UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ANÁLISIS DE LA RELACIÓN ENTRE LA ADHERENCIA AL TRATAMIENTO ANTIRRETROVIRAL (TAR) DE PACIENTES CON VIH+ Y EL CONSUMO DE ALCOHOL Y OTRAS DROGAS, EN BASE A TÉCNICAS DE MACHINE LEARNING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

NATHALIE BEATRIZ ECHEVERRÍA SOLÍS

PROFESORA GUÍA:
ROCÍO RUIZ MORENO

MIEMBROS DE LA COMISIÓN:
FELIPE VERA CID
FELIPE VILDOSO CASTILLO

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniera Civil Industrial
POR: Nathalie Beatriz Echeverría Solís
FECHA: 2022
PROF. GUÍA: Rocío Ruiz Moreno

ANÁLISIS DE LA RELACIÓN ENTRE LA ADHERENCIA AL TRATAMIENTO ANTIRRETROVIRAL (TAR) DE PACIENTES CON VIH+ Y EL CONSUMO DE ALCOHOL Y OTRAS DROGAS, EN BASE A TÉCNICAS DE MACHINE LEARNING

Al 2020 se encuentran 77 mil personas con VIH en Chile, de los cuales tan solo 54 mil conocen su diagnóstico [3]. Es bajo este contexto que nace el proyecto del WIC de desarrollo de una Plataforma informática basada en inteligencia artificial para la identificación y caracterización del grado de adherencia en pacientes VIH+. Este proyecto contempla la construcción de 4 predictores centrados en distintas aristas de información del paciente: Datos farmacológicos, de depresión, socioeconómicos y de consumo de alcohol y otras drogas, teniendo en común para todos los datos clínicos del paciente. Es en este último predictor, el de consumo de alcohol y otras drogas, que se da el trabajo de memoria, teniendo en cuenta que el consumo de alcohol y drogas juega un rol importante en Chile donde 1 de cada 10 personas posee consumo de alcohol de riesgo. El trabajo realizado se ejecuta en conjunto con la Fundación Arriarán, la contraparte del proyecto. En esta, en base a una encuesta preliminar, se observa un total de 15% de pacientes que presentan problemas de consumo.

La hipótesis de este trabajo considera que existe una relación que modifica la adherencia al tratamiento TAR y los pacientes VIH+ que consumen alcohol y otras drogas. Para ello se tiene el siguiente objetivo: Desarrollar un modelo predictivo de adherencia al tratamiento TAR de VIH+ con relación al consumo de alcohol y otras drogas. El trabajo realizado comienza con la búsqueda bibliográfica en donde se encuentran las principales variables a utilizar, divididas en datos clínicos, de caracterización de paciente y niveles de consumo. Luego del análisis de las variables de interés y de la información disponible dentro de las tablas de información de la Fundación Arriarán, se construye la base del predictor, el cual cuenta con 58 variables correspondientes a la información de 2484 pacientes.

El predictor elegido por su desempeño corresponde a SVM (*support vector machine*) el que entrega un recall de 0,75. Con este modelo se genera la clasificación de adherencia de los pacientes, lo cual permite la identificación de aquellos pacientes en riesgo de abandono de tratamiento. La correcta identificación de esto y el análisis tanto del perfil del paciente, en base a las variables del predictor, como el peso de estas, permite generar políticas de seguimiento con optimización de recursos, lo cual no solo trae un bienestar al paciente y su recuperación, sino también a la Fundación como un todo. El proyecto contempla a futuro la creación de una plataforma que ayude no solo a la visualización de los resultados de los predictores sino también una estandarización de ingreso de la información, lo cual soluciona problemas encontrados en el trabajo de memoria y abre aún más las posibilidades de replicación del trabajo.

A cada uno de los miembros de mi familia, los pilares de mi vida

Agradecimientos

Gracias al equipo del WIC, a los profesores Juan y Rocío quienes me abrieron las puertas al centro en donde pude realizar esta memoria. Especialmente quiero agradecer al equipo “Core”, Raúl y Cristián quienes siempre tuvieron la disposición de ayudarme a mejorar. A Flavia, en quién encontré no solo una gran mentora y profesional sino también un gran ser humano, su dedicación, empatía y constante apoyo me ayudaron a perseverar en este viaje, me llevo de ella aprendizajes que trascienden lo científico e ingenieril. Prometo encontrar mi *gate out.*

Cristián, Pablo, Coni, Amanda, gracias por ser parte de mi crecimiento durante estos años, por estar en los momentos en que los necesite y hacer más amena la vida.

Agradezco a los pilares de mi vida, mi familia. Sin ellos nada sería posible. Claudia, Coni gracias por ser un apoyo incondicional incluso desde la distancia. Papá por querernos tanto y permitirme tener siempre un entorno de apoyo. A mis hermanos:
Alvaro, cuya sabiduría, enseñanzas, fuerza y dulzura han sido una constante en mi vida. Alex, hermano, gracias por ser el mejor compañero que la vida me podría haber entregado, por cada consejo y risa, las cuales me han salvado en los momentos más obscuros y han mejorado los días más alegres. Gracias a ambos por ser mis hermanos.

Mamá, sentaste las bases de lo que soy, te llevo cada día en mi corazón, sin ti nada de lo que soy sería posible.

Gracias Rodrigo, mi compañero de vida, por estar en los momentos más difíciles de mi vida, apoyarme y creer en mí incluso en momentos en que yo no lo hago. Alegras cada segundo.

Gracias a cada persona que de alguna manera u otra me ayudaron a llegar a este momento, terminó esta etapa con la disposición de seguir aprendiendo de la vida.

Tabla de contenido

Índice de Figuras

Índice de Tablas

1. Introducción

1.1 Antecedentes Generales

1.1.1 El VIH-SIDA

El VIH (virus de inmunodeficiencia humana) es un virus que afecta al sistema inmune (sistema de defensa del organismo) al destruir cierto tipo específico de glóbulos blancos (linfocitos T CD4). Una vez debilitado el sistema inmune permite la aparición de enfermedades, que son denominadas oportunistas. La etapa avanzada de la infección por VIH se denomina Síndrome de Inmuno Deficiencia Adquirida (SIDA), la cual se caracteriza por un conjunto de síntomas (síndrome) que aparece por una insuficiencia del sistema inmune (inmunodeficiencia) causada por un virus que se transmite de persona a persona (adquirida) [1].

En 1981 fue declarada pandemia por la OMS (Organización Mundial de la Salud) y desde entonces y hasta el 2020 se han reportado 79,3 millones de personas que han contraído el virus, de las cuales 36,3 millones han fallecido a causa de enfermedades relacionadas a esta. Desde una perspectiva actual, para el año 2020, a nivel global, se tiene que existían alrededor de 37,7 millones de personas viviendo con la infección, teniendo para ese mismo año un aproximado de 680.000 personas fallecidas debido a complicaciones. De hecho, sólo en 2020 1,5 millones de personas contrajeron el virus a nivel mundial [2]. Si se habla de Chile, para 2020, se estima que existen 77 mil personas viviendo con el virus día a día de la que la gran mayoría son hombres mayores de 15 años (63,000) y 14 mil corresponden a mujeres [3]. De ellos, sólo 54.000 conocen el estado de su condición.

Teniendo en cuenta el estado de propagación del virus, la ONUSIDA, un programa conjunto con las naciones unidas dedicada específicamente a VIH/Sida, estableció el llamado plan 90-90-90, en este se planteaba tener para el año 2020 al 90% de la población con conocimiento de su diagnóstico, de esta 90% en tratamiento, y de aquellos con terapia antirretrovírica, que el 90% tenga supresión viral. Esta meta lamentablemente no ha sido cumplida para la fecha [4].

Los estados de la infección de VIH han sido clasificados por la Organización Mundial de la Salud en 4 grandes etapas [5]:

-**Etapa 1:** El conteo de linfocitos CD4 es por lo menos 500 células por microlitro ($cel/\mu L$).

-**Etapa 2:** El conteo de linfocitos CD4 es entre 350 a 499 ($cel/\mu L$).

-**Etapa 3** (etapa avanzada de VIH): el conteo de linfocitos CD4 es de 200 a 349 ($cel/\mu L$).

-**Etapa 4** (SIDA): el conteo de linfocitos es menor a 200 ($cel/\mu L$) o el porcentaje de células CD4 es menor al 15% de todos los linfocitos. Una cantidad tan baja de CD4 hace probable la aparición de enfermedades oportunistas.

1.1.2 Tratamiento de VIH

El principal tratamiento en la actualidad corresponde a la terapia antirretroviral (TAR), la cual corresponde a una serie de medicamentos que impide la replicación del virus del VIH en el cuerpo humano de manera de disminuir la concentración del virus en la sangre (lo cual se conoce como carga viral). Existe una serie de medicamentos agrupados en distintas clases los cuales son suministrados al paciente tomando en cuenta sus antecedentes con la enfermedad, así como condiciones basales del paciente (edad, enfermedades previas, etc.) [6,7]. Estos medicamentos deben ser tomados en un régimen diario estricto y mantener un constante monitoreo de la carga viral del paciente.

Cabe destacar que el tratamiento TAR no elimina el VIH por completo del cuerpo, solo lo reduce a niveles lo suficientemente bajos para que el sistema inmune pueda volver a funcionar, así como para evitar la transmisión del virus de una persona a otra [6]. Es por esto que mantener el tratamiento contra el VIH es sumamente relevante para la salud del paciente, así como de la comunidad mundial.

1.1.3 Adherencia

Hay varios factores que pueden determinar el fracaso del tratamiento TAR en pacientes con VIH. Este trabajo se centrará en la baja adherencia al tratamiento. Esto es particularmente desafiante dado que puede variar caso a caso y puede darse por múltiples causas.

Como se mencionó anteriormente, los beneficios del tratamiento están comprobados: un paciente con un alta nivel de adherencia logra llegar a niveles de supresión o carga virales indetectable. Este nivel se considera cuando las muestras sanguíneas del paciente presentan niveles inferiores a 200 copias del virus por milímetro de sangre. En estos niveles, según ONUSIDA, el paciente no puede transmitir el virus a través de intercambio sexual por lo que fomentar una adherencia alta al tratamiento genera un inmenso beneficio en el control de esta pandemia.

Por lo mismo, identificar qué factores afectan la adherencia al tratamiento se vuelve crítico para poder avanzar a nivel mundial con esta pandemia. Dado lo anterior, este proyecto se inserta como parte de las iniciativas del Web Intelligence Center (WIC) que buscan

estudiar los factores, efectos y tomar medidas dentro del contexto del tratamiento de los pacientes VIH+.

1.1.4 Factores que influyen en una baja adherencia

Los factores que pueden desembocar en una baja adherencia al tratamiento TAR son principalmente 2: económicos y sociales. En Chile, desde el 2018 que los tratamientos contra VIH se encuentran en el plan auge y pueden ser entregados a toda la población [8], por lo cual el no acceso a tratamiento por no poder costearlo no aplica en el caso de estudio de este trabajo. Dentro de las causas sociales se encuentran razones personales tales como olvido, encontrarse fuera de casa y no tener las pastillas o estar muy ocupados durante el día [9]. Otras causas que evitan adherencia óptima están relacionadas con desórdenes psiquiátricos tales como depresión, abuso de alcohol y drogas y efectos secundarios de otras medicaciones [9].

El uso de alcohol y drogas es de los más mencionados y que puede tener mayor impacto. En Chile, se tiene que 1 de cada 10 personas presenta consumo de riesgo de alcohol [10], esto quiere decir que poseen un consumo regular de 20 a 40 grs. diarios de alcohol en mujeres y de 40 a 60 grs. diarios en hombres [11]. Además, se tiene que, con un consumo promedio de 9,6 litros de alcohol puro al año, en comparación a los 8,4 litros del resto de América Latina, Chile es el país de esta con un mayor consumo anual [12].

Por otra parte, se debe entender que los pacientes que usan drogas intravenosas (UDI) son identificados como una población de riesgo de transmisión de VIH. Se estima que para el año 2003, el 10% de los casos de VIH en el mundo eran debido al uso de drogas y el hecho de compartir jeringas entre ellos. Igualmente, este indicador varía mucho de acuerdo con el país donde se observe y monitoree. Por ejemplo, en China el 71% de los casos reportados de VIH son debido a esta práctica [13].

Estos datos son importantes si se tiene en cuenta que una mayor adherencia al tratamiento se traduce en:

a) Disminución de la mortalidad y morbilidad

b) Mejoría de la calidad de vida,

c) Restablece y preserva la función inmunológica,

d) Logra una supresión viral total (el ARN del virus es indetectable en el plasma),

e) Disminución del contagio-transmisión

f) Reducción del gasto público salud ya que, por ejemplo, se disminuyen las hospitalizaciones frecuentes debido a infecciones oportunistas y cánceres asociados [14,15].

Cabe destacar además que existen estudios que plantean una relación entre consumo de alcohol y baja adherencia al tratamiento TAR, además de una correlación entre consumo de cocaína y un abandono del tratamiento [16,17].

1.1.5 El *Web Intelligence Centre* (WIC)

El trabajo desarrollado se realiza en el *Web Intelligence Centre* (WIC) un centro de investigación que nace en 2008 en el departamento de Ingeniería Civil Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Siendo dirigido y fundado por el profesor Juan Velásquez.

El WIC nace con la visión de “Ser un centro de referencia en investigación, desarrollo y transferencia de conocimientos en soluciones basadas en TICs, DS e IA para Chile y el mundo” la cual cumple a través de su misión de “Poner a disposición de la sociedad soluciones innovadoras basadas en TICs, Ciencia de los Datos e Inteligencia Artificial que apunten a resolver problemas de base científico-tecnológico.” [42]

El centro está conformado por 18 trabajadores full-time los cuales se separan en las distintas áreas, las cuales se pueden observar en el organigrama.

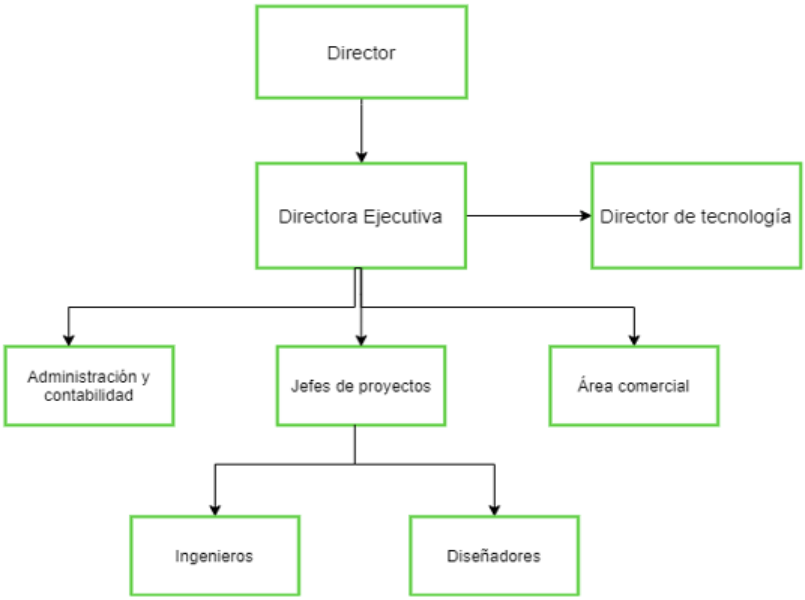

Figura 1. Organigrama del WIC.

Actualmente se encuentra como Director Juan Velásquez y como Directora Ejecutiva Rocío Ruiz.

El WIC se encuentra dividido en tres áreas de trabajo:

- Tecnológicas
- Consultoría e Investigación

- Salud Digital

La mayoría de los proyectos son financiados por fondos públicos. El trabajo de esta memoria se da específicamente en esta última área, salud digital, en la cuales se encuentran proyectos como [43]:

- **Kefuri:** Proyecto de generación de un sistema de aviso automático de presencia de posibles donadores de órganos. El aviso se genera desde el personal de urgencias hacia la unidad de procuramiento y UCI.
- **SONAMA:** Plataforma informática que estudia la prevalencia de marihuana y alcohol en Chile según un análisis de redes sociales.
- **Neutrokid:** Plataforma informática basada en Machine Learning para optimizar el manejo de los episodios de fiebre y neutropenia en niños con cáncer.
- **Proyecto Trastorno del ánimo:** Diseño e implementación de un sistema predictor de riesgo de depresión unipolar y bipolar en la población chilena a través del uso de inteligencia artificial.
- **Proyecto Delirium:** Proyecto de desarrollo de software para prevenir *Delirium* (alteración de las capacidades mentales, suele causar problemas de memoria) en adultos mayores hospitalizados.
- **Proyecto Melanoma:** Desarrollo de un modelo predictivo para la detección de melanomas en la población chilena a través de Inteligencia Artificial.
- **Proyecto VIH:** Desarrollo de una plataforma informática que utiliza inteligencia artificial para la caracterización e identificación del grado de adherencia al tratamiento retroviral para la población con infección de VIH.

Es en este último proyecto donde se enmarca este trabajo de memoria. Cabe destacar que para el desarrollo del proyecto VIH, se cuenta con la colaboración de la Fundación Arriarán, quienes facilitan la información con la cual se desarrolla el trabajo de investigación. La fundación es un centro de atención integral para pacientes con VIH positivo siendo una fundación sin fines de lucro desde que fue fundada en 1992 y actualmente está conformada por [44]:

- Unidad médica, con un total de 14 profesionales entre los cuales se encuentran infectólogos, inmunólogos, e internistas/diabetistas.
- Unidad de Enfermería y salud de la mujer, con 5 profesionales, comprendidos por enfermeras, matronas y técnicos paramédicos.
- Unidad de servicio social: conformado por una asistente social.
- Unidad de salud mental: conformado por 2 psicólogas.
- Unidad de farmacias: conformado por 3 químicos farmacéuticos y 2 técnicos de farmacias.
- Secretarias administrativas.

1.2 Definición y Justificación del Problema

1.2.1 Contexto

En base a la situación descrita de VIH y su impacto en Chile, el WIC está desarrollando el proyecto FONDEF ID20I1074 que tiene por título: “Plataforma informática basada en inteligencia artificial para la caracterización e identificación del grado de adherencia” cuya directora y codirectora son Dra. Claudia Cortes, miembro de la Fundación Arriarán y la Dra. Flavia Guiñazú, *medical advisor* y miembro del WIC, respectivamente. Este proyecto ve la oportunidad de generar una fuente de información para monitorear y predecir la adherencia al tratamiento antirretroviral de pacientes con VIH+ para disminuir la epidemia a nivel nacional. En base a lo mismo se ve como solución que Chile pueda contar con un modelo predictor para determinar cuáles de los pacientes que se encuentran en tratamiento son más propensos a: a) abandonar y retomar el tratamiento o, b) directamente abandonar el mismo. Esto permitiría una mejor distribución de recursos y herramientas, al poder poner un énfasis en aquellos que se encuentren en peligro de generar resistencia a la TAR o de fallecer (en caso de abandonar definitivamente el tratamiento), lo cual repercutiría de forma positiva en el paciente.

El proyecto antes mencionado tiene como tarea principal la generación de cuatro sub predictores, con la finalidad de considerar distintos grupos de variables de relevancia al momento de analizar adherencia en pacientes con VIH+. Las variables de interés se agruparon en base a cuatro sub predictores:

- Datos de farmacia
- Datos relacionados con depresión
- Datos relacionados con ecosistema (ambiente familiar y laboral, contexto sociocultural, etc.) del paciente
- Datos relacionados con alcohol y otras drogas

En este último se enmarca el presente trabajo de memoria, siendo este contexto un predictor de alcohol y otras drogas que no solo funciona como un predictor independiente, sino que también, forma parte de una plataforma que funciona en base a los cuatro sub predictores mencionados. En la Figura 2 se pueden apreciar los distintos sub predictores del proyecto, los cuales toman distintos sets de datos, pero todos tienen un grupo de datos que comparten y son los denominados, datos clínicos del paciente.

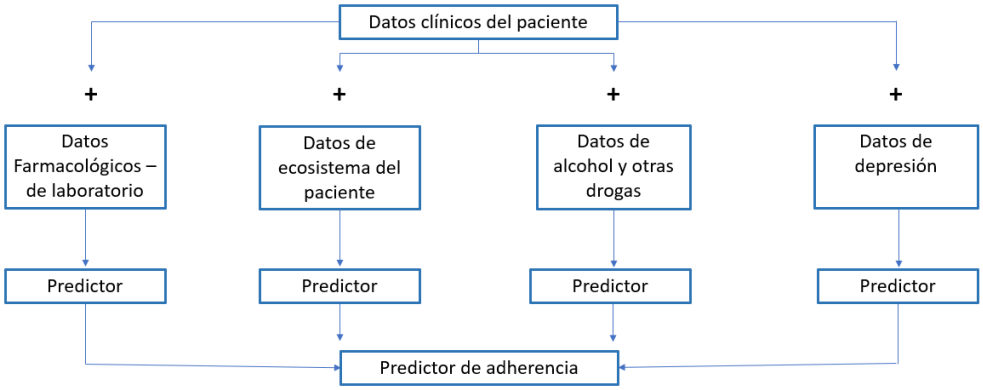

Figura 2. Predictores proyecto VIH. Elaboración propia

Por tanto, el presente trabajo de memoria se desarrolla en el contexto de un análisis que permita entender la adherencia al tratamiento antirretroviral (TAR por sus siglas en español) en un subconjunto específico de la población: pacientes que presentan consumo de alcohol y otras drogas por parte del WIC (Web Intelligence Centre) de la Universidad de Chile. Para esto, se utilizan diversas técnicas de *Machine Learning* para poder corroborar (o descartar) dichas correlaciones para así diseñar herramientas gráficas de apoyo a los equipos de investigación y seguimiento que permitan observar de manera efectiva la adherencia al tratamiento y ser preventivos en los casos que exista desviación de los parámetros relevantes.

En cuanto al impacto, como contexto, el predictor global permitirá tener un predictor de adherencia al tratamiento según una categorización, esto permitirá el enfocar recursos y proyectos en aquellos que se encuentran en riesgo de abandono. El predictor que se habla como trabajo de memoria entregará un *output* dado de la misma forma, pero su importancia está en que permitirá observar el rol de variables que en un pool tan grande como el global podrían haberse perdido. Además, guarda especial importancia estudiar la relación de consumo de alcohol y otras drogas con la adherencia en un país existe una prevalencia de consumo.

La relevancia de esta investigación viene de diversos factores: La pandemia del VIH ha estado presente en el mundo desde inicios de 1980 con un total estimado de 37.6 millones de personas a nivel mundial. Adicionalmente, de dichos afectados, al menos en USA existe un alto sesgo hacia personas de ciertos grupos socioeconómicos y étnicos, usualmente de menor poder adquisitivo y socialmente más vulnerables por lo que el acceso a tratamiento y su posterior adherencia tienden a ser más precarios.

1.2.2 Definición del problema

Enmarcado en el proyecto VIH, en el que nace este trabajo de memoria, el cual consistirá en el estudio de una de las aristas que contempla el predictor global: el consumo de

alcohol y otras drogas, de tal manera de poder entender si existe una relación entre estas y modificaciones de la adherencia al tratamiento de VIH. Para ello se utilizarán tanto variables seleccionadas de la ficha clínica del paciente como variables que guarden directa relación con el consumo de drogas. La hipótesis de donde nace el trabajo es: Existe una relación que modifique la adherencia al tratamiento TAR y los pacientes VIH+ que consumen alcohol y otras drogas.

1.2.3 Enfoque de solución

Al realizar el diseño de predictor se tiene en cuenta el observar al paciente como un todo. Se busca un modelo que relaciona la integración de los datos clínicos de un paciente con los datos de su entorno, así se logra a la vez ver el paciente de forma más completa y entregar una atención médica según las características individuales de estos.

En base a esto y, teniendo en cuenta que el trabajo se plantea en un escenario de salud pública en donde no se cuenta con las facilidades para tener datos genómicos, se plantea la recolección de una variedad de datos (obtenidos de la Fundación Arriarán) que se agrupan en 3 grandes categorías:

- Datos relacionados con el paciente (clínica, laboratorio, farmacológicos y psicológicos).
- Datos socioeconómicos (datos demográficos, educacionales, laborales, etc.).
- Datos del equipo asistencial (servicios, programas educativos, campañas, etc.).

Varios predictores serán desarrollados a partir de esta información. Dentro de este, el trabajo de esta memoria se enfocará en el uso de datos sobre consumo de alcohol y otras drogas, de tal manera de poder entender si existe una relación entre estas y modificaciones de la adherencia al tratamiento de VIH. Para ello se utilizarán tanto variables seleccionadas de la ficha clínica del paciente como variables que guarden directa relación con el consumo de drogas. Por último, en el estudio se busca ver una componente temporal, como se ve modificada la adherencia en base al consumo de alcohol y otras drogas.

2. Marco Conceptual

2.1 VIH y Tratamiento Antirretroviral (TAR)

Se realizará un estudio sobre el impacto del consumo de alcohol y otras drogas en la adherencia al tratamiento antirretroviral, para ello es importante entender qué significa y cómo actúan el retrovirus y el tratamiento TAR.

Para comprender cómo funciona el tratamiento antirretroviral se debe entender primero cómo opera el VIH. Este es un retrovirus humano, el cual tiene la capacidad de infectar y replicarse en una variedad de células del sistema inmune, en especial los linfocitos CD4. EL VIH se transmite de las siguientes formas [1]:

1. Por relaciones sexuales sin preservativo de tipo barrera.
2. Contacto entre piel no intacta y mucosas a sangre contaminada. Dentro de esta categoría cabe el intercambio de jeringas y agujas entre personas con VIH positivo.
3. Vía placentaria cuando el bebe se está gestando o a través de la lactancia materna cuando la madre es portadora.

Al infectarse el individuo con VIH, lo primero que ocurre es una rápida replicación del virus en el organismo, lo cual genera la muerte de células CD4. Posteriormente a ello el organismo comienza a generar anticuerpos para controlar el virus (este proceso tiene una duración aproximada de 3 a 6 semanas después de la infección) [1].

Suele existir un periodo de ventana entre la infección y la aparición de los síntomas de esta, donde el organismo aún no genera suficiente cantidad de anticuerpos como para detectar el virus a través de pruebas de laboratorio. El siguiente paso, el de portador asintomático, se caracteriza por la replicación viral en aproximadamente 1010 viriones por días y una reducción de los linfocitos en alrededor de 50 a 70 células por mm al año [1].

Cuando la concentración de linfocitos CD4 disminuye a cifras menores de 200 mm3 se dice que el paciente padece SIDA [1].

El tratamiento antirretroviral actúa impidiendo la replicación del VIH dentro del organismo, bajando la carga viral, lo cual permite mantener niveles adecuados de linfocitos CD4 [1]. Para esto se utilizan distintos grupos de medicamentos, los cuales varían en cantidades y dosis dependiendo del paciente. Típicamente estos medicamentos se clasifican en 7 grupos: inhibidores de la transcriptasa inversa análogos de los nucleósidos (ITIN), los inhibidores de la transcriptasa inversa no análogos de los nucleósidos (ITINN), los inhibidores de la proteasa (IP), los inhibidores de la fusión, los antagonistas de CCR5,

inhibidores de posfijación, y los inhibidores de la transferencia de cadenas de la integrasa [7].

Es importante destacar que una correcta adhesión del tratamiento puede significar:

- Menores complicaciones en su tratamiento gracias a una disminución de la mortalidad y la morbilidad [14].
- Menor gasto público en salud gracias a menores hospitalizaciones dadas por infecciones oportunistas [15].
- Menores probabilidades de generar resistencia al tratamiento [18].

2.2 Consumo de drogas y VIH

El consumo de drogas intravenosas es uno de los responsables del aumento en los casos de VIH alrededor del mundo, especialmente en Sudamérica, Europa y el suroeste de Asia, esto debido al uso habitual de compartir jeringas en este contexto [19].

Se han visto estudios que ligan el consumo de alcohol y drogas con una menor adherencia al tratamiento TAR en personas con VIH+[4] Algunas de las variables utilizadas en estos son tanto la cantidad como la frecuencia con que se consume, se suelen usar indicadores como *Addiction Severity Index* y se segmenta el nivel de consumo en las categorías “no consume”, “consumo moderado” y “caso de riesgo” [16], también se suelen usar indicadores que permitan ver problemas dados por el consumo, un test de interés para ello es el *alcohol use disorders identification tests* o AUDIT donde se realizan preguntas como: “¿Qué tan seguido durante el último año te has dado cuenta de que no puedes dejar de tomar alcohol una vez que comenzaste?”. Además, en este estudio, que es de nivel general y no centrado específicamente en consumo de alcohol y otras drogas se ven otras variables relacionadas a demografía, depresión, entre otras [20].

Por otra parte, se ha visto que existe otro daño colateral del consumo habitual de drogas en el tratamiento de VIH, pues, genera una baja capacidad de tolerar el tratamiento en pacientes que han desarrollado hepatotoxicidad (daño funcional o anatómico del hígado) por el consumo [21].

2.3 Metodología KDD

Por otra parte, el trabajo realizado se dará en base al uso de técnicas de *data mining*, bajo el marco de KDD. La metodología de *Knowledge Discovery in Databases* (KDD)

busca a través de un proceso iterativo el descubrimiento de patrones útiles en la data. Los pasos principales que conlleva son [22]:

- Selección: En donde se genera un conjunto de variables o de las observaciones de la data original.
- Preprocesamiento: Se busca “limpiar la base de datos”, lo cual hace referencia al tratamiento del “ruido” de la base y a decidir estrategias para tratar con datos faltantes.
- Transformación: Se busca en este paso poder realizar transformaciones a la data que permitan normalizar los datos y poder tenerlos en un formato que sirva para ser utilizados como entrada en los algoritmos de *machine learning* (ML).
- *Data mining*: Corresponde a uno de los pasos centrales de KDD, aquí se busca primero acordar cuál es la meta del modelo a generar, ya sea clasificar, realizar *clusters*, o una regresión entre otras. Para luego en base a ello desarrollar un algoritmo que se ajuste y entregue resultados.
- Interpretación/ evaluación: La finalidad de este paso es lograr extraer conocimiento del modelo, gracias a la interpretación de los resultados, normalmente juega un rol en ello el generar visualizaciones y métricas. Puede ser necesario volver a realizar de forma iterativa pasos anteriores del proceso KDD para obtener un mejor resultado.

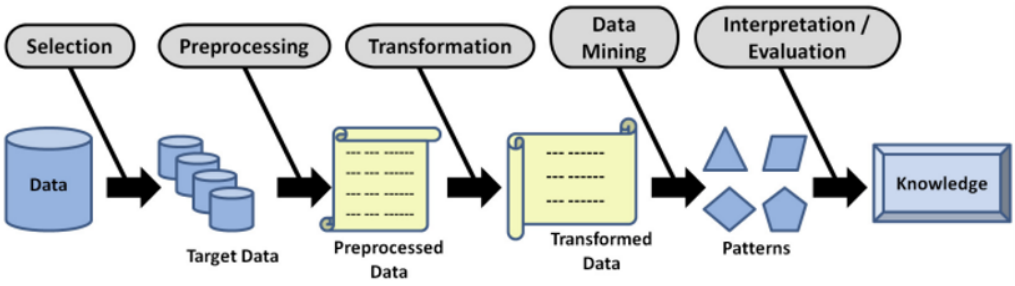

Figura 3. Proceso de The Knowledge Discovery in Databases (KDD) [22].

El método de KDD es un iterativo e interactivo, aunque posee una estructura clara. Suele considerarse como un modelo más preciso y completo con respecto a alternativas como CRISP-DM y SENMA, las cuales son más orientadas a trabajo de empresa [23] bajo estas consideraciones se prevé como una buena metodología de trabajo.

2.4 Algoritmos de aprendizaje supervisados

2.4.1 Regresión logística

La regresión logística corresponde a un método, donde se estima la probabilidad de que un evento ocurra, siendo el evento una variable categórica. El modelo entrega como output valores entre 0 y 1, que corresponde a la probabilidad de pertenecer a la categoría.

La relación que existe entre variables dependientes e independientes no es lineal, por ello se utilizan algoritmos iterativos para estimar los parámetros de la regresión [24].

2.4.2 Arboles de decisión

Los árboles de decisión son un método de aprendizaje supervisado que puede ser usado tanto para clasificación como para regresiones. gráficamente se puede representar los árboles como un conjunto de nodos de decisión, ramas y hojas.

El método genera una segmentación de las data en *subsets* al pasar por nodos de decisión, que actúan como filtros lógicos, los cuales determinan las categorías en las cuales son etiquetadas las observaciones. Cada nodo puede tener una o más ramas que representan distintas categorías en que se puede asignar la data. Al llegar al nodo terminal se obtiene la mejor clasificación (hojas) [25].

Una de las desventajas de este modelo es que puede crear sesgo en el resultado cuando existe una clase dominante en el set de entrenamiento.

2.4.3 *Random Forest*

Algoritmo que mezcla el concepto de *bagging* con árboles de decisión. *Bagging* es una técnica de reducción de la varianza de una función de predicción, en base a ello se construye el modelo de *Random Forrest* como una colección de árboles de decisión con la misma distribución de probabilidad para cada árbol [26]. En base a lo anterior este modelo entrega beneficios como: disminución en la varianza y robustecer el poder predictivo en comparación a árboles de decisión.

2.4.4 *Support Vector Machine* (SVM)

Es un algoritmo de aprendizaje supervisado. Este puede ser utilizado en casos de clasificación y regresión. El modelo está basado en la búsqueda de un hiperplano, dado por el cual se busca que se separe la data en categorías según estén a un lado o el otro de él, bajo el criterio de máxima distancia o margen [27].

2.4.4 Extreme gradient boosting (*XGBoost)*

Algoritmo de aprendizaje integrado que usa aumento de gradiente (*gradient boosting*). Esto quiere decir que está basado en la generación de múltiples modelos secuenciales de predicción específicamente de algoritmos de árboles de decisión (*decisión tree*). Este es un método reconocido por su efectividad en machine learning dada la facilidad que

entrega en cambios de parámetros como la función de perdida, el procesamiento de datos faltantes entre otras [45].

2.5 *Grid Search*

Es un proceso de optimización de algoritmos, trabaja haciendo una búsqueda en *subset* de hiperparámetros con la finalidad de encontrar los mejores resultados posibles dentro de estos. Dado lo anterior los parámetros dentro de los cuales se busca el mejor resultado deben variar según sea el algoritmo de aprendizaje supervizado que se está evaluando [46].

2.6 Evaluación del modelo de clasificación

La matriz de confusión es una forma gráfica de representar el desempeño de un modelo predictivo gracias al uso de los siguientes valores [28]:

- Verdadero positivo (TP): Que corresponde al número de instancias en que se predice o clasifica correctamente una observación en la categoría designada como positiva.
- Verdadero negativo (TN): Corresponde al número de instancias en que se clasifica la observación como negativo y realmente lo era.
- Falso negativo (FN): Corresponde a las instancias clasificadas como negativas cuando la observación es realmente de la categoría positiva.
- Falso Positivo (FP): Corresponde al número de instancias que son clasificadas positivas por el modelo cuando en verdad se ve que las observaciones son negativas.

Tabla 1. Matriz de confusión. Elaboración propia.

		Observación	
		Positiva	**Negativa**
Predicción	**Predice Positivo**	TP	FP
	Predice Negativo	FN	TN

En base a los valores de la matriz de confusión se pueden calcular las siguientes métricas:

- *Accuracy*: Métrica que identifica la precisión del modelo al medir las instancias en que el modelo predice correctamente.

- *Precision:* Corresponde al cociente de instancias clasificadas correctamente como positivas, por tanto, sirve como un indicador de la sensibilidad del modelo para identificar los casos verdaderamente positivos de aquellos clasificados como positivos.
- *Recall:* Corresponde al cociente de instancias clasificadas correctamente en base al total de casos de interés, por tanto, muestra cuantas instancias son clasificadas positivas de todas aquellas que verdaderamente son positivas.

3. Objetivos y Alcances

3.1 Objetivo general

De la hipótesis central: “¿Existe una relación entre la adherencia del tratamiento TAR y los pacientes con VIH positivos que consumen alcohol y otras drogas?” se desprende el trabajo de memoria, con el objetivo general: ***Desarrollar un modelo predictivo de adherencia al tratamiento TAR con relación al consumo de alcohol y drogas para encontrar pacientes de riesgo de tratamiento***

3.2 Objetivos específicos

1. Analizar el estado del arte sobre estudios relacionados a VIH y correlaciones con consumo de alcohol y otras drogas, con la finalidad de entender metodologías de interés y variables usuales de estudio en casos similares.
2. Determinar nuevas variables a estudiar, a las ya establecidas desde la fundación, según el estudio bibliográfico.
3. Seleccionar, implementar y evaluar modelos para la elaboración de un predictor de adherencia de tratamiento de VIH con relación a consumo de alcohol y otras drogas
4. Elaborar un prototipo donde se visualicen los resultados.
5. Evaluar el impacto social y económico del proyecto.

3.3 Alcances

Como se ha explicado anteriormente el proyecto elaborado por el WIC comprende la construcción de varios predictores, donde cada uno toma datos basados en un área de las tres grandes bases (Base clínica, socioeconómica y del equipo asistencial) con información obtenida de la Fundación Arriarán. El trabajo de memoria se centra específicamente en el uso de datos clínicos y de consumo de alcohol y otras drogas, y por tanto no se focalizará en información socioeconómica, depresión o farmacología de forma directa. El trabajo desarrollado se dará con respecto a bases de datos entregados por la *medical advisor* y los ingenieros del WIC, por lo cual no se abordará el proceso de recopilación de información ni el preprocesamiento inicial para obtener los datos en el formato adecuado de trabajo.

Sin embargo, sí es parte del trabajo desarrollado el consolidar la base de “datos de alcohol y otras drogas” como tal, en base a las variables determinadas por la *medical*

advisor y la fundación Arriarán, y el tratamiento de esta para poder desarrollar los predictores de la mejor forma. Se tiene como alcance del proyecto la elaboración de un modelo predictivo en base a algoritmos de aprendizaje supervisado y un prototipo de visualización de los resultados de este, no se contempla la generación de una plataforma funcional de uso. Sin embargo, se espera que los resultados ayuden a elaborar recomendaciones para casos de riesgo de adherencia. Por último, también se ve como alcance del proyecto el medir el impacto mismo de este.

Debido a las posibles diferencias de comportamiento en base a las circunstancias dadas en 2020 y 2021 por la pandemia COVID-19 se trabaja con datos hasta el año 2019 de la fundación.

El WIC ha gestionado la implementación de un formulario, en cual específicamente, de interés para este trabajo de memoria, se encuentran preguntas de nivel de consumo de drogas, capacidad de consumir un nivel deseado para la persona y su tendencia a utilizar el consumo de drogas como un distractor en su vida personal.

Para la fecha de termino de trabajo de memoria no se contaba con los datos de la encuesta por lo cual no forman parte del análisis de la investigación actual. Dichos resultados, su análisis e incorporación de conclusiones se dejan como parte de los trabajos futuros que este trabajo de memoria entrega al WIC para futuros memoristas y ayudantes de investigación.

4. Metodología

4.1 Análisis bibliográfico y selección de variables

4.1.1 Estudio bibliográfico y análisis de datos de la Fundación Arriarán

Para la selección de variables a utilizar en el predictor de alcohol y otras drogas se comienza por la recopilación bibliográfica de estudios de adherencia en pacientes VIH+ que tengan relación al consumo de alcohol u otras drogas. Esta revisión incluye la búsqueda en motores clásicos (Por ejemplo: SCOPUS, Google Scholar) junto con bibliografía entregada tanto desde el grupo de trabajo del WIC como desde la propia Fundación Arriarán.

En base a la bibliografía recopilada, se genera un cuadro resumen con las principales conclusiones y variables utilizadas para el estudio. Realizada la comparación entre las distintas fuentes de información y seleccionadas las variables en común que se observan entre estas, se realiza la corroboración con la *medical advisor* quien entrega un *insight* de las variables observadas y seleccionadas, además de su recomendación de otras variables de interés que se pueden adicionar [Observar Resultados, Ítem 5.1.1, Tabla 2].

El siguiente paso corresponde al análisis de las tablas presentes de la Fundación Arriarán. Se genera una primera inspección de las tablas seleccionando aquellas que cuentan con información relevante y una cantidad de observaciones considerables. En base a este primer acercamiento a las tablas se genera una selección de las tablas a utilizar para el trabajo posterior y las variables de elección de estas. Este proceso de elección de variables es iterado con la *medical advisor.*

4.1.2 Construcción de la base de datos de alcohol y otras drogas

Para la construcción de las variables se comienza por una revisión de cada una de las tablas previamente seleccionadas para el trabajo. En base a ellas se inicia la construcción de la tabla de alcohol y otras drogas. Se evalúa el número de pacientes y observaciones disponibles, posterior a esto se evalúa las variables que se posee y el número de datos faltantes dentro de ellas. Además de la elección de variables, se evalúa en cada tabla la posibilidad de construcción de nuevas variables de interés según lo evaluado en la sección previa (Resultados, ver Ítem 5.1.2, Tabla 3,4,5 y 6).

Paralelamente al trabajo de construcción de la tabla de “alcohol y otras drogas” se realiza el estudio de la tabla master ficha clínica, tabla en común para todos los predictores, la cuál es elaborada por el equipo del WIC y replicada en el trabajo de memoria con la

finalidad de entender a profundidad la creación de variables y la construcción de la misma. (En Resultados, ver en Ítem 5.1.2, Tabla 7).

El estudio de la tabla master ficha clínica va de la mano de la construcción del *label,* “Fallo virológico”. Esto se explicará a profundidad en el capítulo de resultados (En Resultados, ver Ítem 5.1.2.1.2).

Un estudio acabado de la tabla lleva a la observación de pacientes con datos faltantes en rna y cd4 sin estar en situación de abandono de tratamiento, este hecho lleva a tomar distintas decisiones sobre la tabla ficha dadas por el criterio de la *medical advisor* (En Resultados, ver Ítem 5.1.2.1.2).

Para la construcción de la tabla de alcohol y otras drogas no solo se toma en consideración la tabla master ficha clínica, sino que al ser este trabajo parte de un proyecto mayor con múltiples predictores, se consideran de igual forma los demás predictores para la toma de decisiones de pacientes con los cuales hacer los trabajos futuros y, por ende, seleccionar dentro de las tablas.

La tabla master ficha clínica y las tablas de data específica de alcohol y otras drogas son unidas a través de los códigos de los pacientes (llave de paciente, única para cada uno y anonimizada) y el año de la observación. Posterior a esto, se genera una separación de la data según el tipo de variables (categoriza y numéricas). Aquellas categóricas son transformadas a variables *dummy*, lo cual implica la generación de distintas columnas de valores binarios según las categorías de las variables. Por otra parte, las variables numéricas son normalizadas en base a standarscaler. Mayor detalle de estos cambios y el trabajo sobre las tablas se encuentra en la sección de resultados (En Resultados, ver Ítem 5.1.2.2 y 5.2).

4.2 Construcción de modelo

Dentro de la parte de construcción de modelo se decide probar distintos modelos de aprendizaje supervisados y realizar una evaluación posterior en base su desempeño en métricas previamente elegidas. Esto con la finalidad de elegir el modelo para el trabajo posterior de clasificación. Los modelos elegidos son: Logit, Random Forest, decision tree, XGBoost, SVM y Naive Bayes, los cuales fueron detallados en la sección 2.4 y utilizados en Resultados 5.3.1. Los resultados de los modelos en base a las métricas: *recall, precisión, accuracy, log loss* detallados en sección 2.5 *y* utilizados a lo largo de la sección de Resultados 5.3.1

Los datos son divididos entre los *sets* de entrenamiento y testeo con una proporción 75-25.

Para la construcción del modelo se toma en consideración el hecho de estar trabajando con distintas observaciones de un mismo paciente, por lo cual se toman precauciones para la selección de observaciones de un mismo paciente se mantengan en uno de los dos sets (Ver Resultados 5.3.1).

Para la obtención de mejores resultados se decide la utilización de optimización de hiper-parámetros de los modelos, para esto se utiliza *grid search*, una función incluida en las bibliotecas de Python utilizadas, la cual comprueba una grilla de hiper-parámetros, distinta para cada modelo, en busca de la mejor combinación para los datos a utilizar. Mayor detalle de esto se menciona en la sección de Resultados 5.3.1.

Finalmente, los modelos son evaluados bajo las métricas seleccionadas y se realiza una elección del mejor modelo, además de las métricas se utilizan las matrices de confusión que nos dan de una forma rápida la proporción que se está obteniendo de verdaderos positivos y verdaderos negativos. Dada la elección del modelo, se procede a la categorización de adherencia (Ver Resultados 5.3.2). La probabilidad de obtener un fallo es dividida en 5 categorías en base a las probabilidades obtenidas del modelo de que el paciente pueda o no presentar un fallo virológico en aquel año. Posterior a la ejecución del modelo y la categorización de adherencia tomando el *set* de testeo se realiza una exploración de las variables según los datos obtenidos en cada una de las categorías (Ver Resultados 5.3.2).

4.3 Elaboración de prototipo de visualización

Para la construcción del prototipo de solución se tiene en consideración los resultados del modelo. En base a los resultados de la predicción se generan gráficos según algunas variables de interés. Se utiliza Power Bi como herramienta para la construcción de *dashboards* y se realiza un *mock up* de la visualización posible en formato de página web (Ver Resultados 5.4).

La validación del *mock up* se realiza en base a la opinión de usuario, según la cual se realizan cambios pertinentes para una mejor funcionalidad (Ver Resultados 5.4).

4.4 Evaluación de impacto social y económico del proyecto

Para la evaluación del impacto social y económico se plantean las distintas aristas en que la adherencia impacta no tan solo al individuo sino también a la sociedad. Por otra parte, para ver el punto de vista económico se realiza una proyección nacional a 5 años de la población con VIH positivo y se plantean los posibles costos en que se incurriría y los

beneficios del proyecto en torno a este. Los datos son tomados de la ONUSIDA además de plantear ciertas suposiciones en base a la información disponible, estas son detalladas en el avance del análisis (Ver Resultados 5.5).

Por otra parte, se toman los resultados del predictor y la cantidad de pacientes predicho en cada categoría de adherencia para evaluar los costos en base a estos.

5.Resultados

Como fue mencionado anteriormente el trabajo de memoria se enmarca en un proyecto FONDEF donde se plantea la generación de cuatro sub predictores, siendo el relacionado con datos de alcohol y otras drogas en el que se basa la presente memoria. Dicho subpredictor no solo funciona como un predictor independiente, sino que también, forma parte de una plataforma que funciona en base a los cuatro sub predictores mencionados anteriormente (Ver ítem 1.2.1, Contexto). El proyecto busca la predicción de la adherencia en un paciente al tratamiento TAR. La adherencia se mide para cada paciente de forma anual, esto permite obtener información a corto plazo requerida para la toma de decisiones necesarias para un modelo intervencional médico ajustado al paciente, lo cual permite actuar donde, en base a los datos de cada subpredictor, haya que intervenir para evitar el abandono al tratamiento, una adherencia baja e incluso la muerte. Es en base a esto que, el proyecto FONDEF mencionado presenta un enfoque de medicina personalizada, donde cada sub predictor está enfocado de la misma forma.

Se espera que cuando un paciente nuevo llegue a la fundación, se colocarán los datos en la plataforma y esta arrojará, a través de los valores de adherencia, donde se encuentran los posibles factores (variables) que influirán en una adherencia baja, por ejemplo, para la construcción de la base de datos del modelo se utilizan datos en retrospectiva, que incluyan la actividad de los pacientes de la Fundación desde su inicio de actividad en la misma.

5.1 Recolección y consolidación de datos

5.1.1 Análisis bibliográfico y selección de variables en relación con el consumo de alcohol y otras drogas de los pacientes y su efecto en adherencia al tratamiento TAR.

Con la finalidad de conseguir variables que permitan relacionar los datos de los pacientes con su adherencia al tratamiento, se busca en la bibliografía variables utilizadas por autores previos que puedan ayudar con la rectificación de las variables seleccionadas de los datos provistos por la Fundación Arriarán y, por otra parte, en la selección de algunas variables nuevas relacionadas con consumo de alcohol y drogas, sus causas y consecuencias con relación a pacientes VIH+. En la Tabla 2 se presentan las publicaciones más relevantes encontradas, junto con variables extraídas y algunas de las principales conclusiones.

Tabla 2. Bibliografía con información relevante para obtención de variables a utilizar en el modelo de adherencia.

Publicación	**Variables obtenidas**	**Conclusión principal**	**Ref.**
Samet et al. (2004)	Nivel de consumo de alcohol, género, edad, raza, soporte social percibido de amigos, situación de calle, síntomas depresivos, consumo de heroína o cocaína, nivel de CD4, niveles de log VIH RNA (medida usada para describir VIH y expresar el valor de la carga viral como una potencia de diez, lo cual permite ver de forma rápida el número de copias por milímetros), número de medicamentos consumidos por día.	Cualquier nivel de consumo de alcohol en pacientes con VIH está asociado a una baja adherencia al tratamiento.	[16]
Arnsten et al. (2002)	Consumo activo de heroína, cocaína, promedio de consumo de alcohol(categorías), uso de alcohol u droga como *coping mechanism*(una estrategia de adaptación para controlar estrés). situación habitacional, beneficios sociales, depresión y tratamientos médicos.	El uso activo de cocaína está asociado a un 41% de baja en adherencia al tratamiento. No obstante, no existe una diferencia estadísticamente significativa entre niveles de consumo. Se ve una asociación entre el consumo de sustancias como *coping mechanism* y la baja adherencia al tratamiento"	[17]
Malow et al. (2013)	Nivel de consumo de alcohol, capacidad de dejar de consumir, variables demográficas y depresión	Existe una relación vista entre consumo de alcohol y depresión y se ha visto que se puede ligar esta última a adherencia de tratamiento TAR en personas VIH+.	[20]
De Boni et al. (2018)	Variable categórica de nivel de consumo de alcohol, variable continua de ingesta de alcohol, sexo, edad, país de residencia, estado de VIH previo al tratamiento, tiempo que lleva en tratamiento, contacto sexual categorizado por orientación, consumo de marihuana (si o no), consumo de cocaína (si o no), consumo de crack (si o no) y consumo de alguna droga no	El uso de alcohol y drogas no intravenosas en un periodo previo de 7 días de la encuesta incrementa el riesgo de LTFU (*Lost to follow up* o pacientes perdidos durante el seguimiento) durante los próximos 18 meses	[29]

	intravenosa (si o no)		

Como resultado del estudio bibliográfico, resumido en la Tabla 2, se puede concluir que las investigaciones demuestran que existe una relación entre el consumo y la adherencia al tratamiento en pacientes VIH+. Además, se observan variables en común entre los distintos estudios, agrupados de la siguiente manera:

- Caracterización del paciente: género, edad, raza, soporte social percibido de amigos, situación habitacional, beneficios sociales, síntomas depresivos.
- Datos médicos: nivel de CD4, log de RNA, número de medicamentos consumidos por día, estado de VIH previo al tratamiento, tiempo en el tratamiento.
- Niveles de consumo: Promedio de consumo de alcohol (mediciones categóricas y continuas), uso de alcohol u otras drogas como mecanismo de sobrevivencia.

5.1.2 Selección de variables de interés de las fuentes de información (tablas) de la Fundación Arriarán para construir la tabla "Alcohol y otras drogas"

5.1.2.1 Exploración tablas

La información disponible para el caso de estudio se divide en dos fuentes primordiales:

i) Información de la Fundación Arriarán como tal, extraída de la atención al paciente

ii) Información obtenida del área de Farmacia de la Fundación. La data se encuentra en formato *.csv* y es proporcionada por los ingenieros del WIC.

Los nombres de las tablas y los contenidos de estas se encuentran en ingles ya que la Fundación utiliza un protocolo estándar para transferencia de datos llamado "CCASANET *data transfer protocol*" [30].

5.1.2.1.1 Construcción de la tabla final de alcohol y otras drogas

i) Información de la Fundación Arriarán como tal.

La información obtenida de la "Fundación Arriarán" está conformada por 20 tablas, de las cuales 12 contienen información actualizada (más una tabla de diccionario), mencionadas en la Tabla 3:

Tabla 3. Tablas provenientes de “Fundación Arriarán”.

Tabla	Definición de la información
TBL_ART_Auditoria	Contiene los tratamientos asignados a pacientes de la Fundación
TBL_HOSPITAL	Contiene datos de hospitalizaciones de los pacientes
TBL_LAB_CD4	Contiene mediciones de CD4 de pacientes en el tiempo
TBL_LAB_RNA	Contiene mediciones de RNA de los pacientes a través del tiempo
TBL_USER	No contiene información de valor para la investigación
TBLBASIC	Contiene información de caracterización del paciente
TBLCE	Contiene *clinical endpoints* acerca de infecciones oportunistas
TBLCE_CANCER	Contiene información acerca de pacientes con cáncer
TBLCE_TB	Contiene información acerca de los pacientes con tuberculosis
TBLDIS	Contiene información acerca de pacientes con tuberculosis
TBLFOLLOW	Contiene información del paciente acerca de si esta vivo o muerto, si es que se trasladó o no y si es que sigue con su tratamiento o lo abandono
TBLVISIT	Contiene fecha de cada vez que un paciente va presencialmente a la Fundación
TBMASTER	Contiene diccionario de códigos usados en las demás tablas

Una vez identificadas las tablas con las cuales se cuenta, se realiza un primer análisis exploratorio de estas, identificando las variables contenidas en cada una y la cantidad de datos. Tomando en consideración tanto la investigación bibliográfica como la orientación de la *medical advisor*, se obtienen las siguientes tablas a trabajar con las variables especificadas, resumidas en la Tabla 4:

Tabla 4. Variables seleccionadas de “Fundación Arriarán”.

Nombre Tabla	Variables tabla	Descripción variable
TBLBASIC	1. patient	Código único de paciente
	2. education_description	Descripción de nivel de educación

	3. employment_description	Variable categórica de forma de trabajo, Ej: “activo”, “pasivo”, “cesante”
	4. employment_status	Descripción del trabajo del paciente
	5. education	Variable categórica de educación según nivel completa/incompleta
	6. condusexual	Conducta sexual del paciente
TBLCE	1. patient	Código único de paciente
	2. ce_id	Tipo enfermedad oportunista
	3. local_oth	Enfermedad oportunista
TBLDIS	1. patient	Código único de paciente
	2. dis_oth	Enfermedad oportunista

Se debe considerar que la tabla obtenida del trabajo será cruzada con la tabla denominada “ficha clínica máster” elaborada por parte del equipo del WIC, y que contiene las variables significativas que se comparan con los otros tres sub predictores mencionados anteriormente, donde se analiza adherencia (Ver en Capitulo 1.2.1 Figura 3). Por ello se omiten variables que sean consideradas importantes para el estudio pero que se encuentran en “Ficha clínica master” dado que serán incluidas posteriormente.

En un comienzo se cuenta con alrededor de 6000 pacientes. Las distintas tablas, sin embargo, pueden contener un número distinto de pacientes en base a la información que poseen. Decisiones tomadas con respecto a la data, generan también la selección de ciertos pacientes para la confección de la tabla final, estas decisiones serán abordadas en los capítulos futuros.

A continuación, se da una descripción general del trabajo elaborado con cada una de las tablas pre-seleccionadas.

Tabla TBLBASIC

Esta tabla contiene, como se mencionó anteriormente, información de caracterización del paciente, dentro de ella existen múltiples variables que son consideradas en la “tabla ficha clínica master”, por ello en el trabajo realizado se habla por el momento de solo las variables seleccionadas que difieren de esta (ver Tabla 4). Inicialmente esta tabla posee

5923 pacientes y 40 variables de las cuales se consideran 6 de interés para el estudio (ver tabla 4)

Se eliminan las observaciones nulas de las variables 'ART_at_baseline' o 'enrol_d' o 'aids_y' de la tabla “TBLBASIC”, para guardar concordancia con la tabla ficha master, eliminando 357 observaciones, posterior a esto se preprocesa la tabla con las columnas seleccionadas de la tabla “TBLBASIC” para corroborar la razón de datos nulos dentro de las variables. Las variables “patient” y “CODIGO CONASIDA” son utilizadas para unir las tablas.

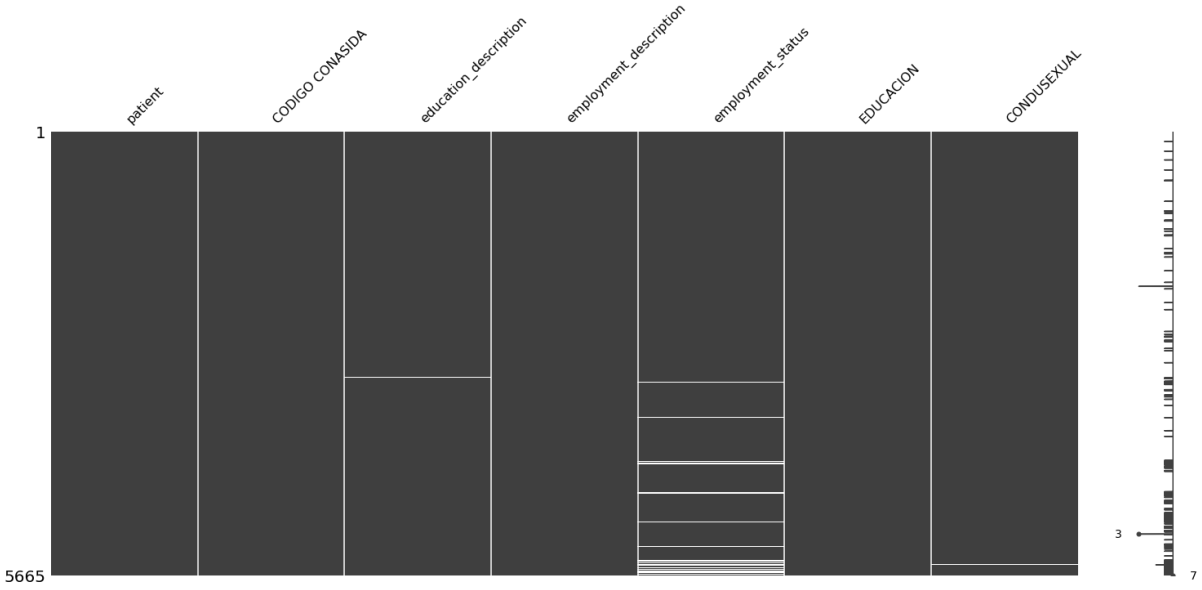

Figura 4. Gráfica de datos nulos por variables escogidas de la tabla TBLBASIC. Elaboración propia

Se genera una nueva variable designada como “nivel educacional” la cual se construye en base a la variable “education description”, la cual nos indica el nivel de educación, tomando los valores: básica, media, universitario y técnico, y “Educación” la cual muestra si el nivel de educación descrito se rindió de manera completa o incompleto. En la nueva variable, “nivel educacional” se mantiene el último nivel educacional completo que posee la persona. Por ejemplo, si la persona pose en “education description” media y en “EDUCACION” incompleta, la nueva variable toma el valor de “básica”. Posterior a la creación de la variable “nivel educacional” se borran las variables “Educación” y “education description”, por otra parte al evaluar la gran cantidad de opciones en la variable “employment_status”(la cual entrega una descripción de cuál es el trabajo del paciente, sin un formato estandarizado) se decide borrar esta de la base a trabajar.

Tabla TBLCE

Posteriormente se preprocesa la tabla TBLCE, en ella se corrobora la razón de datos nulos dentro de las variables. Esta tabla contiene 7 variables que describen las enfermedades oportunistas que posee la persona en el momento del registro, datos importantes para nuestro caso. Específicamente esta tabla posee 2938 pacientes y para

su estudio posterior se planea la unión de esta tabla con TBLDIS, proveniente también de Fundación Arriarán, para la generación de nuevas variables. Esto será explicado más adelante.

Se seleccionan las variables de interés: ce_id, que entrega un código de tipo de enfermedad oportunista y local_oth, la cual entrega información específica de la enfermedad, pero, no se encuentra con un formato estandarizado de ingreso de información. Además, se toman la variable “patient” para generar la unión posterior entre tablas (la selección se puede observar en la Figura 5)

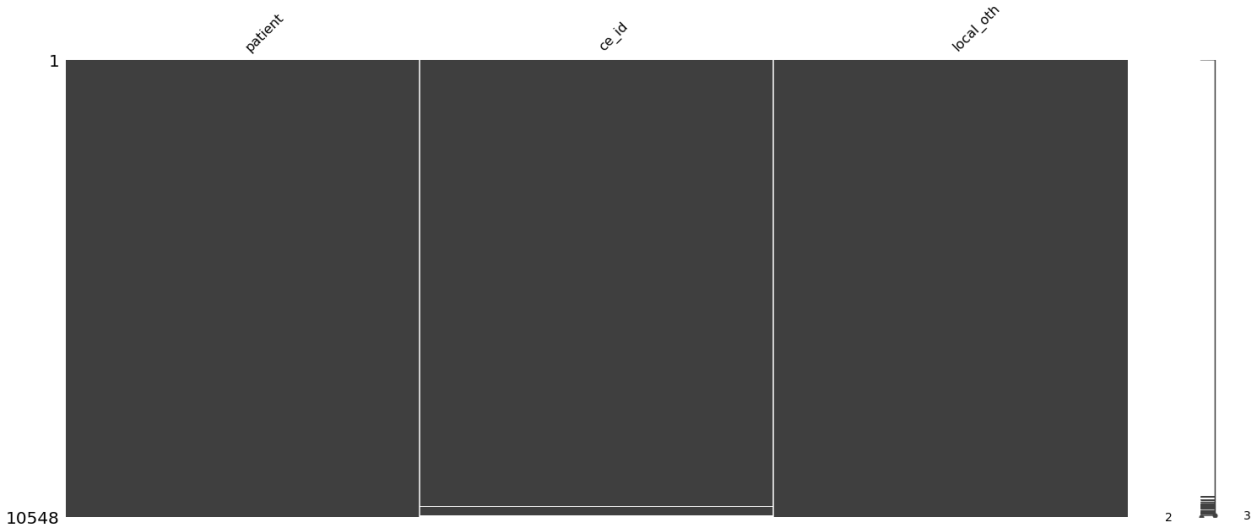

Figura 5. Gráfica de datos nulos por variables escogidas de la tabla TBLCE. Elaboración propia

Tabla TBLDIS

Se realiza el mismo procedimiento con la tabla TBLDIS para ver la razón de datos nulos. Esta tabla contiene información de 1267 pacientes y 9 variables, de las cuales se seleccionan 3 variables de interés, dos para el cruce entre tablas (“patient” y “year”) y dis_oth, que entrega un código de enfermedad oportunista. En la Figura 6 se puede observar las variables seleccionadas de la tabla. Específicamente esta tabla contiene información de 1268 pacientes.

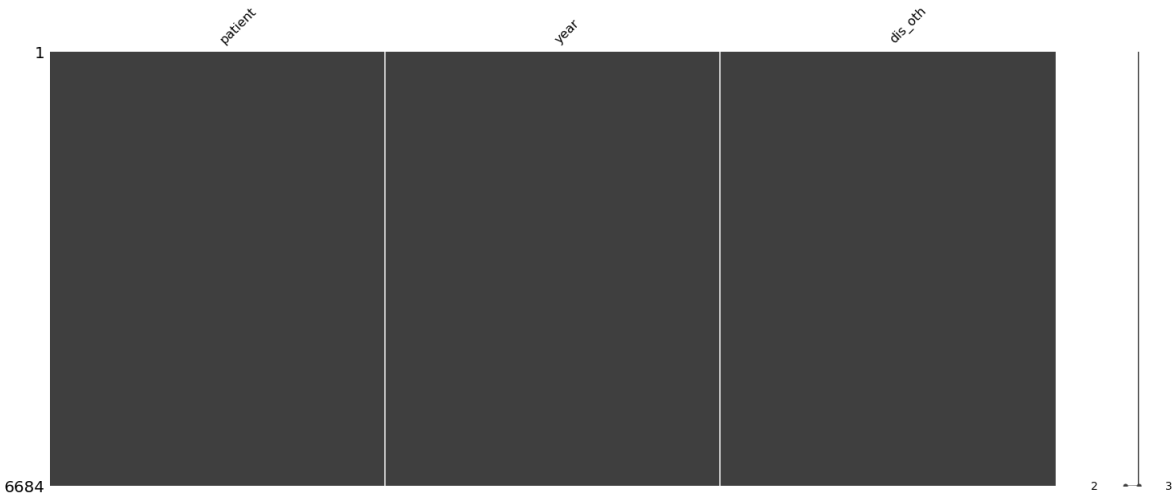

Figura 6. Gráfica de datos nulos por variables escogidas de la tabla TBDIS. Elaboración propia

Posteriormente a estos pasos, se realiza el cruce entre las 3 tablas descritas a través de la variable “patient” para generar la tabla provisoria llamada “FA”. Al realizarse el cruce entre las tres variables, se puede profundizar en el estudio de enfermedades oportunistas de los pacientes de la fundación, a través de las variables “ce_id”, “local_oth” y “Dis_oth”. Esto es útil ya que permite ver si hay enfermedades oportunistas que se relacionan con consumo de alcohol u otras drogas. Se genera así la variable “enfer_op” la cual contiene los valores de “ce_id” para las observaciones en que estas no es data faltante, y de caso contrario obtiene el valor de “local_oth”. De ser ambas un dato faltante, para la observación obtiene el valor de “dis_oth”.

En base a la variables “enfer_op” se crea la variable “enfer_op_consumo” la cual toma los valores: 0 si “enfer_op” es un dato faltante, 1 si “enfer_op” toma valores que caben dentro de las categorías de interés ('nade_renal', 'nade_liver', ‘IRC'(insuficiencia renal crónica), 'DHC'(daño hepático crónico), 'hepatitis por drogas', 'drug abuse'(abuso de drogas), 'cocaína y marihuana (+)', 'cocaína (+)', 'alcohol abuse'(abuso de alcohol), 'oh(+)'(alcohol), 'cocaína (+)/alcoholismo', 'insuficiencia renal', 'NEFROPATIA', 'insuficiencia renal', 'DROGADICCION', 'DRAGADICCION OH(+) (drogadicción y alcohol)', 'COCAINA (+)', 'ABUSO DE DROGAS', 'ABUSO OH'(abuso de alcohol), 'COCAINA (+)/ALCOHOLISMO', 'COCAINA Y MARIHUANA (+)”).

Como se puede observar de lo anterior, el que no haya un ingreso de data estandarizado para todas las variables con las que se genera “enfer_op_consumo” genera una dificultad en la identificación de categorías que entran dentro de las enfermedades de interés. Se debe recalcar que posteriormente en el trabajo se realiza una selección final de pacientes con los cuales se trabajarán los modelos, lo cual se explicara en los siguientes capítulos. Al realizar esto, para los pacientes seleccionados, los valores de esta variable entran en una sola categoría, al no aportar información adicional se decide borrar de las variables utilizadas para la base de “alcohol y otras drogas”

ii) Información obtenida del área de Farmacia de la Fundación

Por otra parte, se realiza un trabajo paralelo en las tablas proporcionadas del área de farmacia de la Fundación. Se cuenta con 50 tablas, de las cuales 16 se encuentran obsoletas (por lo tanto, no se consideran para el trabajo), 19 corresponden a diccionarios, 12 a tablas con información de pacientes y 3 a temas varios (dos tablas administrativas y una con información del paciente). En la Tabla 5 se observan 13 tablas disponibles para el trabajo, las 12 de información del paciente y una de la sección de temas varios.

Tabla 5. Tablas provenientes de “Farmacias”.

Tabla	Definición de la información
AdherenciaCalculo	Se entrega cálculo de adherencia del paciente en formato de “Días sin TAR”
Controles	Presenta control según la entrega de la receta y duración de esta misma
Recetas_Dis	Información en base a la entrega de recetas
Detalle	Entrega información de registro de las entregas de los medicamentos
Información Pacientes	Información básica personal de los pacientes
Observaciones	Observaciones sobre pacientes
Posologías	Indica dosis prescritas
Terapacias_p_op	Posología por remedios por tratamiento de enfermedades oportunistas
Terapacias_pac	Posología de remedios por tratamiento VIH
Validaciones	Registro ingreso de pacientes
Arriaran_Farmacias_mensajes	Información de derivación para complementar la situación de pacientes
Autorizaciones	Registro de terceras personas autorizadas por paciente para retiro de medicamentos
Situación	Situación actual del paciente

De las 13 tablas de Farmacia, mencionadas en la Tabla 6, se realiza un primer trabajo exploratorio de las tablas y las posibles variables de interés dentro de estas. Además, la cantidad y calidad de información de las mismas y se seleccionan 7 tablas de farmacia con las cuales trabajar.

Tabla 6. Variables seleccionadas provenientes de “Farmacias”.

Nombre Tabla	Variable Tabla	Descripción Variable
AdherenciaCalculo	dias_si_tarv	Días sin TAR por paciente
	fecha_de_calculo	Fecha en que se calcula la adherencia
Recetas_dis	c_minsal	Código único de paciente
	fecha_retiro	Fecha retiro de fármaco
	proxima_entrega	Próxima fecha de retiro de fármaco
	rp_oportunista	Categórica de si tiene enfermedad oportunista (Verdadero/Falso)
	retira_autor	Categórica si retira autor (Verdadero/Falso)
Detalle	cantidad	Cantidad de fármaco
	fecha	Fecha del registro
informaciones pacientes	genero	Género
	edad	Edad paciente
	region	Región
	comuna	Comuna
	hla	Antígenos
	previsión	Previsión de paciente
terapacias_p_op	c_minsal	Código único de paciente
	fecha	Fecha del registro
	OpoX	Fármaco X
	PslgX	Dosis asignada para fármaco X
terapias_pac	c_minsal	Código único de paciente
	OpoX	Fármaco X

	PslgX	Dosis asignada para fármaco X
autorizaciones	c_minsal	Código único de paciente
	fecha	Fecha de registro
	rut_autorizada	Rut persona autorizada para retiro de fármacos
	estado	Estado

En las tablas seleccionadas se corrobora la proporción de datos nulos presente, en base a ello se eliminan aquellas columnas en donde existe datos nulos en la columna “Minsal” o “c_minsal”, variable que sirve como llave para la unión entre tablas. Otro cambio importante que remarcar es que en el pre-procesamiento de los datos se transforman las variables al formato adecuado para su trabajo y se crea la variable “año” en cada tabla.

Tabla Terapias_p_op

La tabla presente contiene información de 2183 pacientes, cuenta con 31 variables y de estas se selecciona las variables necesarias para el cruce de tablas, tanto de código del paciente como de año de toma de registro.

Se genera en base a la variables “Opo1” hasta “Opo 4” (variables que declaran el medicamento a consumir para un paciente determinado, cada variable es un remedio distinto) y “Pslg1” a “Pslg4”(variables que declaran la posología, o dosis designada de medicamento para el fármaco correspondiente) dos nuevas variables, que funcionan como contadores, para catalogar la cantidad de pastillas consumidas por cada paciente, según cada posología, y la correspondiente cantidad de miligramos de cada medicamento. La cantidad de pastillas y miligramos es llevada a un consumo anual, tomando la última medicación designada para el año y traspasando esta cantidad al consumo anual del paciente.

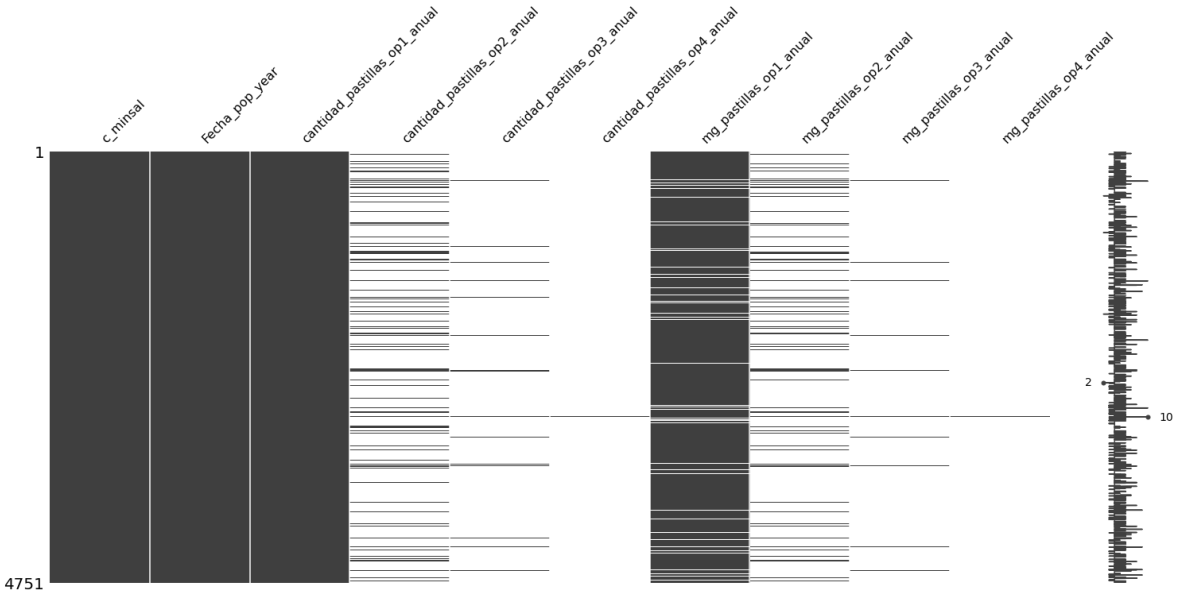

Figura 7. Gráfica de datos nulos por variables escogidas de la tabla Terapias_p_op. Elaboración propia

Tabla AdherenciaCalculo

La tabla presenta información de Días sin Tar para pacientes que han tenido problemas antes, sin embargo, por lo mismo no se encuentra la totalidad de los pacientes ni instancias de forma consistente. Se ha utilizado la variable días sin tar como una forma de control de cálculo, pero esta se ha construido finalmente en base a información presente en “tabla Recetas dis”

Tabla Recetas_dis

De la tabla recetas_dis se seleccionan las variables de interés que se observan en la Figura 8

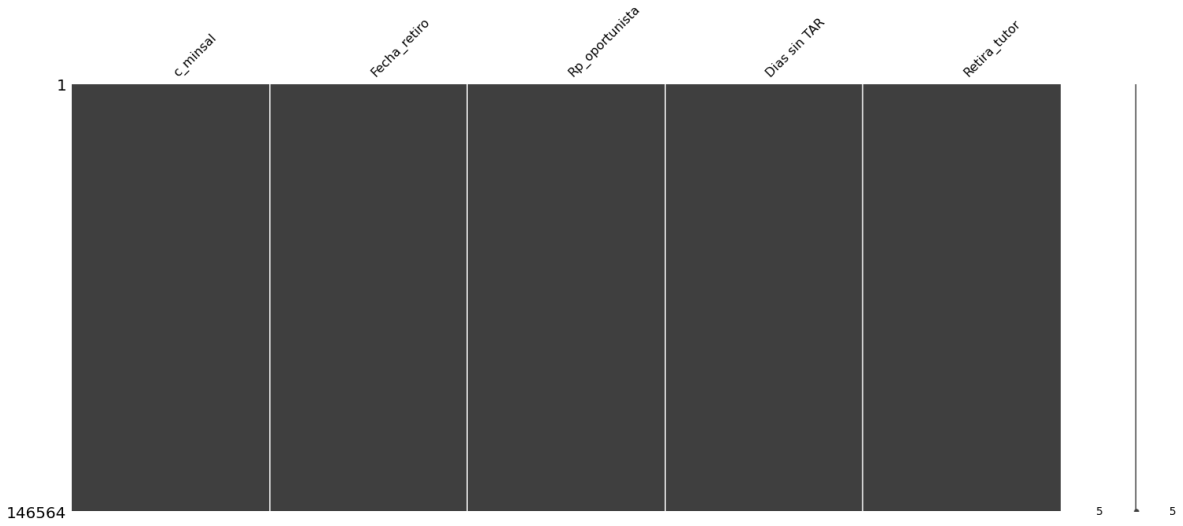

Figura 8. Gráfica de datos nulos por variables escogidas de la tabla Recetas_dis. Elaboración propia

Fecha de retiro y retira tutor nos dan un acercamiento a saber si el paciente tiene una red de apoyo cercana (al saber que tiene alguien que pueda retirar los remedios por él). Por otra parte, rp_oportunista nos da el conocimiento si el paciente posee o no enfermedad oportunista.

Finalmente, días sin TAR, se construye en base a las variables “próxima entrega” (Fecha que se le da al paciente para retirar la siguiente dosis de sus medicamentos TAR) y “Fecha_retiro” (Fecha en que finalmente se retira el medicamento).

Tabla Información pacientes

Originalmente la tabla cuenta con 6652 pacientes y 33 columnas entre las cuales presenta información general del paciente, como la localidad donde vive, nacionalidad, previsión, entre otras. Se seleccionan las variables que se pueden observar en Figura 9.

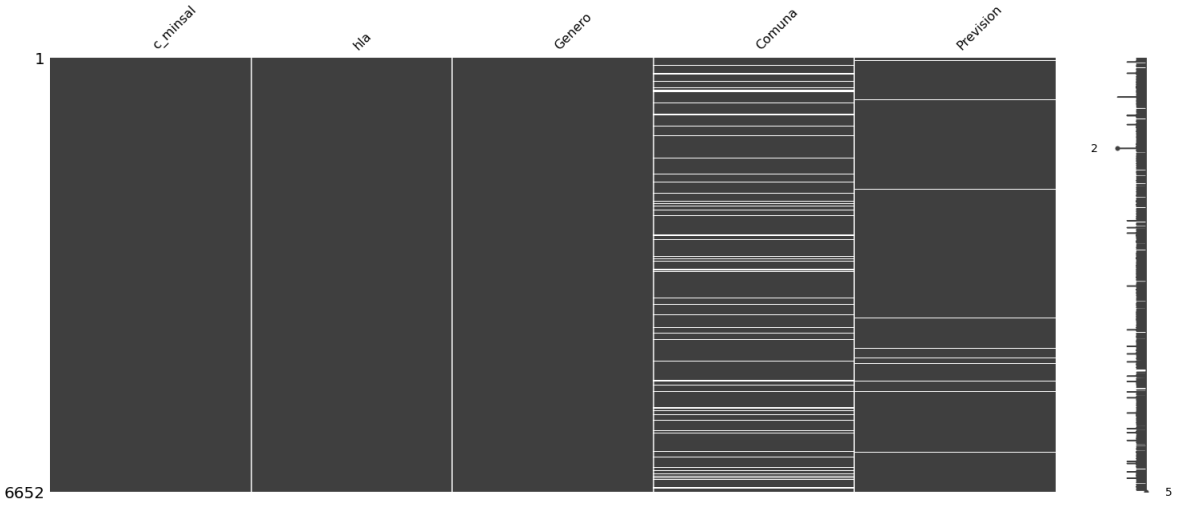

Figura 9. Gráfica de datos nulos por variables escogidas de la tabla Informacion_pacientes. Elaboración propia

Se utiliza la variable Comuna para la construcción de la variable “Segmento comuna” en base al ingreso promedio de estas, teniendo como referencia la encuesta CASEN realizada en 2007.

Tabla Autorizaciones

Esta tabla posee información de si el paciente ha autorizado a un tercero para el retiro de sus medicamentos. La tabla posee 2044 pacientes y las variables seleccionadas se pueden observar en la Figura 10.

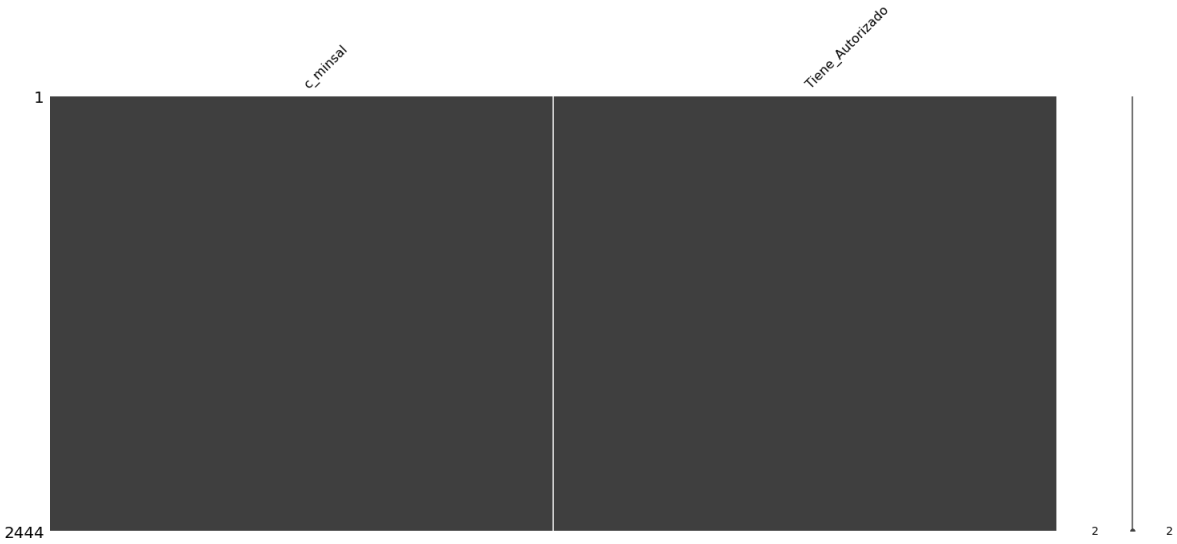

Figura 10. Gráfica de datos nulos por variables escogidas de la tabla Autorizaciones. Elaboración propia

Tiene_autorizado es una variable binaria que toma 1 de tener a un tercero autorizado para retiro de medicamento y 0 si no.

Tabla Terapias_pac

Se realiza un trabajo similar a Terapias_p_op en la tabla de Terapias_pac la cual entrega información sobre fármacos y posología de los pacientes con relación a TAR.

Las tablas seleccionadas se unen formando la tabla provisoria “Farmacia” la cual contiene 6176 pacientes que equivalen a 437914 observaciones. A esta unión de tablas se realiza la selección de los pacientes con los cuales se trabajará.

Finalmente se genera la unión de las tablas para generar la tabla “alcohol y otras drogas”

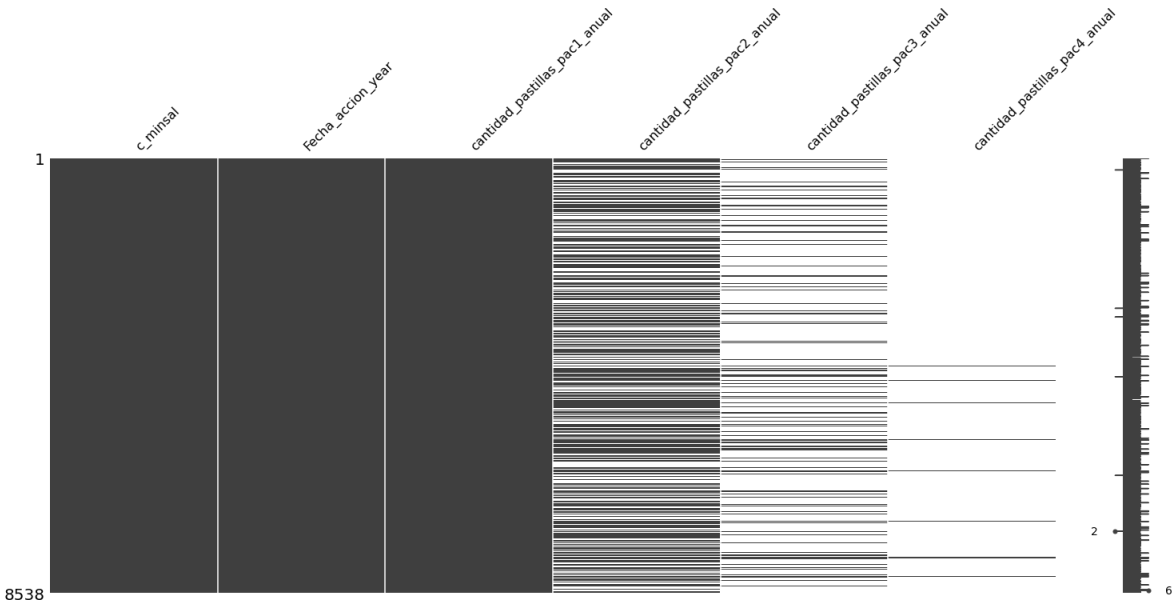

Figura 11. Gráfica de datos nulos por variables escogidas de la tabla Autorizaciones. Elaboración propia

Luego del cruce final de las tablas de farmacia se hace la selección de los 2484 pacientes a trabajar y se obtiene una tabla con 20 variables y 7363 observaciones.

5.1.2.1.2 Estudio de la tabla Master ficha clínica

Se procede a entender la tabla máster, la cual actúa como tabla común para los cuatro subpredictores mencionados anteriormente, incluidos el presente. Para tener un mejor entendimiento de la creación de esta se decide hacer una réplica de la tabla creada por parte del equipo del WIC. Para esta se siguen los siguientes pasos:
Se trabaja desde una table inicial de tabla máster Ficha clínica, la cual contiene 41 columnas, según las variables seleccionadas y la información de un total de 3108 pacientes, lo cual resulta en 197713 observaciones en total.

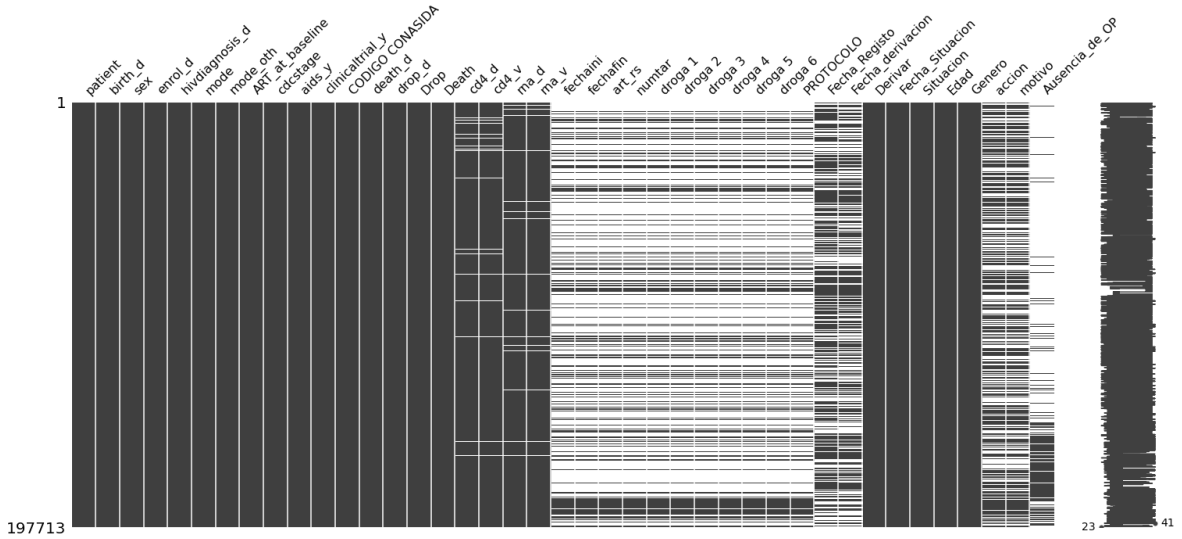

Figura 12. Tabla inicial Máster ficha clínica. Elaboración equipo WIC

Las variables presentes en la tabla máster ficha clínica (Ver Figura 11) y su significado se pueden apreciar en la Tabla 7

Tabla 7. Variables tabla master ficha clínica inicial

Variable	**Descripción variable**
Patient	Código interno de la fundación para el paciente
Birth_d	Fecha de nacimiento del paciente
Sex	Sexo del paciente (F: femenino, M: masculino, T: transexual)
Enrol_d	Fecha de enrolamiento en la Fundación
Hivdiagnosis_d	Fecha de diagnosis del paciente
Mode	Modo de contagio, dado por código
Mode_oth	Modo de contagio, escrito
ART_at_baseline	Indica si se encuentra con tratamiento TAR el paciente
Cdcstage	Etapa de VIH
Aids_y	Año diagnostico
Clinicaltrial_y	Existe estudio clinico
CODIGO CONASIDA	Código del paciente
Death_d	Fecha de fallecimiento si es el caso
Drop_d	Fecha de abandono si es el caso
Drop	Indica si el paciente a abandonado tratamiento
Death	Indica si el paciente se encuentra fallecido
cd4_d	Fecha de toma de cd4
cd4_v	Valor de la medición de cd4 para la fecha cd4_d
rna_d	Fecha de toma de rna
rna_v	Valor de la medición de cd4 para la fecha rna_d
Fechaini	Fecha inicio de tratamiento
Fechafin	Fecha fin de tratamiento
Art_rs	Razón de cambio o suspensión de TAR
Numtar	Número de tratamiento del paciente
Droga X	Medicamento número X
Protocolo	Existencia de un protocolo (toma valores verdadero o falso)
Fecha_registro	Fecha de registro de la observación
Fecha_derivacion	Fecha de derivación del paciente
Derivar	Indica si se debe derivar al paciente
Fecha_situación	Fecha en que se registra la situación del paciente
Situación	Situación del paciente

Edad	Edad del paciente
Género	Género del paciente
Acción	Acción que ocurre con el paciente (cambio, traslado, etc)
Motivo	Motivo de cambio de medicación
Ausencia_de_OP	Refleja el conteo enfermedad oportunista por año

Se debe tener en cuenta que, para el cálculo de la variable dependiente se prevén como variables claves: “cd4_v”, “rna_v”, la situación de adherencia y de abandono o retorno. Las dos primeras variables con cálculos clínicos de laboratorio, las cuales el equipo del WIC decide llevar a mediciones anuales dado que existen distintas cantidades de mediciones para un año en cada paciente, unificando la información. Al tomar esta decisión se lleva por tanto a mediciones anuales las otras variables. El mismo proceso de unificar las variables a un formato anual se sigue con todos los sub predictores del estudio, lo cual hace que se pueda comparar rápidamente los resultados.

Como parte del trabajo realizado en esta tabla se inspecciona las variables rna y cd4. En rna se encuentran valores que no pueden ser catalogados numéricamente, las cuales son eliminadas. Por otra parte, se fija el límite inferior designado en acuerdo por el equipo del WIC, según lo cual los valores menores a este son tomados como 0 en la base. Se realiza un trabajo similar con cd4.

Tal como se señala anteriormente se crean variables iguales al promedio anual de los valores de cd4 y rna para los pacientes a través de los años. Para las distintas variables se toman acciones similares con el objetivo de anualizar la variable, pero, manteniendo la concordancia con el significado de la variable, un ejemplo claro de esto es para la variable “droga X” se toma la moda de cada paciente en el año como el valor de anualizado.

Posterior al trabajo realizado en la ejecución de réplica del trabajo del equipo del WIC en la construcción de la base, estos toman la decisión de cambiar de formato la tabla “ficha clinica máster”. Este trabajo no fue replicado, sin embargo, es necesario dar una pequeña explicación de los cambios.
En la primera iteración de “ficha clínica master” se han descrito las variables a utilizar sin embargo estas se encuentran para cada año del paciente, tomando como referencia el máximo de años que un paciente estuvo en tratamiento, así por ejemplo las columnas de cd4 agrupan a cd4_añoX1, cd4_añoX2 hasta el máximo de año encontrado. El cambio del que se habla acá rádica en la generación de una nueva columna “year” (año) por lo cual se eliminan las columnas por año y pasan a ser estas distintas observaciones dentro de la misma columna.

Paralelamente al trabajo de la tabla se decide como etiquetado, la construcción de la variable “fallo_ virológico”. Aunque previamente se había considerado la utilización de cd4 o rna para el label o incluso una unión entre ambas, sin embargo, se debe tener en

consideración que no existe una relación directa, como se podría esperar entre ambas. Esto ocurre porque son variados los factores que influyen en el momento de la toma de muestra, ya sea de rna o de cd4; el tiempo desde que se ha contagiado el paciente, ya sea recién contagiado o transcurrido una o dos semanas, el tiempo transcurrido desde que se infectó de VIH hasta que se ha enrolado en la Fundación o visitado un médico para su derivación a esta, las particularidades de su sistema inmunológico, entre otras. Este punto será explicado con más detalle en el capítulo de discusión.

Fallo virológico permite una medición fija en base a las mediciones de rna y sin perder por tanto de la base de trabajo las variables relacionadas a cd4, una variable esencial para el estudio. Fallo virológico se toma como la medición de dos medidas de rna sobre 200 copias /mL, lo cual sigue las pautas usadas por la Fundación Arriarán. Esto significa que aquellos pacientes que presenten de forma recurrente valores menores a 200 copias /mL no tienen un nivel alto de carga viral en su sistema, no presentan un fallo virológico y da a entender que el tratamiento se está siguiendo de forma correcta y está funcionando, mientras que aquellos pacientes que presentan dos veces seguidas un valor menos a 200 copias /mL significa que su carga viral se ha mantenido sobre el umbral y por ende el tratamiento se encuentra funcionando.

Para la elaboración de fallo virológico se decide nuevamente evaluar la tabla máster ficha clínica inicial, mostrada anteriormente, ya que con la información presente hasta el momento es difícil tomar una decisión con respecto a las observaciones faltantes de rna para la construcción de la variable fallo virológico.

Se realiza una inspección de la tabla teniendo en consideración dos casos de interés:

- Aquellos pacientes que presentan mediciones de rna pero no mediciones de cd4 para el mismo año
- Aquellos pacientes que presentan mediciones de cd4 pero no mediciones de rna para el mismo año.

Fijarse en estos casos en específicos permite centrarse en aquellos pacientes que realmente les falta la medición de algunas de las dos variables cuando deberían tenerla y no porque haya un abandono del tratamiento.
En base a la revisión previa y con la opinión experta de la *medical advisor* se decide tomar una tolerancia de un año y 3 meses. Aquellos pacientes con datos faltantes en un período mayor a este son eliminados de la base de datos a trabajar. Una vez tomada esta decisión, se rellena la data faltante de la siguiente forma:

- Para aquellos pacientes en que el año que les falta data, ya sea en las mediciones de rna o de cd4, es uno de los años intermedio en su horizonte de tiempo en la fundación, se completa la data con el promedio entre la medición anterior y siguiente a la faltante.

- Para aquellos pacientes en el que el año faltante es su primer año de tratamiento se decide que se eliminen de la tabla máster, y, por consiguiente, de todas las tablas de predictores, ya que no existe una forma coherente de rellenar la data.
- Para aquellos pacientes en el que el año faltante es el último presente, se decide repetir la medición anterior en la data faltante.

Una vez realizado este proceso de rellenado de la data faltante se calcula la variable fallos virológicos, calculando la cantidad de fallos anuales para el paciente por cada año, teniendo en cuenta la definición anterior, es decir, que fallo se considera como una medición superior a 200 copias /mL en dos mediciones consecutivas de rna. De forma paralela, se calcula “fallo virológico acumulado” la cual calcula cuantos fallos virológicos lleva el paciente hasta el año de cálculo.

Al realizar este proceso se debe tener en consideración que se obtiene una tabla de 2619 pacientes con los cuales trabajar. Sin embargo, se decide trabajar de forma separada la información que se posee hasta el 2019 y aquella entre el año 2020 y 2021, ya que se prevé cambios conductuales generados por el contexto de la pandemia, los cuales sería interesante estudiar de forma separada. De aquí en adelante, el trabajo señalado hace relación al trabajo con la base con datos previos al 2020, quedando con una base de 2528 pacientes para el trabajo posterior.

Se debe recalcar que aunque se comenzó analizando los datos provenientes de la Fundación Arriarán desde 1992, debido a la falta de datos para algunas de las variables en el proceso de formación de las tablas finales (para cada sub predictor) para trabajar se debió, junto al grupo del WIC, eliminar 44 pacientes de los 2528, lo cual entrega el total de los 2484 pacientes que se trabajara de aquí en adelante, y en concordancia con lo que se ha expuesto anteriormente, siendo este un trabajo en paralelo de distintos predictores, con data parcialmente en común, en base a la tabla master ficha clínica, estos 2484 pacientes son los que se trabajan en cada uno de ellos.

5.1.2.2 Unión de las Tablas Finales

Una vez realizado el cruce de la tabla máster y la tabla de alcohol y otras drogas, en base a las variables de año y código del paciente. En este punto se decide hacer algunas modificaciones que ayuden al trabajo posterior de los predictores.

Se genera la variable “segmento etario” a partir de la variable “Edad” encontrada en “ficha clínica master”. Este segmento etario se genera en base a los segmentos encontrados en estudios del SENDA (servicio nacional para la prevención y consumo de drogas y alcohol de Chile).

Primero se decide para las variables “nivel educacional”, “condusexual”, “droga X”, “Protocolo”, “employment_status” rellenar las observaciones faltantes con “SIN DATO”. Por otra parte, a las variables numéricas tales como: “cantidad de pastillas opX_anual”, “mg_pastillas_opX_anual”, “cantidad_pastillas_pacX_anual”, “art_rs” se rellena la data faltante con cero.

Se decide dentro del equipo de trabajo del WIC cambiar la variable dependiente, fallo virológico, a una variable binaria, dada la poca cantidad de observaciones en cada una de las categorías, se transforma de forma: 1 si hay fallo y 0 sino. Posteriormente, se separa la base momentáneamente en variables categóricas y numéricas, aquellas variables categóricas son tratadas, creando *dummies*, es decir se crea una columna de 0 o 1 para cada valor que puede tomar la variable. A su vez las variables numéricas son normalizadas, con una normalización de tipo standarscalesr.

Se ocupará por tanto fallo virológico como el *label,* en base a la información que nos entrega relacionada problemas de adherencia

En base a lo anterior se obtiene la tabla final de alcohol y otras drogas, la cual posee 64 variables, de las cuales 39 provienen de tabla máster ficha clínica lo cual genera un resultado de 12738 Observaciones, que corresponde a los 2484 pacientes antes mencionados.

5.2 Transformación y balance de datos.

Una vez construida la tabla final de “Alcohol y otras drogas” y unida está a la tabla máster, se genera una inspección de las variables elegidas. Primero se observa nuevamente los datos faltantes en las variables, ya que, al analizarlas anualmente, ocurre en algunos casos que el dato falta en uno o dos años de 15 años de tratamiento, por ejemplo. Entonces procede a completar la data faltante de la siguiente manera:

- Si la variable a rellenar es de tipo numérica, la data faltante se rellena con 0
- Si la variable a rellenar es de tipo categórico, los datos faltantes son reemplazados por la etiqueta de “Dato faltante”.

Posteriormente se realiza una separación entre aquellas variables que son categóricas y variables numéricas. De esta, se obtiene un total de 34 variables numéricas y 30 variables categóricas. Las variables categóricas son convertidas en variables dummy (Se genera una columna por valor único de la variable o categoría, en cada una de estas columnas la nueva variable toma valores 1 o 0, dependiendo de si cabe dentro de esta categoría o no). De forma paralela las variables numéricas se transforman a través de StandarScaler,

incluido en la librería sklearn de Python. Este método consiste en que los datos se escalen de tal forma que la varianza sea igual a 1.

Se debe tener en consideración que se está trabajando con distintas observaciones de pacientes a lo largo de su horizonte de tiempo en la Fundación. En base a esto, es importante a la hora de generar la separación de la data en las bases de entrenamiento (porcentaje de la base de datos que se utiliza para el entrenamiento de los modelos de aprendizaje supervisado) y testeo (una parte secundaria de la data, de menor cantidad de observaciones con la cual se testea el modelo luego de que este ha sido entrenado con el set de entrenamiento). Por ende, se genera una
separación de tal forma que permita una separación entre se mantienen separados testeo de entrenamiento permitiendo que mantenga las observaciones de un mismo paciente se mantengan en uno de estos set. Para hacer la separación de data se utiliza stratifiedKFolds, disponible en la librería skalearn de Python

Una vez realizado estos cambios se tiene la data preparada, que considera como se ha dicho en las secciones anteriores a 2484 pacientes, correspondiente a 12.738 observaciones y 64 variables, para la ejecución de los modelos de algoritmos de aprendizaje supervisado.

5.3 Desarrollo del modelo de predicción de adherencia de tratamiento VIH con relación a consumo de alcohol y otras drogas

5.3.1 Desarrollo del modelo de predicción

Se decide evaluar el rendimiento de los modelos: logit, KNN, decition tree, random Forrest, SVM, XGBoost y Naive Bayes.

Lo que se explica a continuación, si bien está relacionado con tratamiento de variables (datos) y el *label*, se comenta en este ítem en relación con el desarrollo del modelo ya que el problema se presentó una vez que todos los subpredictores, incluido el de la presente memoria presentaron el mismo problema en relación al rendimiento del modelo y debimos realizar los cambios que se mencionan a continuación. Como se ha señalado anteriormente la elección del *label* es una decisión compleja y se realiza utilizando como referencia el subpredictor de Farmacia. Por lo tanto, como una primera iteración de los modelos, el equipo del WIC, toma como *label* los niveles de cd4 para el sub predictor de farmacia, en base a ello se realizan dos ejecuciones de los modelos, una primera instancia en la cual se elimina de las variables dependientes las variables relacionadas directamente con cd4 (Tabla 8) y una segunda instancia donde estas se mantienen (Tabla 9). Los resultados de los modelos en base a las métricas: *recall, precisión, accuracy, log*

loss no son los esperados en base a rendimiento, se tiene como los tres mejores modelos XGBoost y Random forest, los cuales se pueden observar en la Tabla 8 Y Tabla 9, respectivamente.

Tabla 8. Resultados modelos con mejor rendimiento con label cd4, eliminando variables dependientes directamente relacionadas

Clasificador	*Recall*	*Precision*	*Accuracy*	*Log Loss*
XGBoost	0,49	0,52	0,49	1,06
Random Forest	0,49	0,56	0,51	1,07

Tabla 9. Resultados modelos con mejor rendimiento con label cd4, sin eliminar variables dependientes directamente relacionadas

Clasificador	*Recall*	*Precision*	*Accuracy*	*Log Loss*
XGBoost	0,78	0,81	0,79	0,53
Random Forest	0,75	0,80	0,77	0,61

Posterior a estas iteraciones se realiza una prueba con el *label* de si pertenece a los grupos de baja adherencia reconocidos por la Fundación Arriarán, los resultados se observan en la Tabla 10

Tabla 10. Valores métricas por modelo, predictor Farmacia, caso label según tabla adherencia Fundación Arriarán

Clasificador	*Recall*	*Presicion*	*Accuracy*	*Logloss*
Decision Tree	0,50	0,57	0,30	0,76
KNN	0,52	0,52	0,61	3,07
Naive Bayes	0,56	0,62	0,70	0,67
SVM	0,58	0,56	0,56	0,58
XGBoost	0,54	0,66	0,71	0,57

Posterior al desatisfactorio desempeño de los modelos con el último caso expuesto, se plantea un nuevo tratamiento de los datos, donde se crea el cambio mencionado anteriormente de generar una observación por año que ha estado en tratamiento, para cada paciente. Al realizar esto se generan distintas formas de anualizar los datos según las variables. Los casos más usuales son el cálculo de promedio para las variables numéricas y la selección de la moda anual para las variables de tipo categórica.

Debido a problemas con la cantidad de data disponibles en las variables de drogas (medicamentos TAR consumidos por los pacientes). El año de inicio que se consideró para la toma de droga es de 2011, siendo que la data disponible de la Fundación es desde 1991. Teniendo en consideración que la data de fármacos es esencial para el predictor de Farmacia y que se decide tomar decisiones en cuanto a la base de datos para seguir considerando este conjunto de variables. Se plantean 3 estrategias:

Estrategia 1: Se eliminan de la base de datos aquellos pacientes eliminados en la construcción de la variable fallo virológicos
Estrategia 2: Eliminación de los pacientes enrolados previo al año 1997, se utiliza la variable fallo virológicos para el etiquetado
Estrategia 3: Se eliminan los pacientes enrolados antes del 2011 tanto de tabla master como de farmacia

Luego de la ejecución de los modelos y la interpretación de las métricas de rendimiento se decide seguir con la estrategia 2.

Tomando en consideración todos los puntos anteriores se comienza con construcción de los modelos para el caso del predictor de alcohol y otras drogas. La base de datos es separada en sets entrenamiento y test de tal forma que las observaciones de un mismo paciente se mantengan en el mismo set (esta separación se realiza en una proporción 75-25 respectivamente), ver resultados de predictores en base a distintas métricas para el set de testeo en Tabla 11. De esta tabla se destaca que los valores de recall, excepto para Naive Bayes, dieron alrededor de 0,7 siendo para XGBoost y Random Forest cerca de 0,8. Esto da a entender que el modelo está prediciendo efectivamente los casos positivos, esto es de suma importancia en un proyecto de salud, donde la repercusión de un falso negativo posee un costo alto.

Por parte del equipo del WIC se buscan métodos para mejorar los resultados tales como la selección de variables *backward,* la cual consiste en la búsqueda de las variables optimas a utilizar en el modelo a través de un sistema de prueba, donde se parte el modelo con la totalidad de las variables de la base y se elimina una a una las variables para encontrar el set óptimo de variables. Los resultados de este proceso no mejoran el desempeño de los modelos en las métricas seleccionadas.

Se utiliza *grid search*, un método de optimización de hiper-parámetros, porque a pesar de encontrarse buenos valores en las métricas de accuracy, precisión y recall, las matrices de confusión de los modelos llevan a la decisión de tomar medidas para mejorar la predicción, *grid search* calibra los hiperparametros del modelo mejorando los resultados por lo que se toma la decisión de ocuparlo, los resultados se pueden observar en la Tabla 12.

Tabla 11. Resultado de predictores en base a distintas métricas

Clasificador	*Accuracy*	*Precision*	*Recall*
Decision Tree	0,72	0,72	0,72
XGBoost	0,78	0,79	0,78
logit	0,72	0,72	0,72
KNN	0,72	0,71	0,72

Random Forest	0,77	0,77	0,77
SVM	0,75	0,76	0,75
Naïve Bayes	0,44	0,62	0,44

Tabla 12. Resultado de predictores en base a distintas métricas utilizando *grid search*.

Clasificador	*Accuracy*	*Precision*	*Recall*	*LogLoss*
Decision Tree GS	0,76	0,76	0,73	3,54
XGBoost GS(balanced score)	0,77	0,76	0,74	0,53
KNN GS	0,72	0,7	0,69	0,55
Random Forest GS	0,77	0,8	0,72	0,48
SVM GS	0,76	0,76	0,75	0,5
NB GS	0,51	0,62	0,58	1,98

Hasta este punto del análisis, el valor del *label* representaba si es que, en algún punto anterior a ese año, el paciente había tenido un fallo virológico. Esto, aunque en concepto es correcto, durante la discusión con el equipo intento del WIC se vio más conveniente redefinir el concepto del *label* utilizado hasta este punto. De esta forma, se planteó que desde este punto del análisis en adelante se comenzaría a utilizar una nueva etiqueta binaria para todos los modelos:

- Toma el valor 1, si es que tiene al menos un fallo virológico en el año en análisis.
- Toma el valor 0, si es que el paciente en cuestión no presenta fallos en el año analizado.

Este cambio permite varias mejoras en las componentes de análisis e interpretación de los modelos desarrollados ya que al traducir la información a una base anual, los resultados de los predictores van a estar indicando la probabilidad de un fallo virológico en el paciente en el año predicho, lo que permite planificar mejor los recursos médicos, monetarios y presupuestarios de la institución así como incorporar un componente social muy fuerte: Ahora se puede predecir más claramente la cantidad de personas que durante un año calendario podrían verse afectadas y crear políticas y procesos que permitan aumentar la eficiencia del servicio.

Dado este cambio, los modelos y resultados antes expuestos deben ser actualizados para incorporar este cambio en el *label* ocupado. Por lo mismo, para el proceso de elaboración de los modelos se siguen pasos similares a los previamente enumerados con el caso de la etiqueta previa. Sin embargo, se requiere de la eliminación del set de variables relacionadas con rna, esto pues presentan una correlación directa con la nueva variable ocupada en el "*label"* y el mantenerlas en el set de entrenamiento puede llevar a

un *overfitting* del modelo. Al eliminar estas variables nuestra tabla queda con un total de 58 variables, correspondientes a 2484 pacientes

Se debe recalcar que los modelos ejecutados que se han utilizado son los mismos. Los resultados se pueden apreciar en la Tabla 13. De los valores se destaca que, aunque los valores de *accuracy*, precisión y *recall* para los primeros 6 modelos entregan una buena proporción, las matrices de confusión nos demuestran que no existe un cálculo preciso de verdaderos positivos y verdaderos negativos (Figura 13) lo cual se observa particularmente acentuado en los resultados de SVM, que aunque posee un *accuracy* y precisión alto, da 0 en su proporción de verdaderos negativos.

Tabla 13. Resultado de predictores en base a distintas métricas. Caso con label de fallos virológicos anuales.

Clasificador	*Accuracy*	*Precision*	*Recall*	*LogLoss*
Decision Tree	0,87	0,88	0,87	4,33
XGBoost	0,91	0,90	0,91	0,21
Logit	0,90	0,88	0,90	0,26
KNN	0,90	0,87	0,90	1,35
Random Forest	0,91	0,90	0,91	0,23
SVM	0,90	0,82	0,90	0,25
Naive Bayes	0,12	0,85	0,12	30

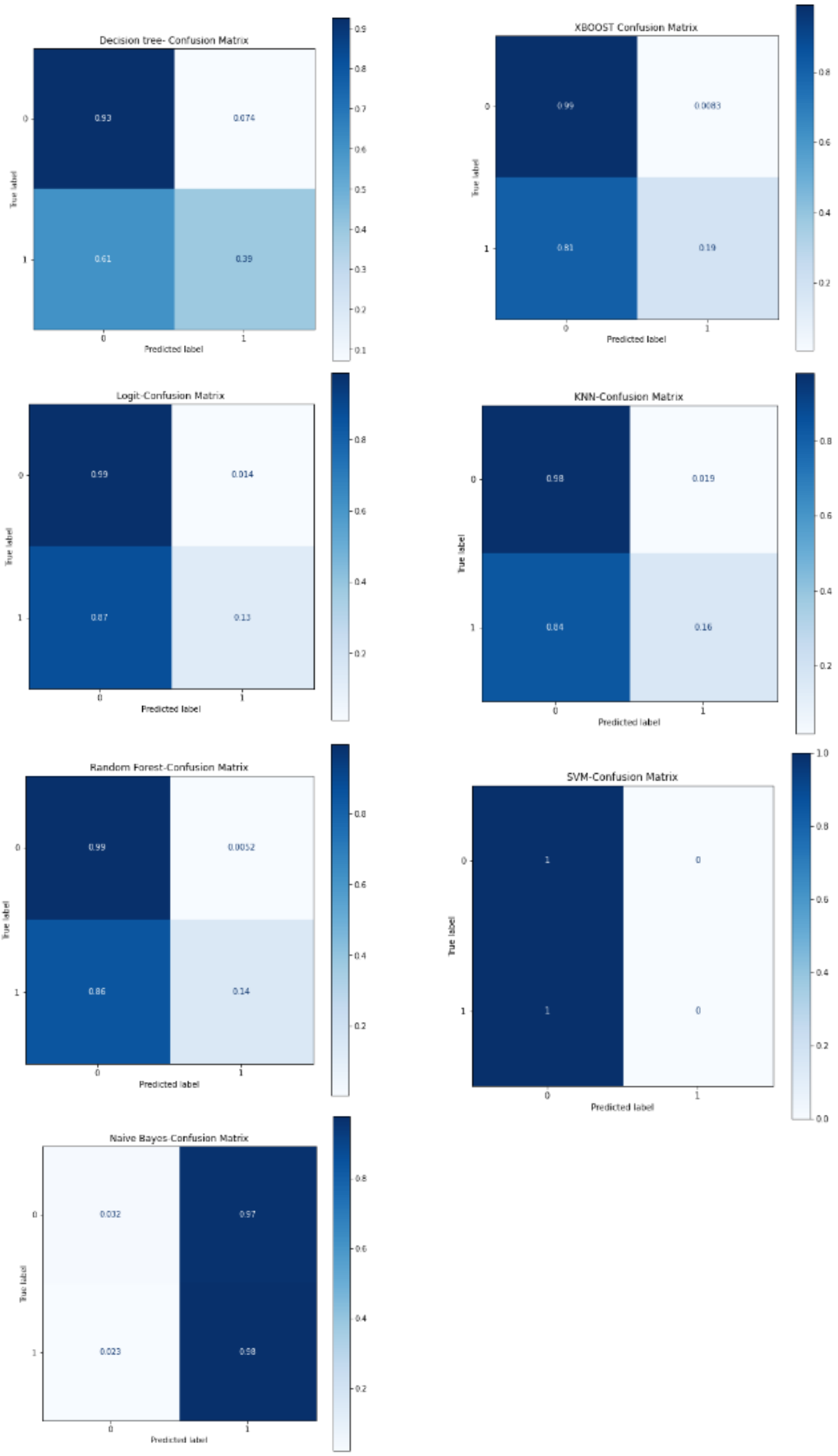

Figura 13. Matrices de confusión por modelo. Elaboración propia

Con la finalidad de mejorar las predicciones de los modelos, y tal y como se hizo en los resultados antes de la incorporación del nuevo label (expuestos en la tabla 11), se decide utilizar también en esta instancia *grid search*. En la Tabla 13 se pueden observar los resultados con la optimización realizada. Se ajusta la grilla de optimización de acuerdo con cada modelo. Los valores de recall dan bajo para KNN y Random Forest, por lo que se descartan estos predictores. Se ve en SVM y en Decisión tree, una baja en *recall* con respecto a la Tabla 11. Sin embargo, existe una mejoría en las predicciones de verdaderos positivos y verdaderos negativos, que se ve reflejado en las matrices de confusión, lo cual nos da en general mejores resultados en los distintos modelos.

De todos los modelos expuestos en la Tabla 14, SVM es el que presenta un mejor valor de *recall*, el cual llega a 0.75 siendo significativamente mayor que el resto de los modelos estudiados. Dado que la métrica de sensibilidad es clave en este tipo de estudios médicos, se prioriza dicho valor a pesar de tener un valor levemente menor de *accuracy* con respecto a los otros modelos. Por lo mismo, SVM será el modelo utilizado para los próximos pasos de este análisis por cumplir con los criterios estadísticos, así como también con las necesidades del equipo investigativo del WIC, así como personal médico en sus objetivos de este trabajo.

Tabla 14. Resultado de predictores en base a distintas métricas utilizando grid search. Label de fallos virológicos anuales.

Clasificador	*Accuracy*	*Precision*	*Recall*	*LogLoss*
Decision Tree GS	0,90	0,70	0,65	2,12
XGBoost GS	0,92	0,79	0,67	0,24
KNN GS	0,90	0,71	0,56	0,26
Random Forest GS	0,91	0,88	0,53	0,23
SVM GS	0,74	0,60	0,75	0,26

En resumen, de la tabla anterior se puede observar una mejoría del modelo, caracterizado por un mejor desempeño en *log loss*, lo cual nos indica que existe una mejor predicción de las categorías. En el caso de los otros tres sub predictores, SVM fue también seleccionado ya que entrega una mejor predicción al ver los resultados de forma global de las métricas.

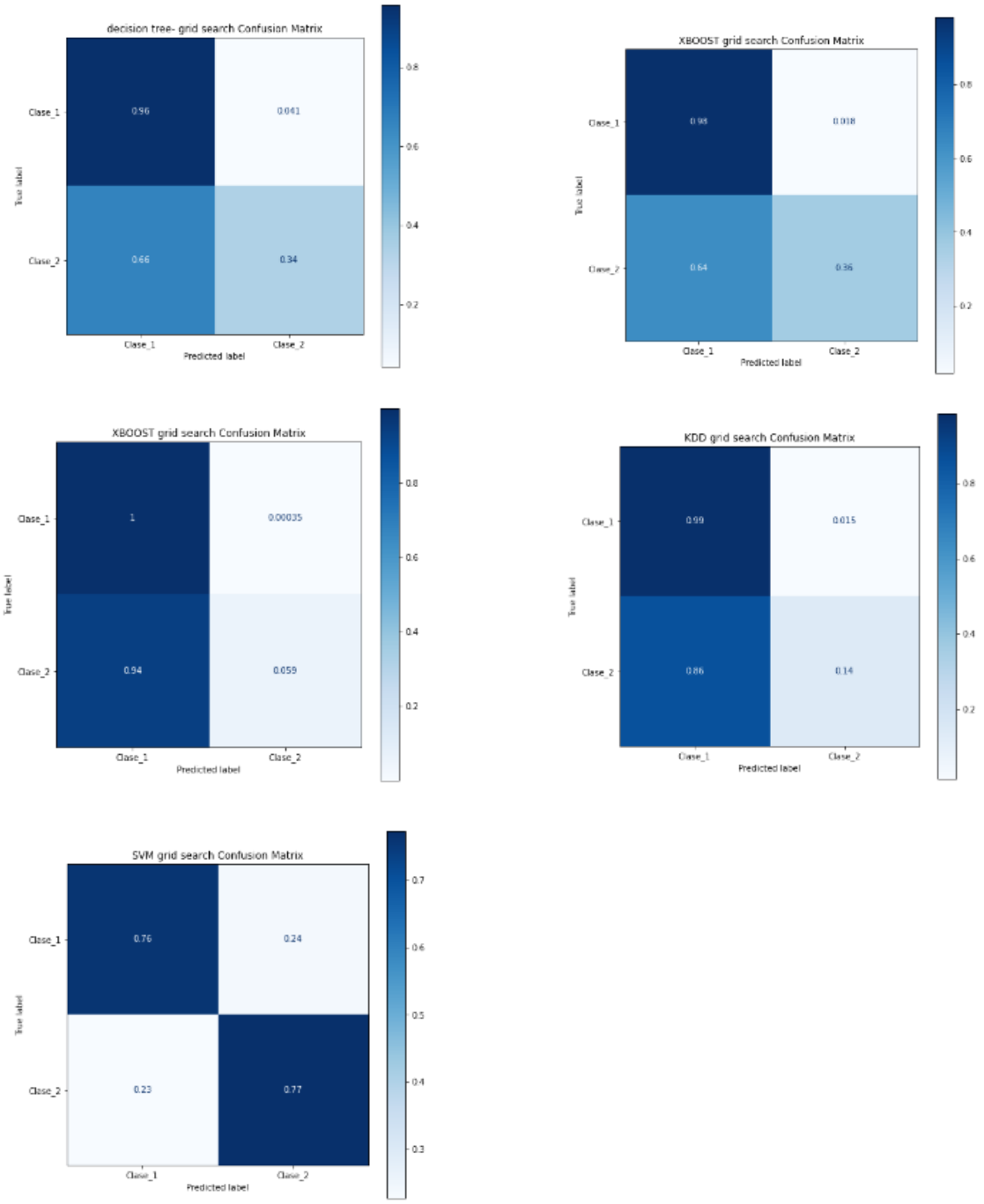

Figura 14. Matrices de confusión por modelo con grid search. Elaboración propia

Por otra parte, de la Figura 14, se puede observar una mejoría en las matrices de confusión, específicamente de la Figura 14 en la posición:

- (0,0) Caso *decision tree* con *grid search,* no muestra un cambio significativo en las proporciones de verdaderos positivos ni de verdaderos negativos, sin embargo, observando la Tabla 10 se ve una reducción en los valores de log loss, lo cual nos indica que ha mejorado la predicción.

- (0,1) Caso *XGBoost* con *grid search*, se aprecia una mejora en la cantidad de verdaderos negativos clasificados con respecto a la Figura 13.
- (1,0) Caso KNN, no se aprecian cambios significativos en la matriz de confusión.
- (1,1) Caso *Random forest* con *grid search*, se ve una mejora a la hora de clasificar verdaderos negativos.
- (2,0) Caso SVM con *grid search*, se ve un aumento tanto en la cantidad de verdaderos positivos como de verdaderos negativos con respecto al grafico de matriz de confusión en la Figura 13, lo cual nos muestra que existe una mejor calificación por parte del modelo, esto sumado a una mejora en la métrica de *recall.*

5.3.2 ¿Cómo medirá adherencia el presente subpredictor?

Una vez elegido el modelo de SVM se prosigue a la categorización de adherencia en base al predictor. Es importante mencionar que se utiliza el mismo algoritmo de categorización para todos los sub-predictores del proyecto Fondef ID20I10174, lo cual implica que los resultados de este trabajo de investigación serán comparables y más fáciles de analizar al momento de realizar comparativas y trabajos futuros al dar un marco común de investigación al equipo del WIC.

Se procede a generar una segmentación de las instancias en base a la probabilidad de cada una las instancias. Se obtiene un umbral de corte en base a la optimización de la instancia de *recall*. Teniendo en cuenta la mayor y menor de las mediciones de probabilidad se hace una segmentación en 5 categorías, que son:

- Categoría 1- Adherencia total 100%
- Categoría 2 Adherencia máxima 75%-99%
- Categoría 3 Adherencia media: 50%-74%
- Categoría 4 Adherencia mínima 25%-49%
- Categoría 5 No adherente: 1-24%

En este caso en particular con un máxima de 0,997 y un mínimo de 0,22 se tiene:

- Categoría 1: 5 instancias [0,15%] (valores mayores a 0,997). Corresponde a 5 pacientes
- Categoría 2: 2765 instancias [86,81%] (valores entre 0,997 a 0,76). Corresponde a 614 pacientes
- Categoría 3: 324 instancias [10,17%] (valores entre 0,766 a 0, 536). Corresponde a 232 pacientes.
- Categoría 4: 85 instancias [2,67%] (valores entre 0,536 a 0,305). Corresponde a 67 pacientes.

- Categoría 5: 6 instancias [0,19%] (valores entre 0,305 a 0,075). Corresponde a 6 pacientes.

Categorías elaboradas según separaciones equitativas de los máximos y mínimos valores de probabilidad dadas por el predictor.

Esto significa a nivel de pacientes que la categoría 2 y 1 son pacientes de una buena adherencia, mientras que aquellos en categoría 3 y 4 tienen un nivel de adherencia menor y se verían beneficiados de medidas de intervención.

Se debe señalar que, de forma complementaria, el WIC con la Fundación Arriarán, han desarrollado una encuesta con la finalidad de obtener más información sobre los pacientes. En especial, para el tema de este trabajo son de interés preguntas tales como: el nivel de consumo del paciente; si ha consumido más de lo planeado o si ha usado este como un *coping mechanism* (mecanismo de evasión) lo que implica un problema a la hora de controlar su nivel de consumo, con lo cual se identifica una adicción a este. A la fecha de esta memoria, no se cuenta con los resultados de la encuesta para todos los pacientes, razón por lo cual no se han podido incluir estas variables de interés en el estudio. No obstante, se plantea como un trabajo futuro.

Posterior a la ejecución del modelo y la categorización de adherencia tomando el *set* de testeo se realiza una exploración de las variables según los datos obtenidos en cada una de las categorías. La exploración de algunas de las variables que se consideran significativas teniendo en cuenta el impacto de estas sobre adherencia, tables como: a) segmento etario, b) años de enrolamiento en la Fundación Arriarán, c) número de pastillas consumidas y d) tener una red de apoyo para el retiro de medicamentos.
Resumidas y visualizadas en las que se pueden observar de forma visual en las siguientes figuras:

a) Segmento etario: Para la categoría 1, observable en la Figura 15, donde se define cada segmento etario como:

    - Segmento 1: menores a 12 años
    - Segmento 2: entre 12 a 18 años
    - Segmento 3: entre 19 a 25 años
    - Segmento 4: entre 26 a 34 años
    - Segmento 5: entre 35 a 44 años
    - Segmento 6: entre 45 a 64 años
    - Segmento 7: mayores a 65 años

No se observa una tendencia clara de edad en base a los resultados, principalmente por la cantidad de observaciones designadas en esta hace pensar que falta información para una conclusión más definitiva y en la Figura 16, clase 3 de adherencia se ve una mayor

dispersión de edades con una concentración en los grupos etarios 4 y 5. Sin embargo, en la Figura 17, para la categoría 5 se puede observar una tendencia de pacientes que se encuentran en un rango de edad mayor. Tanto para la categoría 2, 3 y 4 se observa una distribución con concentraciones en los rangos medios de edad.

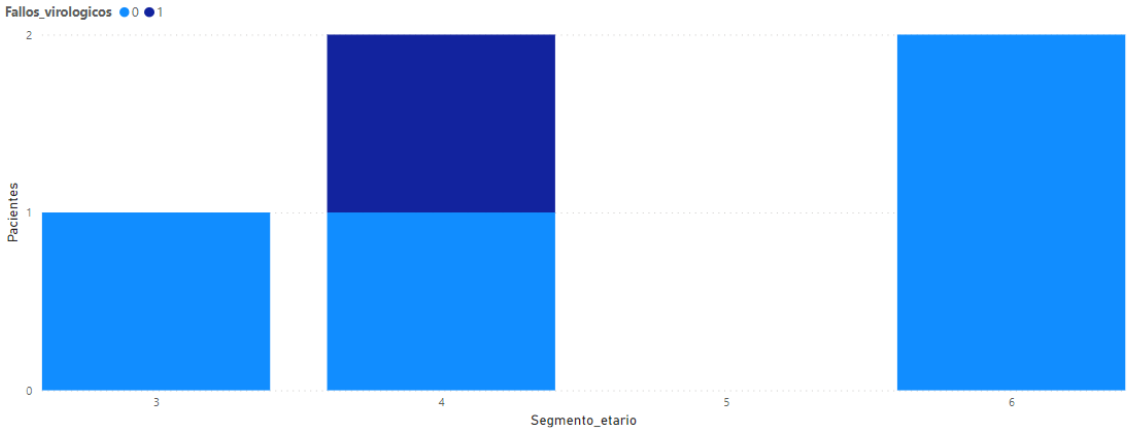

Figura 15. Pacientes por segmento etario, categoría adherencia 1. Elaboración propia

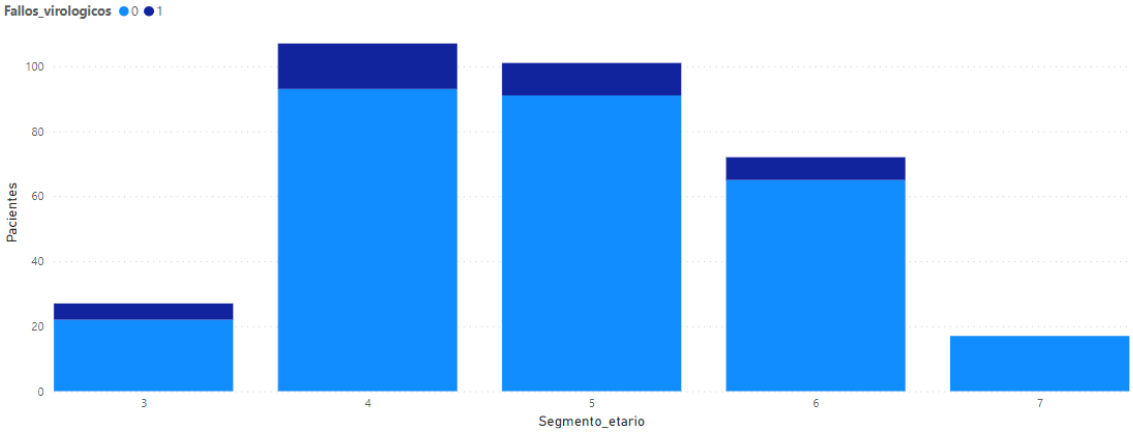

Figura 16. Pacientes por segmento etario, categoría adherencia 3. Elaboración propia

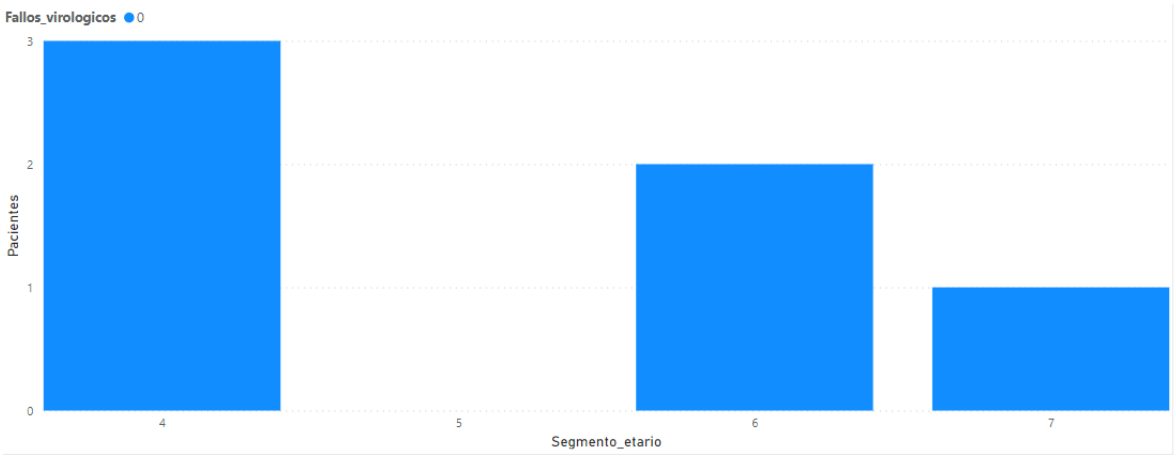

Figura 17. Pacientes por segmento etario, categoría adherencia 5. Elaboración propia

b) Años de enrolamiento en la Fundación Arriarán:

Con respecto a los años de enrolamiento en la Fundación y los años de diagnóstico, no se observa una tendencia clara, pero si se puede ver en el caso de categoría 1(Ver Figura 18), se identifican en esta mayor adherencia caracterizado por dos tipos de pacientes, aquellos que han ingresado hace poco o aquellos que llevan gran tiempo en la Fundación. Tanto en la categoría 2(dato no mostrado) como en la categoría 3(Figura 19) se ve una muestra más grande, pero con una tendencia de pacientes con menor cantidad de años de enrolamiento. En la Figura 20 se ve una concentración en dos grupos, aquellos con bajo tiempo en la Fundación, entre 2 a 6 años y por otra parte aquellos de largo tiempo (15 años).

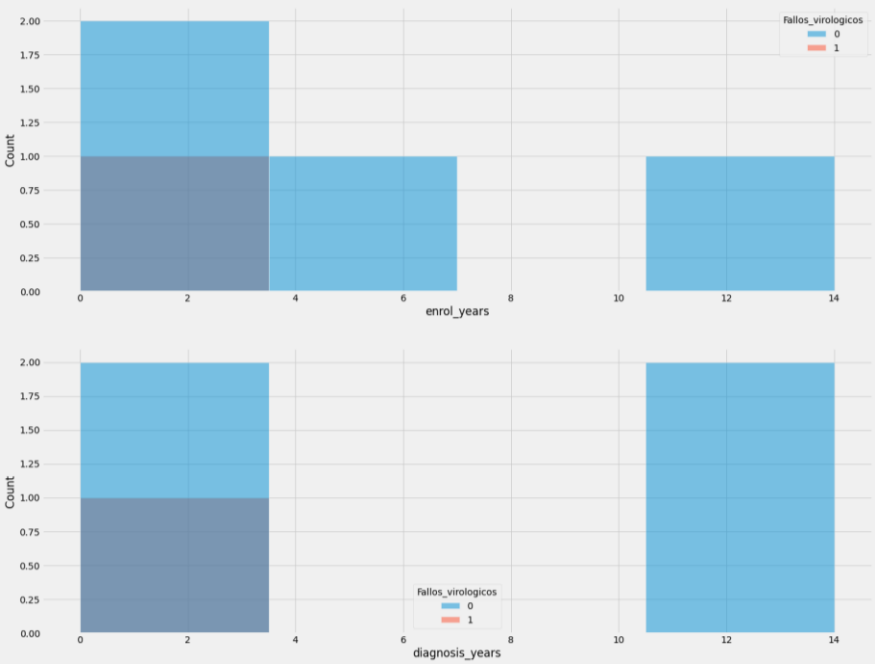

Figura 18. años de enrolamiento y diagnosis. Adherencia categoría 1. Elaboración propia

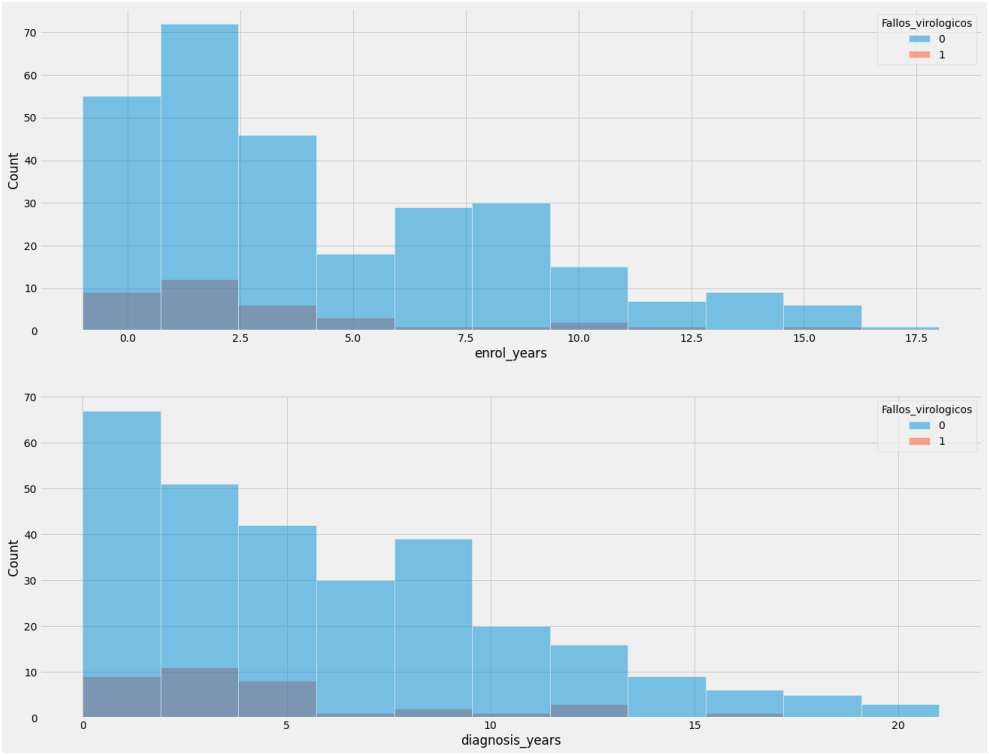

Figura 19. años de enrolamiento y diagnosis. Adherencia categoría 3. Elaboración propia.

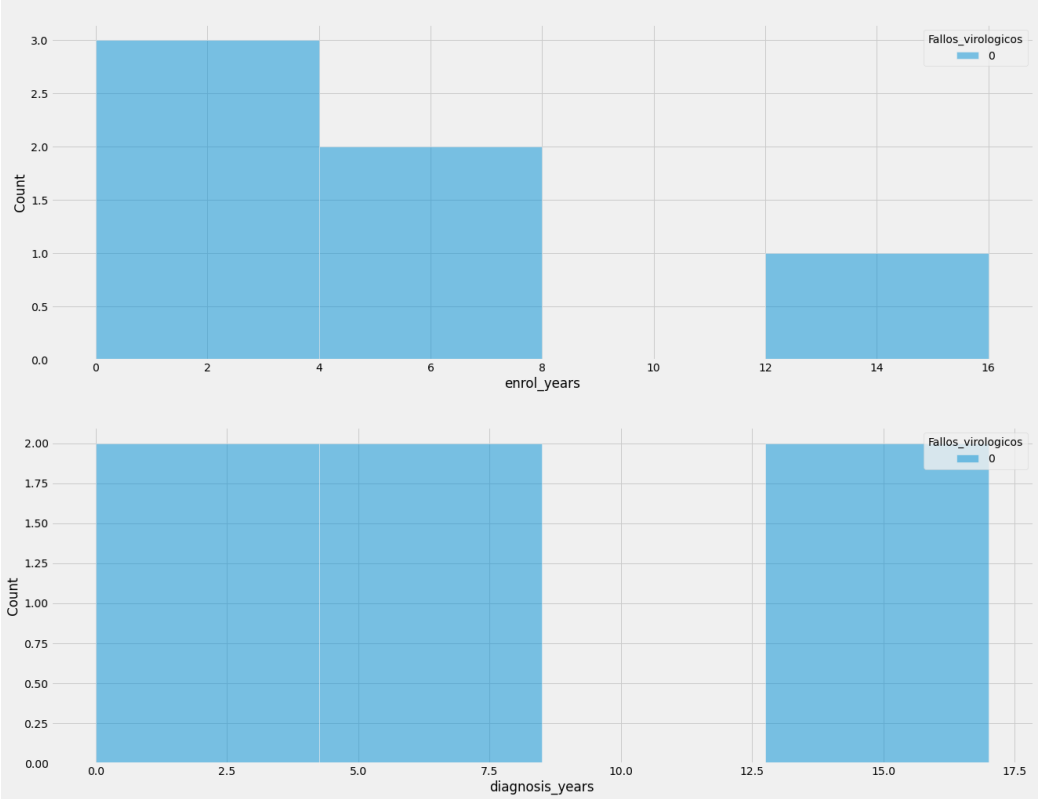

Figura 20. años de enrolamiento y diagnosis. Adherencia categoría 5. Elaboración propia.

c) Número de pastillas

En relación con el número de pastillas, en la categoría 1 (ver Figura 21) se ve pacientes con una baja cantidad de consumo de pastillas por enfermedad oportunistas “sum total op”, lo cual coincide con lo esperado dado que esta es una categoría de alta adherencia. Sin embargo, no se ve un aumento lineal al pasar a categorías de menor adherencia (observar las Figuras 22 y 23)

d) Tener una red de apoyo para el retiro de medicamentos

Realizando el mismo análisis, no se ve una tendencia en los casos extremos con respecto a tener una red de apoyo para el retiro de medicamentos (Observar Figuras 21, 22 y 23). Sin embargo, en el grupo de categoría 3, un grupo de posible intervención se ve una mayoría de pacientes que no cuentan con una tercera persona autorizada para el retiro de medicamentos.

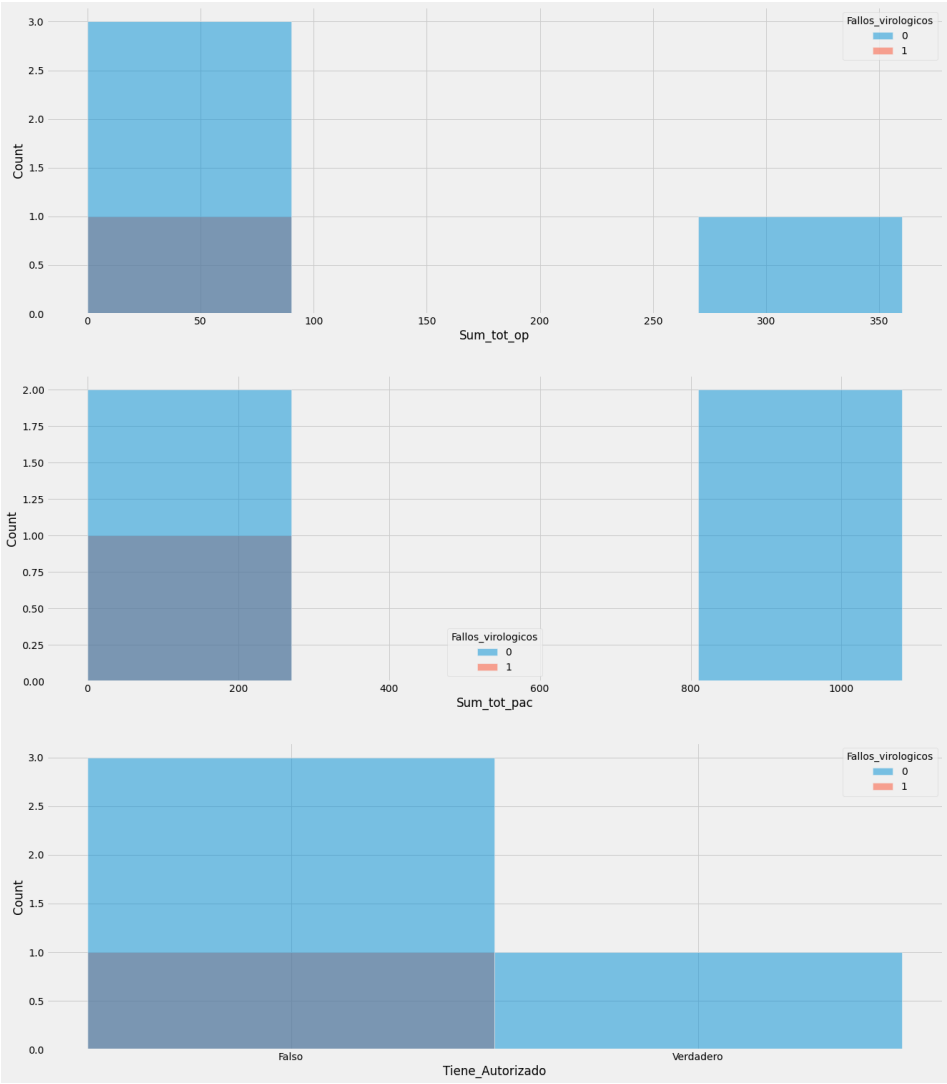

Figura 21. Recuento total pastillas oportunistas, tratamiento TAR y tiene acompañamiento Adherencia categoría 1. Elaboración propia.

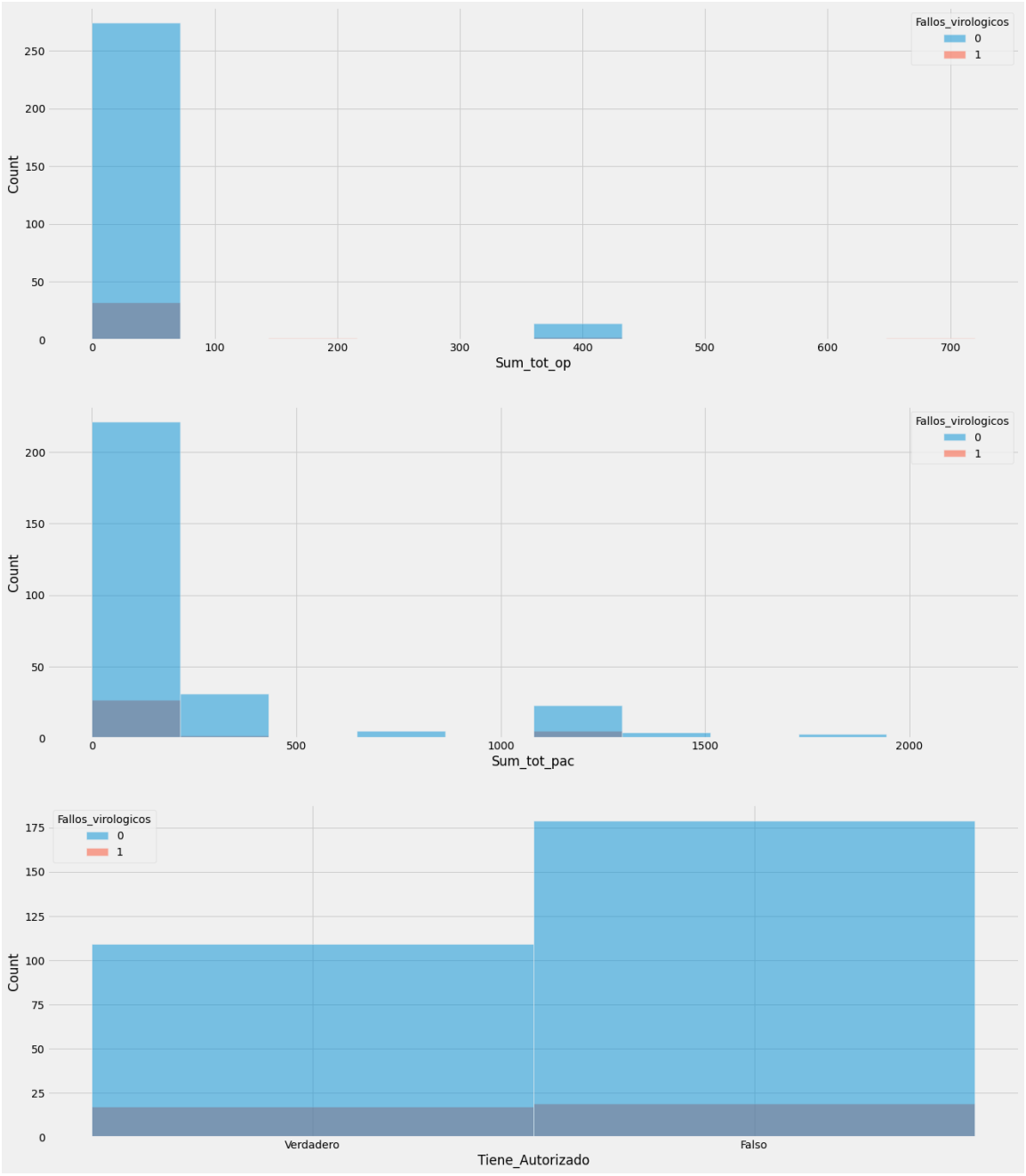

Figura 22. Recuento total pastillas oportunistas, tratamiento TAR y tiene acompañamiento Adherencia categoría 3. Elaboración propia.

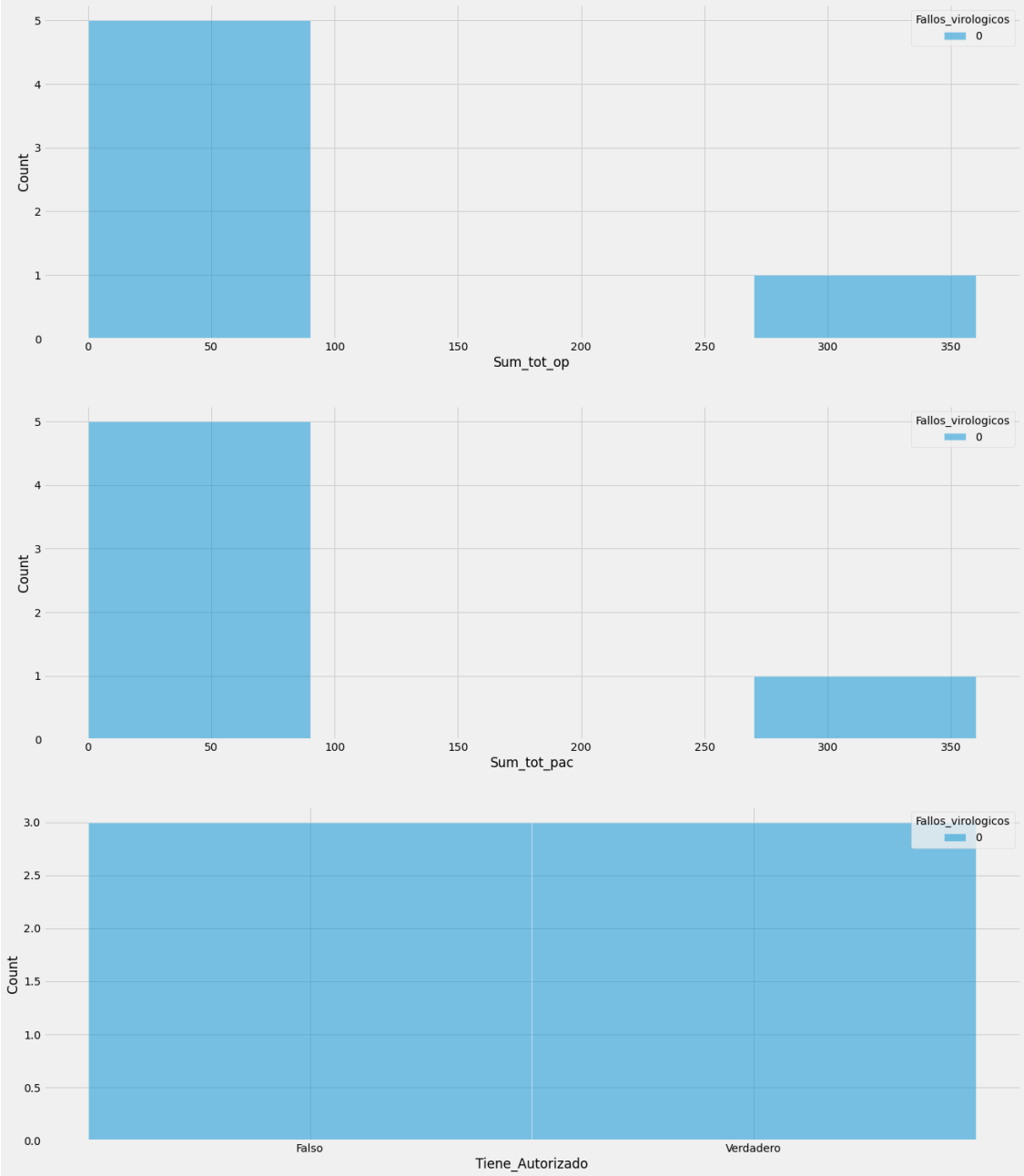

Figura 23. Recuento total pastillas oportunistas, tratamiento TAR y tiene acompañamiento Adherencia categoría 5. Elaboración propia.

Por otra parte, al analizar el sexo de los pacientes y su orientación sexual según la proporción de fallos virológicos que tienen en base a las categorías (Observar Tabla 15 y 16) se aprecia que existe un mayor porcentaje de fallos virológicos en hombres antes que en mujeres y, por otra parte, también existe un mayor caso de fallos en personas con orientación heterosexual. Lo anterior por supuesto omitiendo el caso donde no se da la orientación sexual de la persona para el registro.

Tabla 15. Proporción de fallos virológicos según sexo

Sexo\ Fallos virológicos	0	1	Proporción de fallo
F	24	2	0.077
M	264	34	0.11

Tabla 16. Proporción de conducta sexual según sexo

Conducta sexual\ Fallos virológicos	0	1	Proporción de fallo
Homosexual	66	7	0.095
Heterosexual	17	2	0.105
Bisexual	10	0	0
Sin dato	195	27	0.12

5.4 Elaborar un prototipo donde se visualicen los resultados

Con la finalidad de representar la data de mejor forma, se realiza un *mock up* de la visualización posterior. Entendiendo que la etapa de visualización del proyecto como tal no se da en los plazos en que se enmarca la memoria.

Para el *mock up* se tiene como objetivo mostrar una posible visualización del tema estudiado en la memoria, adherencia al tratamiento TAR en pacientes VIH+ con relación al consumo de alcohol y otras drogas.

Se plantea una visualización web que conste de las siguientes partes:

a) Un ingreso con clave para los usuarios, el personal médico de la Fundación Arriarán. La finalidad de esto recae en mantener la privacidad de los datos al público, hecho por el cual se pedirá una identificación. Específicamente se recomienda el uso de una autentificación de dos pasos, esto quiere decir que al ingresar con el usuario y contraseña a la página se envíe un mensaje a un dispositivo autorizado del usuario con una segunda clave. Esta precaución se toma debido a la sensibilidad de los datos y la recomendación en una etapa de validación con el usuario. (Ver Figura 24)

Figura 24. Página de ingreso al sitio de visualización. Elaboración propia

b) Inicialmente se espera poder tener una selección de cuál es la información a la cual se desea ingresar, ya sea gráficos e indicadores de forma general o específicamente de datos directos de consumo.
Para datos de información general del paciente se tiene gráficos de algunas variables de interés, como el segmento etario en que se encuentra, la orientación y la cantidad de pacientes que se encuentran en cada una de las distintas categorías de adherencia. Además, se debe permitir el cruce de variables en las gráficas a través de filtros tales como sexo, categoría de adherencia, medicamentos con el cual se encuentra y la posibilidad de buscar específicamente a un paciente a través de su código. Estos filtros y el cruce de información que serían más ventajoso en un aspecto visual son discutido con la Dra. Claudia Cortez quien es parte del segmento de usuario de la visualización (Ver Figura 25).

Figura 25. Página 1 de visualización. Elaboración propia

c) Como se ha dicho en los capítulos anteriores para el momento de escritura de esta memoria no se cuenta con los resultados de la encuesta realizada a los pacientes, sin embargo, se sugiere en una posible visualización, mostrar la cantidad de pacientes que están dentro de las categorías de frecuencias según las preguntas específicas de interés para consumo de alcohol y otras drogas. Además de algunas métricas de interés, como la proporción de pacientes que se encuentran en las categorías más críticas de consumo (Ver Figura 26).

   Al igual que con las visualizaciones anteriores se sugiere filtros por código CONASIDA (código del paciente) y categoría de adherencia.

Figura 26. Página 2 de visualización. Elaboración propia

d) Se recomienda un sistema de ingreso de datos en la plataforma que ayude a la renovación de data necesaria para los predictores. Se debe asignar un control de ingreso de la información a través de opciones para los campos que son categóricos y autentificación de tipo de data para los ingresos numéricos. Se propone el tener como campos solo aquellas variables utilizadas en la construcción de las tablas para los 4 predictores con la finalidad de optimizar el ingreso de la información (Ver Figura 27).

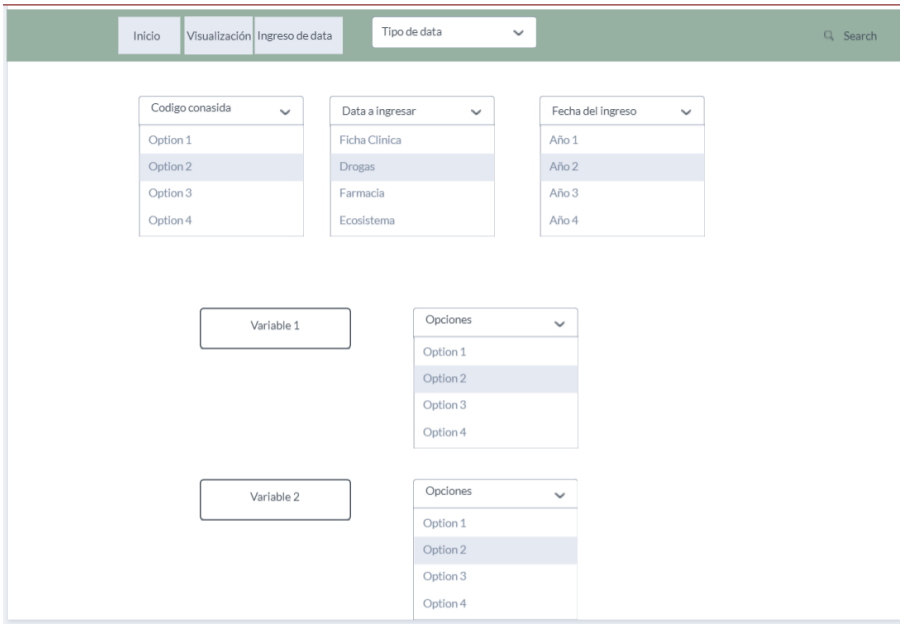

Figura 27. Página 3 de visualización. Elaboración propia

5.5 Análisis del impacto social y económico del proyecto

El trabajo de título presentado ha buscado predecir como la adherencia al tratamiento de VIH/SIDA se ve afectada en pacientes de vulnerabilidad extrema como son aquellos aquejados por la adicción a drogas y alcohol. Por lo mismo, entender como esto tiene un impacto primero en lo social y segundo en lo económico a nivel regional y luego nacional es clave para comprender la motivación de todo el análisis realizado.

Ni el proyecto FONDEF, ni la presente memoria trabaja sobre análisis y medición de variables que generan un impacto social y económico directamente, por lo tanto, los datos a este respecto son limitados en su contenido, así como también en su actualización. Por lo mismo, el análisis que se desarrollará en esta sección presenta diversos supuestos sobre la población a la que se impacta, así como ciertos alcances que deben realizarse para que resultado obtenido tenga sentido en un contexto global. Por lo mismo, algunas precisiones con respecto al trabajo a realizar en esta sección:

- Estimación del impacto social y económico que conllevan hoy para las vidas de estos pacientes el vivir con esta condición, así como el sistema de salud (tanto

público como privado) para tener una dimensión global del problema que se está abordando
- Adicionalmente, se hará un análisis más acabado del costo de transformar un paciente (de este segmento) de no adherencia a adherencia para así entender si es tanto social como económicamente beneficioso, así como presentar presupuestos que el Estado y las diversas instituciones que tratan con este problema, deben contar para definir una política público-privada efectiva,

Finalmente, es importante comentar que no se cuenta con toda la información necesaria para poder estimar con la mayor precisión posible los números antes presentados en esta metodología analítica. Por lo mismo, se tomarán diversos supuestos y se realizarán extrapolaciones de información en base a la información que la fundación Arriaran, ONUSIDA y otras fuentes de datos han entregado a este proyecto. Cada vez que uno de estos supuestos sea presentado al lector, las condiciones y sus implicancias serán comentadas para que sea claro como esto puede ser replicado o debatido cuando se cuente con más información en trabajos futuros.

5.5.1 Impacto social

Dada la naturaleza de los pacientes que son parte de este estudio, su situación de vulnerabilidad. Si se combina con la subyacente adicción a diversas sustancias, se encuentra un grupo de personas que presentan grandes desafíos sociales que es clave poder apoyar como sociedad.

Como se ha comentado en secciones anteriores, la adherencia al tratamiento retroviral tiene excelentes resultados en la sobrevida de los pacientes VIH positivos, ya que con 6 meses continuos de tratamiento se logra frenar el avance del virus en el sistema inmune del paciente eliminando la probabilidad de muerte por SIDA si dicho tratamiento se mantiene. Por lo mismo, la recomendación médica, de la Organización Mundial de la Salud, así como toda la investigación médica sugiere que trabajar en que la adhesión al tratamiento sea alta es uno de los factores claves, junto con el testeo, para asegurarse de entregar una mejorar calidad y experiencia de vida a los pacientes. Esto no es beneficioso solo a un nivel personal, sino que global ya que mantener un tratamiento de manera sistemática permite:

- Llegar a niveles no transmisibles del virus, lo que anula la probabilidad de continuar expendiendo el virus a más personas.
- Una adherencia alta previene el surgimiento de enfermedades oportunistas que repercute en mayor asistencia médica lo que pone más presión a un sistema de salud, tanto privado como público, estresado por la falta de profesionales y recursos.

De la población VIH positiva en Chile, según el estudio de “Calidad de vida en personas viviendo con VIH/SIDA” [31] se define como Alta Adherencia Global cuando se cumple en dosis y horario del tratamiento alcanza a un 69.6% de la población estudiada. Por lo mismo, y tomando los datos más recientes de la población global con VIH en Chile de ONUSIDA [3], que corresponde a 77 mil personas en 2020, la población en riesgo por la no adherencia global es casi de 23,408. Esta estimación es conservadora ya que estima en base al cumplimiento total del tratamiento, tanto en dosis como horario pero común en pacientes con prevalencia en consumo de alcohol y drogas que su seguimiento en ambos indicadores es menor que el resto de la población. Si se analiza el caso donde la dosis es correctamente administrada pero el horario no es cumplido, el mismo estudio de MINSAL indica que dicho número llega al 62.7% de la población, o sea, ya así aumenta la cifra a 28.721 pacientes en 2020. Este dato de por sí ya es alarmante y más aún si se hace la proyección a 5 años considerando los efectos del COVID-19 en el control y seguimiento de la pandemia de VIH/SIDA. Para hacer dichas proyecciones no se cuenta aún con la información suficiente para entender en qué medida cambiarán las tendencias de contagio y muerte asociada a VIH/SIDA pero se espera un aumento en ambos conceptos. De hecho, en el documento de la ONUSIDA sobre prevención de infección de VIH y su disrupción y adaptación durante la pandemia de COVID-19 [35], se comentan de 2 grandes efectos a esperar:

- Mayor cantidad de muertes y complicaciones por interrupciones del tratamiento de VIH/SIDA durante la pandemia de COVID-19: Esto por la falta de recursos en países en vías de desarrollo o de menores recursos, así como falta en el seguimiento, campañas y seguimiento de los pacientes por las condiciones de cuarentena que las ciudades del mundo han estado, así como los cambios de las prioridades sanitarias que han debido adoptar los gobiernos del mundo durante la pandemia. De hecho, se menciona que solamente en el sector sub-arábico de África se espera un aumento de 500 mil personas [36] que mueran por complicaciones relacionadas de VIH/SIDA durante 2021/2022 por este efecto COVID-19 y sus repercusiones en el acceso y seguimiento a tratamiento.
- Aumento de nuevos contagios por condiciones derivadas de COVID-19: Desde 2010 se había observado una disminución de nuevos casos de un 23% a nivel mundial, pero se espera que dicho número se revierta en los próximos años principalmente por efectos de una disminución del testeo preventivo de VIH al existir nuevas prioridades sanitarias, precarización de poblaciones de riesgo, así como barreras adicionales para acceder a atención médica oportuna durante la pandemia actual.

Por lo mismo, el mantener altos niveles de adherencia tiene un impacto enorme a nivel social y en el “*wellbeing*” (bienestar) de las personas. Además, esto repercute directamente en la vida de las personas en el corto plazo respecto al contagio y propagación del VIH/SIDA, así como con el prejuicio y percepción social de la misma. Tal

y como se comenta en [37], numerosos avances se han hecho en los últimos años en la percepción de la vida con VIH/SIDA pero aún existe un largo camino que recorrer. Aún no se consiguen objetivos específicos para controlar la enfermedad, así como la marcada desigualdad en el acceso a tratamiento y prevención en regiones de menores ingresos y en países en vías de desarrollo, como lo es Chile. Esto especialmente es relevante en poblaciones vulnerables donde la discriminación y estigma de vivir con VIH es aún alto y puede verse aún muy exacerbado por el avance de otras pandemias, como COVID-19.

5.5.2 Impacto económico

Tal y como se comentó al inicio de esta sección se trabaja con pacientes VIH+ de vulnerabilidad extrema como son aquellos aquejados por la adicción a drogas y alcohol. El impacto económico puede tener diversas aristas por lo que el análisis está enfocado en las siguientes componentes:

- Para personas VIH positivas que se encuentran en tratamiento: Costo de recibir dicho tratamiento
- Para personas VIH positivas que no se encuentran en tratamiento: Costo médico y de hospitalización derivado de las enfermedades que conlleva dicha condición de no tratamiento en el caso de VIH/SIDA
- Para pacientes VIH positivos con adicción a drogas o alcohol: Costo de mantener a dicho paciente dentro del tratamiento.

5.1.2.1 Costos de recibir TARV

De acuerdo con el informe de Isapres de Chile de noviembre de 2018 [32], para dicho año, estimaron que 75% de los casos de 2018 (7.020) fueron diagnosticadas y tratadas por Fonasa, dejando el 25% restante para las Isapres. En la sección 5.5.1 se comentó sobre las dificultades para estimar las proyecciones de contagios y seguimiento del tratamiento de VIH/SIDA a nivel mundial y nacional dado el incierto impacto de COVID-19 en la población. De hecho, tal y como se comenta en [38], uno de los puntos clave para el control de la pandemia de VIH/SIDA es el control de la desigualdad social y de acceso a salud, tema que se ha visto perjudicado en los últimos 2 años. Igualmente, tal como se menciona al final del punto 5.5.1, si se tomase el supuesto de que 2018 es un año representativo de la situación global de Chile con respecto a VIH, y que no hubiese existido el efecto del COVID-19 se podría suponer un crecimiento normal de contagios, llegando a valores esperables para el 2022 de 77 mil personas contagiadas en Chile. De ellas, sólo 54,000 conocen su condición y están bajo un tratamiento antirretroviral [3,33]. Por lo mismo, los números más realistas de personas que están recibiendo algún nivel de tratamiento para el VIH estarían 40.500 de ellas en FONASA y 13.500 en el sistema de Isapres. Siguiendo las proyecciones hipotéticas de contagio, el total de pacientes en

tratamiento serian de 71.400 personas en 5 años más y por lo mismo, la proporción de FONASA e Isapre pasará a ser 53.550 y 17.850 personas respectivamente.

En el mismo informe de Isapres se comenta que, del total de gastos que hace el Estado para suministrar medicamentos a los centros de salud públicos, el 11% ($58 mil millones) se destina a la compra de las principales terapias contra el VIH. Por ende, el gasto por persona del Estado es de 1.432.098. Por su parte, en el caso de las Isapres, al menos para datos de 2017, desembolsaron un total de 42 mil millones de pesos para sus afiliados, quienes eran un total de 8.200 personas, por ende, un valor 5.121.951 por cada uno de ellos. Dado lo anterior, el gasto total de la sociedad chilena, estimado para año 2020 se observa en la tabla 17, el que asciende a 127.146.307.500 pesos chilenos. Adicionalmente, asumiendo la tasa de contagio actual, reportada por ONUSIDA, de 5 mil personas al año (con un intervalo de confianza de 4.200 a 6.000), en los próximos 5 años dicho monto pasará a ser 168,115,673,250, un aumento del 32% en dicho período de tiempo. Esto claramente es un escenario conservador ya que dicho número potencialmente será mucho mayor debido al efecto del COVID-19 en la población. Dicho efecto se apreciará en los próximos años y se deja como parte de los trabajos futuros el actualizar estas proyecciones.

Tabla 17. Proyección de costos por paciente

	Total pacientes (2020)	Total pacientes (2025)	Costo por paciente	Total (2020)	Total proyectado
Fonasa	40500	53550	$ 1,432,098	$ 57,999,969,000	$ 76,688,847,900
Isapre	13500	17850	$ 5,121,951	$ 69,146,338,500	$ 91,426,825,350
				$ 127,146,307,500	$ 168,115,673,250

Se debe tener en cuenta que el análisis anterior se realiza en torno a la data nacional que se posee.
La pregunta siguiente sería, teniendo en cuenta los datos arrojados por el predictor desarrollado en la presente memoria, que ocurre, Para evaluar específicamente la situación analizada en la presente memoria se debe tener en consideración los resultados del predictor en torno a la categorización de adherencia de los pacientes. Para las 3185 observaciones en el set de testeo se tienen las siguientes poblaciones ya mencionadas en resultados ítem 5.3.2:

- Categoría 1: [0,15%] Corresponde a 5 pacientes
- Categoría 2: [86,81%] Corresponde a 614 pacientes

- Categoría 3: [10,17%] Corresponde a 232 pacientes.
- Categoría 4: [2,67%] Corresponde a 67 pacientes.
- Categoría 5: [0,19%] Corresponde a 6 pacientes.

Para este se tomarán los pacientes en categoría 1 y 2 ya que son los que presentan mejor adherencia. Además, es donde el gasto en el tratamiento TARV se está concentrado mayormente. Dado esto, considerando el gasto promedio de los pacientes FONASA e Isapre, se consigue que el costo total TARV para la población del predictor es de 2.028.478.16. Al ser los pacientes que presentan menor riesgo, si hay escaso dinero para invertir en salud, este total podría disminuirse ya que no es la población que presenta el mayor riesgo, aunque es numéricamente la mayor. Contrariamente, las categorías 4 y 5 representan a la población con un mayor problema de adherencia, por lo tanto, son las que generarán más gastos a largo plazo por paciente al igual que son los pacientes de mayor riesgo. Si realizamos el mismo cálculo para este segmento de pacientes, se tiene un gasto de 239.222.788 pesos chilenos, el cual sería el gasto que aumentaría al invertir en mantener esta población en tratamiento. Entonces, combinando ambos resultados el predictor podría recomendar donde invertir para que la perdida sea menor y el éxito sea mayor. Es decir, invertir el dinero en la población de mayor riesgo a mediano plazo es la opción más económica y la mejor para la sobrevida de las personas VIH+.

5.1.2.2 Costos médicos y hospitalarios derivados del VIH/SIDA

Si se enfoca primero el análisis en las hospitalizaciones y sus costos se puede referir al estudio de gastos en VIH/SIDA de 2005 [31] el que detalla que directamente las enfermedades oportunistas representaban el 0.9% de las Cuentas Nacionales. En particular, se da el caso del hospital San Borja con 30 incidencias en el año de estudio y los costos que ello conlleva por tipo de enfermedad.

Morbilidad	Nº casos
Acidosis Lactica(*)	0
Candidiasis orofaringea	2
Linfoma	4
LEMP	3
MAC	1
Meningitis por criptococcus.	3
PCP	7
Sarcoma de Kaposi	4
Toxoplasmosis	1
Herpes Zoster	1
CMV	4
Tuberculosis	0
Nº hospitalización por I.O	**30**

Figura 28. Casos de enfermedades oportunistas en el hospital San Borja (2005)

Para cada uno de esos casos, se incurrieron en gastos de días cama, interconsultas, exámenes tanto de rutina como especiales. Estos gastos se observan en detalle en la figura 29, los cuales son 307.708.000 pesos chilenos.

Si es llevado a una base por paciente, da un costo de 10.256.933 por cada uno de ellos. Este costo está sobreestimado ya que considera todas las prestaciones necesarias para un paciente que presenta complicaciones médicas derivadas de enfermedades oportunidad. Si se considera el promedio de un caso que se presenta se puede obtener el ejemplo más representativo de la situación. Esto sería un paciente que necesita exámenes de rutina y parcialmente días cama. En el reporte se considera para enfermedades oportunistas con mayor complejidad, que es correcto asumir que 30% del costo de día cama representa más precisamente la condición de un paciente promedio. Por lo mismo, el costo por paciente queda definido por 3.541.000 pesos chilenos.

Prestaciones	Toxoplasmosis	MAC	Candidiasis orofaringea	Linfoma	Meningitis Criptococcus	Neumonia Neumocistis Carinni	Sarcoma Kaposi	CMV	Herpes Zoster	LEMP	TOTAL
Días cama	245	931	319	173	254	81	184	306	172	1.059	7.020
Interconsultas	19	0	13	22	15	8	19	30	0	38	464
Exámenes de rutina	34	43	41	37	31	28	42	66	15	157	1.201
Exámenes especiales	31	60	35	69	68	17	78	94	0	126	1.571
Costo total	**328**	**1.034**	**407**	**301**	**367**	**134**	**323**	**496**	**187**	**1.379**	**10.257**
Nº casos año	1	1	2	4	3	7	4	4	1	3	30
Costo total casos año	**328**	**1.034**	**814**	**1.204**	**1.101**	**935**	**1.290**	**1.985**	**187**	**4.138**	**307.708**

Figura 29. Detalle de gastos y prestaciones en Hospital San Borja en base a enfermedades oportunistas (2005). Valores en miles

Igualmente, en el informe final de evaluación del programa nacional de prevención y control de VIH/SIDA y ETS [34] se observa que en dicho año de estudio (2009) sólo el 3% de las personas VIH positivas. Dadas las estimaciones históricas, en dicho año existían 34.000 personas viviendo con VIH en dicho momento, por lo que corresponden a 1.020 personas que tuvieron alguna incidencia asociada a enfermedades oportunistas u otras complicaciones por la enfermedad viral. Es importante mencionar que no solo no se cuentan con datos más recientes de dicha incidencia, sino que sabemos por lo que ha ocurrido en otros países, que los valores de la actualidad y a futuro próximo incrementaran exponencialmente. La situación presentada por el COVID19 ha generado no solo una reducción del acceso a test de VIH sino también una demora en el tratamiento debido a colapsos en hospitales, un ejemplo claro de las disyuntivas generadas por la situación global se ve reflejada en China donde de la gente viviendo con VIH un 32,6% se encuentra en riesgo de discontinuar el tratamiento y un 48,6% no sabe dónde encontrar los médicos de tratamiento TAR en el futuro predecible [40]. No obstante, si a modo de mera especulación se asume que para la actualidad dicho valor continúa en el mismo monto para los datos de 2020, del total de personas que saben que cuentan con

VIH/SIDA (54.000), 1.620 personas han tenido alguna complicación hospitalaria o médica durante el año en curso. Tomando el costo por persona estimado en la sección anterior, se llega a un monto total de 16.616.231.460 pesos chilenos y proyectados a 5 años llega a 21.970.350.486. En la realidad este número será mayor quizá el doble o más.
La pregunta siguiente sería, teniendo en cuenta los datos arrojados por el predictor desarrollado en la presente memoria: ¿Qué ocurre? Llevando el análisis a la situación particular del trabajo de memoria y nuevamente llevándolo a los predictores realizados en este trabajo, así como la población estudiada a través de la sección de análisis y resultados, se tiene que en las categorías 4 y 5 se encuentran aquellos pacientes con un menor porcentaje de adherencia, y consecuentemente, son aquellos más propensos a una atención hospitalaria. Teniendo el estimado de un costo de 10.256.933 por cada persona hospitalizada con relación a VIH/SIDA, nos da un costo estimado de 748.756.109 pesos chilenos para la población de estudio de los predictores.

5.1.2.3 Costos de programas para adherencia al tratamiento en pacientes con adicciones

Dado el universo de personas en tratamiento retroviral y quienes no han podido mantener dicho tratamiento en el tiempo, se han diseñado diferentes políticas públicas, así como ONGs y privadas para apoyar al paciente y aumentar la adherencia de ellos a dichos tratamientos. En particular, si el análisis se enfoca en la población vulnerable que tiene adicción a drogas y alcohol debe entenderse el gasto total en los esfuerzos de adherencia de la población particular.

Si se observa el informe final de evaluación del programa nacional de prevención y control de VIH/SIDA, en sus apartados de “Estrategia de Prevención” y “Atención integral” existen subprogramas específicos para la farmacovigilancia de los pacientes en el programa nacional, así como actividades de consejería individual y de prevención de dejar los fármacos. En particular, se utilizan los puntos 1.2, 1.3, 1.5, 2.3 y 3.1 de la figura 30, con lo que se alcanza un monto 828.520.000 pesos chilenos. Esto ha tenido un crecimiento de 34% promedio en 4 años en el período de 2006 a 2009 y si se toma el supuesto hipotético de un crecimiento similar hasta 2020, con lo que corresponde a 1.993.505.000 pesos y en 2025 debería estar en 2.671.297.000 si la progresión continúa en la misma tasa en los próximos años. Esto es sólo para ejemplificar este impacto en las proyecciones de los próximos años ya que el efecto real puede ser potencialmente más alto por el efecto de la pandemia COVID-19 en las tasas de contagio de VIH, así como en su tratamiento dentro de la población. Dicho efecto no es aún medible, pero es clave tener en cuenta dicha condición para poner en perspectiva las proyecciones hipotéticas planteadas en esta sección.

Componentes y Subcomponentes		2006	2007	2008	2009	Var 2006-2009
1.- Estrategia de Prevención		**909.581**	**1.086.540**	**732.082**	**1.215.102**	**34%**
1.1	Campañas de Comunicación Social Educativa	476.651	461.077	123.587	709.295	49%
1.2	Proyectos de Prevención Grupal	261.807	267.159	297.209	140.791	-46%
1.3	Consejería Individual (Corresponde sólo a cara cara, RM)	41.390	34.285	55.703	s/i	35%
	Consejería Individual (Estimado nacional)	125.698	182.873	180.689	221.495	76%
1.4	Disponibilidad Condones	4.035	139.428	74.894	143.521	3%
1.5	Capacitación y asesoría técnica	0	1.718	0	0	
2. Atención Integral		**15.420.702**	**15.206.721**	**11.400.995**	**21.164.774**	**37%**
2.1	Detección y Diagnóstico VIH- Sífilis. Establecimientos de Servicios Salud	1.411.322[46]	2.988.592	2.294.868	2.622.250	86%
2.1	Detección VIH –FONASA (embarazadas)	415.705	465.619	319.789	353.910	-15%
	Controles de Salud Sexual realizadas por matrona o enfermera en establecimientos de nivel secundario de atención	26.498	24.231	26.498	19.832	-25%
	Manual de Procedimientos y Sistema Registro Único	0	0	0	63.160	
2.2	Atención Integral PV-VIH/SIDA-FONASA (medicamentos ARV, Exámenes (cargas virales, linfocitos y genotipos) y medicamentos especiales para tratamiento enfermedades oportunistas en niños y adultos).	12.810.975	10.974.880	7.973.714	17.160.170	34%
	Atención integral PVVIH en Establecimientos-Servicios de Salud CON TARV (Médico y Mat./Enf.) / 6 consultas anuales c/u	330.792	378.354	386.304	475.984	44%
	Atención integral PVVIH en Establecimientos-Servicios de Salud SIN TARV (Médico y Mat./enf.) / 3 consultas anuales c/u	66.290	60.474	63.218	62.002	-6%
2.3	Atención Integral Personas consultantes de ITS en Establecimientos- Servicios de Salud (Medico y matrona consultas promedio p/paciente)	359.120	314.571	326.098	322.713	-10%
2.4	Capacitación y Asesoría Técnica en VIH/SIDA- Plan 90 días, Nivel Central CONASIDA	0	0	10.506	82.821	
2.4	Capacitación y Asesoría Técnica en VIH/SIDA-Programa Formación y Capacitación 2009, Subsecretaría Redes Asistenciales	0	0	0	84.753	
3.Vigilancia Epidemiológico y Estudios						
3.1	Seguimiento Epidemiológico	s/i	s/i	s/i	s/i	
3.2	Estudios en VIH/SIDA e ITS	s/i	s/i	s/i	s/i	
4 Estrategias de Cooperación						
4.1	Acuerdos de Cooperación	s/i	s/i	s/i	s/i	
Gastos administración (Considera gasto operacional nivel central y Secretarías Regionales Ministeriales de Salud)		126.025	120.752	234.239[47]	123.392	-2%
Otros gastos de Administración (Considera gasto operacional nivel central y Secretarías Regionales Ministeriales de Salud)		4.873	0	0	0	
Recurso Humano no distribuido por componente(Considera RRHH CONASIDA, FONOSIDA y Secretarías Regionales Ministeriales de Salud)		94.558	138.871	305.730	445.823	371%
Total		**16.555.739**	**16.552.884**	**12.673.046**	**23.031.911**	**39%**
	Gasto Total sin considerar estimaciones (corresponde a gasto de la Subsecretaría de Salud Pública y FONASA)	**14.236.040**	**12.603.790**	**9.359.371**	**19.222.733**	**35%**

Figura 30. Detalle de gastos del programa nacional de prevención y control de VIH/SIDA. Los valores están en miles de pesos.

Los pacientes que dichos programas alcanzaron durante 2009, en el apartado de capacitaciones, asesorías y seguimientos fue un total de 18.057 personas por lo que entrega un costo por persona de 110.400.

5.1.2.4 Estimación hipotética de costos totales y beneficio económico del estudio

Dado los puntos anteriormente calculados, el trabajo de memoria presentado en este informe busca entregar herramientas predictivas para entender cuando un paciente, influenciado por su consumo de drogas y alcohol puede abandonar dicho tratamiento retroviral. Por esta razón, el impacto económico de tener esta información y evitar que ocurra es esencial.
A modo de proyección para 2020 en relación con costo por paciente:

- Costo de un paciente que mantiene su tratamiento retroviral:
    - $2.354.561 anuales si se toma en promedio entre FONASA e Isapre. Esto se llamará desde ahora *CDi*
    - $110.400 anuales por actividades de seguimiento y consultoría. Esto se llamará desde ahora *CSCi*
- Costo de un paciente que no sigue el tratamiento retroviral:
    - $3.541.000 asociados a gastos hospitalarios por enfermedades oportunistas. Esto será llamado desde ahora *GHi*

Como fue comentado en los puntos anteriores, es muy probable que se esté analizando un escenario optimista de dichos costos y beneficios ya que no es posible medir el efecto de COVID-19 en esto, pero de todas maneras los costos serán muy superiores a los analizados en esta sección. También cabe mencionar que existen efectos de largo plazo en tener una alta adherencia al tratamiento retroviral como es la disminución de la transmisión del virus. Dichos efectos quedan excluidos de este análisis ya que son de más de largo plazo y tienen otros factores a analizarse que no son parte del alcance de este análisis.

Dado lo anterior, el beneficio de convertir un paciente que no sigue su tratamiento retroviral a uno que si lo hace sigue el patrón de beneficios siguiente:

$$Bi = GHi - CSCi - CDi$$

De dicha fórmula se obtiene que el beneficio por paciente que se mantiene en su tratamiento es de 1,076,039 pesos chilenos. Como se comentó en la sección anterior, la adherencia actual en dosis y horario es de 69.6% en el país por lo que 16,416 personas conviven y conocen su condición de VIH+ y no tienen un tratamiento retroviral correctamente administrado.

De este 30.4% de personas que no han seguido su tratamiento existen diversos motivos para los mismos. En este trabajo de memoria en predecir como el alcohol y otras drogas afectan a dicha adherencia y el impacto total se debe medir de cómo dicho grupo se ve beneficiado de predecir su no adherencia y potenciar su salud. Para ello, se cuenta con una encuesta realizada por la fundación Arriaran y el WIC sobre 800 pacientes VIH+ se observó que el 15% de ellos tienen un problema de consumo de alguna droga y alcohol [Anexo B]. Es importante recalcar que unos 200 pacientes estan dentro de los 2484 utilizados para realizar las predicciones en la presente memoria.

Por lo mismo, del total de pacientes que no está siguiendo su tratamiento (16.416 personas) se tomará el supuesto de que el 15% de consumo de drogas y alcohol distribuye homogéneo tanto en pacientes adheridos y no a su tratamiento, por lo que 2,462 es el resultante de personas que no siguen su tratamiento por causa de dicho motivo. Si esto se lleva a un beneficio económico, corresponde a 2,649,208,018 de pesos

si lograra convertir a todos y cada uno de ellos en pacientes con adherencia a su tratamiento. Si se estiman escenarios donde este modelo es más o menos exitoso en predecir la salida de un paciente de su tratamiento por el efecto de drogas y alcohol, se pueden simular diferentes escenarios de cómo este beneficio distribuye

Tabla 18. Beneficios proyectados según porcentaje de pacientes

Porcentaje de pacientes	Beneficio 2020 (millones)	Beneficio 2025 (millones)
25%	662	932
50%	1,324	1,864
75%	1,986	2,796
100%	2,649	3,728

Dado todo lo anterior, el beneficio económico de este proyecto se puede concluir que está estimado entre los 662 a los 2.650 millones de pesos en la actualidad a una proyección entre los mil y 3.730 millones para 2025. Estos valores en la realidad serán mayores por todas las razones que se viene mencionando a lo largo de esta sección 5.5.

La pregunta siguiente sería, teniendo en cuenta los datos arrojados por el predictor desarrollado en la presente memoria, qué ocurre. Teniendo nuevamente en consideración el caso de interés, en base a el predictor desarrollado se va a considerar los pacientes que tienen una baja adherencia al tratamiento (categorías 4 y 5) y se les buscará estimar el costo y beneficio de moverlos a la categoría 1 y 2, los tramos que dice el predictor donde existe una mayor adherencia al tratamiento retroviral. Esta población es de un total de 73 personas estudiadas. Al considerar que el beneficio económico por paciente que se mantiene en su tratamiento es de 1,076,039 pesos chilenos tal como se dijo anteriormente, entonces el beneficio social especifico que entrega el predictor sobre la base de usuarios estudiada es de 78.550.847 pesos chilenos.

Con esta información, se pueden definir políticas públicas específicas, así como iniciativas privadas que permitan mantener dichos beneficios al utilizar la información de los predictores para enfocar el esfuerzo en aquellos pacientes que realmente necesitan intervenciones efectivas tanto en forma como en tiempo para mantener su adherencia alta y así combatir los efectos negativos del VIH/SIDA en sus vidas, así como el impacto social que esto conlleva.

6.Discusión

6.1 Estandarización trabajo de datos

En las secciones 5.1 y 5.2 del presente trabajo se muestra como se ha construido una base de datos para el predictor de alcohol y otras drogas. Es importante recalcar en primera instancia los pasos generados para la concepción de las variables a utilizar. El proyecto por su naturaleza, liga componentes tanto de medicina como de ingeniería lo cual genera la necesidad de no solo un médico guiando el entendimiento de las variables sino también tener un *medical advisor* que permita en profundidad entender el cómo se relacionan estas entre sí.

Dicho lo anterior el primer paso necesario para la construcción de la base de datos es una recopilación bibliográfica, pues permite entender si se ha ligado previamente el nivel de consumo de drogas a la adherencia del tratamiento en pacientes con VIH. Como se ve en la tabla 2 (en ítem 5.1.1 de resultados) el tener este primer acercamiento lleva a ver variables en común en la bibliografía que sirve para el estudio. En el caso ~~de~~ del estudio particular de alcohol y otras drogas una de las variables más transversales y que se comparte en los estudios revisados es el nivel de consumo de alcohol y/o drogas, ya que nos entregan información directa de si el paciente tiene un problema de adicción. Es importante, sin embargo, recalcar la difícil obtención de datos de consumo, los cuales suelen obtenerse por un cuestionamiento directo al paciente en cuyo caso la data puede estar sesgada por lo que el paciente quiera compartir o si la información que proporciona sea verídica o no. De Boni et al., menciona la importancia de conocer los niveles de consumos y no solo el hecho de consumir o no, por ejemplo, a través de la cantidad de copas de alcohol consumida, ya que a un mayor nivel de consumo se observa un aumento en la probabilidad de no seguir en el tratamiento [29].

Por otro lado, se sabe que para el estudio de los niveles de consumo de alcohol y su impacto en la adherencia al tratamiento en una persona con VIH+, se requiere el obtener otras variables que entreguen un panel de información más completo de la persona, dentro de estas se destacan: la edad, niveles socioeconómicos, el grado de acompañamiento del paciente, entre otras, las cuales nos dan una idea del estado social del paciente.

Por último, es importante recalcar que cuando se realiza un estudio de adherencia, no solo para el estudio específico del caso de consumo de alcohol y otras drogas, se requiere de la obtención de data clínica como niveles de rna, niveles de cd4, los cuales entregan información de cuál es el estado del paciente con respecto a VIH. En este punto es importante entender nuevamente lo complejo de medir la adherencia del paciente y la

importancia de tener algunas métricas que más que observables sean un registro sistemático.

Se debe recordar que el trabajo no considera la data generada de las encuestas que se elaboraron por parte del WIC en conjunto con la Fundación Arriarán, ya que al día de la fecha aún no han sido respondidas por todos los pacientes incluidos en el estudio (diseño de subpredictores), por lo cual no contempla información directamente relacionada al consumo de drogas. En un trabajo posterior se recomienda el uso de variables como:

- Nivel de consumo de alcohol y otras drogas
- Tipo de drogas de consumo
- Frecuencia del consumo de alcohol y otras drogas

En base a la obtención de estas variables se aconseja el estudio de nuevas variables que nazcan de la interacción, un ejemplo claro seria la construcción de una variable que demuestre si el paciente posee en conjunto un nivel de consumo de pastillas más alto a la vez que muestra un problema de adicción. Y con relación a este último identificar los tipos de adicciones ya que no resulta lo mismo si la persona consume marihuana a si es cocaína, por ejemplo

La calidad de los datos por otra parte juega un rol importante en la construcción de una base de datos, se puede observar de las tablas iniciales que se entregaron por parte del WIC, con información proveniente de la Fundación Arriarán, que la cantidad de pacientes con la cual se comienza a estudiar supera los 5000. Sin embargo, resulta difícil caracterizar de forma completa a todos los pacientes debido a: i) la cantidad de data faltante en cada variable y ii) en las tablas, ya sea por la naturaleza de información que contienen o por la calidad del registro, se tiene información de distintos grupos de pacientes en cada una de ella. Para resolver esto, se toma la decisión de eliminar pacientes en base a la información disponible que se tiene para los pacientes luego de unir las tablas y ver cuáles de estos contaban con la información necesaria para los predictores. De esto se llega a un número preliminar de 3108 pacientes que se encuentran en la tabla ficha clínica máster, el cual fue posteriormente reducido hasta 2484 pacientes. Tener una data completa de una gran cantidad de pacientes mejora la calidad de las predicciones de los modelos.

Otro aspecto importante es la temporalidad de los datos, viendo específicamente las mediciones de cd4 y rna se encuentra que existe una discrepancia en la cantidad de mediciones que se tiene para cada paciente en un año. Esto se ve reflejado de la misma forma en otras variables que son de relevancia para el predictor, en base a esto se destaca la metodología utilizada de anualizar los datos, lo cual se hace ya sea tomando la media o la moda anual según la variable lo amerite, con la finalidad de obtener la mayor cantidad de información disponible para la construcción de la base para los modelos. El

ideal sería que estos datos eventualmente logren ser medidos de forma sistemáticamente semestral o acortar el horizonte de tiempo en que se miden, lo cual permitiría tener información aún más precisa para la toma de decisiones.

También se debe recalcar la construcción de la variable fallo virológicos, la cual nace en base a las mediciones de rna. Para la construcción de la variable se observa la variable rna, desde la cual se toman decisiones de la base en sí. Como se menciona en el capítulo de resultados, la variable de fallos virológicos se genera en base a dos observaciones seguidas con un valor mayor a 200 copias /mL, el tener este método de cálculo hace que sea necesario la exploración de los datos faltantes en la variable. Es esta revisión se observa la existencia de pacientes con datos tomados de cd4 en una fecha en específico, pero no de rna y viceversa, lo cual nos demuestra que el paciente no es alguien que ha abandonado el tratamiento, sino que no se ha registrado la medición. En base a la cantidad de pacientes con datos faltantes y la relevancia de estas variables, es la *medical advisor* quien toma la decisión de eliminar aquellos pacientes con más de un año y tres meses de data faltante en alguno de los campos.

Como se menciona en 5.1.2.1.2 se había considerado en un inicio construir el *label* como una variable que incluyese la relación entre los valores de cd4 y rna. Sin embargo, se tuvo en consideración al momento de la discusión el hecho de que no existe una relación directa como se podría esperar entre ambos valores, lo esperado es que al haber un aumento de los linfocitos cd4 vaya acompañado de una disminución de los niveles de rna, sin embargo, existen variados fatores que pueden incidir en el aumento de los niveles de rna sin necesariamente significar un cambio en cd4 tales como: una resistencia a medicamentos, los años que lleva en el tratamiento, el tiempo desde que se contagió. En la figura se puede observar cómo los valores de cd4 y rna no son directamente proporcionales.

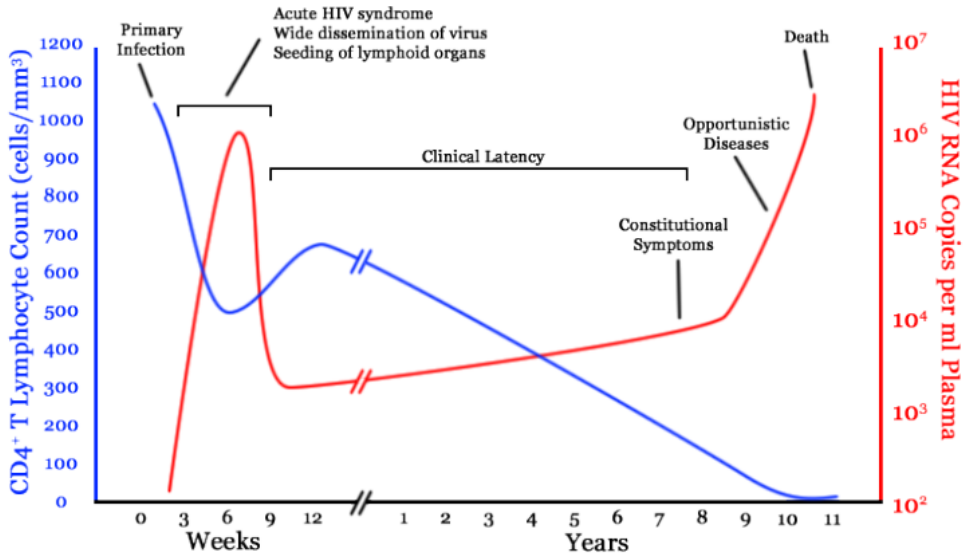

Figura 31. Cambios de cd4 y rna. Fuente: "Oscillations in a Model for HIV Infection with Three Intracellular", Mohamed Omari [41]

Finalmente se debe señalar los problemas a la hora de trabajar en Python variables como los nombres de los fármacos o enfermedad oportunista existente. Como equipo se decide la construcción de un manual para el uso de la plataforma de visualización, que tenga la forma estándar de ingreso de información para la estandarización de la data disponible.

6.2 Desarrollo de modelo de predicción

Se debe considerar como parte fundamental del estudio realizado el hecho de que no se cuenta con un *label* dado, sino que este debió ser construido en base a la información disponible y la evaluación de una experta. Existen problemas claros a la hora de elección de este, una primera iteración del trabajo realizado por el equipo del WIC contemplaba la construcción de una segmentación de los niveles de cd4 para este, sin embargo, el problema con ello es que implicaba la eliminación de la variable de registro de cd4, una variable que se plantea esencial para el estudio. Este problema fue solucionado con la construcción de la variable “fallos virológicos”

Luego de observar los resultados de las métricas en los modelos para el caso del primer *label*, “fallo virológico acumulado” en Tabla 11 y Tabla 12 (Ver resultados ítem 5.3.1), se ven mejores resultados en base a las métricas en los modelos de *XGBoost* y SVM, el primero es consistente con lo visto en estudios de salud digital, donde normalmente este modelo suele tener un buen rendimiento ya que procesa de forma eficiente la data faltante además de ser flexible en la capacidad de modificar sus parámetros [39].

A pesar de los buenos resultados obtenidos para el *label,* la decisión de que se genere una lectura anual del paciente lleva al cambio de *label* como se explica en el capítulo de

resultados (Ver ítem 5.3.1), esto permitiría una mejora a la hora de tomar decisiones por parte de la Fundación Arriarán, ya que entregaría información más precisa.

Los resultados de los modelos con el nuevo *label* observados en la Tabla 13 y 14 (Ver ítem 5.3.1), nos muestran problemas en una primera instancia de forma general en los modelos para la clasificación de las instancias, reflejado en las matrices de confusión, en la Figura 13, teniendo modelos como SVM la cual no posee una correcta medición de instancias verdaderamente negativas. Estos resultados se ven mejorados al ocupar el mecanismo de optimización de hiper-parámetros *grid search*, donde se ve una mejoría tanto de los casos verdaderamente negativos como en *recall,* haciendo que este modelo sea nuestra elección para el trabajo posterior de adherencia.

Finalmente, se debe recalcar la forma de medir adherencia en base al predictor, la categorización de esta en 5 grupos permite una fácil lectura por parte del usuario, personal médico, lo cual genera una mejora a la hora de tomar decisiones, ya que se identifica de forma rápida los pacientes que pertenecen a un grupo de intervención. No obstante, se analizan cada una de las variables y es importante mencionar que a modo de ejemplo se presentaron la exploración de algunas de las variables que se consideraron más significativas teniendo en cuenta el impacto de estas sobre adherencia, tales como:

- Segmento etario
- Años de enrolamiento en la Fundación Arriarán
- Número de pastillas consumidas
- Tener una red de apoyo para el retiro de medicamentos
- Analizar el sexo y orientación sexual de los pacientes según la proporción de fallos virológicos que tienen en base a las categorías de adherencia (Ver ítem 5.3.1, las Tablas 15 y 16)

Estos datos son importantes no solo porque nos entregan relaciones entre las variables, sino que ayudan a vislumbrar como se puede diseñar ciertos aspectos de la plataforma al combinar variables y encontrar patrones y segmentos de interés.

Finalmente se debe mencionar que en el punto 5.4 se plantea la elaboración de un prototipo donde se visualicen los resultados del predictor desarrollado durante la memoria. Este punto no se ahondará en discusiones ya que primero, es parte del objetivo FONDEF del cual forma parte el trabajo de memoria el elaborar una plataforma y segundo, ha sido discutido en el mismo ítem los resultados presentados. No obstante, el primer punto planteado, el prototipo diseñado será tenido en consideración a la hora de diseñar la plataforma final.

Respecto al análisis del impacto social y económico, se ha discutido ampliamente este punto en el ítem 5.5, lo que sería redundante repetirlo en esta sección.

7.Conclusiones y recomendaciones

El problema de adherencia a tratamiento en personas con VIH es crucial, no solo a nivel de salud individual, ya que como se ha explicado anteriormente una correcta adherencia al tratamiento permite una mejor respuesta inmunológica y menores tasas de mortalidad, lo cual genera mejores condiciones a nivel país al disminuir la población en riesgo y con ello disminuir además el gasto público.

En este contexto, se evalúa la generación de un predictor de adherencia en base a los datos de pacientes con VIH + en la Fundación Arriarán. Se generó un predictor con información específica de consumo de alcohol y otras drogas, en donde se utilizaron 58 variables, con información proveniente tanto de la consulta e historial del paciente como de laboratorio, generando un perfil acabado del paciente. Al comprobar distintos modelos de predicción se concluye que SVM es el mejor modelo para el caso, obteniendo un predictor con un *recall* de 0,75 y *accuracy* de 0,74. Los valores obtenidos son similares a los demás predictores del proyecto general. Estos valores son suficientes, considerando que esto es una predicción inicial, la cual mejorará a futuro con trabajos posteriores de ingreso de data estandarizada. Sin embargo, un *recall* de 0,75 permite la categorización de pacientes necesaria para que se generen modelos de intervención a corto plazo, permitiendo intervenir aquellos grupos de un nivel de riesgo de adherencia mayor y generar políticas de seguimiento a estos dentro de la fundación.

Luego del estudio de las interacciones entre las variables que conforman la base de datos de alcohol y otras drogas y su relación con tener o no fallos virológicos por parte del paciente, se puede ver una relación entre la acción de estas variables y la adhesión al tratamiento de VIH. Sin embargo, se debe considerar que, de la cantidad inicial de pacientes, se debió seleccionar una muestra de 2484 pacientes definitivos para el trabajo, que eran aquellos que contaban con un estado más completo de su data. Teniendo en cuenta que la cantidad de observaciones en el set de testeo es más reducida y, por tanto, con una menor cantidad de pacientes para la categorización son los utilizados, las interacciones que se observan son un primer acercamiento para ver la relación de variables de alcohol y otras drogas, mas no definitivas, las cuales han sido seleccionadas por su importancia en el tema. El trabajo posterior de generación de estandarización e ingreso de data permitirían tener una mayor cantidad de observaciones y permitiría ver de mejor forma la interacción y peso de las variables. Dado esto, se entiende que en un futuro tanto la estandarización de la data como el uso de la aplicación permitirá tener más información y reevaluar el estudio.

Haciendo una evaluación más puntual de los pasos seguidos para la construcción de los modelos se tienen:

Se analizan más de 50 variable simultáneamente, las cuales se encuentran tanto a nivel anual como por paciente. Esto permite un conocimiento acabado del paciente y un cálculo prometedor de la adherencia. Realizar un cálculo anual por paciente permite exitosamente que el médico genere modelos de intervención a corto plazo, con los cuales puede intervenir en los grupos de un nivel de adherencia riesgoso y evitar el abandono e incluso la muerte del paciente. Además, tal como lo mencionamos en el análisis del impacto social y económico, el uso del predictor genera una mejora social en un grupo vulnerable, además de la optimización de recursos y disminución de costos. No debemos olvidar que cada sub predictor analiza un número similar de variables, lo que hace no solo del presente sub predictor, sino de la plataforma, un sistema muy confiable a la hora de medir adherencia.

Se concluye que el modelo SVM el mejor modelo posible para la data disponible con la cual se cuenta, teniendo en cuenta el desempeño en las distintas métricas.

El trabajo bibliográfico permite tener una visión general de que variables son necesarias para una noción del trabajo a desarrollar a la hora de generar bases de datos para un modelo predictivo, sin embargo, para la adecuación de esto a un proyecto se requiere el conocimiento de expertos dentro del área para una correcta elección de las variables a utilizar, que resultaron mucho mayor en número que las sugeridas por la bibliografía analizada.

Los resultados obtenidos y la posterior categorización de adherencia realizada en base a estos es lo que permitirá a la Fundación Arriarán:

- Realizar intervenciones médicas por paciente según su categoría de adherencia y características particulares para un año determinado.
- Realizar intervenciones en los grupos más vulnerables, pudiendo priorizar aquellos segmentos de pacientes más propensos a presentar problemas de adherencia.
- Reducir el gasto público ya que se pueden optimizar los recursos en grupos específicos además de poder actuar de forma preventiva para el paciente, disminuyendo los costos hospitalarios por enfermedades oportunistas
- Disminuir la probabilidad de que el paciente tenga enfermedades oportunistas, mejorando el bienestar de estos.
- Mejora la atención y disponibilidad del personal de salud para actuar dónde y cuándo se requiera, optimizando la disponibilidad de recursos.
- Generar proyectos específicos a nivel comuna, región o país según corresponda sobre poblaciones más vulnerables, esto permite campañas con un foco más específico y por ende una mejor utilización de los recursos.
- Mejora social al lograr un mejor control de adherencia y con ello la potencialidad de frenar el contagio entre la población.

Es importante recalcar que se está trabajando en un tema sensible y causa de estigmatización social por lo que algunos datos son más difíciles de obtener que otros. Como mencionamos anteriormente, se debieron tomar numerosas decisiones en base a la data disponible tomando en consideración tanto el juicio médico como ingenieril, específicamente en base a la cantidad de data faltante en algunas de las variables de interés para los distintos sub predictores, lo que llevó a eliminar pacientes. Se recomienda en base a lo mismo y, teniendo en consideración la gran cantidad de información con la cual cuenta de la Fundación, enfocar esfuerzos futuros en la recolección de las variables seleccionadas para el uso de los modelos en los cuatro sub predictores.

Particularmente, respecto a los datos de consumo de alcohol y otras drogas se plantea la dificultad de la obtención de tres variables específicas de los cuales no se tienen todos los datos completos de todos los pacientes a la fecha: cantidad y nivel de consumo, necesidad de consumo y frecuencia de consumo. Estos datos se recolectan mediante la generación de un cuestionario mencionado anteriormente. Sin embargo, a la fecha de la memoria esta no ha sido contestada por todo el espectro de los pacientes. Se recomienda mantener registro de estos datos debido a la importancia en el predictor de alcohol y otras drogas, no solo por el valor de la data en forma directa sino también por las interacciones que se pueden lograr con las demás variables para un estudio más profundo. Aparte del cuestionario, se podrían realizar pruebas respiratorias para conocer los niveles de consumo dentro de los pacientes, sin embargo, se debe considerar que al igual que la encuesta, la implementación de estas debe de carácter voluntario. No obstante, con las variables que están incluidas en el predictor, se tiene que más de la mitad están asociadas directamente al consumo.

El trabajo generado en base al predictor debe pasar por dos prontas etapas siguientes para el trabajo futuro:

1.- Se debe unir las variables relacionadas a consumo de drogas a la base de “alcohol y otras drogas”, con lo que se pueden tener estudios más acabados en torno a adherencia en base a consumo.

2.- Realizar una contraprueba en un inicio con un paciente VIH+ nuevo con características similares a los pacientes de la Fundación Arriarán y luego con grupos de pacientes de las mismas características que se tengan datos previos de abandono y retorno a tratamiento y otros que no hayan tenido problemas de adherencia para apreciar el comportamiento del predictor.

3.- La generación del prototipo de la plataforma a poner a disposición en la Fundación Arriarán, para el personal de este, de forma que la información se encuentre disponible de una manera sencilla y didáctica para el seguimiento del estado del paciente. Todos los nuevos datos de los nuevos pacientes serán manejados directamente desde la

plataforma. De esta manera también se eliminarán problemas en base a la falta de estandarización del proceso de ingreso de data, lo cual ralentizó el proceso de construcción de las tablas del presente trabajo, ya que el equipo resolvió generar a futuro (durante el corriente año) un protocolo de manejo de la plataforma que incluye un primer capítulo sobre ingreso de datos a esta. En la sección 5.4 del presente escrito, se plantea generar métodos de corroboración de la información ingresada en la pronta visualización que se va a construir.

Específicamente para el caso del predictor de “alcohol y otras drogas” se puede obtener más información con la estandarización de enfermedades oportunistas y hospitalizaciones del paciente, además de variables de información general como nivel educacional y comuna.

Además, se ve la oportunidad de usar este modelo no tan solo en distintas partes en Chile, gracias a la metodología de trabajo y su capacidad de ser replicado teniendo la data disponible, sino también a nivel Latinoamericano gracias a las similitudes que se comparten entre los distintos países de habla hispana y el interés a nivel global de controlar el contagio y tener una mejor adherencia en pacientes VIH+.

Bibliografía

[1] Yuri Arnold-Domínguez, Manuel Licea-Puig, Lizet Castelo-Elías-Calles, VIH/Sida y terapia antirretroviral: efectos endocrino-metabólicos, REVISTA PERUANA DE EPIDEMIOLOGÍA. 2012.

[2] ONUSIDA. Hoja Informativa 2021: Estadísticas mundiales sobre el VIH. [en línea] <https://www.unaids.org/es/resources/fact-sheet> [Consulta: 5 de diciembre de 2021].

[3] ONUSIDA. Datos: Chile. [en línea] <https://www.unaids.org/es/regionscountries/countries/chile> [Consulta: 8 de noviembre de 2021].

[4] ONUSIDA. 90-90-90: avanzamos, pero el mundo sigue lejos de conseguir los objetivos para 2020. [en línea] 21 de septiembre de 2020. <https://www.unaids.org/es/resources/presscentre/featurestories/2020/september/20200921_90-90-90> [Consulta: 4 de diciembre de 2021].

[5] Thompson, E. Gregory & Shalit, Peter. HIV: Stages of Infection [en línea] <https://www.uofmhealth.org/health-library/hw182771> [Consulta: 15 de noviembre de 2021].

[6] HIVInfo. Tratamiento para la infección por el VIH: Conceptos básicos. HIVInfo.NIH [en línea] <https://hivinfo.nih.gov/es/understanding-hiv/fact-sheets/tratamiento-para-la-infeccion-por-el-vih-conceptos-basicos> [consulta: 29 de octubre de 2021].

[7] HIVinfo. Glosario de Términos Relacionados con el VIH/SIDA. 9° Ed. 2021.

[8] Minsal. Comenzó a regir nuevo decreto AUGE: Más de 70 mil personas podrán ser beneficiadas con nuevas mejoras para 10 problemas de salud. [en línea] <https://www.minsal.cl/comenzo-a-regir-nuevo-decreto-auge-mas-de-70-mil-personas-podran-ser-beneficiadas-con-nuevas-mejoras-para-10-problemas-de-salud/> [Consulta: 29 de octubre de 2021].

[9] BARTLETT, John A. Addressing the challenges of adherence. Journal of acquired immune deficiency syndromes. 29(2002):2-10, 1999.

[10] Ministerio de salud, Prevalencia de consumo de alcohol en Chile, ENS 2016-2017.

[11] Felipe Leyton, Pamela Arancibia, SENDA-MINSAL. El consumo de alcohol en Chile. Situación epidemiológica. Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol. 2016.

[12] Morgozzinni, Paula et al. Facultad de medicina, pontificia universidad de católica de Chile, departamento de salud pública, Estudio del costo económico y social del consumo de alcohol en Chile. 2018.

[13] Catalina Sánchez Álvarez, José Julián Acevedo Mejía, Miguel González Vélez, Factores de riesgo y métodos de transmisión de la infección por el Virus de la Inmunodeficiencia Humana, Revista CES Salud Pública. 3(1):28-37, 2012.

[14] Sax PE et al. Adherence to antiretroviral treatment and correlation with risk of hospitalization among commercially insured HIV patients in the United States. PLoS One. 7(2):, 2012.

[15] Luga Ao et al. Adherence and health care costs. Risk Manag Healthc Policy. 20(7):35-44, 2014.

[16] Samet JH, Horton NJ, Meli S, Freedberg KA & Palepu A. Alcohol consumption and antiretroviral adherence among HIV-infected persons with alcohol problems. Alcoholism, Clinical & Experimental Research. 28:572-577, 2004.

[17] Arnsten JH, Demas PA, Grant RW, Gourevitch MN, Farzadegan H, Howard AA, Schoenbaum EE. Impact of active drug use on antiretroviral therapy adherence and viral suppression in HIV-infected drug users. J Gen Intern Med. 17(5):377-81, 2002.

[18] Belmar, J., & Stuardo, V. Adherencia al tratamiento anti-retroviral para el VIH/SIDA en mujeres: una mirada socio-cultural. Revista chilena de infectología, 34(4): 352-358, 2017.

[19] Mathers, B. M., Degenhardt, L., Phillips, B., Wiessing, L., Hickman, M., Strathdee, S. A. & Mattick, R. P. Global epidemiology of injecting drug use and HIV among people who inject drugs: a systematic review. The Lancet. 372,9651:1733-1745, 2008.

[20] Malow, R., Dévieux, J.G., Stein, J.A. et al. Depression, Substance Abuse and Other Contextual Predictors of Adherence to Antiretroviral Therapy (ART) Among Haitians. AIDS Behav. 17:1221–1230, 2013.

[21] Krakower, D.S., Gruber, S., Hsu, K., Menchaca, J. T., Maro, J. C., Kruskal, B. A. & Klompas, M. Development and validation of an automated HIV prediction algorithm to identify candidates for preexposure prophylaxis. The Lancet HIV. 2019.

[22] Francesco Gullo, From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. Physics Procedia. 62:18-22, 2015.

[23] Shafique, U., & Qaiser, H. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1):217-222, 2014.

[24] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1):3-14, 2002.

[25] Weber, Richard. *Introducción a la minería de datos*. 2020.

[26] Trevor Hastie, Robert Tibshirani, & Jerome Friedman. Overview of supervised learning. In The elements of statistical learning. Springer. 9–41, 2009.

[27] Kecman, V. Support vector machines–an introduction. In Support vector machines: theory and applications. Springer. 1-47, 2005.

[28] Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, Lee MJ, Asadi H. Peering into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. AJR Am J Roentgenol. 212(1):38-43, Jan. 2019.

[29] DE BONI, Raquel B., et al. Is substance use associated with HIV cascade outcomes in Latin America?. PloS one. 13(3):194-228, 2018.

[30] CCCASANET Data Transfer Protocol. Standard Procedure for Data Transfer. [en línea] <https://www.ccasanet.org/wp-content/uploads/2014/12/CCASANET_DTP_20141021_fhw.pdf> [Consulta: 15 de noviembre de 2021].

[31] Daigre, Maria Luisa, Alvarez, Patricia, Flores, Ingrid & Andrade, Carolina. Informe Final: Estudio cuentas nacionales VIH/SIDA y ETS. Comisión Nacional del SIDA. 2005.

[32] Isapres de Chile. INFORME VIH/SIDA: COBERTURA GES Y ESTADÍSTICAS. [en linea] <https://www.isapre.cl/PDF/2018.11%20Informe%20VIH.pdf> [Consulta: 8 de noviembre de 2021].

[33] Ministerio de Salud. Plan Nacional de prevención y control del VIH/SIDA E ITS. 2018-2019. 2019.

[34] Romero, María Inés, Palma, Irma & Belmar, Christian. INFORME FINAL DE EVALUACIÓN: PROGRAMA NACIONAL DE PREVENCIÓN Y CONTROL DEL VIRUS DE INMUNODEFICIENCIA HUMANA/SINDROME DE INMUNODEFICIENCIA ADQUIRIDA (VIH/SIDA) Y DE LAS INFECCIONES DE TRANSMISIÓN SEXUAL (ITS). [en línea] <https://www.dipres.gob.cl/597/articles-141173_informe_final.pdf> [Consulta: 9 de noviembre de 2021]

[35] UNAIDS. Preventing HIV infections at the time of a new pandemic: A synthesis report on programme disruptions and adaptations during the COVID-19 pandemic in 2020. [en línea] <https://www.unaids.org/sites/default/files/media_asset/Status%20of%20HIV%20Prevention%20Services%20in%20the%20Time%20of%20COVID-19_web.pdf> [Consulta: 27 de marzo de 2022].

[36] Jewell BL, Mudimu E, Stover J, ten Brink D, Phillips A, Smith JA et al. Potential effects of disruption to HIV programmes in sub-Saharan Africa caused by COVID-19: results from multiple mathematical models. Lancet HIV. 7(9):629-640, 2020.

[37] The Lancet HIV. Marking 40 years of the HIV/AIDS response. The lancet. HIV. 8(6):311, 2021.

[38] The Lancet HIV. 40 years of HIV/AIDS: a painful anniversary. Lancet (London, England). (21):140-6336, 2021.

[39] Wang, Xingchen, et al. Predicting the Prognosis of Patients in the Coronary Care Unit via Machine Learning Using XGBoost. Available at SSRN 3801858. 2021.

[40] Hongbo Jiang, Yi Zhou, Weiming Tang et al. Maintaining HIV care during the COVID-19 pandemic. The Lancet. HIV. 7(5):308-309, 2021.

[41] Mohamed Omari, Oscillations in a Model for HIV Infection with Three Intracellular. African Institute for mathematical Sciences.

[42] WIC. [en línea]<https://wic.uchile.cl/>[consulta: 01 de julio de 2022]

[43] Diego Ignacio Cornejo Barriga, Rediseño del proceso de toma de decisiones en municipios para la prevención del consumo de drogas a través de herramientas de machine learning (Ingeniero Civil Industrial). Santiago, Chile. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, 2021. 105 h.

[44] La Fundación. [en línea]< https://www.fundacionarriaran.cl/fundacion-arriaran/ >[consulta: 01 de julio del 2022]

[45] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[46] Menke, W., & Menke, J. (2016). *Environmental data analysis with Matlab*. Academic Press.

Anexos

Anexo A

Gráficos según variables abandono (drop) y muerte (death)

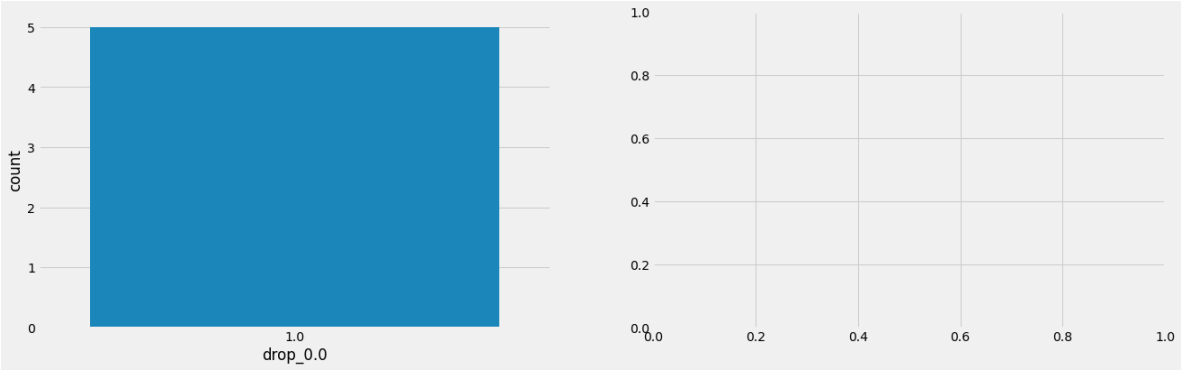

Figura A.1: Pacientes por situación de abandono y por situación de muerte, categoría adherencia 1. Elaboración propia

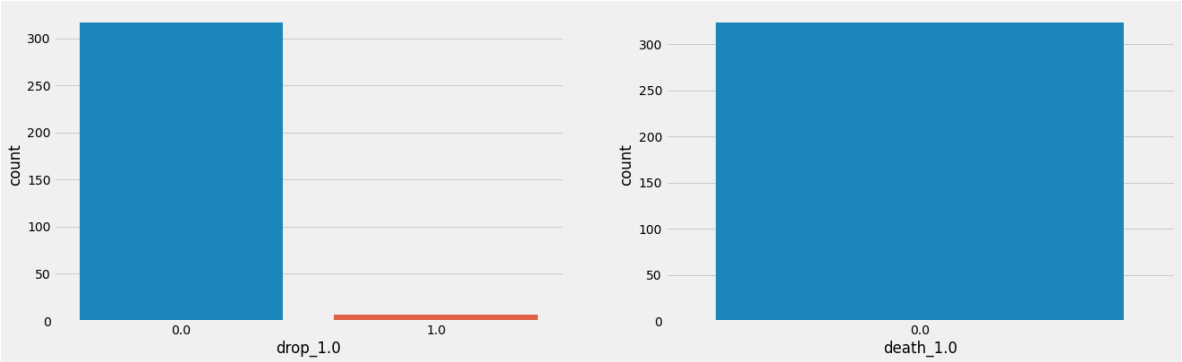

Figura A.2: Pacientes por situación de abandono y por situación de muerte, categoría adherencia 2. Elaboración propia

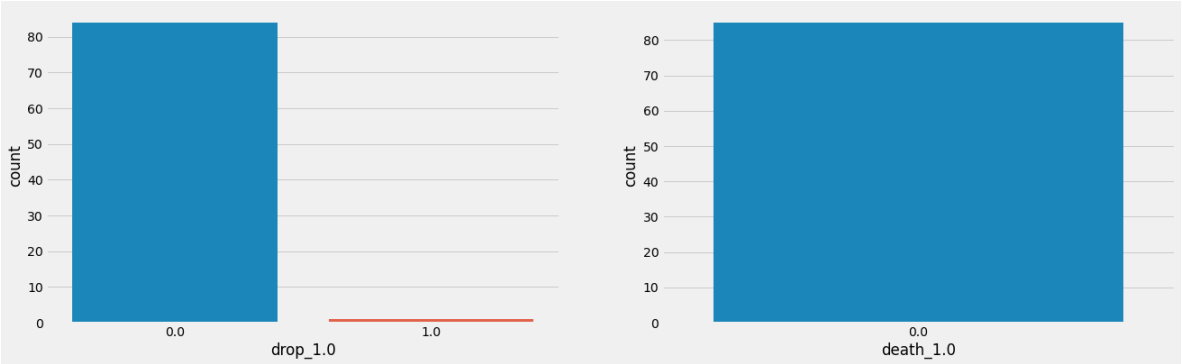

Figura A.3: Pacientes por situación de abandono y por situación de muerte, categoría adherencia 3. Elaboración propia

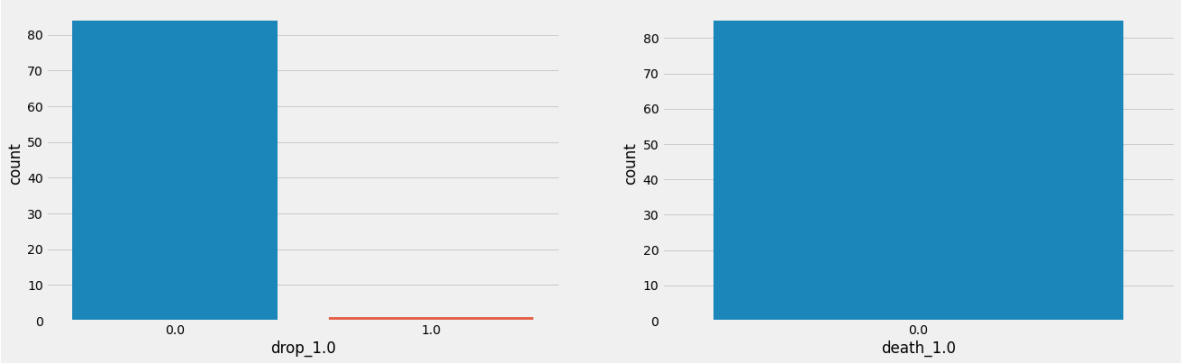

Figura A.4: Pacientes por situación de abandono y por situación de muerte, categoría adherencia 4. Elaboración propia

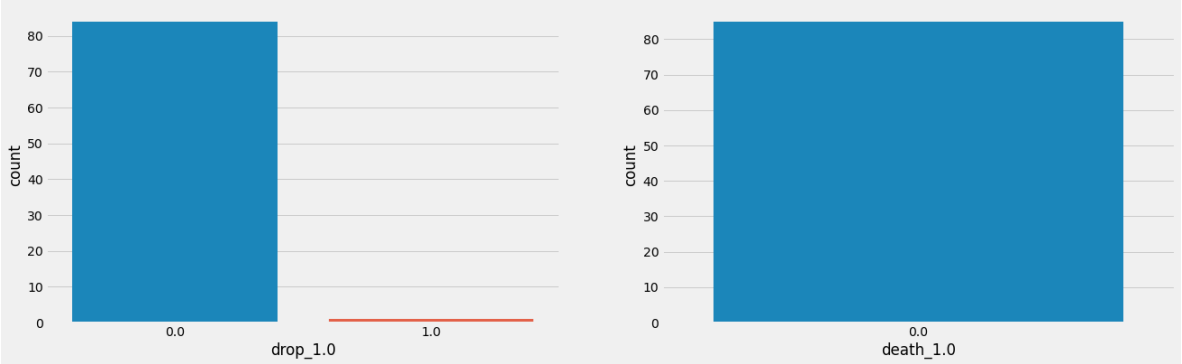

Figura A.5: Pacientes por situación de abandono y por situación de muerte, categoría adherencia 5. Elaboración propia

Gráficos según variables edad, años en la fundación(enrol_years) y años de diagnóstico (diagnosis_years)

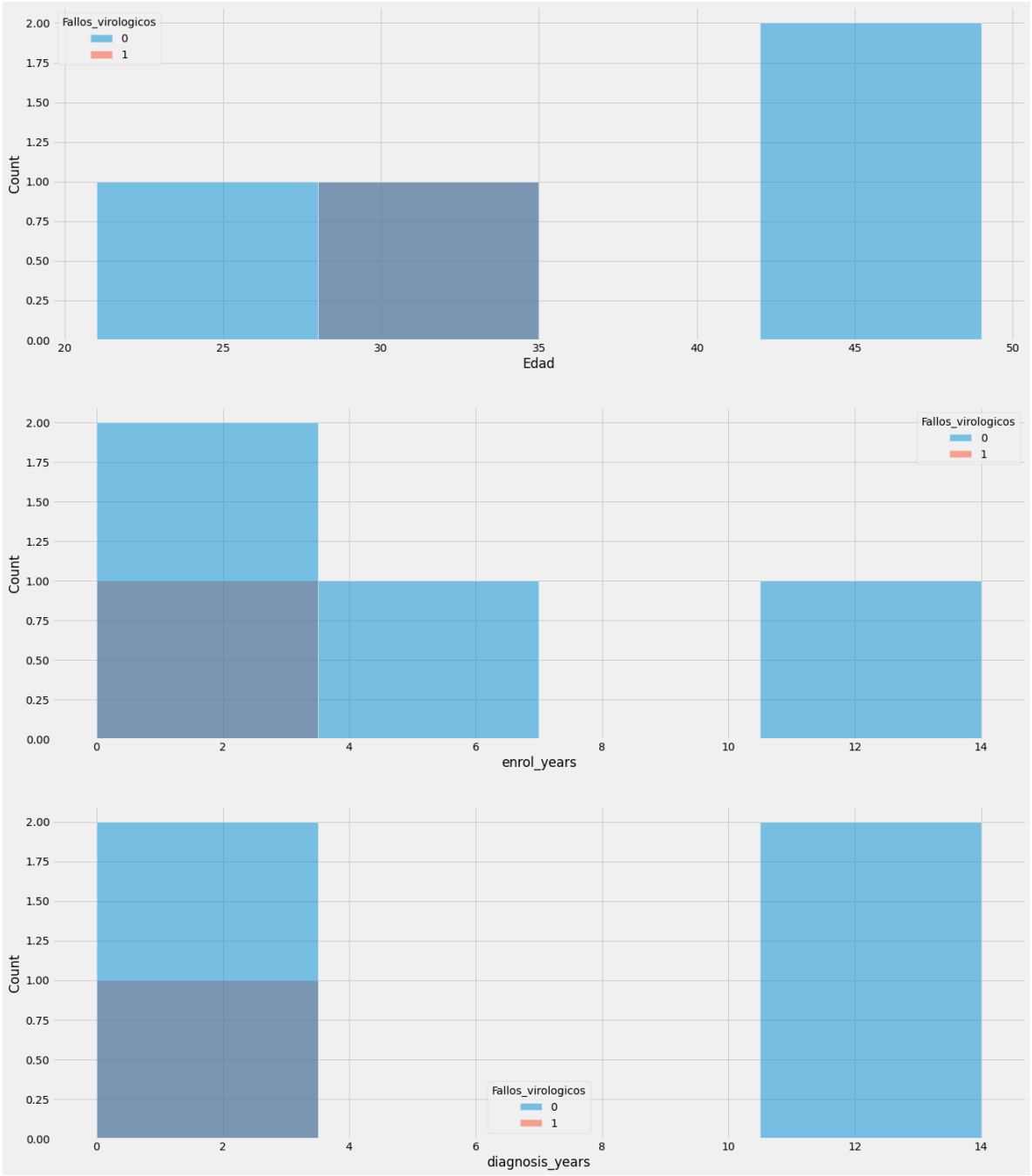

Figura A.6: Pacientes por edad, enrol_years y diagnosis_years, categoría adherencia 1. Elaboración propia

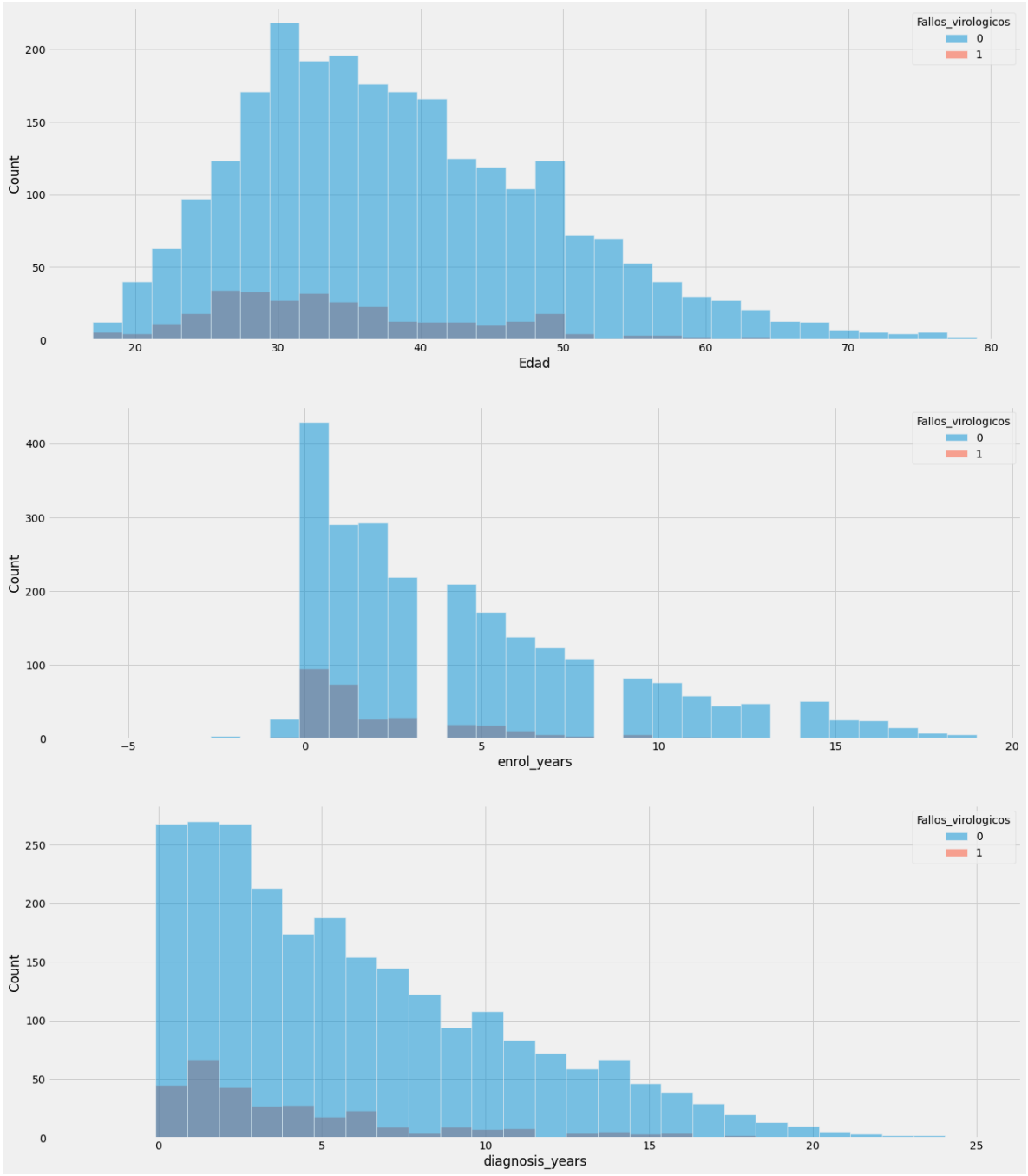

Figura A.7: Pacientes por edad, enrol_years y diagnosis_years, categoría adherencia 2. Elaboración propia

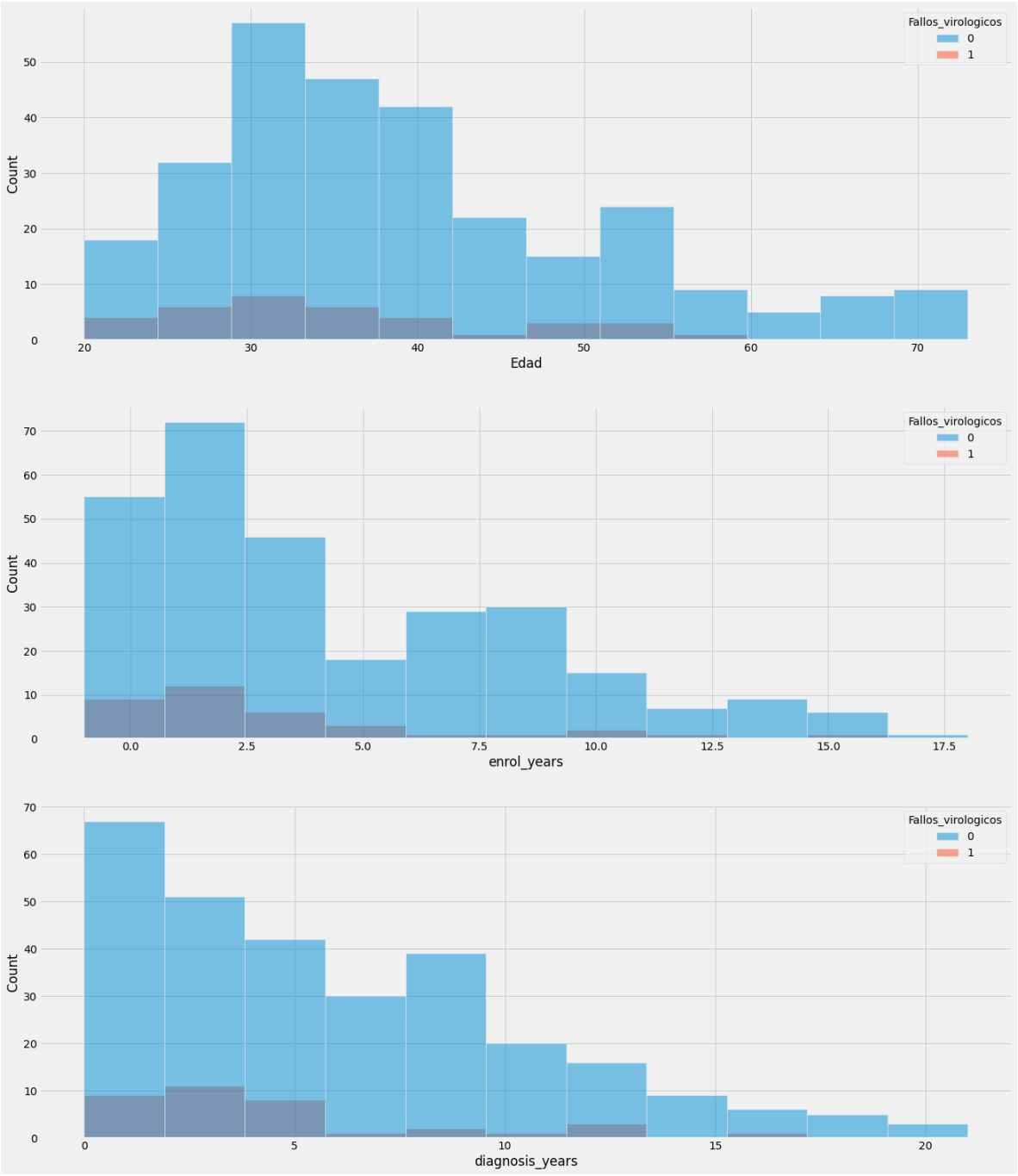

Figura A.8: Pacientes por edad, enrol_years y diagnosis_years, categoría adherencia 3. Elaboración propia

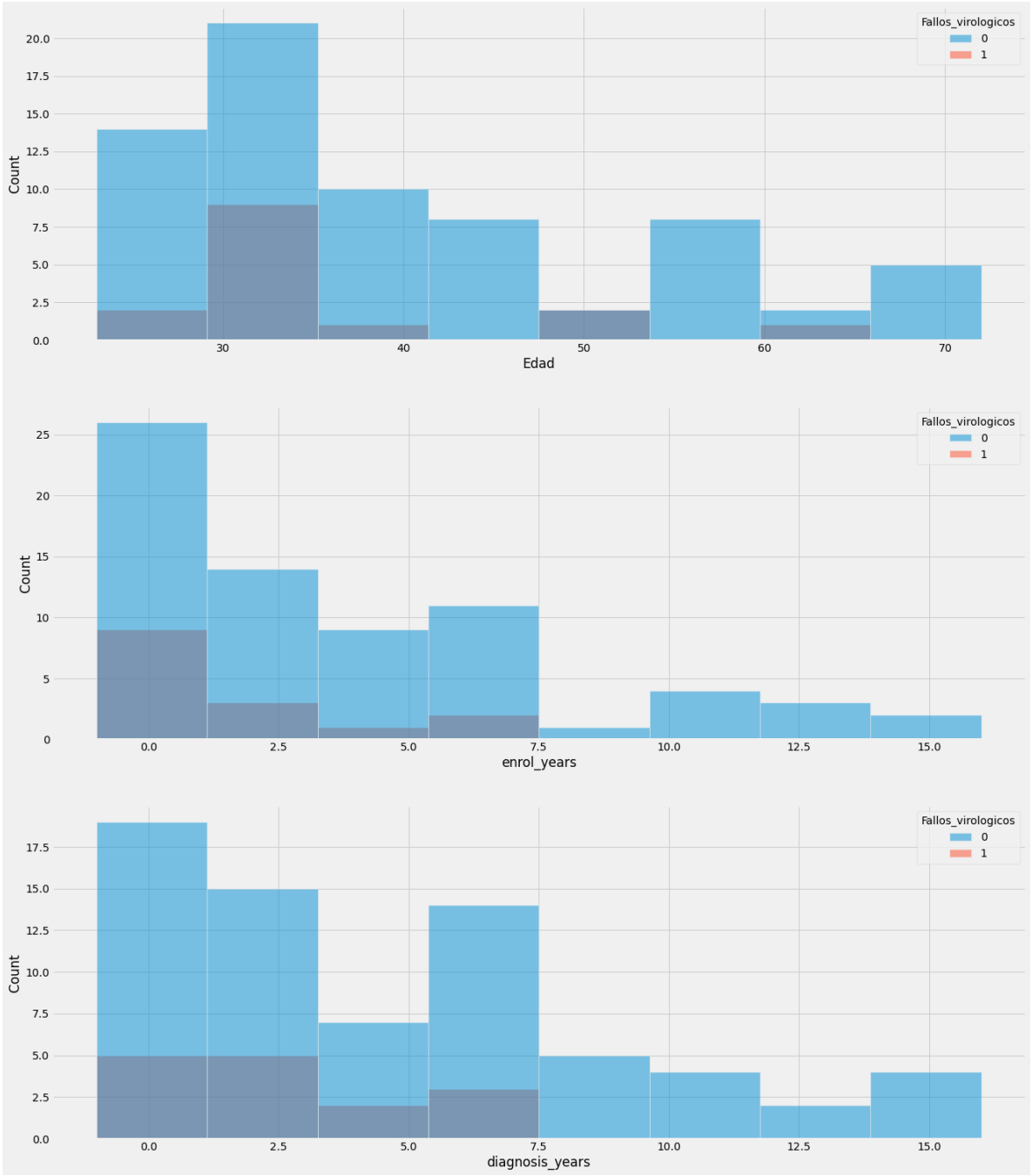

Figura A.9: Pacientes por edad, enrol_years y diagnosis_years, categoría adherencia 4. Elaboración propia

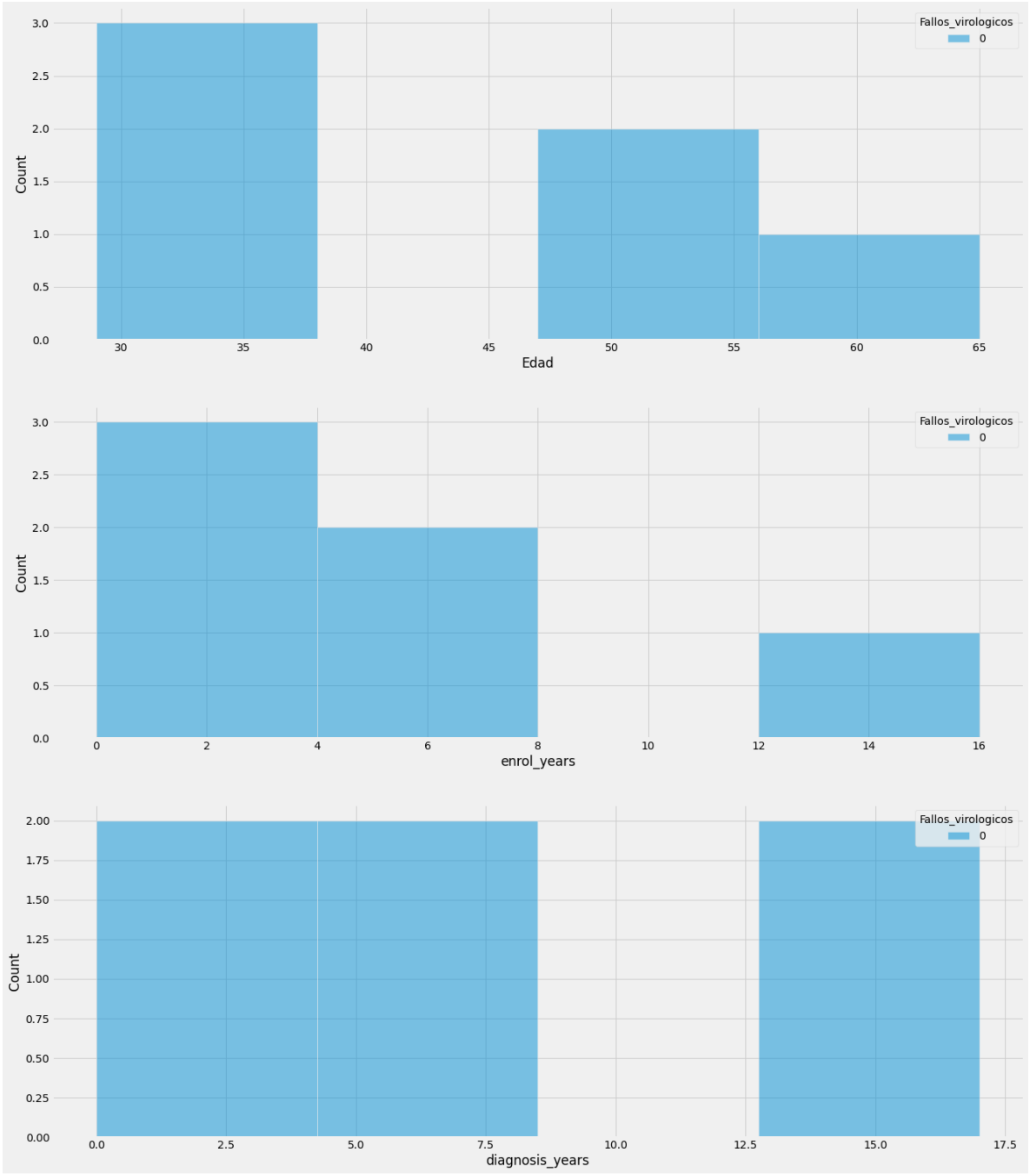

Figura A.10: Pacientes por edad, enrol_years y diagnosis_years, categoría adherencia 5. Elaboración propia

Gráficos según variables nivel educacional, segmento etario, segmento comuna

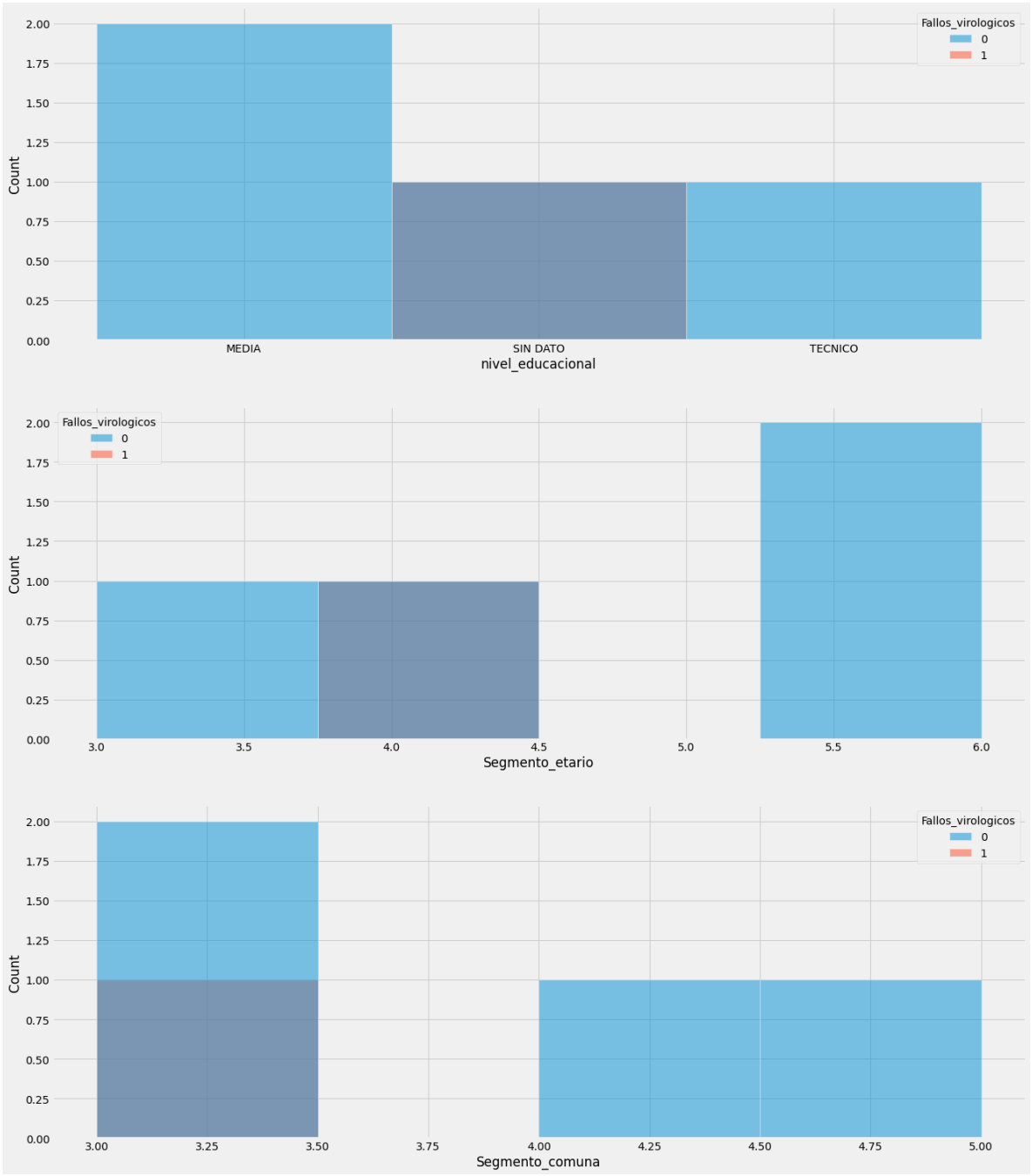

Figura A.11: Pacientes por nivel educacional,segmento etario, segmento comuna categoría adherencia 1. Elaboración propia

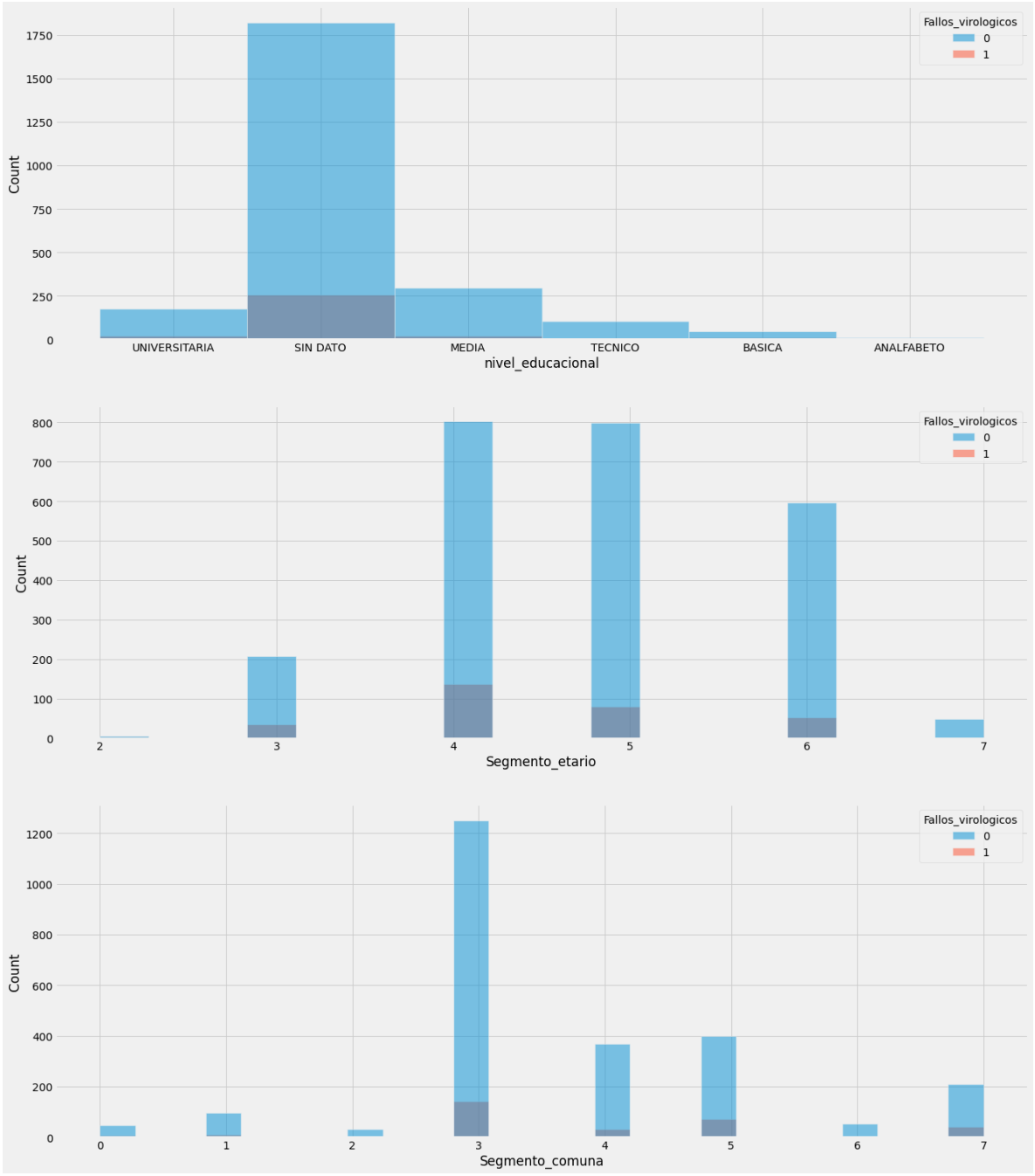

Figura A.12: Pacientes por nivel educacional,segmento etario, segmento comuna categoría adherencia 2. Elaboración propia

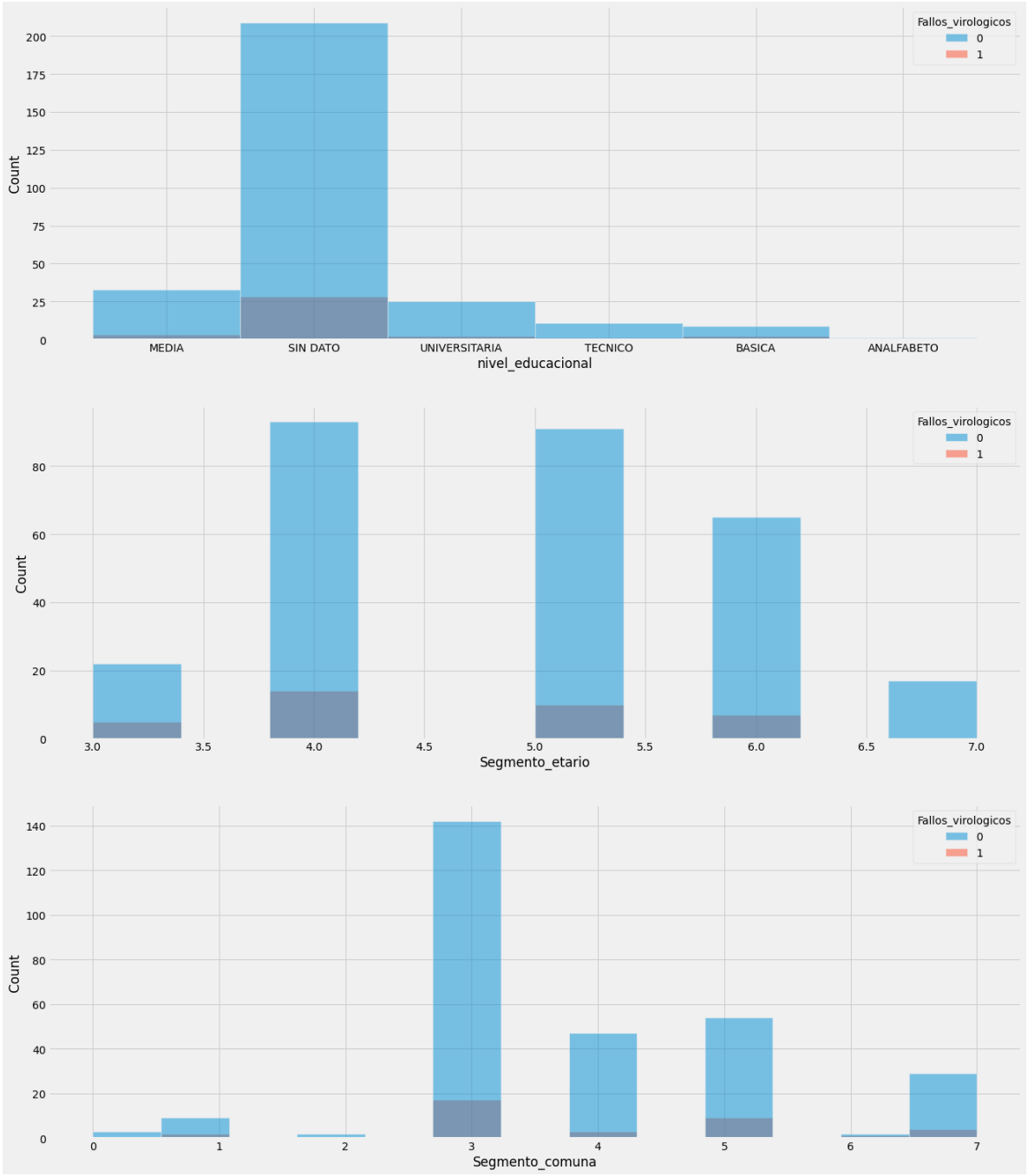

Figura A.12: Pacientes por nivel educacional,segmento etario, segmento comuna categoría adherencia 3. Elaboración propia

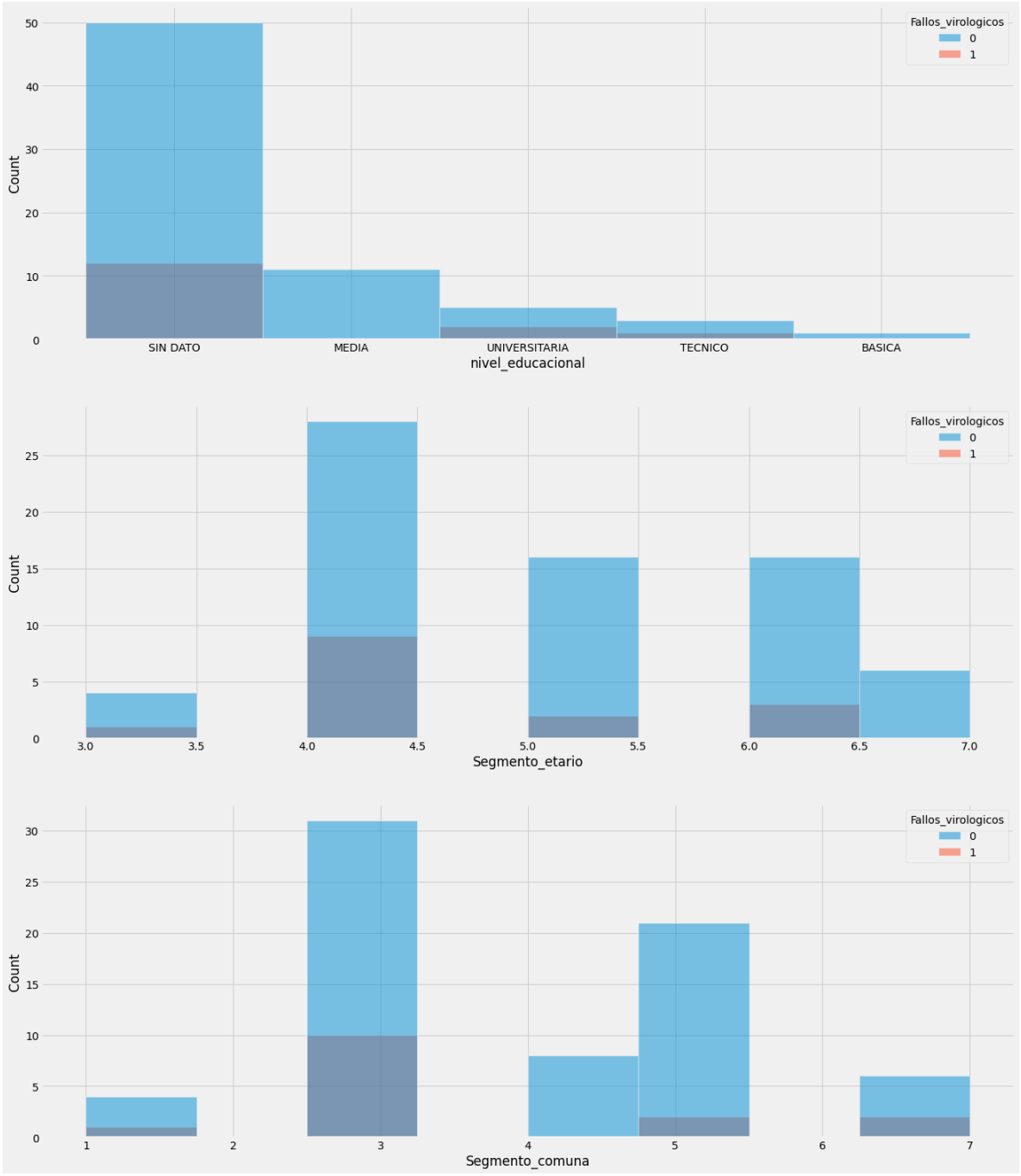

Figura A.13: Pacientes por nivel educacional,segmento etario, segmento comuna categoría adherencia 4. Elaboración propia

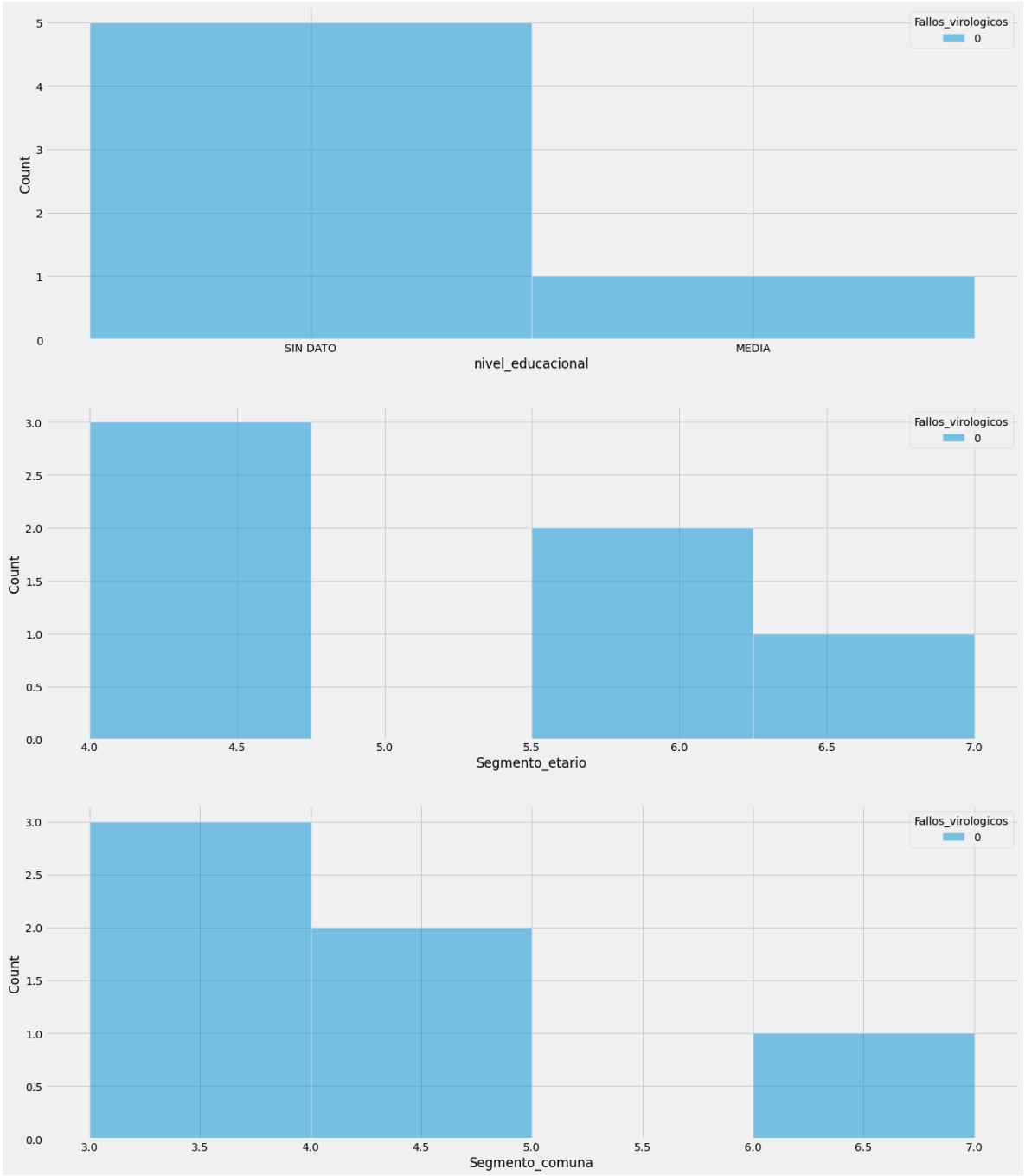

Figura A.14: Pacientes por nivel educacional,segmento etario, segmento comuna categoría adherencia 5. Elaboración propia

Anexo B

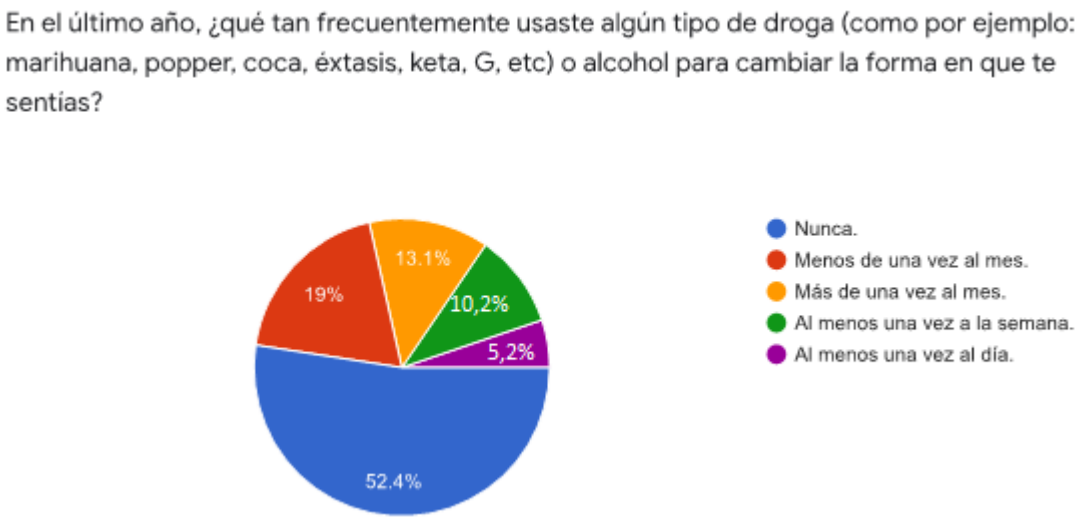

Figura B.14: Porcentajes de pacientes según problemas de consumo