psicoterapéutico. Para esto, se aplicó un nivel de significación de 0.05. La media y la desviación estándar fueron calculado en base a los reportes diarios entre los periodos noviembre-2015 y agosto-2016. La muestra total de reportes tomados en el tiempo fue de 264. El cálculo para verificar la aceptación se muestra a continuación:

$$z_{\alpha=1-0.05} = 1.645$$

$$z_c = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt[2]{n}}} = 0.99$$

Como resultado se obtuvo que el $z_c \leq z_\alpha$, por lo tanto, se acepta la hipótesis nula. La Figura 3.4, que contiene la curva de Gauss, detalla gráficamente la aceptación de

la hipótesis nula. Se concluyó que para el quinto mes de iniciado el proceso, las probabilidades de abandono son muy bajas, por lo que se puede tomar como punto de referencia la cita efectiva número 16 como la meta para tener cierto grado de certeza que el paciente proseguirá con el proceso. Los datos del análisis se encuentran en el Anexo 2.

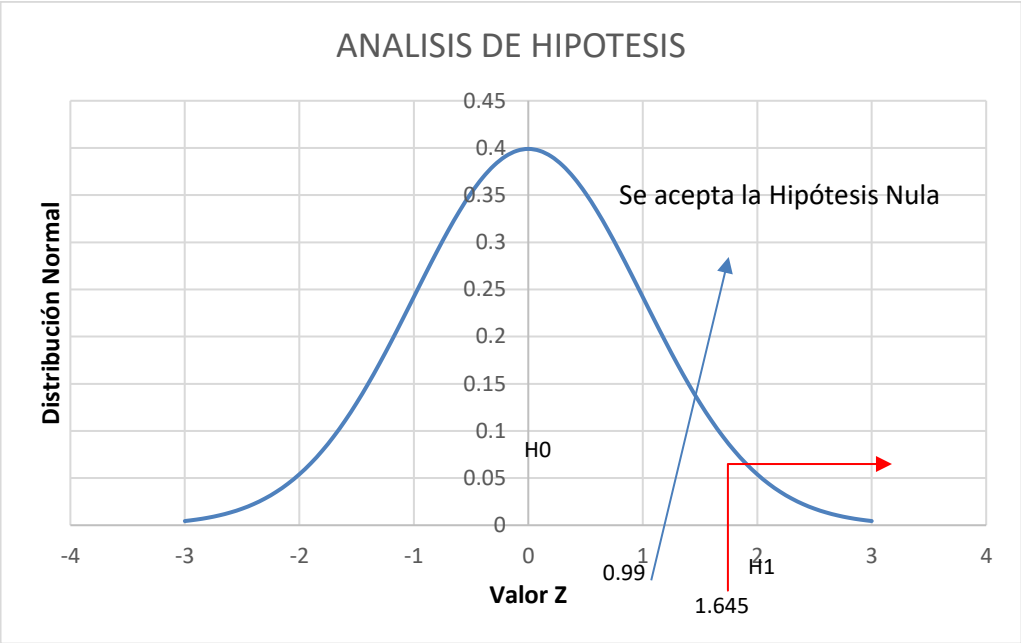


Figura 3.4: Grafica del Análisis de Hipótesis en la curva de Gauss [Elaboración propia].

El segundo modelo a desarrollar permitió, en caso de fracaso, identificar en qué etapa del proceso psicoterapéutico el paciente abandonó el proceso. Para poder definir las clases para el segundo conjuntos de datos, primero se acotó seleccionar a todos los pacientes que abandonaron el proceso en las primeras 15 citas efectivas.

En el capítulo 2, se hizo mención sobre las etapas del proceso psicoterapéutico. En la institución, el cual es el estudio, la fase de indicación corresponde a la primera cita efectiva y la fase inicial, en la mayoría de los casos, corresponde a las siguientes 3 citas efectivas. En la última cita de la fase inicial se le entrega los resultados de las primeras citas y se le propone iniciar el proceso de Psicoterapia para tratar su caso, es decir, para iniciar la fase intermedia. Siendo la fase inicial la etapa previa al inicio de un proceso con tiempo indeterminado, muchos pacientes optan por no continuar el proceso. Las razones de abandono en esta transición pueden ser múltiples, desde el factor económico, el cual implica un gasto mensual de manera indeterminada, hasta no sentir confianza en los resultados que podría obtener a lo largo del proceso.

En la Tabla 3.9 se visualiza la distribución de la cantidad de pacientes que llegaron a un determinado número de citas efectivas antes de abandonar el proceso. En la misma Tabla se puede observar que el 63% del total de pacientes que pertenecen a este análisis no llego a la fase intermedia. Siendo la fase de indicación y la fase inicial las fases previas a un proceso indeterminado, el segundo modelo a desarrollar permite predecir si el paciente abandona el proceso antes de iniciar la fase intermedia o durante la misma.

Tabla 3.9: Cantidad de pacientes por cantidad de citas efectivas durante el proceso [Elaboración propia].

CITA														
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
657	176	138	234	122	120	96	82	66	47	45	33	27	25	36
1205				699										
63%				37%										

Para el tercer modelo, se analizó las primeras 4 citas efectivas con el fin de predecir en cuál de estas citas el paciente abandonó el proceso. Según la Tabla 3.9, la fase de indicación o primera cita tiene un alto nivel de abandono, aproximadamente el 55% de los pacientes que no pasan a la fase intermedia y el 29% del total de pacientes que no llegaron a las 16 citas efectivas. Siendo significativa la cantidad de pacientes de la fase de indicación, así como los pacientes que no pasaron a la fase intermedia, el tercer modelo permite predecir cuantas citas, dentro de las primeras 4 citas, el paciente abandonó el proceso psicoterapéutico.

En la Tabla 3.10 se puede observar los atributos y clases asignados a cada uno de los modelos a desarrollar. La diferencia entre los modelos, a parte de las clases, radica en el tamaño de los conjuntos de datos. M1, M2 y M3 corresponden a los atributos que son tomados por cada modelo. Para el modelo 1, se toma el conjunto completo de datos, los otros dos, según la predicción a realizar.

Tabla 3.10: Atributos y clases del conjunto de datos de entrenamiento [Elaboración propia].

ATRIBUTO	VALORES	M 1	M 2	M 3
Terapeuta	Valor entero	x	x	x
Genero del Terapeuta	1: Femenino, 2: Masculino	x	x	x
Edad del Terapeuta	Valor entero	x	x	x
Experiencia del Terapeuta	Valor entero	x	x	x
Edad del Paciente	Valor entero	x	x	x
Genero del Paciente	1: Femenino, 2: Masculino	x	x	x
Estado Civil del Paciente	1: Soltero, 2: Casado, 3: Conviviente, 4: Divorciado, 5: Viudo	x	x	x
% Deserción Mes 1	Valor real	x	x	x
% Deserción Mes 2	Valor real	x	x	x
% Deserción Mes 3	Valor real	x	x	x
% Deserción Mes 4	Valor real	x	x	x
Tipo de familia del Paciente	1: Nuclear, 2: Extendida, 3: Monoparental, 4: Ensamblada, 5: Conviviente, 6: Solo	x	x	X
Profesión del Paciente	1: Ingeniero, 2: Licenciados, 3: Técnico, 4: Empleado, 5: Oficio, 6: Ama de casa, 7: Estudiante, 8: Independiente, 9: Desempleado	x	x	X
¿Independiente?	0: No, 1: Si	x	x	X
Situación Económica del Paciente	1: A, 2: B, 3: C, 4: D, 5: E	x	x	X
¿Cómo se enteró?	1: Amigo, 2: Familiar, 3: Radio, 4: Internet, 5: Facebook	x	x	X
¿iniciativa propia?	1: Si, 2: No	x	x	X
Día	Valor entero	x	x	X
Hora	Valor entero	x	x	X

Éxito o Fracaso (Clase)	0: Fracaso, 1: Éxito	**x**		
Etapa del Proceso (Clase)	1: Citas entre 1-4, 2: Citas entre 5-15		**x**	
Citas (Clase)	1: Primera cita, 2: Segunda cita, 3: Tercera cita, 4: Cuarta cita			**X**

En la Tabla 3.11 se puede observar el tamaño de cada uno de los conjuntos de datos para cada modelo.

Tabla 3.11: Cantidad de registros por modelo [Elaboración propia].

MODELO 1	MODELO 2	MODELO 3
2256	1904	1205

En la tabla 3.12, se detalla la distribución de las clases para cada uno de los conjuntos de datos de cada modelo establecido. Según se observa, las 3 clases se encuentran desbalanceadas, por lo que se debió aplicar un tratamiento especial sobre las mismas para poder hacer su análisis.

Tabla 3.12: Distribución de clases para cada modelo a analizar [Elaboración propia].

CLASES	VALOR	DESCRIPCIÓN	CANTIDAD
Éxito o Fracaso	1	Éxito	352
	2	Fracaso	1904
Etapa del Proceso	1	Citas entre 1-4	1205
	2	Citas entre 5-15	699
Citas	1	Primera cita	657
	2	Segunda cita	176
	3	Tercera cita	138
	4	Cuarta cita	234

3.2.2 *Análisis de Componentes Principales*

Luego de haber seleccionado los atributos, se procedió a verificar si era posible reducir los atributos para los modelos que fueron desarrollados en el presente Proyecto. Para esto se utilizó el Análisis de Componentes Principales. La Figura 3.5 detalla de manera gráfica como se distribuyen los componentes en el conjunto de

datos de entrenamiento del modelo 1, debido a que este contiene de manera completa todos los registros de los pacientes.

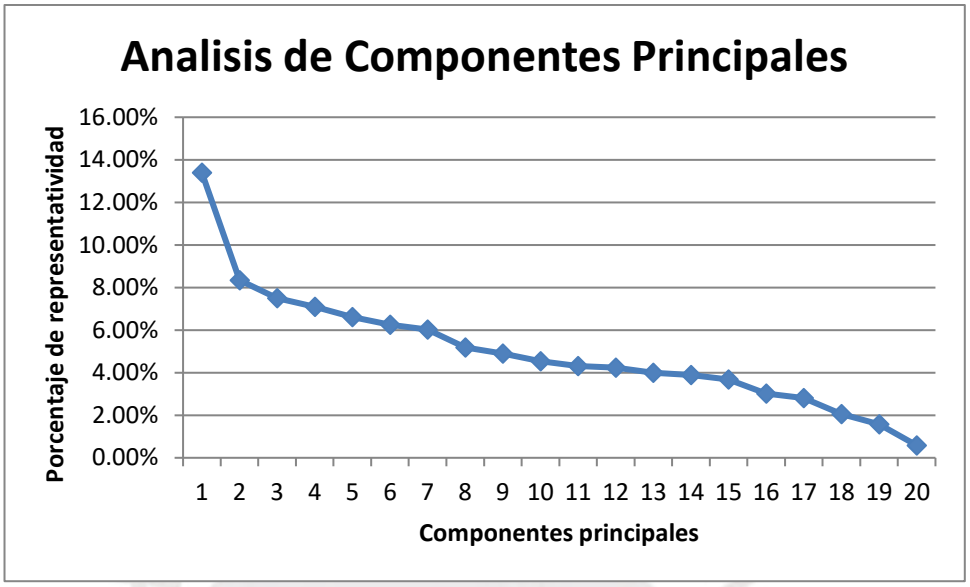


Figura 3.5: Grafica de distribución de los Componentes Principales de los datos del Modelo 1 [Elaboración propia].

La Tabla 3.13 muestra, tanto la distribución de componentes, así como la varianza y el porcentaje de aporte del componente con respecto a todos los datos. Las 8 componentes principales describen el 60.43% de la varianza del conjunto de datos, por lo que podemos concluir que su aporte no es muy bueno, sin embargo, se pudo identificar que atributos tienen cierta importancia en la composición de la varianza (no necesariamente para la tarea predictiva) y cuáles podrían ser eliminados.

Tabla 3.13: Atributos divididos por componentes [Elaboración propia].

ATRIBUTO	VARIANZA	COMPONENTE	PORCENTAJE
EdadP	0.830		
Profesion	0.853		13,39%
Independiente	0.834	COMPONENTE 1	
Desercion1	0.544		
Desercion2	0.562		8,35%
Desercion3	0.647		
Desercion4	0.527	COMPONENTE 2	
Hora	0.618	COMPONENTE 3	7,49%
Terapeuta	0.671	COMPONENTE 4	7,09%

Experiencia	0.667		
GeneroT	0.600	COMPONENTE 5	6,62%
GeneroP	0.522		
TipoFamilia	0.521	COMPONENTE 6	6,26%
EstadoCivil	0.514	COMPONENTE 7	6,03%
Dia	0.514	COMPONENTE 8	5,19%
TOTAL			**60,43%**

Para la asignación de los atributos a cada componente, se utilizó la varianza que superaba el 0.5 en valor absoluto en cada componente obtenida en la Matriz Factorial sin rotar. En caso la varianza de un atributo supera en más de una componente principal, se le asigna a la componente en donde tenga mayor valor de varianza.

Los atributos no considerados son: EdadT, SituacionEconomica, Enteraste, IniciativaPropia. Sin embargo, luego de aplicar los 4 algoritmos de clasificación con los 4 filtros para clases desbalanceadas, se obtuvieron resultados muy similares y mínimamente inferiores, por lo que se decidió tomar el conjunto de datos completo. El análisis comparativo se encuentra en el Anexo 3.

3.2.3 *Estructura de los datos de entrenamiento*

Para el presente proyecto de fin de carrera, para el análisis de los datos de entrenamiento se usó un formato de archivo “arff”. Este tipo de archivo contiene una estructura definida para poder ser leído por WEKA. La estructura es la siguiente:

@RELATION <nombre-de-la-relación>:
@ATTRIBUTE <nombre-del-atributo> <TIPO-DE-DATO>: Nombre del atributo para que pueda ser identificado por WEKA.
@DATA: En esta sección se colocan todos los datos en orden.

3.3 **Conclusiones**

En el presente Capítulo, se ha podido analizar todo el proceso de pre-procesamiento de datos y análisis para la obtención del conjunto de datos de entrenamiento. Después de obtener los resultados esperados de los Objetivos 1 y 2, se puede mencionar la importancia del análisis sobre el conjunto global de datos. La aplicación de métodos estadístico permitió verificar y justificar los supuestos establecidos.

4. CAPÍTULO 4: MODELO DE CLASIFICACIÓN EN CASO DE ÉXITO O FRACASO

En este capítulo se describe el desarrolló el segundo objetivo específico el cual consiste en aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción del éxito o fracaso en el proceso psicoterapéutico. Para desarrollar este objetivo, se cuenta con un conjunto de datos, el cual fue analizado, pre-procesado y seleccionado en base a la investigación realizada para obtener un conjunto de datos de entrenamiento óptimo y de calidad. Como resultado de este objetivo específico, se obtuvo dos resultados esperados: resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción del éxito o fracaso del proceso psicoterapéutico y modelo de clasificación optimizado para la predicción del éxito o fracaso del proceso psicoterapéutico.

4.1 Criterio de Evaluación y Validación de datos

A continuación, se presentan los criterios de evaluación y validación de datos para la obtención de los modelos de predicción.

4.1.1 *Evaluación de los datos de entrenamiento*

Los datos de entrenamiento, luego de haber sido pre-procesados y estructurados, son evaluados usando la herramienta Weka. Esta herramienta presenta varios criterios de precisión, los cuales permiten su análisis y evaluación de los datos de entrenamiento con los distintos algoritmos disponibles. En la Tabla 4.1, se puede observar la estructura de la Matriz de Confusión, el cual permitió evaluar los algoritmos para los modelos de predicción. A continuación, se presentan los principales criterios de precisión extraídos de Powers (POWERS, 2011).

Tabla 4.1: Representación de una matriz de confusión [Elaboración propia].

PREDICCIÓN			
Fracaso (Negativos)	Éxito (Positivos)		
Predicciones correctas (Verdaderos Negativos [TN])	Predicciones incorrectas (Falsos Positivos [FP])	Fracaso	**CLASES**
Predicciones incorrectas (Falsos negativos [FN])	Predicciones correctas (Verdaderos Positivos [TP])	Éxito	

- **Instancias correctamente clasificadas:** Es la proporción de instancias correctamente clasificadas dividida por el total de instancias.

$$ICC = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Instancias incorrectamente clasificadas:** Es la proporción de instancias incorrectamente clasificadas dividida por el total de instancias.

$$IIC = \frac{FP + FN}{TP + TN + FP + FN}$$

- **Tasa TP:** Es la tasa de casos positivos que están correctamente identificados o proporción de casos que la prueba declara positivos y que son positivos.

$$Tasa\ TP = \frac{TP}{TP + FN}$$

- **Tasa FP:** Es la tasa de casos negativos que fueron erróneamente clasificados como positivos o proporción de casos que la prueba declara positivos y que en realidad son negativos.

$$Tasa\ FP = \frac{FP}{FP + TN}$$

- **Precisión:** Es la proporción del total de predicciones positivas correctas sobre el total de instancias clasificadas como esa clase.

$$Precisión = \frac{TP}{TP + FP}$$

- **F-Mesure:** Es la media harmónica de la Precisión con el *Recall*.

$$FMesure = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

- **Recall:** Es la proporción del total de predicciones positivas sobre el total de instancias de esa clase.

$$Recall = Tasa\ TP = \frac{TP}{TP + FN}$$

- ***Accuracy*:** Es la proporción de las predicciones correctas sobre el total de instancias.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Sensibilidad:** Es la proporción de las predicciones positivas sobre el total de instancias que pertenecen a esa clase. La fórmula es igual a la tasa TP o Recall.

$$Sensibilidad = Tasa\,TP\ o\ Recall = \frac{TP}{TP + FN}$$

- **Especificidad:** Es la proporción de las predicciones negativas sobre el total de instancias que pertenecen a esa clase.

$$Especificidad = \frac{TN}{TN + FP}$$

- **Área ROC:** Es un método estadístico el cual permite determinar la exactitud de las pruebas realizadas, para el presente caso, con los datos de entrenamiento. La evaluación de la misma se realiza utilizando el criterio de sensibilidad y especificidad.

- ***Precision-Recall Curve*** **(PRC):** Es la representación gráfica entre la precisión y el *recall.* Permite observar el comportamiento de los algoritmos en las clases desbalanceadas.

- **Matriz de Confusión:** Es una matriz con "n" columnas y "n" filas, donde cada columna representa las predicciones por clase y cada fila representa la clase. En la tabla 15, se puede observar la distribución de una matriz de confusión.

4.1.2 *Criterio de Validación*

Para el presente Proyecto de fin de carrera se utilizó la validación cruzada usando k grupos (*k-folds Cross-validation*). La razón por la cual se usó este criterio de validación es porque mediante este método de prueba se puede verificar la precisión de los resultados de la predicción con diferentes conjuntos de datos de entrenamiento y datos de prueba. Adicionalmente, se iteró 10 veces este método de prueba y se calculó el promedio de todas las iteraciones totales por algoritmo procesado.

Según KOHAVI (KOHAVI, 1995), la validación cruzada de k-iteraciones divide el conjunto de datos en k subgrupos, donde (k-1) subgrupos son los datos de entrenamiento y solo un subgrupo corresponde a los datos de prueba. Este proceso se repite k-veces con cada diferente subgrupo de datos. Para calibrar el valor de los k-*folds*, se realizó un experimento, el cual consistió en evaluar el conjunto de datos

de entrenamiento con los siguientes valores de *folds*: 10, 15, 20, 25, 30 por cada uno de los modelos a usar.

4.2 Análisis comparativo de algoritmos de predicción del Éxito o Fracaso del Proceso Psicoterapéutico.

Para esta sección se aplicó los criterios de precisión usando cuatro métodos de clasificación. Antes de aplicar estos criterios, se debió tener en cuenta que los datos de entrenamiento de los conjuntos de datos a analizar se encuentran desbalanceados, por lo que fue necesario realizar algunos experimentos adicionales para poder determinar cuál fue el mejor método de a usar. Para el presente objetivo, el análisis se realizó para el modelo de predicción del éxito o fracaso del proceso psicoterapéutico.

4.2.1 *Clases No Balanceadas*

Según PETROSINO (PETROSINO et al., 2016), el análisis regular se realiza con datos de entrenamiento balanceados, es decir, tienen una cantidad de instancias o registros similares por cada clase. Cuando se realiza el análisis sobre datos de entrenamiento no balanceados, es decir, sobre datos de entrenamiento que tiene una gran cantidad de datos de una clase, en comparación con otra u otras clases, el procedimiento es distinto.

El presente Proyecto, tiene como datos de entrenamiento dos clases, los cuales según el análisis que se realizó en el capítulo 3, tiene una clase mayoritaria clasificada como fracaso y una clase minoritaria clasificada como éxito. En la Tabla 4.2, se puede observar el resumen de la distribución de estas dos clases.

Tabla 4.2: Relación de distribución entre las clases de los datos de entrenamiento [Elaboración propia].

	Cantidad	**Porcentaje**
Fracaso	1904	84%
Éxito	352	16%

Para abordar este caso, existen algunas propuestas que buscan balancear los datos para poder ser analizados. Para este Proyecto, se analizó algunas técnicas aplicadas al desbalanceo de clases.

4.2.2 *Técnicas de Balanceo de Clases*

Las clases desbalanceadas terminan siendo un problema complejo al ser analizados debido a que la precisión de las mismas dependerá de la cantidad de datos en cada clase. En la Tabla 4.3, se observa que la clase "Fracaso" tiene un alto valor de *recall* por cada algoritmo aplicado, por lo que se puede deducir que la cantidad de datos de la clase mayoritaria favorece en la correcta predicción de la clase.

Tabla 4.3: Comparación de Recall por clase y por algoritmo [Elaboración propia].

CLASE	J48	LMT	*Random Forest*	*Random Tree*
Fracaso	0.95	0.95	0.99	0.93
Éxito	0.58	0.59	0.65	0.72

Para abordar el problema de desbalanceo de clases, existen métodos de re-muestreo para balancear las clases. Los métodos aplican diferentes técnicas, entre las que se tienen:

- **Sobre-muestreo:** Agrega objetos de la clase minoritaria.
- **Sub-muestreo:** Elimina objetos de la clase mayoritaria.
- **Hibrido:** Combinación de técnicas de sobre-muestreo y sub-muestreo.

Para el presente Proyecto, se aplicaron los siguientes filtros sobre el conjunto de datos desbalanceado: ***Spread Subsample***, ***SMOTE***, ***Resample***, ***ClassBalancer***.

Para poder aplicar cada uno de estos filtros sobre el conjunto de datos de entrenamiento, fue necesario usar el meta clasificador ***filteredClassifier***, el cual permitió incorporar estos filtros sobre los datos de entrenamiento, pero no sobre los datos que evalúan el clasificador. Para evaluar el clasificador y el filtro a aplicar, se usó la validación cruzada con *k-folds*.

4.2.3 *Justificación de la elección de los algoritmos de árboles decisión*

Para la justificación en la elección de los algoritmos de árboles de decisión, se utilizó el análisis comparativo. Los algoritmos que fueron comparados son: DecisionStump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree y REPtree. Los algoritmos seleccionados para la comparación son los principales y más utilizados para la selección de algoritmos de predicción basados en arboles de decisión. La razón por la cual se hizo esta comparación fue porque no es posible definir el mejor algoritmo sin realizar pruebas sobre el mismo conjunto de datos. El análisis se realizó

utilizando los filtros ClassBalancer, Resample, SMOTE y Spread Subsample con k-fold=10 (Ver Anexo 4). Del analisis se concluyó que los algoritmos con mejores resultados fueron: J48, LMT, Random Forest y Random Tree. Estos algoritmos serán usados para su posterior análisis, el cual se muestra en la Tabla 4.4

De la Tabla 4.4, se observa que el filtro que tiene mejores resultados es el *Spread Subsample.* Con respecto al k-fold, se obtuvo buenos resultados con un k=30, por lo que se escogió este valor de "k" para evaluar los clasificadores. Finalmente, se observa que, para todos los filtros aplicados, el algoritmo *Random Forest* presentó los mejores resultados.

Tabla 4.4: *Accuracy* de validación cruzada usando usando filteredClassifier [Elaboración propia].

FILTRO	ALGORITMO	% instancias correctamente clasificadas				
		K=10	K=15	K=20	K=25	K=30
ClassBalancer	J48	86.04	**87.19**	86.61	86.4	87.01
	LMT	89.41	89.36	**89.54**	89.1	89.41
	Random Forest	93.44	92.86	93.04	93.22	**93.62**
	Random Tree	90.56	90.38	**90.74**	90.6	90.51
Resample	J48	85.51	85.51	**86.52**	86.35	85.15
	LMT	85.68	85.86	86.75	**86.79**	85.55
	Random Forest	91.31	91.4	90.78	**91.8**	91
	Random Tree	86.17	86.88	**87.19**	86.08	86.66
SMOTE	J48	88.34	87.59	88.65	88.52	**88.79**
	LMT	**89.85**	89	89.14	88.83	89.81
	Random Forest	93.53	93.48	93.57	**93.97**	93.53
	Random Tree	89.18	89.81	89.98	89.45	**90.43**
Spread Subsample	J48	88.56	87.77	88.7	89.32	**89.36**
	LMT	89.45	89.05	89.05	**89.72**	**89.72**
	Random Forest	93.57	93.75	93.93	94.1	**94.06**
	Random Tree	90.34	90.43	**91.09**	90.82	90.56

4.2.4 *Resultado de los criterios de Evaluación*

A continuación, se presentan los resultados de los criterios de precisión de cada uno de los algoritmos utilizados en el presente proyecto de fin de carrera. Para esto, se

escogió como filtro el *Spread Subsample*, el cual obtuvo mejores resultados en comparación a los otros filtros. Los resultados de los criterios de precisión se muestran en la Tabla 4.5.

Tabla 4.5: Resultados de los criterios de evaluación con K=30 y Spread Subsample [Elaboración propia].

Algoritmo	% Instancias Correctamente Clasificadas	% Instancias Incorrectamente Clasificadas	Tasa TP	Tasa FP	Precisión	Recall	F-Mesure	Área ROC	Área PRC
J48	89.36	10.64	0.89	0.36	0.89	0.89	0.89	0.82	0.88
LMT	89.72	10.28	0.90	0.33	0.89	0.90	0.90	0.82	0.88
Random Forest	94.06	5.94	0.94	0.30	0.94	0.94	0.94	0.93	0.96
Random Tree	90.56	9.44	0.91	0.25	0.91	0.91	0.91	0.83	0.88

J48

Tabla 4.6: Matriz de confusión de J48 [Elaboración propia].

Fracaso	Éxito	
1811	93	**Fracaso**
147	205	**Éxito**

LMT

Tabla 4.7: Matriz de confusión de LMT [Elaboración propia].

Fracaso	Éxito	
1808	96	**Fracaso**
136	216	**Éxito**

Random Forest

Tabla 4.8: Matriz de confusión de Random Forest [Elaboración propia].

Fracaso	Éxito	
1894	10	**Fracaso**
124	**228**	**Éxito**

Random Tree

Tabla 4.9: Matriz de confusión de Random Tree [Elaboración propia].

Fracaso	Éxito	
1793	111	**Fracaso**
102	250	**Éxito**

De las Tabla 4.6 a 4.9 se observan las matrices de confusión de los cuatro clasificadores. En todos los casos los verdaderos positivos y negativos son superiores a los falsos negativos y positivos. Los resultados de estas matrices resultan satisfactorios para el análisis.

4.3 Optimización del modelo de clasificación del Éxito o Fracaso del Proceso Psicoterapéutico.

Para poder comparar los clasificadores se usó el AUC (Área bajo la curva). Para esto se calculó un valor escalar del área bajo la curva ROC, el cual representó el rendimiento del clasificador. La justificación de usar este método, es debido a que el valor del AUC es equivalente a la probabilidad de que un clasificador clasifique una instancia positiva elegida al azar más alta que una instancia negativa elegida al azar (Fawcett, 2004).

Para el presente modelo, el cual contiene solo dos clases, aplicar este método solo consiste en calcular el AUC. Según los resultados del AUC en la Tabla 4.10, se puede observar que el clasificador con mejor rendimiento es ***Random Forest*.**

Tabla 4.10: Resultados de AUC por cada clasificador [Elaboración propia].

	AUC
J48	0.82
LMT	0.82
Random Forest	0.93
Random Tree	0.83

4.4 Conclusiones

El desbalanceo de clases ha resultado ser un problema resaltante en este Capítulo. El tratamiento de los datos utilizando filtros de pre-procesamiento, así como un meta-clasificador ha permitido realizar un buen análisis para escoger el mejor clasificador para el modelo algorítmico a desarrollar. El método AUC permitió identificar al mejor clasificador, por lo que se escogió ***Random Forest***.

5. CAPÍTULO 5: MODELOS DE CLASIFICACIÓN EN CASO DE FRACASO

En este Capítulo se desarrolla el tercer objetivo específico el cual consiste en aplicar y analizar los métodos de clasificación basados en árboles de decisión para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso. Para el desarrollo de este objetivo, se cuenta con dos conjuntos de datos, los cuales fueron analizados, pre-procesados y seleccionados en el Capítulo 3. Como resultado de este objetivo específico, se obtuvo dos resultados esperados, los cuales buscan predecir a que fase del proceso terapéutico el paciente permanece con el terapeuta asignado en caso de predecir el fracaso con el modelo analizado en el Capítulo 4. Los resultados esperados son: resultados de los criterios de precisión de los cuatro métodos de clasificación para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso y modelo de clasificación optimizado para la predicción de la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.

5.1 Análisis comparativo de la cantidad de citas efectivas en el caso de Fracaso.

En el Capítulo 4, se realizó un análisis sobre un conjunto de datos desbalanceado y se aplicaron métodos para contrarrestar el mismo. En la presente Sección, se analizan y aplican métodos de clasificación sobre dos conjuntos de datos, los cuales tiene como propósito predecir, en caso de fracaso, en qué etapa de las primeras 15 citas efectivas, el paciente abandona el proceso.

5.1.1 *Modelo en caso de Fracaso*

A continuación, se analiza el conjunto de datos correspondiente a la predicción de la permanencia entre las primeras 15 citas efectivas, con dos clases correspondientes a las primeras 4 citas efectivas para la primera clase y las siguientes para la segunda clase. La elección de los algoritmos se realizó por medio del análisis comparativo (Ver Anexo 5). Los algoritmos seleccionados fueron: **J48, LMT, Random Forest y Random Tree**.

5.1.1.1 Análisis del Modelo en caso de Fracaso

El presente modelo cuenta con dos clases, las cuales fueron analizadas en el Capítulo 3. Este modelo también presentó el problema de las clases desbalanceadas, por lo que se aplicaron los filtros de desbalance para determinar el mejor filtro a

aplicar y el k-fold a usar para evaluar los clasificadores. La Tabla 5.1, muestra los resultados del análisis usando FilteredClassifier.

De la Tabla 5.1, se observa que el filtro que tiene mejores resultados es el *Spread Subsample.* Con respecto al k-fold, se obtuvo buenos resultados con un k=30, por lo que se escogió este valor de "k" para evaluar los clasificadores para el presente modelo. Finalmente, se observó que, para todos los filtros aplicados, el algoritmo *Random Forest* presentó los mejores resultados.

5.1.1.2 Resultado de los criterios de Precisión del Modelo en caso de Fracaso

En la Tabla 5.2, se presentan los resultados de los criterios de precisión de cada uno de los algoritmos utilizados en el presente proyecto de fin de carrera. Para esto, se escogió como filtro el *Spread Subsample*, el cual obtuvo mejores resultados en comparación a los otros filtros.

Tabla 5.1: *Accuracy* de la validación cruzada usando *filteredClassifier* sobre el Modelo en caso de fracaso [Elaboración propia].

FILTRO	ALGORITMO	% instancias correctamente clasificadas				
		K=10	K=15	K=20	K=25	K=30
ClassBalancer	J48	77.1	78.41	79.57	79.57	**79.78**
	LMT	81.41	82.51	82.93	**83.98**	83.19
	Random Forest	85.4	86.34	86.45	**86.76**	**86.76**
	Random Tree	82.98	82.62	**83.88**	83.82	83.66
Resample	J48	73	74.21	**75**	73.9	73.11
	LMT	74.74	75.16	**77.47**	77.26	76.21
	Random Forest	79.52	80.46	**81.46**	80.62	80.62
	Random Tree	77.52	75.74	**78.26**	76.1	77.15
SMOTE	J48	78.1	79.1	78.41	**79.25**	78.94
	LMT	80.41	82.04	**82.77**	82.51	82.62
	Random Forest	85.24	86.19	86.45	**86.82**	86.61
	Random Tree	81.15	**84.87**	83.93	83.88	83.51
Spread Subsample	J48	79	80.62	79.52	80.15	**80.78**
	LMT	81.78	82.3	82.46	**83.3**	82.83
	Random Forest	85.87	86.4	87.08	86.76	**87.29**
	Random Tree	83.51	84.45	84.51	82.62	**85.08**

Tabla 5.2: Resultados de los criterios de precisión con K=30 y Spread Subsample aplicado sobre el modelo en caso de fracaso [Elaboración propia].

Algoritmo	% Instancias Correctamente Clasificadas	% Instancias Incorrectamente Clasificadas	Tasa TP	Tasa FP	Precisión	Recall	F-Mesure	Área ROC	Área PRC
J48	80.78	19.22	0.81	0.23	0.81	0.81	0.81	0.81	0.78
LMT	82.83	17.17	0.83	0.19	0.83	0.83	0.83	0.82	0.81
Random Forest	87.29	12.71	0.87	0.18	0.87	0.87	0.87	0.94	0.95
Random Tree	85.08	14.92	0.85	0.17	0.85	0.85	0.85	0.84	0.80

J48

Tabla 5.3: Matriz de confusión de J48 [Elaboración propia].

Etapa 1	Etapa 2	
1034	171	**Etapa 1**
195	504	**Etapa 2**

LMT

Tabla 5.4: Matriz de confusión de LMT [Elaboración propia].

Etapa 1	Etapa 2	
1035	170	**Etapa 1**
157	542	**Etapa 2**

Random Forest

Tabla 5.5: Matriz de confusión de Random Forest [Elaboración propia].

Etapa 1	Etapa 2	
1135	70	**Etapa 1**
172	**527**	**Etapa 2**

Random Tree

Tabla 5.6: Matriz de confusión de Random Tree [Elaboración propia].

Etapa 1	Etapa 2	
1054	151	**Etapa 1**
133	566	**Etapa 2**

De las Tablas 5.3 a 5.6 se observa las matrices de confusión de los cuatro clasificadores. Se observa que los verdaderos positivos y negativos son superiores a los falsos negativos y positivos. Los resultados de estas matrices resultan satisfactorios para el análisis.

5.1.2 *Modelo de la etapa de evaluación*

A continuación, se analiza el conjunto de datos correspondiente a la predicción de la permanencia entre las primeras 4 citas efectivas, con cuatro clases correspondientes a cada una de las citas. La elección de los algoritmos se realizó por medio del análisis comparativo (Ver Anexo 6). Los algoritmos seleccionados fueron: **J48, LMT, Random Forest y Random Tree**.

5.1.2.1 Análisis del Modelo de la Etapa de Evaluación en caso de Fracaso

En la Tabla 5.7, se muestran los resultados del análisis que se realizó para determinar el filtro y el k-fold que presento mejores resultados.

Tabla 5.7: *Accuracy* de la validación cruzada usando *filteredClassifier* sobre el Modelo de la Etapa de Evaluación en caso de fracaso [Elaboración propia].

FILTRO	ALGORITMO	% instancias correctamente clasificadas				
		K=10	K=15	K=20	K=25	K=30
ClassBalancer	J48	59.34	59.5	59.17	**62.16**	61.58
	LMT	**74.35**	72.86	73.69	74.28	73.61
	Random Forest	76.43	76.85	77.1	77.84	**78.09**
	Random Tree	73.2	73.69	74.69	76.18	**77.51**
Resample	J48	59.83	62.99	**63.24**	62.32	61.24
	LMT	65.64	66.64	**66.81**	64.65	64.73
	Random Forest	70.54	**73.2**	71.7	72.28	71.95
	Random Tree	63.9	**66.31**	65.06	63.98	64.32
SMOTE	J48	67.39	68.63	69.79	**70.79**	69.38
	LMT	72.12	73.53	74.61	**75.02**	**75.02**
	Random Forest	79.83	80.41	**80.91**	80.75	**80.91**
	Random Tree	74.36	75.44	**76.76**	75.52	75.68
Spread Subsample	J48	68.63	69.05	68.71	69.21	**71.37**
	LMT	74.69	74.94	75.52	74.85	**75.93**
	Random Forest	78.76	80	**80.75**	80.66	**80.75**
	Random Tree	73.61	74.36	76.02	**76.1**	75.68

De la Tabla 5.7, se observa que el filtro que tiene mejores resultados es el filtro S*MOTE.* Con respecto al k-fold, se obtuvo buenos resultados con un k=20, por lo que se escogió este valor de "k" para evaluar los clasificadores. Finalmente, se observó

que, para todos los filtros aplicados, el algoritmo *Random Forest* presenta los mejores resultados.

5.1.2.2 Resultado de los criterios de Precisión del Modelo de la Etapa de Evaluación en caso de Fracaso

En la Tabla 5.8, se presentan los resultados de los criterios de precisión de cada uno de los algoritmos utilizados en el presente proyecto de fin de carrera. Para esto, se escogió como filtro *SMOTE*, el cual obtuvo los mejores resultados en comparación a los otros filtros.

Tabla 5.8: Resultados de los criterios de precisión con K=20 y SMOTE [Elaboración propia].

Algoritmo	% Instancias Correctamente Clasificadas	% Instancias Incorrectamente Clasificadas	Tasa TP	Tasa FP	Precisión	Recall	F-Mesure	Área ROC	Área PRC
J48	69.79	30.21	0.70	0.19	0.69	0.70	0.69	0.80	0.66
LMT	74.61	25.39	0.75	0.16	0.75	0.75	0.75	0.81	0.69
Random Forest	80.91	19.09	0.81	0.18	0.82	0.81	0.80	0.92	0.87
Random Tree	76.76	23.24	0.77	0.13	0.77	0.77	0.77	0.82	0.67

J48

Tabla 5.9: Matriz de confusión de J48 [Elaboración propia].

1	2	3	4	
541	37	27	52	**1**
45	103	14	14	**2**
50	6	69	13	**3**
66	19	21	128	**4**

LMT

Tabla 5.10: Matriz de confusión de LMT [Elaboración propia].

1	2	3	4	
541	39	28	49	**1**
35	124	9	8	**2**
31	11	86	10	**3**
62	16	8	148	**4**

Random Forest

Tabla 5.11: Matriz de confusión de Random Forest [Elaboración propia].

1	2	3	4	
631	4	11	11	**1**
49	117	7	3	**2**
48	1	83	6	**3**
75	6	9	144	**4**

Random Tree

Tabla 5.12: Tabla 5.12: Matriz de confusión de Random Tree [Elaboración propia].

1	2	3	4	
552	32	29	44	**1**
31	127	9	9	**2**
30	5	88	15	**3**
45	15	16	158	**4**

De las Tablas 5.9 a 5.12 se observa las matrices de confusión de los cuatro clasificadores. Tal como el caso anterior, los verdaderos positivos y negativos son superiores a los falsos negativos y positivos. Los resultados de estas matrices resultaron satisfactorios para el análisis.

5.2 Optimización del modelo de clasificación para la predicción de la cantidad de citas en caso de Fracaso

En esta Sección se desarrolla el Resultado Esperado 6.

5.2.1 *Elección del mejor Clasificador para el Modelo en caso de Fracaso*

Para el presente modelo, el cual contiene solo dos clases, aplicar este método solo consiste en calcular el AUC. Según los resultados del AUC en la Tabla 5.13, se pudo observar que el clasificador con mejor rendimiento fue ***Random Forest*.**

Tabla 5.13: Resultados de AUC por cada clasificador [Elaboración propia].

	AUC
J48	0.81
LMT	0.82
Random Forest	0.94
Random Tree	0.84

5.2.2 *Elección del mejor Clasificador para el Modelo de la Etapa de Evaluación en caso de Fracaso*

Para el presente modelo, el cual contiene cuatro clases, aplicar este método consiste en calcular el AUC. Según los resultados del AUC en la Tabla 5.14, se puede observar que el clasificador con mejor rendimiento fue ***Random Forest*.**

Tabla 5.14: Resultados de AUC por cada clasificador [Elaboración propia].

	AUC
J48	0.80
LMT	0.81
Random Forest	0.92
Random Tree	0.82

5.3 **Conclusiones**

En este Capítulo se ha aplicado también las técnicas para el desbalanceo de clases el cual ha permitido obtener dos modelos para los casos de fracaso. Para ambos casos, se han encontrado buenos resultados aplicando el algoritmo Random Forest. Finalmente, por el método AUC se permitió identificar al mejor clasificador, por lo que se escogió ***Random Forest***.

6. CAPÍTULO 6: PROTOTIPO FUNCIONAL

En este Capítulo se desarrolla el cuarto objetivo específico el cual consiste en desarrollar un prototipo funcional que permita predecir el mejor vínculo terapeuta-paciente en un proceso psicoterapéutico usando el método de clasificación seleccionado, así como la cantidad de citas efectivas en caso de fracaso. Para este objetivo, se desarrolló un prototipo funcional en Java con funcionalidades básicas, los cuales permitieron al usuario interactuar con la herramienta y hacer uso de los modelos obtenidos en los Capítulos anteriores para la predicción de la permanencia de un paciente en un proceso psicoterapéutico. Como resultado de este objetivo específico, se obtuvo dos resultados esperados, los cuales tienen como fin realizar la predicción desde un conjunto amplio de resultados a un conjunto reducido de citas según sea el caso. Los resultados esperados son: prototipo funcional que permita predecir el éxito o fracaso del proceso psicoterapéutico por medio de la información del paciente y prototipo funcional que permita predecir la cantidad de citas efectivas en la asignación paciente-terapeuta en el caso de fracaso.

Para obtener los resultados esperados 7 y 8, se usó la herramienta NetBeans como IDE para programar el prototipo funcional. El lenguaje de programación usado fue Java y el prototipo desarrollado fue de escritorio. Los algoritmos ***Random Forest***, ***filteredClassifier*** y los filtros ***Spread Subsample*** y ***SMOTE*** han sido obtenidos mediante la API de Weka. El desarrollo de cada uno de los resultados esperados se explica en las secciones del capítulo.

6.1 Módulo de predicción del Éxito o Fracaso

Para el desarrollo del prototipo funcional del Resultado Esperado 7, primero se desarrolló la interfaz gráfica. Para esto se usó la IDE Netbeans en modo diseño y haciendo uso de las componentes graficas se armó el contenedor principal, los paneles para cada funcionalidad, los formularios y una tabla de datos. Posteriormente, se implementó las librerías, métodos y el flujo algorítmico para obtener el resultado esperado.

6.1.1 *Diseño*

Para el proceso de diseño, se estableció mantener una interfaz simple y funcional el cual permita la interacción con los usuarios de manera básica, por lo que no se consideró un módulo de control de usuarios. Para cada una de las interfaces, se hizo uso de las componentes de Java incorporados en la IDE.

Para el contenedor principal se incorporó un ***JFrame***. Sobre este se incorporaron cuatro paneles con pestaña, debido a que se quería una interfaz simple, intuitivo y amigable. Por cada panel se incorporó una funcionalidad distinta, de tal forma que cada una de estas no se mezclen en una misma interfaz. El orden en el cual fueron colocados cada uno de los paneles fue de acuerdo a su importancia.

6.1.1.1 Interfaz del módulo Clasificación

Para el panel principal, se armó un formulario que contenía los siguientes campos: Edad Paciente, Genero Paciente, Estado Civil, Tipo Familia, Profesión, Independiente, Nivel Socioeconómico, Enteraste, Iniciativa propia, Terapeuta, Día, Hora y Resultados. El fin de este formulario es ser una interfaz de entrada para la predicción por parte del usuario. Con el fin de facilitar el uso y evitar el error al ingresar los datos, se colocaron todas las casillas como ***JComboBox***, excepto las casillas "Edad Paciente" y "Resultados".

La Tabla 6.1 resume los nombres de las casillas, atributos, tipos de componentes y los valores que toman cada una de estas. La casilla "Día" es dinámico y dependiente de la casilla "Terapeuta" debido a que cada terapeuta puede trabajar de 7:30 am – 2:30 pm o 2:30 pm – 9:30 pm. La atención por cada sesión es de una hora, por lo que la última cita asignada puede ser a la 1:30 pm o a las 8:30 pm según corresponda. Debido a que todas las atenciones empiezan a las medias horas, en esta casilla solo se está tomando el valor numérico correspondiente a la hora y no a los minutos.

Tabla 6.1: Características de las casillas del formulario del módulo de clasificación [Elaboración propia].

CASILLA	**ATRIBUTO**	**TIPO DE CAMPO**	**VALORES**
Edad Paciente	txtEdadP	JTextField	Valores Enteros
Género Paciente	comboGeneroP	JComboBox	1: Femenino 2: Masculino
Estado Civil	comboEstadoCivil	JComboBox	1: Soltero 2: Casado 3: Conviviente

			4: Divorciado 5: Viudo
Tipo Familia	comboTipoFamilia	JComboBox	1: Nuclear 2: Extendida 3: Monoparental 4: Ensamblada 5: Conviviente 6: Solo
Profesión	comboProfesion	JComboBox	1: Ingeniero 2: Licenciado 3: Técnico 4: Empleado 5: Oficio 6: Ama de Casa 7: Estudiante 8: *Freelance* 9: Desempleado
Independiente	comboIndependiente	JComboBox	1: Si 2: No
Nivel Socioeconómico	comboSituacionEconomica	JComboBox	1: A 2: B 3: C 4: D 5: E
Enteraste	comboEnteraste	JComboBox	1: Amigo 2: Familiar 3: Radio 4: Internet 5: Facebook
Iniciativa Propia	comboIniciativa	JComboBox	1: Si 2: No
Terapeuta	comboTerapeuta	JComboBox	Lista de 18 terapeutas

Día	comboDia	JComboBox	Días de la semana
Hora	comboHora	JComboBox	Lista de Horas de 7-20 dinámico y dependiente del horario del terapeuta
Resultados	txtResultado	JTextArea	Texto
Clasificar	btnClasificar	JButton	Clasificar

La Figura 6.1, muestra la interfaz gráfica correspondiente al formulario, el cual tiene como nombre “CLASIFICACIÓN”. La alineación de las componentes correspondientes a las casillas que deben ser llenadas y seleccionadas ha sido distribuida al lado izquierdo, mientras que la casilla en donde el algoritmo muestra los resultados de la predicción al lado derecho.

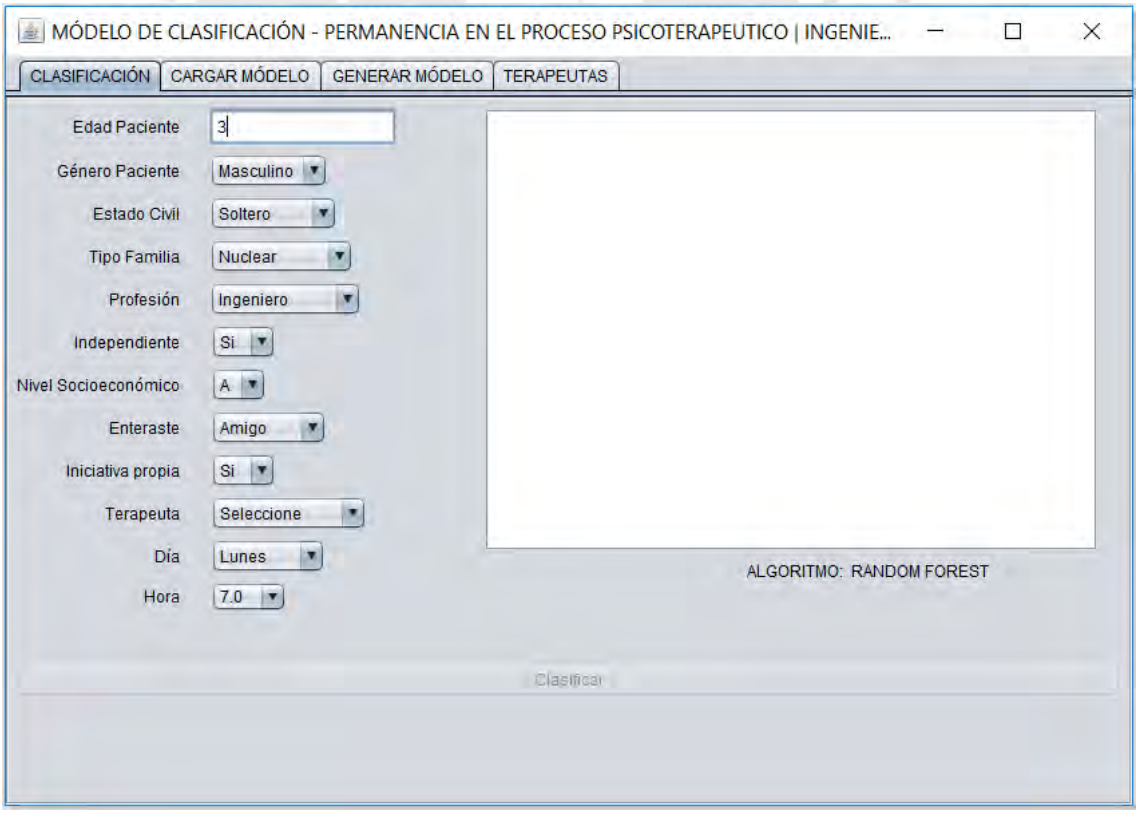


Figura 6.1: Interfaz gráfica del módulo de Clasificación [Elaboración propia].

6.1.1.2 Interfaz del módulo Cargar Modelo

El segundo panel corresponde al módulo que permite la carga del Modelo. Mediante este módulo, el usuario carga los datos de entrenamiento y la ubicación donde se almacena el modelo generado. El modulo contiene un formulario con dos casillas por cada modelo, cuyo fin tiene ser los atributos de entrada para el método de clasificación. La Tabla 6.2 detalla las características de cada una de los componentes del formulario.

Tabla 6.2: Características de las casillas del formulario del módulo Cargar Modelo [Elaboración propia].

MODELO	CASILLA	ATRIBUTO	TIPO DE CAMPO	VALORES
Modelo 1	Ruta del Corpus	txtRutaCorpusModelo1	JTextField	Texto
	Ruta del Modelo	txtRutaModelo01	JTextField	Texto
Modelo 2	Ruta del Corpus	txtRutaCorpusModelo2	JTextField	Texto
	Ruta del Modelo	txtRutaModelo02	JTextField	Texto
Modelo 3	Ruta del Corpus	txtRutaCorpusModelo3	JTextField	Texto
	Ruta del Modelo	txtRutaModelo03	JTextField	Texto

La Figura 6.2 muestra la interfaz con sus componentes. Las casillas han sido agrupadas por cada modelo y los botones permiten interactuar con la clase *JFileChooser* de Java.

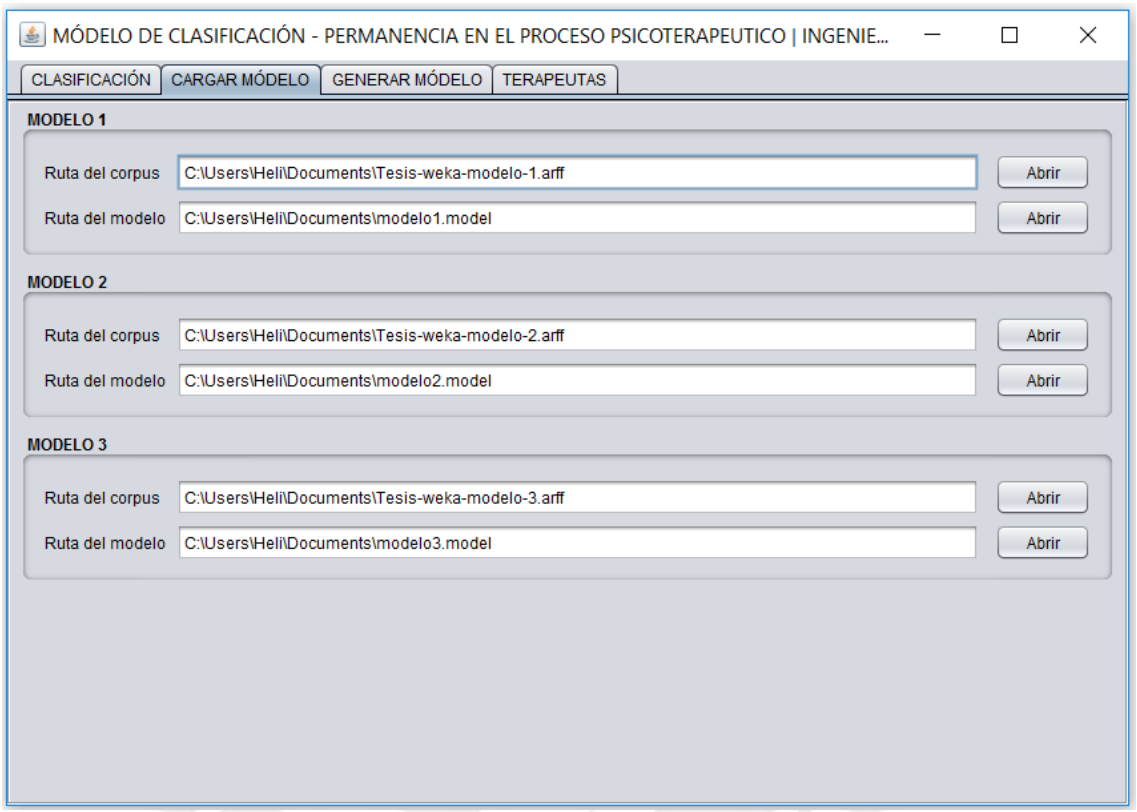


Figura 6.2: Interfaz gráfica del módulo Cargar Modelo [Elaboración propia].

6.1.1.3 Interfaz del módulo Generar Modelo

El tercer panel corresponde al módulo que genera los modelos. Para esta interfaz se desarrolló un formulario que contenga la ruta de cada uno de los modelos y una casilla con el algoritmo a usar. Debido a los resultados que fueron obtenidos en los capítulos 4 y 5, el algoritmo seleccionado fue "***Random Forest***", por tanto, por defecto el modulo tiene seleccionado este algoritmo. El fin de este módulo es generar los modelos y almacenarlo en la ruta de destino especificado por la casilla "Ruta de Destino" por cada modelo. La Tabla 6.3, detalla las características de los componentes del formulario del módulo.

Tabla 6.3: Características de las casillas del formulario del módulo Generar Modelo [Elaboración propia].

CASILLA	**ATRIBUTO**	**TIPO DE CAMPO**	**VALORES**
Ruta de destino del Modelo 1	txtRutaDestino1	JTextField	Texto

Ruta de destino del Modelo 2	txtRutaDestino2	JTextField	Texto
Ruta de destino del Modelo 3	txtRutaDestino3	JTextField	Texto
Algoritmo	comboAlgoritmo	JComboBox	Random Forest
Generar Modelo	btnGenerar	JButton	Generar Modelo

La Figura 6.3, muestra la interfaz del formulario con sus componentes. En la parte inferior contiene un botón que permite la generación del modelo. La generación del modelo permite la interacción con el botón “Clasificar” del módulo “Clasificación”, debido a que, si el modelo no ha sido generado, el botón “Clasificar” permanece deshabilitado.

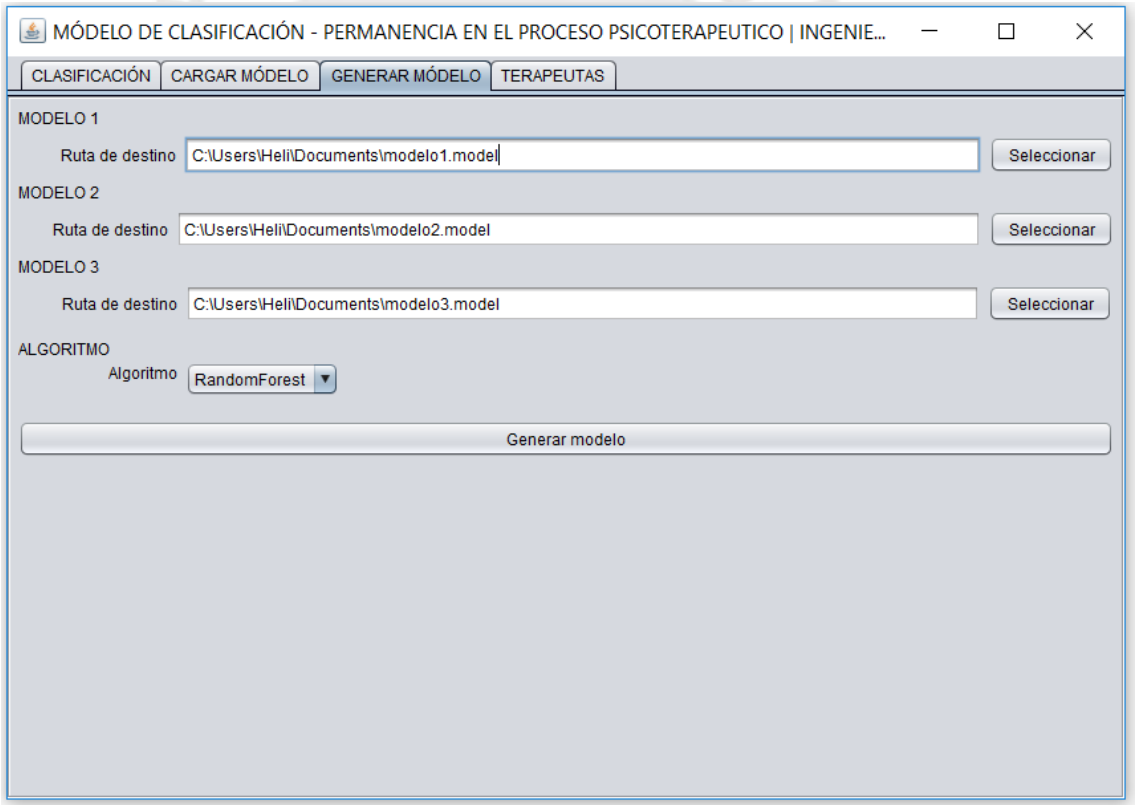


Figura 6.3: Interfaz gráfica del módulo Generar Modelo [Elaboración propia].

6.1.1.4 Interfaz del módulo Terapeutas

El cuarto panel corresponde al módulo de terapeutas, el cual contiene una lista de los terapeutas con sus atributos obtenidos de la base de datos del prototipo funcional. Los valores mostrados han sido obtenidos de la base de datos de la institución y en

algunos casos calculados para ser tomados como datos de entrada para la clasificación. La Tabla 6.4 describe los atributos mostrados en el ***JTable*** del módulo.

Tabla 6.4: Atributos de la tabla con la información de los Terapeutas [Elaboración propia].

ATRIBUTO	**DESCRIPCIÓN**
Terapeuta	Corresponde al nombre del terapeuta, para el caso de estudio los nombres se manejan de manera anónima, por lo que los nombres tienen la forma “TERAPEUTA” seguido de un número correlativo.
Edad	Corresponde a la edad actual del terapeuta.
Genero	Corresponde al género del terapeuta.
Turno	Corresponde al turno en el que labora el terapeuta.
Experiencia	Corresponde a la experiencia del terapeuta calculado cuantitativamente en base a la experiencia de todos los terapeutas, el cual fue explicado en el capítulo 3.
Deserción 1	Corresponde al nivel de deserción al primer mes con respecto a todos los pacientes que le fueron asignados.
Deserción 2	Corresponde al nivel de deserción al segundo mes con respecto a todos los pacientes que le fueron asignados.
Deserción 3	Corresponde al nivel de deserción al tercer mes con respecto a todos los pacientes que le fueron asignados.
Deserción 4	Corresponde al nivel de deserción al cuarto mes con respecto a todos los pacientes que le fueron asignados.

La Figura 6.4 muestra la interfaz del módulo de Terapeutas con los datos de los terapeutas, los cuales corresponden a la última actualización de la base de datos.

TERAPEUTA	EDAD	GENERO	TURNO	EXPERIENCIA	DESERCION...	DESERCION...	DESERCION...	DESERCION...
TERAPEUTA1	29.0	MASCULINO	MANANA	78.0	37.58	24.84	8.92	7.64
TERAPEUTA2	27.0	MASCULINO	MANANA	68.0	31.45	20.97	16.13	10.48
TERAPEUTA3	26.0	MASCULINO	TARDE	69.0	27.22	27.85	17.09	6.96
TERAPEUTA4	27.0	FEMENINO	TARDE	0.0	36.77	27.94	12.25	2.45
TERAPEUTA5	25.0	FEMENINO	TARDE	69.0	35.23	26.14	14.77	4.55
TERAPEUTA6	26.0	MASCULINO	MANANA	67.0	27.81	39.07	9.93	3.97
TERAPEUTA7	28.0	FEMENINO	TARDE	63.0	28.16	25.86	10.92	10.35
TERAPEUTA8	27.0	MASCULINO	MANANA	55.0	37.5	30.88	11.03	8.09
TERAPEUTA9	28.0	MASCULINO	MANANA	67.0	26.6	42.2	8.26	5.5
TERAPEUTA...	29.0	FEMENINO	MANANA	36.0	0.0	0.0	0.0	0.0
TERAPEUTA...	26.0	MASCULINO	TARDE	48.0	22.98	36.65	11.8	6.21
TERAPEUTA...	30.0	MASCULINO	TARDE	47.0	30.0	22.5	13.33	7.5
TERAPEUTA...	28.0	MASCULINO	TARDE	47.0	23.85	28.44	19.27	0.92
TERAPEUTA...	28.0	FEMENINO	TARDE	45.0	25.0	35.37	12.2	6.1
TERAPEUTA...	28.0	FEMENINO	MANANA	31.0	25.51	44.9	8.16	9.18
TERAPEUTA...	27.0	MASCULINO	TARDE	43.0	18.33	29.17	15.83	6.67
TERAPEUTA...	24.0	MASCULINO	TARDE	29.0	32.98	34.04	10.64	6.38
TERAPEUTA...	29.0	FEMENINO	MANANA	20.0	32.56	27.91	4.65	9.3

Figura 6.4: Interfaz gráfica del módulo Generar Modelo [Elaboración propia].

6.1.2 *Implementación*

Para el proceso de implementación, se estableció usar la arquitectura MVC (Modelo-Vista-Controlador). Se crearon dos paquetes: “tesis” y “proyectoClasificacion”. El paquete “tesis” sirvió para la implementación de la interfaz gráfica, así como los métodos que permitan la interacción de las componentes con los métodos de clasificación y el paquete “proyectoClasificacion” para las clases y métodos que hagan posible la clasificación. El detalle de la distribución de clases en los paquetes se muestra en el diagrama de clases del prototipo funcional.

6.1.2.1 Diagrama de clases

Para la elaboración del diagrama de clases, se tomó como base la arquitectura MVC. La Figura 6.5 muestra la distribución de las clases en los paquetes “tesis” y “proyectoClasificacion”. La Tabla 6.5 detalla la distribución de las clases, métodos y atributos de las clases de cada paquete.

La clase “frmPrincipal” corresponde a la vista, mediante esta el usuario interactúa con el prototipo funcional para solicitar la clasificación de acuerdo a los datos ingresados. La clase “ProyectoClasificacion” corresponde al controlador, mediante esta permite

la interacción entre la vista y el modelo. Las clases “clsClasificacion”, “clsModelo” y “clsInstanciaWeka” corresponden al modelo, mediante estas se permite la clasificación de los datos.

Tabla 6.5: Descripción de las clases por paquete [Elaboración propia].

PAQUETE	**CLASES**	**ATRIBUTOS**	**METODOS**
tesis	frmPrincipal	txtGeneroP txtEstadoCivil txtTipoFamilia txtProfesion txtIndependiente txtSituacionEconomica txtEnteraste txtIniciativaPropia txtDia txtTerapeuta	frmPrincipal() rellenaCombo 1() rellenaCombo 2() validarTurno() llenar() procesando() abrirArchivo() main() run()
proyectoClasificac ion	ProyectoClasificac ion		main()
	clsClasificacion	Entrenamiento Clasificador data	clsClasificacio n() clasificar()
	clsInstanciaWeka		crearInstancia ()
	clsModelo	clasificadorRandomFo rest	clsModelo() generarModel o()

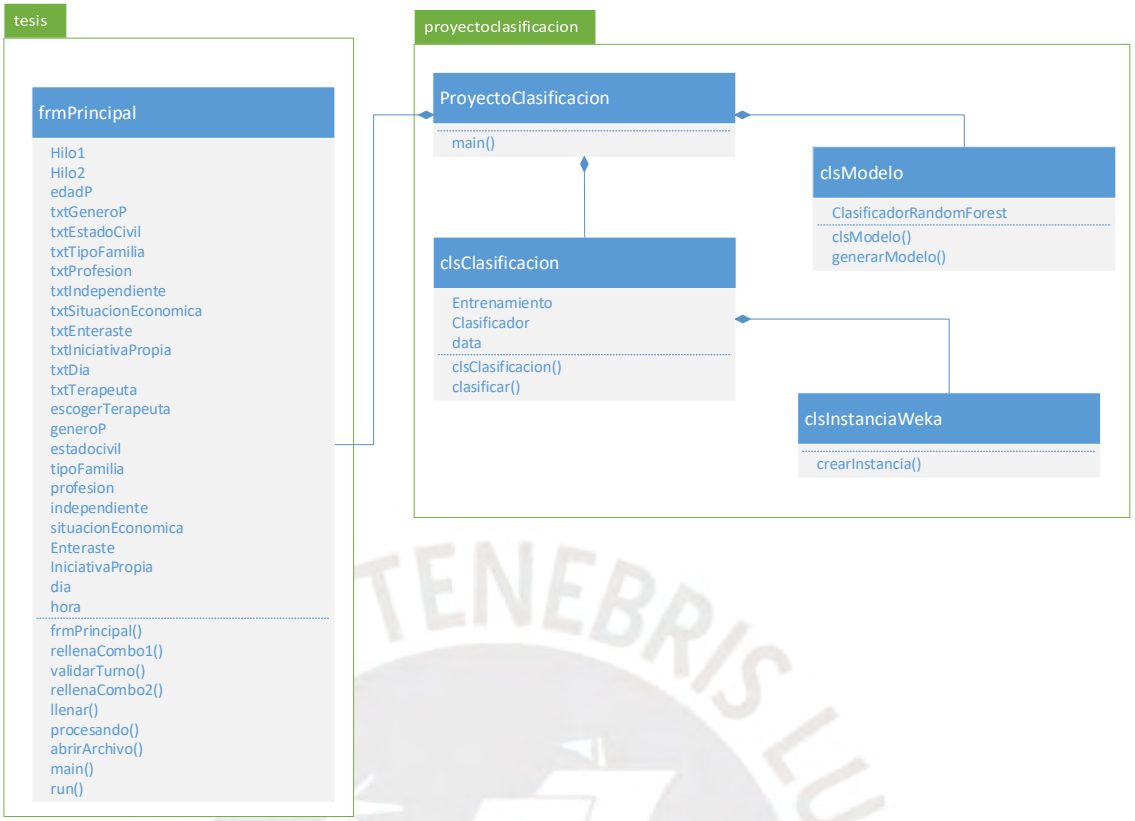


Figura 6.5: Diagrama de clases [Elaboración propia].

6.1.2.2 Implementación Algorítmica

Para la implementación algorítmica de predicción, se hizo uso de las librerías de la API de Weka. Las librerías que se usaron fueron:

- *weka.classifiers.meta.FilteredClassifier*
- *weka.filters.supervised.instance.SpreadSubsample*
- *weka.classifiers.trees.RandomForest*

Estas librerías fueron usadas en la clase "clsClasificacion" para instanciar los algoritmos ***Random Forest, Filtered Classifier*** y ***Spread Subsample.*** El método "clasificar" fue implementado para encargarse de realizar el proceso de clasificación. Para esto, el método tiene como parámetros los datos que fueron ingresados por el formulario, los cuales son transformados a un tipo ***Instance*** por medio de la clase "clsInstanciaWeka". La Figura 6.6 contiene un extracto del código implementado en el método "clasificar".

```
instance = instancia.crearInstancia(terapeuta,generoT,edadT,experiencia,edad
SpreadSubsample SSubSample=new SpreadSubsample();
Classifier = (RandomForest) weka.core.SerializationHelper.read(modelo);
FilteredClassifier fc = new FilteredClassifier();
fc.setFilter(SSubSample);
fc.buildClassifier(data);
predicted = fc.classifyInstance(instance);
return train.classAttribute().value((int)predicted);
```

Figura 6.6: Extracto del código implementado en el método “clasificar” [Elaboración propia].

La implementación del método “clasificar” da como resultado la predicción del éxito o fracaso en el proceso psicoterapéutico. Los resultados de la predicción son mostrados de la siguiente manera:

- Para el caso de Éxito: “PERMANENCIA MAYOR O IGUAL A 16 CITAS”.
- Para el caso de Fracaso: “PERMANENCIA MENOR A 16 CITAS”.

6.2 Modulo para predecir la cantidad de citas efectivas en el caso de fracaso

Para el desarrollo del prototipo funcional del Resultado Esperado 8, se basó en el diseño y las clases ya implementadas en la obtención del Resultado Esperado 7. Por tanto, su implementación consistió en aplicar cambios sobre los métodos ya desarrollados, así como en las funcionalidades. En la sección anterior, se hizo referencia a las interfaces “Cargar Modelo” y “Generar Modelo” los cuales contienen atributos que permitieron la integración de los dos modelos a usar para este Resultado Esperado. La carga de archivos como la generación de los modelos, fueron modificados para poder hacer uso de los tres modelos a usar en la predicción de la permanencia del paciente.

Para la incorporación en el uso de los dos modelos de predicción, se modificó el flujo del algoritmo que hace la llamada al método “clasificar”. La Figura 6.7, representa el flujo de predicción implementado. El modelo 1, corresponde al modelo de predicción del éxito o fracaso. El modelo 2, corresponde al modelo de predicción de permanencia en el caso de fracaso. El modelo 3, corresponde al modelo de predicción de la cantidad de citas efectivas en la fase de evaluación.

El flujo de predicción empieza por predecir el éxito o fracaso del proceso. En caso la predicción tiene como resultado 1, el cual es equivalente al éxito en el proceso, devuelve el resultado, caso contrario, predice la permanencia del paciente entre las

primeras quince primeras citas efectivas. En caso la predicción de este último tiene como resultado 1, el cual es equivalente a la permanecía en el proceso entre la cita 5-15, devuelve el resultado, caso contrario, predice la permanencia del paciente entre las 4 primeras citas efectivas.

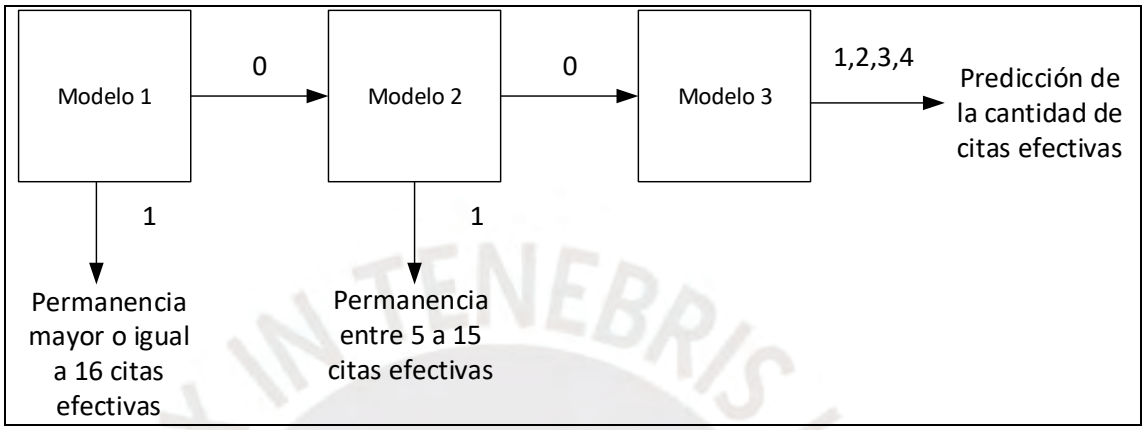


Figura 6.7: Flujo de predicción utilizando los 3 modelos [Elaboración propia].

Para la modificación, se incluyó la librería *weka.filters.supervised.instance.SMOTE* a la clase "clsClasificacion" y se aplicó una modificación al método "clasificar". Los modelos 1 y 2 utilizan el filtro Spread Subsample y el modelo 3 utiliza el filtro SMOTE, por tanto, las modificaciones se dieron sobre el setFilter del objeto filteredClassifier. La Figura 6.8 contiene un extracto de la modificación del método "clasificar".

```
instance = instancia.crearInstancia(terapeuta,generoT,edadT,experiencia,edadP,gene
SpreadSubsample SSubSample=new SpreadSubsample();
Classifier = (RandomForest) weka.core.SerializationHelper.read(modelo);
FilteredClassifier fc = new FilteredClassifier();
if(numModelo==3){
    fc.setFilter(Smote);
}else{
    fc.setFilter(SSubSample);
}
fc.buildClassifier(data);
predicted = fc.classifyInstance(instance);
return train.classAttribute().value((int)predicted);
```

Figura 6.8: Extracto del código implementado en el método "clasificar" modificado [Elaboración propia].

La modificación del método "clasificar" da como resultado la permanencia del paciente en el proceso psicoterapéutico. Los resultados de la predicción son mostrados de la siguiente manera:

- Para el caso de Éxito: “PERMANENCIA MAYOR O IGUAL A 16 CITAS”.
- Para el caso de Fracaso:
    - Predicción=1 del Modelo 2: “PERMANENCIA ENTRE 5 A 15 CITAS”.
    - Predicción=0 del Modelo 2: “PERMANENCIA EN EL PERIODO DE EVALUACIÓN”.
        - Predicción=1: “TERAPEUTA” (Cantidad de Citas: 1).
        - Predicción=2: “TERAPEUTA” (Cantidad de Citas: 2).
        - Predicción=3: “TERAPEUTA” (Cantidad de Citas: 3).
        - Predicción=4: “TERAPEUTA” (Cantidad de Citas: 4).

6.3 Pruebas del Prototipo Funcional

Para las pruebas del algoritmo, se usó información histórica almacenada en la base de datos de la institución correspondiente al periodo posterior a la extracción de los datos. Debido a que los datos tenían que cumplir ciertas condiciones, la cantidad de datos para las pruebas se limitó a 147. Las condiciones que debían cumplir eran las siguientes:

- El periodo mínimo de antigüedad del paciente debe ser mayor a seis meses.
- Solo se tomaron en cuenta los pacientes que abandonaron el proceso.
- Solo se tomaron a pacientes cuyos terapeutas habían sido considerados en el análisis por su antigüedad.

Los datos de prueba han sido incorporados en el Anexo 7. En el mismo se detalla la cantidad de citas real, así como, una columna adicional con la respuesta de predicción arrojada por el prototipo. El porcentaje de precisión fue del 72.7%, obteniendo 107 predicciones correctas de 147. El resumen de los resultados se presenta a continuación:

- De 65 predicciones = “TERAPEUTA” (Cantidad de Citas: 1), 60 fueron acertadas.
- De 11 predicciones= “TERAPEUTA” (Cantidad de Citas: 2), 4 fueron acertadas.
- De 9 predicciones= “TERAPEUTA” (Cantidad de Citas: 3), 3 fueron acertadas.
- De 14 predicciones= “TERAPEUTA” (Cantidad de Citas: 4), 7 fueron acertadas.

- De 45 predicciones= “PERMANENCIA ENTRE 5 A 15 CITAS”, 32 fueron acertadas.
- De 3 predicciones= “PERMANENCIA MAYOR O IGUAL A 16 CITAS”; 1 fue acertada.

Debido a que el algoritmo Random Forest utiliza una gran cantidad de árboles para realizar la predicción, no es posible utilizar la ventaja de los árboles de decisión, el cual permite la fácil comprensión e interpretación de resultados (BERTHOLD, 2010). Por lo tanto, no es posible precisar las características que permiten la permanencia en el proceso de los pacientes.

Tabla 6.6: Análisis comparativo de los resultados de Precisión de Random Forest con los datos de prueba y de entrenamiento

CANT. CITAS	PRUEBA CON DATOS REALES DISTINTO A LOS DATOS DE ENTRENAMIENTO				ANALISIS CON DATOS DE ENTRENAMIENTO REAL			
	ACIERTOS	TOTAL	PRECISION POR	PRECISION POR MODELO	ACIERTOS	TOTAL	PRECISION POR	PRECISION POR MODELO
1	60	65	92.31		631	657	96.04	
2	4	11	36.36		117	176	66.48	
3	3	9	33.33		83	138	60.14	
4	7	14	50.00	74.75	144	220	65.45	81.86
1 - 4	74	99	74.75		1135	1205	94.19	
5 - 15	32	45	71.11	73.61	527	699	75.39	87.29
1 - 15	74	99	74.75		1894	1904	99.47	
16 +	1	3	33.33	73.53	228	352	64.77	94.06

De la Tabla 6.6, se observa que los niveles de precisión por modelo de las pruebas superan el 70%, por lo que se puede definir como aceptable, sin embargo, dista de la precisión por modelo de los análisis realizados con los datos de entrenamiento. La explicación de este distanciamiento se puede dar debido a la cantidad de pruebas realizadas con datos reales, los cuales fueron limitados debido a la rotación de los terapeutas que habían sido tomados para el análisis y las condiciones que debían

cumplir los pacientes para poder ser evaluados (datos de pacientes con una antigüedad mayor a 6 meses).

6.4 Conclusiones

El desarrollo del prototipo funcional tomo algunas dificultades, debido a que era necesario investigar sobre la API de Weka y los métodos a usar para poder realizar el proceso de clasificación correctamente. La construcción de cada una de las funcionalidades utilizando un diseño sencillo y práctico, permitió que esta sea intuitiva y fácil de usar. La arquitectura MVC permitió que el proceso de modificación para obtener el Resultado Esperado 8 sea más sencillo. Las funcionalidades incorporadas correspondientes a la carga y generación de los modelos no han sido explicadas al detalle debido a que no resulta relevante para el proceso de clasificación, sin embargo, si resultan ser importantes si en trabajos futuros se desean cargar datos de entrenamiento con mayor cantidad de registros.

La cantidad de datos de prueba obtenidos tuvieron que pasar por un proceso de pre-tratamiento de datos, debido a que se extrajo de la base de datos de origen el cual tenía problemas de normalización de datos. Después de haber desarrollado el prototipo funcional y realizar las pruebas del mismo, podemos concluir que los resultados son aceptables y se encuentran dentro de lo esperado. Esta última afirmación se basa en los resultados obtenidos en cada modelo por clase analizados en los capítulos 4 y 5 en comparación con los resultados del prototipo funcional obtenidos de los datos de prueba y que se resumen en la Tabla 6.6. Se debe notar que la precisión de los resultados es determinada por la cantidad de datos por clase, por lo tanto, mientras más datos pertenezcan a una determinada clase, el nivel de precisión aumentara.

7. CAPÍTULO 7: CONCLUSIONES Y RECOMENDACIONES

En este Capítulo se desarrolla las conclusiones y recomendaciones para trabajos futuros referentes al proyecto de fin de carrera.

7.1 Conclusiones

El proyecto de fin de carrera tenía como finalidad determinar la permanencia de los pacientes en el proceso psicoterapéutico, a partir de los datos obtenidos desde el primer contacto entre el paciente y la institución. Después de todo el desarrollo del proyecto se concluye lo siguiente:

- Se verifico, por medio de la investigación del Estado del Arte, que existen investigaciones relacionadas, sin embargo, los niveles de precisión son aspectos que se buscan mejorar con el incremento de los datos de entrenamiento y los atributos a analizar.
- Se observó que los datos almacenados en la base de datos de la institución carecen de estandarización y normalización, por lo que su extracción y pre-tratamiento de datos resulto ser una de las labores que tomo más tiempo dentro de todo el proyecto.
- Se observó que la cantidad de pacientes que sobrepasan las 15 citas efectivas resulta ser el 15.6% del total de pacientes del análisis, por lo que se verifica la necesidad de entender y analizar la deserción de los pacientes durante el proceso.
- Se comprobó, por medio del análisis de componentes, que los atributos seleccionados de los pacientes han sido seleccionados de manera adecuada de tal forma que se pudo determinar un patrón para determinar la predicción de la permanencia en el proceso Psicoterapéutico, sin embargo, para mejorar los niveles de precisión se optó por tomar algunos atributos, los cuales, según el análisis, no colaboraban de manera significativa en la descripción de los datos.
- Se comprobó, por medio de la Hipótesis nula sobre los niveles de deserción al quinto mes, que el mismo tiende a ser menor o igual a 3.56%. Por tanto, para el análisis se definió la clase “Éxito” como los pacientes que permanecieron en el proceso durante 16 citas efectivas a más, debido a la baja probabilidad de deserción.
- Se comprobó que el desbalanceo de clases resulta ser un problema grave al realizar el análisis algorítmico, por lo que fue necesario investigar sobre el

tratamiento de clases desbalanceadas y aplicar filtros sobre los conjuntos de datos.

- Se observó que el flujo algorítmico establecido para la predicción en base a los resultados previos, permitió ser más específico para determinar la permanencia del paciente en el proceso psicoterapéutico.
- Se observó que las interfaces establecidas en el prototipo funcional para la carga de los datos de entrenamiento y generación de los modelos permiten tener un prototipo funcional flexible para actualizar e incrementar los datos de entrenamiento para la predicción.
- Se verifico que la precisión de la predicción de las clases depende del tamaño del conjunto de datos al cual pertenece la clase, en consecuencia, se obtuvo bajos niveles de precisión en las clases minoritarias.
- Se verifico que los niveles de precisión obtenidos del prototipo funcional son aceptables en base a los datos de entrenamiento y datos de pruebas extraídos de la base de datos de la institución, el cual tuvo que pasar por un proceso de pre-tratamiento de datos para poder trabajar sobre estos datos.

7.2 Recomendaciones

Las siguientes recomendaciones han sido recogidas a partir del desarrollo del presente proyecto de fin de carrera con el fin de que puedan ser consideradas en trabajos futuros.

- Se recomienda re-estructurar la base de datos de la institución, de tal forma que los datos de entrenamiento puedan ser obtenidas de manera automática y pueda retro-alimentarse de manera permanente.
- Se recomienda incrementar el periodo de análisis de tal forma que la cantidad de datos permita incrementar el nivel de precisión en las fases del proceso.
- Se recomienda incrementar de manera significativa los datos de entrenamiento, con el fin de mejorar los niveles de precisión en todo el análisis.
- Se recomienda no considerar en los datos de entrenamiento a terapeutas con un tiempo de permanencia en la institución menor a un año.
- Se recomienda no considerar en los datos de entrenamiento a pacientes con un tiempo de antigüedad desde su primera cita menor a 6 meses.
- Se recomienda realizar el proceso de predicción por cantidad de citas en el rango de 1-15 citas efectivas con 15 clases, con el fin de ser más específico en la predicción.

- Se recomienda añadir factores conductuales al análisis con el fin de investigar su impacto y mejoría sobre los niveles de precisión, teniendo en cuenta que el análisis abarcaría el proceso terapéutico.

Referencias bibliográficas

- ALCALA-FDEZ, Jesus. KEEL: A data mining software tool integrating genetic fuzzy systems. 2008, p. 83-88.

- BARRENO-VEREAU, E. (2012). Análisis Comparativo de modelos de clasificación en el estudio de la deserción universitaria. Interfases, (5), 45-82.

- BENÍTEZ, Ángela Patricia Rondón; BASTIDAS, Iván Leonardo Otálora; CAMARGO, Yenny Salamanca. Factores que influyen en la deserción terapéutica de los consultantes de un centro universitario de atención psicológica. *International Journal of Psychological Research*, 2009, vol. 2, no 2, p. 137-147.

- BERNARDI, Ricardo, et al. Guía clínica para la psicoterapia. Rev. Psiquiátrica Uruguaya, 2004, vol. 68, no 2, p. 99-146.

- BERTHOLD, M. R., BORGELT, C., HÖPPNER, F., & KLAWONN, F. (2010). Guide to intelligent data analysis: how to intelligently make sense of real data. Springer Science & Business Media.

- BLEGER, José. Psicoanálisis del encuadre psicoanalítico. *Revista de Psicoanálisis*, 1967, vol. 24, no 2, p. 241-258.

- BREIMAN, L. (2001). Random forests. Machine learning, 45(1), 5-32.

- BUZAI, G. D. (2011). Modelos de localización-asignación aplicados a servicios públicos urbanos: análisis espacial de Centros de Atención Primaria de Salud (CAPS) en la ciudad de Luján, Argentina. Cuadernos de Geografía-Revista Colombiana de Geografía, 20(2), 111-123.

- CAMARGO, H., SILVA, M. Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP. Rev. Tecnol, 9(1), 2011.

- CARRASCO Ochoa, Jesus Ariel, & MARTÍNEZ Trinidad, Jose Francisco. Reconocimiento de patrones. de patrones. Komputer Sapiens, 2011, Año 3, Vol.2. p. 5-9.

- CASTRO SOLANO, Alejandro. ¿Son eficaces las psicoterapias psicológicas? 2002.

- CHAWLA, Nitesh; BOWYER, Kevin; HALL, Lawrence; and KEGELMEYER, Philip. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16(1):321–357, 2002.

- DE RIVERA, JL González. La psicoterapia multidimensional. Psiquis. 1990. vol. 11, p. 246-254.

- DUBIAU, Luciana; ALE, Juan M. Análisis de Sentimientos sobre un Corpus en español: Experimentación con un Caso de Estudio. 2013.

- FAWCETT, Tom, "ROC graphs: Notes and practical considerations for researchers," Machine learning, vol. 31, pp. 1-38, 2004.

- GARRE, Miguel, et al. Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. Revista Española de Innovación, Calidad e Ingeniería del Software, 2007, vol. 3, no 1, p. 6-22.

- GERVILLA García, Elena, et al. The methodology of Data Mining. An application to alcohol consumption in teenagers. Adicciones, 2009, vol. 21, no 1.

- GUZMÁN REYES, Daniel. Bases de datos distribuidas con una solución LAMP (Linux, Apache, MySQL y PHP). México: Universidad Autónoma del Estado de Hidalgo, Instituto de ciencias básicas e Ingeniería. 2006.

- HALL, Mark; FRANK, Eibe; HOLMES, Geoffrey; PFAHRINGER, Bernhard; REUTEMANN, Peter; and WITTEN Ian H. The WEKA data mining software: an update. SIGKDD Explorations, 11(1):10–18, 2009.

- HAN, J., PEI, J., & KAMBER, M. Data mining: concepts and techniques. Elsevier, 2011.

- HAN, Jianchao; RODRIGUEZ, Juan C.; BEHESHTI, Mohsen. Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. 2008, p. 96-99.

- HERRERO, Félix de Moya Anegón Víctor; BOTE, Solana Vicente Guerrero. La aplicación de Redes Neuronales Artificiales (RNA): a la recuperación de la información. 1998, p. 147-164.

- INFANTE, Marta, et al. Minería tecnológica para el análisis de oportunidades de publicaciones en la universidad. Revista CENIC. Ciencias Biológicas, 2010, vol. 41.

- JUÁREZ, O., & CASTELLS, E. (2010). Modelos de árbol de regresión bayesiano: un estudio de caso. Investigación Operacional, 31(2), 109-125.

- KALMEGH, Sushilkumar. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. International Journal of Innovative Science, Engineering & Technology, 2015, vol. 2, no 2, p. 438-446.

- KOHAVI, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

- KRAUSE, M. S., HOWARD, K. I., & LUTZ, W. Exploring individual change. Journal of Consulting and Clinical Psychology. 1998, 66, 838–845.

- LANDWEHR, Niels; HALL, Mark; FRANK, Eibe. Logistic model trees. En European Conference on Machine Learning. Springer Berlin Heidelberg, 2003. p. 241-252.

- LIRIA, Alberto Fernández; VEGA, Beatriz Rodríguez. La práctica de la psicoterapia. *Descleé de Brouwer, Bilbao*, 2001.

- LÓPEZ, José Manuel Molina; HERRERO, Jesús García. Técnicas de análisis de datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA. 2006.

- LUTZ, W.; SAUNDERS, S. M.; LEON, S. C.; MARTINOVICH, Z.; KOSFELDE, J.; SCHULTE, D.; GRAWE, K.; THOLEN, S. Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. Psychological Assessment, 2006, 18(2), 133-141. doi:10.1037/1040-3590.18.2.133.

- LUTZ, Wolfgang; LAMBERT, Michael J.; HARMON, S. Cory; TSCHITSAZ, Armita; SCHÜRCH, Eva; STULZ, Niklaus. The probability of treatment success, failure and duration—what can be learned from empirical data to support decision making in clinical practice? 2006, Volume 13, Issue 4, p.223–232.

- MARCANO Cedeño, A. E., Chausa Fernández, P., Cáceres Taladriz, C., García, A., López, R., Tormos Muñoz, J. M., & Gómez Aguilera, E. J. (2011). Análisis comparativo de algoritmos de aprendizaje para predecir la evolución de pacientes con Daño Cerebral Adquirido.

- MARTÍNEZ, Rocío Erandi Barrientos, et al. Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 2009, vol. 9, no 2, p. 19-24.

- MORENO, Jaime; ROZO, Margarita; CANTOR, Martha. Permanencia y abandono terapéutico en un centro de servicios psicológicos. *Psychologia: avances de la disciplina*, 2012, vol. 6, no 2, p. 23-34.

- MOSQUERA, R., PARRA-OSORIO, L., & CASTRILLÓN, O. D. (2016). Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos. Información tecnológica, 27(6), 259-272

- OROZCO, Ricardo; ROSETY, Carmen Hernández. *Flores de Bach Recursos y estrategias terapéuticas*. El Grano de Mostaza, 2014.

- PARRA, Juan Eduardo Matamala. Estudio y modelamiento de un shallow parser de textos en lenguaje natural utilizando técnicas de computación

evolutiva. Concepción: Universidad de Concepción, Facultad de Ingeniería, Departamento de Ingeniería Civil en Informática, 2007.

- PERVERSI, Ignacio. Aplicación de minería de datos para la exploración y detección de patrones delictivos. 2007. p. 3-30.

- PETROSINO, Alfredo; LOIA, Viscenzo; & PEDRYCZ, Witold (Eds.). (2017). Fuzzy Logic and Soft Computing Applications: 11th International Workshop, WILF 2016, Naples, Italy, December 19–21, 2016, Revised Selected Papers (Vol. 10147). Springer.

- PORTILLO, María Teresa Escobedo; MENDOZA, Jorge A. Salas Plata. P. CH. Mahalanobis y las aplicaciones de su distancia estadística. CULCyT. 2015, no 27.

- POWERS, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

- RIVERO, José Matías, et al. From mockups to user interface models: an extensible model driven approach. En Current Trends in Web Engineering. Springer Berlin Heidelberg. 2010. p. 13-24.

- ROKACH, Lior; MAIMON, Oded. Data mining with decision trees: theory and applications. World scientific, 2014.

- SEIJAS, Leticia María. Reconocimiento de dígitos manuscritos mediante redes neuronales: una técnica híbrida. Jornada Argentina de informática e investigación operativa, 2003, p. 1-21.

- SUÁREZ, Armando; MONTOYO, Andrés. Estudio de cooperación de métodos de desambiguación léxica: Marcas de Especificidad vs. Máxima Entropía. Procesamiento Lenguaje Natural. 2001, vol. 27, no 1, p. 207-214.

- TRUJILLANO, J., SARRIA-SANTAMERA, A., ESQUERDA, A., BADIA, M., PALMA, M., & MARCH, J. (2008). Aproximación a la metodología basada en

árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. Gaceta Sanitaria, 22(1), 65-72.

- VÁZQUEZ, Sonia. Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN. Alicante: Universidad de alicante, Depto. de Lenguajes y Sistemas Informáticos. 2009.

- VERA, Miguel; BUSTAMANTE, J. Modelo dinámico para la generación de pronóstico usando redes neurales artificiales (RNA). Visión Gerencial, 2007 p. 130-142.

- VILLMANN, Thomas; HESSEL, Aike. Analyzing Psychotherapy Process Time Series Using Neural Maps. Universidad Leipzig. Alemania, 1999.

- Xue, Ming. A Study and Application on Machine Learning of Artificial Intellligence. 2009, p. 272-274.

- ZUKERFELD, Rubén. Alianza terapéutica, cambio psíquico y encuadre analítico. *Aperturas psicoanalíticas: Revista de psicoanálisis*, 2001, no 7, p. 8.