FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2018/2019

Trabajo de Fin de Máster

Como influye la información personal, social en el consumo de drogas (Estudio de predicción en la población escolar de Chile)

***Alumno*: Carlos Andrés Jarrín Vivar**

***Tutor*: Antonio Sarasa Cabezuelo**

Septiembre de 2019

UNIVERSIDAD COMPLUTENSE
MADRID

AGRADECIMIENTO

A Dios por la fortaleza que me brinda cada día y durante este proceso.

Gracias.

"TU MEJOR MAESTRO ES TU ULTIMO ERROR". Ralph Nader.

Índice

Índice de tablas.

Índice de figuras.

1. INTRODUCCIÓN

El Consumo de drogas en la actualidad ha entrado en un debate entre la prohibición y la legalización. Algunos expertos y estudios consideran que es uno los principales problemas de salud en Chile y a nivel global. Otros profesionales como académicos consideran que se debe optar por políticas orientadas a la reducción de daños y consumo responsable. Además el consumo de drogas constituye en la actualidad uno de los principales problemas que abarcan todos los ámbitos de convivencia social; su relación con la enfermedad atribuible a su consumo y por su asociación con otras consecuencias sociales y económicas (Fernández, 2012; Ministerio de Salud de Chile, 2007; SENDA, 2012) [1] [2]. Este importante fenómeno social involucra principalmente a los adolescentes, ya que es una edad crítica para la formación de la persona. En la misma tienen que desarrollarse a nivel emocional, social, académico, entre otras. Los procesos de socialización, con familiares, amigos, centros educativos y medios de comunicación son importantes en ello. Así su percepción de riesgo, tiempo libre y vida recreativa también son elementos que se debe considerar para comprender este debate sobre el consumo de drogas.

El presente trabajo pretende analizar a través de técnicas de Machine Learning (construcción de modelos explicativos y predictivos), algunos aspectos de la población escolar tales como ámbitos personales, sociales, incluyendo el entorno familiar y escolar, percepciones y patrones de consumo del estudiante con respecto al alcohol y tabaco; además de alguna información adicional como características sociodemográficas que puedan llevar a explicar si existe alguna relación o considerar que aspectos son los más importantes al momento de detectar si los estudiantes han consumido drogas. Dicho trabajo pondrá énfasis en el consumo de marihuana para la explicación del estudio.

En la primera parte del estudio en la sección de depuración de datos una vez definidos las secciones del conjunto de datos y metodología del trabajo, se realiza el proceso de exploración y modificación de las muestras con el fin de corregir y discernir la información del estudio. Luego en la sección de construcción de modelos de machine Learning se procede a la aplicación de las distintas técnicas de minería de datos, algunas de ellas con el objetivo de que aporten información del estudio y así observar la importancia e influencia de las variables. Dado que el estudio se subdivide en tres variables objetivos que abarcan distintas drogas, para todas ellas se realiza modelos de árboles de clasificación que además de aportar un modelo de predicción también aportan una relación y muestran información de una serie de reglas sobre las decisiones tomadas; para cuando la variable objetivo trate del consumo de marihuana se realiza también modelos de Regresión y otras técnicas como Redes Neuronales, Random Forest, Gradient Boosting y métodos de Ensamblado con el propósito de encontrar un buen modelo predictivo de esta droga. En la última parte del trabajo se realiza el análisis de resultado tanto desde las perspectiva de interpretación como de predicción para evaluar la relación de los aspectos de variables de estudio y la de su capacidad predictiva respectivamente; finalmente se expone una sección de discusión y conclusiones.

2. PRESENTACIÓN DEL PROBLEMA

2.1 PLANTEAMIENTO DEL PROBLEMA

En la actualidad, el consumo de drogas de la población escolar de enseñanza secundaria del país de Chile se ha incrementado, se posiciona como el más alto de América según revelo el último informe desarrollado por la organización de Estados Americanos (OEA); se destaca por ejemplo una tasa de consumo de marihuana sobre el 30%, seguido por Antigua y Barbuda y Estados Unidos con registros que van por alrededor del 24% y 23% respectivamente, respecto a la cocaína, los chilenos de estudios secundarios superan el 4%. De acuerdo con la investigación realizada entre octubre y diciembre de 2017 en jóvenes entre 13 y 17 años, uno de cada tres adolescentes declara haber consumido marihuana; se puede considerar una de las principales preocupaciones de la sociedad Chilena y la comunidad internacional. Siendo este un tema social de gran importancia merece su análisis desde infinidad de perspectivas, en el estudio lo que se pretende es evaluar el poder predictivo que tiene la información personal y social del estudiante sobre el haber experimentado con drogas; además de proporcionar modelos para la detección de consumo en nuevos estudiantes. Dentro de este fenómeno social se considera importante encontrar, cuáles son los principales aspectos y su grado de influencia para que haya existido el consumo de estas sustancias en los jóvenes de estudios secundarios.

En este trabajo se pretende dar respuesta a la siguiente pregunta:

¿Es posible analizar y encontrar una relación entre los aspectos sociodemográficos, características personales, sociales y percepciones de la población escolar sobre el consumo, que puedan llegar a explicar si los estudiantes han consumido alguna vez droga?

2.2 RESOLUCIÓN DEL PROBLEMA

Existen técnicas que permite encontrar información de datos que no siempre resulta ser obvia dado la gran cantidad de información esta nunca suele ser analizada, con estas técnicas se espera realizar un proceso de identificación de información relevante con el objetivo de descubrir relaciones, patrones y tendencias acerca del consumo de drogas de la población escolar, tomando en consideración la información obtenida de las encuestas sobre los estudiantes sobre los aspectos mencionados. Una vez consolidada la información, se estudiará y aplicara las diferentes técnicas de minería de datos, principalmente con el objetivo de encontrar buenos modelos predictivos que puedan lograr explicar cómo influyen cada uno de los aspectos en el consumo de drogas (haciendo énfasis en el consumo de marihuana).

2.3. ANTECEDENTES

Diversos son los estudios que se han realizado sobre este fenómeno social, la mayoría de ellos buscan encontrar relaciones o asociación de ciertos aspectos de carácter individual, familiar, estudiantil, entre otros con respecto al consumo de drogas. Pero estos análisis son más de carácter estadístico como la encuesta sobre drogas a la población escolar de Cantabria (D. G. de S. P, Real María, 2016)[16]; además que la mayoría de estos buscan encontrar como afectan diversos aspectos en el consumo de drogas como el estudio descriptivo en estudiantes universitarios de primer curso de España (Franco, Agustín, Baile, Valero, & Puerta, 2009)[9]; donde muestran el conocimiento de algunos factores que pueden influir en el consumo de drogas en los estudiantes donde se considera relevante que el inicio del tabaquismo temprano puede actuar como la puerta de entrada para el consumo de alcohol y drogas.

Existe un estudio realizado por el Ministerio del Interior y Seguridad Publica donde utilizan el mismo formulario del presente estudio, este proviene del año 2015 y aplica modelos de regresión con el propósito de describir el nivel de involucramiento de los padres con sus hijos y establecer su asociación con el consumo de alcohol y marihuana (*Valencia-Recabarren*, *Boletin 19 Involucramiento parental y consumo de drogas en escolares de Chile.pdf*, s. f.)[1] ; sus resultados sobre el consumo de marihuana indican que a medida que aumenta el involucramiento parental disminuye en cerca de un 20% el riesgo de consumir marihuana y también el tener amigos que consumen marihuana incrementa en los estudiantes el riesgo de consumo de esta sustancia.

(García & Pol, 2009)[6], presenta un estudio sobre la predicción del consumo de cocaína en adolescentes mediante árboles de decisión, el cual busca evaluar el poder predictivo de la impulsividad y la búsqueda de sensaciones sobre el consumo de cocaína en la población adolescente. En la que los resultados obtenidos muestran la importancia de rasgos de personalidad asociados a la búsqueda de sensaciones y la impulsividad en relación con el consumo de sustancias; tomando en consideración también que el consumo de algunas sustancias legales e ilegales en la adolescencia guarda una relación para clasificar a un adolescente como consumidor o no consumidor. Otro estudio similar utilizando análisis de regresión determina el nivel de predicción de características psicosociales (búsqueda de sensaciones, bienestar subjetivo, permisividad y la orientación escolar) sobre el consumo de alcohol, tabaco y drogas en adolescentes) (*Delgado, & Martínez, 2016, Características Psicosociales Asociadas al Consumo*) [12] donde se muestra que la permisividad es el primer factor que influye en el consumo de drogas, seguido de la búsqueda de sensaciones.

Otros estudios interesantes y similares muestran los factores asociados al inicio del consumo de drogas ilícitas en la educación secundaria de Perú mediante regresión logística (*Saravia, Gutiérrez, & Frech*, 2014)[13] donde buscan establecer una relación entre factores demográficos, escolares, familiares y sociales, y el inicio de consumo de drogas ilegales en escolares peruanos; mostrando que los factores familiares tienen un gran peso en el consumo de drogas ilegales como el no convivir con los padres y la poca constancia en la crianza; con respecto al factor social características como la accesibilidad a las sustancias psicoactivas y el nivel de pobreza llegan hacer determinantes en la experimentación del consumo.

Un estudio determina la influencia del factor consumo de drogas de familiares como factor de riesgo de consumo de jóvenes y adolescentes (*Victoria de Girón*, 2014)[8] concluyendo que el consumo de drogas por los familiares presenta un factor de riesgo de consumo para adolescentes, en particular si se trata de la figura paterna.

Otro trabajo desarrollado en la Universidad de Oviedo analiza a través de Regresión Múltiple y Logística, los factores relacionados con las actitudes juveniles hacia el consumo de alcohol y otras sustancias psicoactivas (Rodríguez, De la Villa, Sirvent, 2006)[11]; donde comprueban que el consumo de drogas es más favorable cuando menor sea la percepción de riesgo de consumo, así como donde más permisiva sea la disposición para el consumo de alcohol y otras drogas ilegales. Otros factores que consideran determinantes son la relación con el grupo de iguales consumidores de drogas.

Finalmente citar una investigación utilizando modelos discriminatorios donde analizan los factores de riesgo predictores del patrón de consumo de drogas durante la adolescencia en centros públicos de la provincia de Alicante España (Alfonso, Huedo, Medina, & Espada, 2009)[7], donde mencionan que las variables de grupos de amistades, vulnerabilidad familiar y consumo de drogas han tenido mayor pesos específico en los modelos que mejor modelan el consumo; así por ejemplo la mayor frecuencia de consumo de cannabis está relacionada con una también mayor frecuencia de consumo de tabaco, de alcohol, el fácil acceso a las drogas y grupos de amigos con actitudes y comportamientos favorables a las mismas.

3. OBJETIVOS

El Objetivo principal del estudio es proporcionar y evaluar a través de modelos de predicción la relación de la información personal, social de la población escolar de Chile en el consumo de drogas[1]. Este objetivo se puede refinar en los siguientes objetivos más específicos:

- Determinar cuáles son los aspectos más importantes de la población escolar que influyen en la predicción para detectar si han consumido drogas.
- Evaluar el poder predictivo de los aspectos de la población escolar sobre el consumo de drogas.
- Proporcionar modelos predictivos para detectar si los estudiantes han consumido alguna vez drogas.
- Realizar una comparación y análisis de los modelos de predicción con el propósito de observar, cual es el mejor y evaluar su capacidad predictiva a través de sus diferentes medidas de clasificación.

4. CONJUNTO DE DATOS

La unidad de investigación proviene del Ministerio del Interior y Seguridad pública de Chile, y está constituida por los resultados de las encuestas realizadas por SENDA (Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol) a la población escolar de Chile en los periodos 2013 y 2015, estos estudios se realizan cada dos años durante el segundo semestre de los años impares en 121 comunas urbanas seleccionadas a lo largo de las 15 regiones del país en diferentes instituciones educativas. Las muestras consideradas para el presente trabajo fueron de alrededor de 110000 estudiantes; la encuesta cuenta 111 preguntas que se han desagregado en el proceso de depuración y formarán parte del análisis 87, las cuales corresponderán a las variables y metadatos de estudio. Dentro de la información recopilada se encuentran aspectos de monitoreo parental, entorno familiar (preguntas acerca del padre y la madre), escolar (preguntas que tienen relación con cómo se sienten en el colegio), relación de amistades, de convivencia (preguntas que tiene relación con las personas con quienes viven), percepción y patrones de consumo de algunas drogas licitas e ilícitas, aspectos sociodemográficos y otra información adicional.

El diseño muestral de estudio consiste en un muestreo probabilístico, donde la mayoría de las variables son de tipo categóricas que toman como argumentos una clasificación de diversas opciones de respuestas de los estudiantes sobre las preguntas de la encuesta; existen pocas variables que inicialmente participaran como tipo cuantitativas, tomaran argumentos de cantidades numéricas discretas entre ellas se podría mencionar la edad. La información de la encuesta se encuentra en dos bases datos, dado que contienen las mismas variables, se unificará la información en tres conjuntos de datos para los distintos estudios con sus diferentes variables dependientes que se pueden observar en la tabla 4.1. Se toma en consideración para el análisis el consumo de marihuana en la que se pone énfasis

[1] DROGAS: Consumo de Marihuana, Cocaína o Pasta Base, Otras drogas (Considerando otras drogas como: crack, éxtasis, heroína, alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos).

para encontrar el mejor modelo predictivo que ayude a pronosticar si los estudiantes han consumido esta sustancia, además de evaluar sobre este consumo, el impacto que tienen los aspectos mencionados anteriormente de la población escolar. Se estudia además el consumo de cocaína, pasta base y otras drogas en las que está considerado crack, éxtasis, heroína, alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos, en estos últimos se evalúa el poder o influencia de los aspectos de los estudiantes sobre haber consumido estas drogas.

Conjunto de Datos	Tema de Estudio o Variable Dependiente	Número de Observaciones	Número de variables
Conjunto 1	**Consumo de Marihuana**	110021	87
Conjunto 2	**Consumo de Cocaína o Pasta Base**	110420	87
Conjunto 3	**Consumo Otras Drogas**	109720	87

Tabla 4.1. Conjuntos de Datos.

DICCIONARIO DE DATOS En el siguiente cuadro se identifican el conjunto de variables de estudio.		
Aspecto	Nombre	Descripción
	P1	1. Sexo
	P2	2. Edad
Monitoreo Parental	P3	3. Después de que sales del colegio o durante los fines de semana, ¿cuántas veces ocurre que tu madre, padre, apoderada o apoderado no saben dónde estás? Ya sea por un período de una hora o más. ① Nunca o casi nunca saben dónde estoy ② A veces no saben ③ Siempre o casi siempre saben dónde estoy
	P4	4. ¿Cuán atentos están tu padre, madre, apoderado o apoderada (o alguno de ellos) respecto de lo que haces en el colegio? ① Mucho ② Bastante ③ Poco ④ Nada
	P5	5. En general, ¿cuánto crees que tu padre, madre, apoderado o apoderada (o alguno de ellos) conocen a tus amigos y amigas más cercanos/as? ① Bastante ② Más o menos ③ Poco
Percepción de riesgo tabaco y alcohol	P6a	6. ¿Cuál crees tú que es el riesgo que corre una persona que hace alguna de estas cosas? Fumar cigarrillos de vez en cuando (ocasionalmente): Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6b	Fumar cigarrillos frecuentemente: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6c.	Fumar una o más cajetillas de cigarros al día : Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6d.	Tomar bebidas alcohólicas de vez en cuando (ocasionalmente): Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6e.	Tomar alcohol frecuentemente: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6f.	Emborracharse con alcohol: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
	P6g.	Tomar uno o dos tragos de alcohol todos o casi todos los días: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨
Las siguientes preguntas son acerca de la madre y padre	P73	¿Quién es el jefe de tu hogar? Jefe de hogar se define como la persona, hombre o mujer, reconocida como tal por los integrantes del hogar: ① Padre ② Madre ③ Abuela o Abuelo ④ Otro
	P74_a	¿Qué educación alcanzaron tu padre? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta ⑤ Técnica superior completa ⑥ Universitaria incompleta ⑦ Universitaria completa ⑧ No sé o No aplica ⑨
	P74_b	¿Qué educación alcanzaron tu madre? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta ⑤ Técnica superior completa ⑥ Universitaria incompleta ⑦ Universitaria completa ⑧ No sé o No aplica ⑨

	P74_c	¿Qué educación alcanzaron el jefe/a de hogar? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta ⑤ Técnica superior completa ⑥ Universitaria incompleta ⑦ Universitaria completa ⑧ No sé o No aplica ⑨
	P75	¿Con qué personas vives actualmente? ① Padre y madre ② Padre y su pareja ③ Madre y su pareja ④ Sólo con el padre ⑤ Sólo con la madre ⑥ Sólo con Hermana(s) o hermano(s) ⑦ Sólo con Abuelo(s) o Abuela(s) ⑧ Otro adulto responsable
Las siguientes preguntas tienen relación con las personas con quien vive	P76	¿Quién es tu apoderado/apoderada? Apoderado/apoderada es quien se responsabiliza por ti ante las autoridades del colegio ① Padre ② Madre ③ Abuela o Abuelo ④ Otro
	P77	¿Has conversado con tu padre, madre o apoderado/a acerca de las consecuencias del consumo de drogas? ① Sí ② No
	P78	¿Crees tú que tu padre, madre o apoderado/a sabe que has probado o consumido alguna droga? (no consideres alcohol, cigarrillos o tranquilizantes) ① Si ② No ③ Nunca he probado drogas
	P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven? (no consideres alcohol, cigarrillos o tranquilizantes) ① Sí ② No
	P80	Hasta donde tú conoces ¿alguno de tus hermanos o hermanas consume alguna droga ilícita (ilegal)? ① Estoy seguro que no lo ha(n) hecho ② Creo que no lo ha(n) hecho ③ Creo que lo hace(n) ④ Estoy seguro que lo hace(n) ⑤ No tengo hermanos o hermanas
	P81_a	¿Cómo describirías el hábito que tiene tu padre respecto al alcohol (vino, cerveza, licor)? Nunca toma alcohol ① Solo en ocasiones especiales ② Solo en fines de semana, pero nunca en días de semana ③ Toma alcohol diariamente, uno o dos tragos ④ Toma alcohol diariamente, más de dos tragos ⑤ No aplica, no tiene padre o madre vivo, no lo ve nunca ⑨
	P81_b	¿Cómo describirías el hábito que tiene tu madre respecto al alcohol (vino, cerveza, licor)? Nunca toma alcohol ① Solo en ocasiones especiales ② Solo en fines de semana, pero nunca en días de semana ③ Toma alcohol diariamente, uno o dos tragos ④ Toma alcohol diariamente, más de dos tragos ⑤ No aplica, no tiene padre o madre vivo, no lo ve nunca ⑨
	P82_a	¿Cómo crees que estaría tu papá y tu mamá en estas situaciones? Si tu papá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥
	P82_b	Si tu mamá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥
	P82_c	Si tu papá descubriera que fumas marihuana: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥
	P82_d	Si tu mamá descubriera que fumas marihuana: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥
Las siguientes preguntas tienen relación con cómo se siente en el colegio en el que está actualmente	P83	Durante este año, ¿Te ha tocado asistir o participar en el colegio en actividades específicamente destinadas a prevenir el consumo de drogas, como por ejemplo charlas o talleres? ① No ② Sí, una vez ③ Sí, más de una vez
	P84	Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste del colegio una parte importante de la jornada o en toda la jornada ① Nunca ② Casi nunca ③ Pocas veces ④ Varias veces ⑤ Muchas veces
	P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ① Menos de 4,5 ② Entre 4,5 y 4,9 ③ Entre 5,0 y 5,4 ④ Entre 5,5 y 5,9 ⑤ Entre 6,0 y 6,4 ⑥ Entre 6,5 y 7,0
	P86	¿Cuántos cursos has repetido en tu vida escolar? ① Ninguno ② Uno ③ Dos o más
	P87	En general, ¿consideras que en tu colegio hay estudiantes que traen, toman o comparten alcohol dentro del colegio? ① Sí ② No
	P88	En general, ¿consideras que en tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos dentro del colegio? ① Sí ② No
	P89	¿Y consideras que en los alrededores de tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos en las afueras o cercanías del colegio? ① Sí ② No
	P90_a	Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio? 90a. Participado en un grupo que molesta a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P90_b	90b. Participado en un grupo que ha agredido físicamente a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P90_c	90c. Participado en un grupo que ha comenzado una pelea con otro grupo : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P90_d	90d. Comenzado una pelea solo con otro/a compañero/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P90_e	90e. Has robado algo a alguien en el colegio: ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces

	P91_a	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio 91a. Has sido molestado/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P91_b	91b. Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P91_c	91c. Has estado en un grupo que ha sido atacado por otro grupo ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P91_d	91d. Alguien solo/sola ha iniciado una pelea contigo ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P91_e	91e. Te han robado algo en el colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
	P92	¿Cuán probable es que pases de curso este año? ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable ⑤ Imposible
	P93	¿Cuán probable es que termines cuarto medio? ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable ⑤ Imposible
	P94	¿Cuán probable es que sigas estudiando después del colegio? (en la Universidad, Instituto Profesional, Centro de Formación técnica u otro) ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable ⑤ Imposible
Las siguientes preguntas son acerca de sus amistades y la relación que mantienes con ellos y ellas	P95_a	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alguna de estas sustancias con el evidente propósito de volarse, drogarse o embriagarse? 95a. Marihuana: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤
	P95_b	95b. Cocaína: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido ⑤
	P95_c	95c. Pasta base: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido ⑤
	P95_d	95d. Inhalables: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido ⑤
	P95_e	95e. Alcohol : Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido ⑤
	P96	Si en tu grupo de amigas y amigos cercanos supieran que fumas marihuana ¿tú crees que: ① Te harían algún reproche o te dirían algo para que no lo hicieras ② Algunos te harían reproches y otro no ③ No te harían ningún problema ④ Te alentarían a que lo siguieras haciendo
	P97	¿Si en tu grupo de amigas y amigos más cercanos supieran que has probado una droga distinta a la marihuana como cocaína, pasta base, éxtasis, ácidos o cosas parecidas, tú crees que: ① Te harían algún reproche o te dirían algo para que no lo hicieras ② Algunos te harían reproches y otro no ③ No te harían ningún problema ④ Te alentarían a que lo siguieras haciendo
	P98	¿Cuántos de tus amigas y amigos toman regularmente alcohol? Digamos, todos los fines de semana o más seguido ① Ninguno ② Menos de la mitad ③ Como la mitad ④ Más de la mitad ⑤ Todos o casi todos
	P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana? Digamos, todos los fines de semana o más seguido ① Ninguno ② Menos de la mitad ③ Como la mitad ④ Más de la mitad ⑤ Todos o casi todos
Las siguientes preguntas son acerca del consumo de tabaco	P7	¿Has fumado cigarrillos alguna vez en la vida? ① Sí ② No
	P8_a	8a. ¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez? No consideres si tus padres o algún adulto te dieron a probar siendo niño. Edad en años: Marca"0" en la hoja de respuestas si no has fumado
	P8_b	8b. ¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días? Edad en años: Marca"0" en la hoja de respuestas si no has fumado todos o casi todos los días
	P9	9. ¿Cuándo fue la primera vez que fumaste cigarrillos? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P10	10. ¿Cuándo fue la última vez que fumaste un cigarrillo? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P11	11. ¿Cuántos días has fumado cigarrillos en los últimos 30 días? N° de días: Marca "0" en la hoja de respuestas si no has fumado en los últimos 30 días
	P12	12. Considerando sólo los días que fumaste en el último mes. ¿Aproximadamente, cuántos cigarrillos fumaste al día? N° de cigarrillos: Marca"0" en la hoja de respuestas si no has fumado en los últimos 30 días
Las siguientes preguntas son acerca del	P15	15. ¿Has probado alcohol alguna vez en la vida (cerveza, vinos o tragos fuertes)? ① Sí ② No
	P16	16. ¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? No consideres si tu padre, madre o una persona adulta te dieron a probar siendo niño/niña. Edad en años: Marca"0"en la hoja de respuestas si no has probado

consumo de bebidas alcohólicas	P17	17 ¿Cuándo fue la primera vez que probaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P18	18. ¿Cuándo fue la última vez que tomaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P19	19 ¿Cuán difícil te sería comprar alguna bebida alcohólica, si quisieras hacerlo? ① Me sería muy fácil ② Me sería fácil ③ Me sería difícil ④ Me sería muy difícil ⑤ No podría comprarla ⑨ No sé
	P20_a	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Marca"0" en la hoja de respuestas si no has consumido alcohol en los últimos 12 meses 20a. Amigos, amigas o familiares te han sugerido o mencionado que disminuyas el consumo de alcohol ① Sí ② No
	P20_b	20b. Consumir alcohol estando solo o sola ① Sí ② No
	P20_c	20c. Peleas con golpes, empujones o patadas ① Sí ② No
	P20_d	20d. Tener relaciones sexuales sin condón ① Sí ② No
	P21	21. Piensa en los últimos 30 días ¿Cuántos días has consumido algún tipo de alcohol? N° de días: Marca"0" en la hoja de respuestas si no has consumido
	P22	22. ¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol? Guíate por la siguiente tabla para saber cuántos tragos consumes 1 trago (una botella o lata individual de cerveza (333 cc.); Un vaso de vino (140 cc.); Un trago de licor (40 cc. de pisco, ron, vodka o w hisky, sólo o combinado) 1 trago y medio (medio litro de cerveza) 3 tragos (un litro de cerveza) 6 tragos (una botella de vino (750 cc.) 8 tragos (una caja de vino (1 litro) 18 tragos (una botella de licor (750 cc.) ① 1 a 2 tragos ② 3 a 4 tragos ③ 5 a 6 tragos ④ 7 a 8 tragos ⑤ 9 o más tragos ⑨ Nunca o casi nunca consumo alcohol
	P23_a	¿Cuántas veces te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? 23a. En tu vida Nunca ⓪ 1-2 veces ① 3-5 veces ② 6-9 veces ③ 10-19 veces ④20-39 veces ⑤ 40 o más veces ⑥
	P23_b	23b. En los últimos 12 meses Nunca ⓪ 1-2 veces ① 3-5 veces ② 6-9 veces ③ 10-19 veces ④20-39 veces ⑤ 40 o más veces ⑥
	P23_c	23c. En los últimos 30 días Nunca ⓪ 1-2 veces ① 3-5 veces ② 6-9 veces ③ 10-19 veces ④20-39 veces ⑤ 40 o más veces ⑥
	P25	25. Pensando en el último día que consumiste alcohol ¿cuál de las siguientes bebidas alcohólicas fue la que más tomaste ese día? Marca aquella bebida (o tipo de alcohol) que más consumiste ① Cerveza ② Vino ③ Espumantes (champaña, Manquehuito, vinos con sabores u otros) ④ Tragos fuertes solos o combinados (piscola, roncola, vodka naranja u otro) ⑤ No consumo alcohol
	P26	26. Indica, de 1 a 10, qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol, donde 1 equivale a "tomé alcohol pero no sentí ningún efecto" y 10 equivale a "estaba tan borracho que no me acuerdo de nada". No me hizo efecto 1 2 3 4 5 6 7 8 9 10 No me acuerdo de nada; Marca "99" en la hoja de respuestas si no has consumido
	P27	27. Pensando en una SALIDA DE SÁBADO POR LA NOCHE ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ① Nunca he tomado alcohol ② Ninguno ③ Menos de 1 ④ Uno ⑤ Entre 2 y 5 ⑥ Entre 6 y 10 ⑦ Más de 10
sección aborda información adicional	P105	105 ¿Con qué religión te identificas? ① Católica ② Evangélica/Protestante ③ Otra religión ④ Ninguna religión ⑨ No lo sé
	P106	106. ¿De cuánto dinero al mes dispones generalmente para tus gastos? Haz un cálculo mensual ⓪ No dispongo de dinero para mis gastos ① Menos de $5.000 ② Entre $5.000 y $10.000 ③ Entre $10.001 y $20.000 ④ Entre $20.001 y $30.000 ⑤ Entre $30.001 y $40.000 ⑥ Entre $40.001 y $60.000 ⑦ Más de $60.000
	P108	108. ¿Trabajas regularmente además de estudiar? ① Sí ② No
	P109	109. ¿Cuál es el estado conyugal actual de tus padres? ① Casados ② Convivientes ③ Separados, anulados, divorciados o no viven juntos ④ Viudo o viuda ⑤ Soltero o soltera ⑥ Otra situación
	P110	110. Pensando en los últimos 7 días, ¿cuántos días hiciste ejercicio o actividad física, fuera del horario de clases, durante al menos 20 minutos y que te haya hecho transpirar o respirar fuertemente? N° de días: Marca "0" en la hoja de respuestas si no has realizado actividad física
Consumo Drogas	**P35**	35. ¿Has consumido marihuana alguna vez en la vida? ① Sí ② No → Es nuestra variable Objetivo
	P47_53	47_53. ¿Has consumido cocaína o pasta base alguna vez en la vida? ① Sí ② No → Es nuestra variable Objetivo
	P66_f_g_i_j	66f_g_i_j.¿Has consumido las siguientes sustancias alguna vez en la vida Crack, Éxtasis, Heroína, Alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos ① Si ② No → Es nuestra variable Objetivo

Tabla 4.2. Diccionario de Datos.

5. METODOLOGÍA DE TRABAJO

5.1 PLANTEAMIENTO DE SOLUCIÓN

De acuerdo con los objetivos que se pretenden conseguir en los diferentes estudios, se realizará una unión de las bases datos para formar los diferentes conjuntos de datos; donde se analizará las diversas categorías y valores de las variables para realizar en primer instancia una exploración y depuración de los datos, identificando también que información es relevante para participar inicialmente en las construcción de los modelos de machine Learning. Se evitará usar transformaciones de variables para no perder interpretación de nuestros metadatos a la hora de encontrar la información más importante que explique las distintas variables dependientes del consumo de drogas; una vez realizada las modificaciones de nuestros conjunto de datos se construirán diferentes técnica de modelados para observar la capacidad predictiva del modelo en general y de las variables de los diferentes aspectos mencionados anteriormente.

El estudio que se desarrolla a continuación tiene información particular propia de la encuesta que no poseen otros estudios; lo que analizará es si los diversos aspectos de los estudiantes influyen para que hayan consumido alguna vez drogas, además, de proporcionar un modelo de predicción para la detección del consumo en nuevos estudiantes. Además hace referencia por separado el consumo de diferentes drogas.

5.2 METODOLOGÍA SEMMA

Para la realización de la solución del trabajo, se utilizará la metodología SEMMA, una metodología que lleva una organización lógica de las tareas más importantes del proceso de construcción de técnicas de minería de datos (*Montequín et al.(2019).-METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE D.pdf*, s. f.) [3]; además está metodología provee una guía general del trabajo a realizar en cada fase proponiendo trabajar con un muestreo de datos originales (en el caso de tener un gran volumen de datos como en el presente estudio). SEMMA fue desarrollada por el SAS Institute y su nombre es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Evaluación).

Con esta metodología se realiza el proceso de exploración y modelado de grandes cantidades de datos para descubrir relaciones y patrones desconocidos acerca del consumo de marihuana y otras drogas, que tenga que ver con ciertos aspectos tomados en cuenta de la población escolar de Chile que es el objetivo del trabajo. El nombre de esta terminología corresponde a las siguientes fases de la figura 5.1, las cuales se mencionan a continuación y se irán describiendo acorde a la implementación del trabajo.

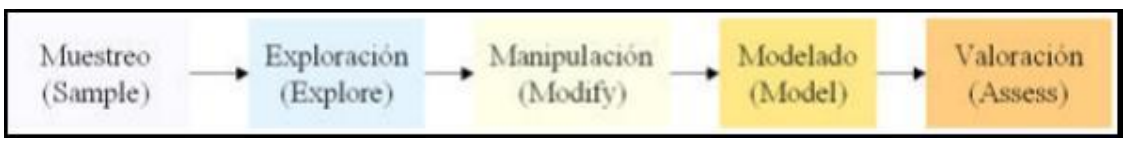


Figura 5. 1. Metodología SEMMA.

(S)ample. – Las muestras de los datos de estudio será la información proporcionada y depurada sobre la encuesta del Estudio Nacional de Drogas en Población Escolar de Chile, en la cual, dependiendo de la importancia de las variables o la mejora de estas en los modelos a emplear (modelos de predicción para predecir el consumo de drogas y ver cómo se comportan los aspectos relacionados a la población escolar), se ira agregando o discerniendo la información. Con el conocimiento de que esta metodología permite que cada una de estas fases se vaya modificando en orden o repitiendo según el estudio lo requiera, se han incorporado y segregado las variables a utilizar y por ende las muestras del estudio han variado para las diferentes técnicas. Para las técnicas de árboles de clasificación y regresión se considera la muestra total (previo al proceso de depuración), para los otros modelos predictivos se considera un subconjunto de la muestra.

(E)xplore.– Una vez obtenida las muestras, se procede a la exploración de la información disponible con el fin de simplificar en lo posible el problema y optimizar la eficiencia de los modelos, refiriéndonos a la exploración con detectar errores, ausentes, atípicos en los datos o también relaciones entre variables. Estudios anteriores como el del Involucramiento parental y consumo de drogas (Valencia-Recabarren, 2015.) [1] ponen en manifiesto, que es posible que existan variables acerca de los aspectos del estudiante (datos de la encuesta) que puedan estar relacionadas con el consumo de drogas, en esta fase con la ayuda de herramientas de visualización se detecta algunas posibles relaciones de variables de clase, que en algunos casos ponen en manifiesto similitudes en varias categorías y bajo el índice de ocurrencias en otras variables, se han considerado para su modificación.

(M)odify.– Se realiza el proceso de modificación, unión de varias categorías en los datos, correspondientes a un número de significativo de variables de estudio detectado en la fase de exploración, además se emplea artificios y técnicas de selección de variables, eliminando variables de poca o mínima importancia, con el fin de un mejor estudio y performance en las técnicas empleadas de minería de datos.

(M)odel.– Una vez determinadas las entradas del modelo, con su formato adecuado para la aplicación de técnicas de minería de datos, se procede a la elaboración y análisis de estas . El objetivo que tendrá dicha fase es establecer una relación entre las variables explicativas tomadas en cuenta y la variable objetivo de estudio (Ha consumido droga alguna vez), que tienen varias medidas a analizar por ejemplo la asociación de los aspectos de monitoreo parental, amistades, percepciones y consumo del estudiante de ciertas sustancia en el consumo de drogas, además de su impacto o influencia que tiene cada una de ellas para ese consumo. Las técnicas empleadas en los datos incluirán métodos estadísticos tradicionales y otras técnicas muy utilizadas de gran poder predictivo. El estudio de los modelos predictivos pone énfasis cuando la variable objetivo es el consumo de marihuana, se emplean técnicas tales como análisis de regresión, árboles de clasificación, redes neuronales, random forest, gradient boosting y técnicas de ensamblado, cuya documentación y explicación se encuentran en la siguiente referencia bibliográfica (Hastie, Tibshirani. (2009). Elements of Statistical Learning)[4]; ya que además de buscar las relaciones mencionadas anteriormente, se busca el mejor algoritmo o modelo de predicción para nuevas observaciones. Para cuando la variable

objetivo sea el consumo de cocaína y pasta base, así como el consumo de otras drogas se emplean la técnica de árboles de decisión, que además de proporcionar un modelo predictivo, son capaces de proporcionar gráficamente relaciones entre variables; así se podrá evaluar el poder predictivo de la información respecto a los factores de la población secundaria de Chile sobre el consumo de estas sustancias.

(A)ssessment. – Finalmente en la última fase de la metodología, se realizará la valoración de los resultados y modelos obtenidos mediante medidas de evaluación de dichos modelos y especialmente tablas o gráficos donde se visualicen claramente los resultados de los aspectos influyentes con respecto al consumo de drogas; así como las medidas de avaluación de predicción de los modelos.

6. DESARROLLO DEL TRABAJO.

En este apartado se realiza el de desarrollo de cada una de las fases de la metodología SEMMA para cumplir el objetivo principal del estudio, donde se detallan los pasos necesarios para su elaboración, los mismos que van desde la exploración y modificación de la información hasta la construcción y evaluación de los modelos implementados.

6.1 DEPURACIÓN DE DATOS

Siguiendo la metodología SEMMA, se realiza la depuración de los conjuntos de datos, contemplando este proceso las fases de exploración y modificación.

La mayoría de la depuración de las Bases de Datos ese la realiza con el Software EM 14.1 de SAS; inicialmente se contaba con dos bases de datos, en la que las preguntas de las encuestas del periodo 2013 y 2015 sobre el Estudio Nacional de Drogas en Población Escolar de Chile no contemplaban el mismo orden, existe información adicional que no participara en el estudio, mal estructurada en cuanto a sus valores; además según los objetivos a evaluar de las distintas drogas esta se tuvo que reorganizar en tres DataSet diferentes, agrupando respuestas, corrigiendo valores, además de realizar la unión correspondiente de los periodos a analizar.

Una vez realizado el proceso previo de exploración y modificación de datos, a continuación se explica con más detalle el proceso posterior de depuración llevado a cabo con la herramienta de EM. Dada la similitud de los conjuntos de datos y por fines prácticos se explica el proceso de depuración de uno de ellos y el que se pondrá énfasis en los análisis, es decir el Dataset con la variable objetivo Consumo de Marihuana, para los otros conjuntos de datos las tablas que se mostraran a continuación se indicaran como Anexos. Para su mejor comprensión se ha dividido en los siguientes apartados.

6.1.1 Conjunto de datos, asignación de Roles y Tipos de Variables.

Se cuenta con un total 110021 observaciones(obs) en el caso de la variable objetivo consumo de marihuana, 110420 (obs) consumo de cocaína o pasta base y 109720 (obs) consumo de otras drogas; se trabaja con 86 variables de entrada (input) de las cuales ID es de tipo identificadora, no participara en el estudio, siete variables son de tipo intervalo o continuas, nueve son de tipo binario o dicotómicas y 69 son de tipo nominal o categóricas; se contempla además las tres variables objetivos.

Para cuando el objetivo sea encontrar el mejor modelo predictivo para detectar el consumo de marihuana se considera una muestra del 10% del total de observaciones, se trabajará 11004 observaciones, al igual que el número de variables se consideran las más importantes, que se describe más adelante.

La tabla 6.1.1.1 que se muestra a continuación, contiene la asignación de roles y tipos de variables de los conjuntos de datos, vale recalcar que se especifica tres variables objetivos, las cuales participaran individualmente en su estudio, con las mismas variables de entrada. Para cuando la variable objetivo sea el consumo de (Cocaína y Pasta Base), la variable (P35)→ el haber consumido marihuana alguna vez participara como variable input; cuando el objetivo sea el consumo de (Otras drogas), (P35) y

P47_53 → el haber consumido cocaína o pasta base participarán en el estudio. Para conocer las descripción de cada uno de los nombres de las variables observar el diccionario de datos (ver aquí).

NOMBRE	ROL	TIPO		NOMBRE	ROL	TIPO
ID	ID	N/A		P74_c	INPUT	NOMINAL
P1	INPUT	BINARY		P75	INPUT	NOMINAL
P10	INPUT	NOMINAL		P76	INPUT	NOMINAL
P105	INPUT	NOMINAL		P77	INPUT	BINARY
P106	INPUT	NOMINAL		P79	INPUT	BINARY
P108	INPUT	BINARY		P80	INPUT	NOMINAL
P109	INPUT	NOMINAL		P81_a	INPUT	NOMINAL
P11	INPUT	INTERVAL		P81_b	INPUT	NOMINAL
P110	INPUT	NOMINAL		P82_a	INPUT	NOMINAL
P12	INPUT	INTERVAL		P82_b	INPUT	NOMINAL
P15	INPUT	BINARY		P82_c	INPUT	NOMINAL
P16	INPUT	INTERVAL		P82_d	INPUT	NOMINAL
P17	INPUT	NOMINAL		P83	INPUT	NOMINAL
P18	INPUT	NOMINAL		P84	INPUT	NOMINAL
P19	INPUT	NOMINAL		P85	INPUT	NOMINAL
P2	INPUT	INTERVAL		P86	INPUT	NOMINAL
P20_a	INPUT	NOMINAL		P87	INPUT	BINARY
P20_b	INPUT	NOMINAL		P88	INPUT	BINARY
P20_c	INPUT	NOMINAL		P89	INPUT	BINARY
P20_d	INPUT	NOMINAL		P8_a	INPUT	INTERVAL
P21	INPUT	INTERVAL		P8_b	INPUT	INTERVAL
P22	INPUT	NOMINAL		P9	INPUT	NOMINAL
P23_a	INPUT	NOMINAL		P90_a	INPUT	NOMINAL
P23_b	INPUT	NOMINAL		P90_b	INPUT	NOMINAL
P23_c	INPUT	NOMINAL		P90_c	INPUT	NOMINAL
P25	INPUT	NOMINAL		P90_d	INPUT	NOMINAL
P26	INPUT	NOMINAL		P90_e	INPUT	NOMINAL
P27	INPUT	NOMINAL		P91_a	INPUT	NOMINAL
P3	INPUT	NOMINAL		P91_b	INPUT	NOMINAL
P35	**OBJETIVO**	BINARY		P91_c	INPUT	NOMINAL
P4	INPUT	NOMINAL		P91_d	INPUT	NOMINAL
P47_53	**OBJETIVO**	BINARY		P91_e	INPUT	NOMINAL
P5	INPUT	NOMINAL		P92	INPUT	NOMINAL
P65_f_g_i_j	**OBJETIVO**	BINARY		P93	INPUT	NOMINAL
P6_a	INPUT	NOMINAL		P94	INPUT	NOMINAL
P6_b	INPUT	NOMINAL		P95_a	INPUT	NOMINAL
P6_c	INPUT	NOMINAL		P95_b	INPUT	NOMINAL
P6_d	INPUT	NOMINAL		P95_c	INPUT	NOMINAL
P6_e	INPUT	NOMINAL		P95_d	INPUT	NOMINAL
P6_f	INPUT	NOMINAL		P95_e	INPUT	NOMINAL

P6_g	INPUT	NOMINAL		P96	INPUT	NOMINAL
P7	INPUT	BINARY		P97	INPUT	NOMINAL
P73	INPUT	NOMINAL		P98	INPUT	NOMINAL
P74_a	INPUT	NOMINAL		P99	INPUT	NOMINAL
P74_b	INPUT	NOMINAL				

Tabla 6.1.1.1. Roles y Tipos de Variables.

6.1.2 Análisis descriptivo del conjunto de datos, detección de errores y posibles relaciones entre variables.

Se ha realizado un análisis exploratorio de los datos para observar información importante en cada una de las variables, esto dará una idea susceptible de los posibles errores de estas; para las variables de intervalo se identifican los valores como el mínimo, máximo, media, desviación típica, asimetría, curtosis, datos totales y faltantes; para las variables de tipo categóricas, se identifican el número de niveles y datos faltantes. La información mencionada se puede visualizar en la tabla 6.1.2.1 y 6.1.2.2 respectivamente. El análisis descriptivo para las otras variables objetivo se encuentra en el Anexo I: Depuración del conjunto de datos.

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
P11	P11	1358.0	108663.0	0.0	30.0	3,13	7,42	2,66	6,01
P12	P12	862.0	109159.0	0.0	40.0	1,86	5,61	4,90	26,35
P16	P16	882.0	109139.0	0.0	21.0	10,27	6,10	-0,92	-0,85
P2	P2	1476.0	108545.0	10.0	25.0	15,60	1,59	0,22	0,02
P21	P21	1398.0	108623.0	0.0	30.0	2,13	4,64	3,82	16,89
P8_a	P8_a	1869.0	108152.0	0.0	21.0	7,21	6,92	-0,002	-1,86
P8_b	P8_b	1899.0	108122.0	0.0	21.0	3,25	6,14	1,41	0,10
Missing		0	110021	0.0	35.0	1.59			

Tabla 6.1.2.1. Consumo Marihuana: Análisis Descriptivo Variables de Intervalo.

Variable	Etiqueta	Tipo	Número de niveles	Ausente		Variable	Etiqueta	Tipo	Número de niveles	Ausente
P1	P1	N	2.0	707.0		P77	P77	N	2.0	645.0
P10	P10	N	4.0	2067.0		P79	P79	N	2.0	1023.0
P105	P105	N	6.0	2372.0		P80	P80	N	5.0	932.0
P106	P106	N	8.0	2535.0		P81_a	P81_a	N	6.0	1121.0
P108	P108	N	2.0	2141.0		P81_b	P81_b	N	6.0	725.0
P109	P109	N	6.0	5321.0		P82_a	P82_a	N	6.0	1038.0
P110	P110	N	8.0	1850.0		P82_b	P82_b	N	6.0	795.0
P15	P15	N	2.0	754.0		P82_c	P82_c	N	6.0	1095.0
P17	P17	N	4.0	935.0		P82_d	P82_d	N	6.0	924.0
P18	P18	N	4.0	1009.0		P83	P83	N	3.0	759.0

P19	P19	N	6.0	660.0		P84	P84	N	5.0	673.0
P20_a	P20_a	N	3.0	924.0		P85	P85	N	6.0	858.0
P20_b	P20_b	N	3.0	965.0		P86	P86	N	3.0	705.0
P20_c	P20_c	N	3.0	2531.0		P87	P87	N	2.0	751.0
P20_d	P20_d	N	3.0	2686.0		P88	P88	N	2.0	796.0
P22	P22	N	6.0	936.0		P89	P89	N	2.0	1554.0
P23_a	P23_a	N	7.0	641.0		P9	P9	N	4.0	1324.0
P23_b	P23_b	N	7.0	689.0		P90_a	P90_a	N	5.0	951.0
P23_c	P23_c	N	7.0	658.0		P90_b	P90_b	N	5.0	815.0
P25	P25	N	5.0	1511.0		P90_c	P90_c	N	5.0	948.0
P26	P26	N	11.0	16512.0		P90_d	P90_d	N	5.0	1053.0
P27	P27	N	7.0	403.0		P90_e	P90_e	N	5.0	996.0
P3	P3	N	3.0	322.0		P91_a	P91_a	N	5.0	1102.0
P35	P35	N	2.0	0.0		P91_b	P91_b	N	5.0	1167.0
P4	P4	N	4.0	374.0		P91_c	P91_c	N	5.0	1324.0
P5	P5	N	3.0	654.0		P91_d	P91_d	N	5.0	1565.0
P6_a	P6_a	N	5.0	885.0		P91_e	P91_e	N	5.0	1655.0
P6_b	P6_b	N	5.0	798.0		P92	P92	N	5.0	869.0
P6_c	P6_c	N	5.0	844.0		P93	P93	N	5.0	957.0
P6_d	P6_d	N	5.0	1103.0		P94	P94	N	5.0	1221.0
P6_e	P6_e	N	5.0	948.0		P95_a	P95_a	N	5.0	1791.0
P6_f	P6_f	N	5.0	957.0		P95_b	P95_b	N	5.0	1310.0
P6_g	P6_g	N	5.0	1268.0		P95_c	P95_c	N	5.0	1310.0
P7	P7	N	2.0	495.0		P95_d	P95_d	N	5.0	1529.0
P73	P73	N	4.0	2402.0		P95_e	P95_e	N	5.0	1497.0
P74_a	P74_a	N	9.0	1411.0		P96	P96	N	4.0	1563.0
P74_b	P74_b	N	9.0	1008.0		P97	P97	N	4.0	1662.0
P74_c	P74_c	N	9.0	2245.0		P98	P98	N	5.0	1712.0
P75	P75	N	8.0	2267.0		P99	P99	N	5.0	2175.0
P76	P76	N	4.0	1057.0		-	-	-	-	-

Tabla 6.1.2.2. Consumo Marihuana: Análisis Descriptivo Variables Categóricas.

Analizando las tablas anteriores de descripción de variables y estadísticos; dado que se realizó una depuración previa, no parece existir mayores inconvenientes. La tabla 6.1.2.1 ofrece información estadística muy útil para conocer en cierto modo la variabilidad de los datos, su tipo de distribución, los valores máximos, mínimos, además se evidencia la presencia de missings que se tratarán más adelante . La tabla 6.1.2.2 que contiene la descripción de las variables de clase, dado el conocimiento de los datos si bien no evidencia errores en los niveles de sus variables, se han identificado en algunos casos categorías con significado similar, además observando el bajo porcentaje de observaciones en algunas de ellas se ha considerado reducir y agrupar convenientemente varias categorías. Las variables P8_a, P8_b y P16 se agruparán en rangos y serán consideradas categóricas.

A continuación, en la figura 6.1.2.1 a partir de una muestra aleatoria de los datos se presenta una seria de gráficos de interés sobre la variable objetivo consumo de

marihuana. Para observa de mejor manera estas estadísticas así como los gráficos de las demás variables objetivo se pueden acceder a través del Anexo VII que contiene dichos gráficos en formato HTML.

Figura 6.1.2.1. Estadísticas sobre la muestra objetivo Consumo de Marihuana.

6.1.3 Corrección de los errores.

Una vez identificado los errores y posibles relaciones con respecto a las variables de clase, se realiza convenientemente varias agrupaciones de categorías, con lo cual se realiza un mejor estudio. El resumen de las modificaciones se puede observar en la Tabla 6.1.3.1.

Para observar todas las agrupaciones que se llevaron a cabo ver el Anexo I: Depuración del conjunto de datos (Corrección de Errores).

AGRUPACIÓN DE CATEGORIAS
- **P4 ¿Cuán atentos están tu padre, madre, apoderado o apoderada (o alguno de ellos) respecto de lo que haces en el colegio? ①Mucho ②Bastante ③Poco ④Nada**

Se unen las categorías mucho y bastante, al tener significados semejantes, **(1) Y (2) se unen = (1)** → Mucho o bastante.

- **P6c. Fumar una o más cajetillas de cigarros al día : Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨**

Se unen la categoría de ningún riesgo y riesgo leve, **(1) Y (2) se unen = (1) → riesgo leve o ninguno**

- **P74_a: ¿Qué educación alcanzaron tu padre?**

Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta ⑤ Técnica superior completa ⑥ Universitaria incompleta ⑦ Universitaria completa ⑧ No sé o No aplica ⑨

Se unen la categoría **5 y 7 = 5 → (Técnica superior incompleta o universitaria Incompleta)**
todas bajan un número de identificador, 9 se mantiene

- **P76 ¿Quién es tu apoderado/apoderada? Apoderado/apoderada es quien se responsabiliza por ti ante las autoridades del colegio ①□Padre ②□Madre ③□Abuela o Abuelo ④□Otro**

Se unen la categoría **(3) y (4) = (3) → Abuelo(a) u otro familiar**

¿Cómo crees que estaría tu papá y tu mamá en estas situaciones?

- **P82_a: Si tu mamá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥**

Se unirán las categorías **(1 y 2)=(1)→ extremada o baste molest**o(a) **y (3 y 4)=(2)→ algo o poco molesto(a)**
todas bajan 2 puestos en el número de identificador, 6 se mantiene

- **P84: Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste al colegio una parte importante de la jornada o en toda la jornada ①□Nunca ②□Casi nunca ③□Pocas veces ④□Varias veces ⑤□Muchas veces**

Se unirán las categorías **(2 y 3)=2 y (4 y 5)=3**

- **Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio?**

P90_a.	Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio? 90a. Participado en un grupo que molesta a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P90_e.	90e. Has robado algo a alguien en el colegio: ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_a.	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio 91a. Has sido molestado/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_b.	91b. Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces

Se unirán las categorías **(4 y 5)=(4)→ 3 o 4 veces con 5 o más veces**

- **P94¿Cuán probable es que sigas estudiando después del colegio? (en la Universidad, Instituto Profesional, Centro de Formación técnica u otro) ①□Es seguro ②□Muy probable ③□Más o menos probable ④□Poco probable ⑤□Imposible**

Se unen las categorías **(1 y 2) = 1 Seguro o Muy probable**
Se unen las categorías **(4 y 5) = 3 Poco probable o Imposible**

Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alguna de estas sustancias con el evidente propósito de volarse, drogarse o embriagarse?

- **95a. Marihuana: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95b. Cocaína: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95c. Pasta base: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**

Se unen las categorías **(2 y 3)=2 Casi Nunca o de vez en cuando y (4 y 5)=3 Bastante o muy seguido**

- **P8a: ¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez? No consideres si tus padres o algún adulto te dieron a probar siendo niño. Edad en años: Marca"0" en la hoja de respuestas si no has fumado todos o casi todos los días**

Se consideran los siguientes rangos **(5 a 12)=1 , (13 a 17)=2, (18 a 21)=3**

- **P16. ¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? No consideres si tu padre, madre o una persona adulta te dieron a probar siendo niño/niña. Edad en años: Marca"0"en la hoja de respuestas si no has probado**

Se consideran los siguientes rangos **(5 a 12)=1 , (13 a 17)=2, (18 a 21)=3**
¿Cuántas veces te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? **Nunca ⓪ 1-2 veces ① 3-5 veces ② 6-9 veces ③ 10-19 veces ④20-39 veces ⑤ 40 o más veces ⑥** - **P23a: En tu vida** - **P23b: En los últimos 12 meses** - **P23c: En los últimos 30 días**
Se unen las categorías **(2,3,4,5,6)=2→ más de 3 veces** Se unen las categorías **(1 a 6)=1→ una o más de una vez**
- **P26: Indica, de 1 a 10, qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol, donde 1 equivale a "tomé alcohol pero no sentí ningún efecto" y 10 equivale a "estaba tan borracho que no me acuerdo de nada".** **No me hizo efecto 1 2 3 4 5 6 7 8 9 10 No me acuerdo de nada, Marca "99" en la hoja de respuestas si no has consumido**
Se unen las categorías **(1 a 3)=1 → Poco o casi nada** Se unen las categorías **(4 a 7)=2 → Medio tomado** Se unen las categorías **(8 a 10)=3 → Bien tomado** Se considera una nueva categoría **"no responde" (88), debido al gran número de observaciones.**

Tabla 6.1.3.1 Agrupación de variables Categóricas.

6.1.4 Tratamiento de datos atípicos y faltantes

Luego del proceso de estudio y corrección de los datos, se realiza el proceso de identificación de atípicos y datos faltantes, para observar si es necesario algún tipo de tratamiento (eliminación o imputación), considerando también los modelos a emplear. En el caso de la identificación de atípicos para las cuatro variables de intervalo se utiliza varios métodos dependiendo de su distribución; como buena práctica se considera dos métodos de detección y observando que estos dos métodos los detecten, tomando en cuenta los intervalos menos restrictivos. A continuación, se explica brevemente este procedimiento, en resumen no se han detectado atípicos en las variables.

Para la variable P2 → Edad que muestra una distribución simétrica, se utiliza el método desviación estándar y rangos intercuartílicos (solo se detectan por un método, no se consideraran missing).

Para las variables P11→ días que has fumado cigarrillos en los últimos 30 días, P12→ cuántos cigarrillos fumaste al día (considerando el último mes), P21→ días que has consumido algún tipo de alcohol, muestran una distribución asimétrica, como su mediana es cero, se utiliza el método de percentiles y rangos intercuartílicos (solo se detectan missing por un método, no se consideraran missing).

En el caso de valores ausentes ya se identificaron en las tablas anteriores, dado el gran número de observaciones con respecto a los datos ausentes, no se considera eliminar ninguna variable. Tomando en consideración los modelos que se van a emplear, se opta por tener varios conjuntos de datos, en algunos se considerara la imputación de los valores ausentes ya que pueden tener gran influencia en los resultados como son modelos de Redes Neuronales, Regresión; para técnicas basada en árboles como Árboles de Clasificación, Random Forest, Gradient Boosting, no será necesario trabajar con un conjunto de datos que no contenga missing, ya que tiene un tratamiento automático para estos y son modelos robustos. En este proceso se aumenta una variable missing, que contenga el número de datos ausentes por fila (No se eliminan observaciones).

6.1.5 Transformación y selección inicial de variables.

Dado el gran número de variables, conociendo además que la mayor parte de ellas son de tipo categóricas y principalmente considerando que el objetivo principal es encontrar la influencia de los aspectos de la población escolar representados en las variables de estudio, no se pone énfasis en el proceso transformación de variables (se perdería interpretación).

Se considera inicialmente una preselección de las variables ya que se trabaja con gran número ellas y se requiere utilizar las más útiles en la fase de modelización (Los resultados son similares para variable con missing o imputadas). Para ello, en primera instancia dentro de la taxonomía de métodos de selección de variables se utiliza métodos filters, los cuales ordenan las variables por importancia, utilizando algunos estadísticos y su representación gráfica como el X^2, V de Cramer, un criterio R^2 mínimo en los que se tiene en cuenta la relación con la variable objetivo. Esta preselección se realiza con el objetivo de implementar técnicas de Regresión Logística y árboles de clasificación donde se puedan comprender los resultados y evaluar el poder predictivo que tienen los aspectos de estudio (variables de la encuesta) con el haber consumido alguna vez las distintas drogas.

En esta primera selección se opta por la creación de una variable aleatoria, para tener una referencia de la utilidad de las variables. En la tabla 6.1.5.1 se encuentran las variables que ya no participaran en el estudio, debido a las siguientes razones: el valor del estadístico R^2 está por debajo de la variable aleatoria, lo que indica que no existe relación alguna con la variable dependiente, además del criterio de un R^2 mínimo conjuntamente observando que sea de poca importancia con respecto a la variable objetivo de Consumo de marihuana. En el Anexo II se encuentran las variables excluidas de los otros análisis.

P76	Quién es tu apoderado/apoderada?
P77	¿Has conversado con tu padre, madre o apoderado/a acerca de las consecuencias del consumo de drogas?
P73	¿Quién es el jefe de tu hogar?
P6_c	¿Cuál crees tú que es el riesgo que corre una persona que Fuma una o más cajetillas de cigarros al día?
P6_g	¿Cuál crees tú que es el riesgo que corre una persona que toma uno o dos tragos de alcohol todos o casi todos los días
P91_a	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio, Has sido molestado/a estando solo/sola, por un grupo del colegio?
P83	Durante este año, ¿Te ha tocado asistir o participar en el colegio en actividades específicamente destinadas a prevenir el consumo de drogas, como por ejemplo charlas o talleres?
P93	¿Cuán probable es que termines cuarto medio?
P1	Sexo
P87	En general, ¿consideras que en tu colegio hay estudiantes que traen, toman o comparten alcohol dentro del colegio?
P110	Pensando en los últimos 7 días, ¿cuántos días hiciste ejercicio o actividad física, fuera del horario de clases, durante al menos 20 minutos y que te haya hecho transpirar o respirar fuertemente?
P94	¿Cuán probable es que sigas estudiando después del colegio?
P91_e	Te han robado algo en el colegio
P74_c	¿Qué educación alcanzaron el jefe/a de hogar?
P74_a	¿Qué educación alcanzo tu padre?
P5	En general, ¿cuánto crees que tu padre, madre, apoderado o apoderada (o alguno de ellos) conocen a tus amigos y amigas más cercanos/as?

P74_b	¿Qué educación alcanzo tu madre?
P92	¿Cuán probable es que pases de curso este año?
P91_b	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio, Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio ?

Tabla 6.1.5.1. Consumo de Marihuana: Variables excluidas del estudio.

En la figura 6.1.5.1 y tabla 6.1.5.2, se puede observar la importancia de las variables en relación con la variable dependiente de consumo de marihuana. Los gráficos y tablas de "importancia de variables" de las otras variables objetivos se encuentran en el Anexo II. Se encuentran resaltadas las variables que coinciden como importantes en los diferentes estudios. Además, se observa que la variable de entrada consumo de marihuana en el análisis de Consumo de cocaína y pasta base se incluyen dentro de las importantes.

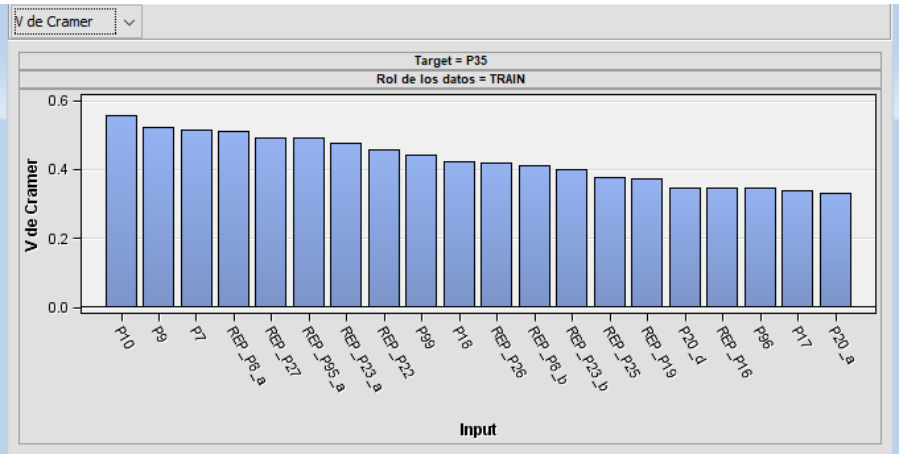


Figura 6.1.5.1. Consumo de Marihuana: Importancia de Variables V de Cramer.

P10	¿Cuándo fue la última vez que fumaste un cigarrillo?
P9	¿Cuándo fue la primera vez que fumaste cigarrillos?
P7	¿Has fumado cigarrillos alguna vez en la vida?
P8_a	¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar?
P95_a	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo marihuana con el evidente propósito de volarse, drogarse o embriagarse?
P23_a	¿Cuántas veces en tú vida te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P22	¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol?
P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana?
P18	¿Cuándo fue la última vez que tomaste alcohol?
P26	Indica, de 1 a 10, qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol,
P8_b	¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días?
P23_b	¿Cuántas veces en los últimos 12 meses te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P25	Pensando en el último día que consumiste alcohol ¿cuál de las siguientes bebidas alcohólicas fue la que más tomaste ese día?
P19	¿Cuán difícil te sería comprar alguna bebida alcohólica, si quisieras hacerlo?
P20_d	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Tener relaciones sexuales sin condón
P16	¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica?
P96	Si en tu grupo de amigas y amigos cercanos supieran que fumas marihuana ¿tú crees
P17	¿Cuándo fue la primera vez que probaste alcohol?

P20_a	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Amigos, amigas o familiares te han sugerido o mencionado que disminuyas el consumo de alcohol

Tabla 6.1.5.2. Consumo de Marihuana: Importancia de las variables.

En definitiva, para las técnicas de Árbol de clasificación y Regresión que proporcionan información relevante sobre la información de nuestras variables, se cuenta con dos conjuntos de datos, uno en lo que se mantiene los valores ausentes (Conjunto A) y otro conjunto de datos donde se ha realizado una imputación de los datos, sustituyendo por valores válidos teniendo en cuenta su distribución (Conjunto B). Para las distintas variables dependientes se tiene su set de variables que participaran en el estudio (la mayoría de las variables coinciden en los diferentes estudios). A continuación, se resumen en la siguiente tabla 6.1.5.3. Para su mejor comprensión se utilizarán los identificadores prima y doble al referirse a los conjuntos de datos con variable objetivo P47_53 y P66_f_g_i_j respectivamente.

Variable Dependiente	Número y set de variable que participaran en el estudio
P35: Consumo de Marihuana **Conjunto de Datos A y B**	**66 variables:** P26 P10 P105 P106 P108 P109 P15 P17 P18 P20_a P20_b P20_c P20_d P3 P6_a P6_b P6_d P6_e P7 P75 P79 P80 P81_a P81_b P85 P86 P88 P89 P9 P96 P97 P98 P99 P16 P19 P22 P23_a P23_b P23_c P25 P27 P4 P6_f P82_a P82_b P82_c P82_d P84 P8_a P8_b P90_a P90_b P90_c P90_d P90_e P91_c P91_d P95_a P95_b P95_c P95_d P95_e P11 P12 P2 P21
P47_53: Consumo de Cocaína o Pasta Base **Conjunto de Datos A' y B'**	**73 variables:** P2 P3 P6_b P6_e P73 P75 P78 P79 P80 P81_b P85 P86 P96 P97 P98 P99 P7 P9 P10 P11 P12 P17 P18 P20_a P20_b P20_c P20_d P21 P105 P108 P109 P35 P16 P19 P22 P23_a P23_b P23_c P25 P26 P27 P4 P6_c P6_f P6_g P74_a P74_b P76 P82_a P82_b P82_c P82_d P84 P8_a P8_b P90_a P90_b P90_c P90_d P90_e P91_a P91_b P91_c P91_d P91_e P92 P93 P94 P95_a P95_b P95_c P95_d P95_e
P66_f_g_i_j: Consumo Otras Drogas **Conjunto de Datos A'' y B''**	**70 variables:** P2 P3 P6_b P6_e P73 P75 P78 P80 P81_b P85 P86 P96 P97 P98 P99 P7 P9 P10 P11 P12 P17 P18 P20_a P20_b P20_c P20_d P21 P105 P108 P109 P35 P16 P19 P22 P23_a P23_b P23_c P25 P26 P27 P4 P6_c P6_f P6_g P76 P82_a P82_b P82_c P82_d P84 P8_a P8_b P90_a P90_b P90_c P90_d P90_e P91_a P91_b P91_c P91_d P91_e P92 P93 P94 P95_a P95_b P95_c P95_d P95_e

Tabla 6.1.5.3. Conjunto de Datos y Variables de los estudios.

Para la construcción de los modelos predictivos en el consumo de marihuana se ha tenido en cuenta modelos de selección de variables que ordenan la importancia de las variables y otros métodos Wrapper de búsqueda secuencial, en el estudio se considera las variables de selección y las variables más importantes que resulten de la ejecución del mejor modelo de regresión y árboles de clasificación del objetivo principal, se incluye

además las variables que resulten de la importancia en un árbol de gradient boosting, siendo la intersección de los resultados el punto de partida, se toma en cuenta las variables que mayor coincidencias tengan por los cuatro diferentes métodos mencionados; las cuales se encuentran ordenadas en la tabla 6.1.5.3. La selección de variables por las diferentes técnicas se encuentra en el Anexo III.

Variable	Coincidencias		Variable	Coincidencias
P95_a	4		P19	3
P7	4		P8_a	2
P99	4		P8_b	2
P9	4		P22	2
P27	4		P98	2
P23_a	4		P97	2
P96	4		P82_c	2
P79	3		P82_d	2
P10	3		P80	2
P84	3		P26	2
P85	3		P16	2
P25	3		P20_d	2

Tabla 6.1.5.4. Coincidencias Selección de variables.

Teniendo en cuenta los datos con missing, imputados y la última tabla se han definido otros dos conjuntos de datos, con diferentes set de variables, las cuales se describen a continuación en la tabla 6.1.5.5.

Conjunto de Datos	Número y Set de variable
Conjunto "C"	**13 variables:** **P95_a P7 P99 P9 P27 P23_a P96 P79 P10 P84 P85 P25 P19 P2**
Conjunto "D"	**25 variables:** **P95_a P7 P99 P9 P27 P23_a P96 P79 P10 P84 P85 P25 P19 P8_a P8_b P22 P98 P97 P82_c P82_d P80 P26 P16 P20_d P2**

Tabla 6.1.5.5. Conjunto de Datos para otras técnicas de predicción

La descripción de las variables con sus categorías corregidas se encuentra en el Anexo IV. En los siguientes apartados, utilizaremos "A", "B", "C" y "D", para referirnos a los conjuntos de datos que se describieron anteriormente.

6.2 CONSTRUCCIÓN DE MODELOS DE MACHINE LEARNING

Posteriormente al proceso de depuración, se procede a la construcción y ejecución de técnicas de problemas predictivos supervisados de Machine Learning; en las que inicialmente se busca resultados simples y que se puedan comprender a través de modelos de Árboles de Clasificación y Regresión logística.. Así se podrá encontrar asociaciones y evaluar el poder predictivo de los aspectos de los estudiantes en el consumo de drogas, además de contar con modelos predictivos que puedan dan lugar a soluciones suficientemente buenas; luego con el objetivo de predecir de mejor manera el haber consumido alguna vez marihuana para nuevos estudiantes, se procede al modelado y evaluación de varias técnicas predictivas como son Redes Neuronales, Random Forest, Gradiente Boosting y técnicas de ensamblado.

Se utilizan los softwares de lenguaje Sas y R, aprovechando las ventajas de estos dos distintos lenguajes de programación estadística. Para las técnicas de Data Mining de árboles de clasificación y regresión logística, donde se utilizan un mayor número de set de variables se utilizan Enterprise Miner de Sas, para la técnica de Redes Neuronales se utiliza Sas Base, para los modelos de Random Forest, Gradient Boosting y métodos de ensamblado se implementa con R-Studio.

A continuación, se describen los diferentes modelos empleados en él estudio, donde se han definido y variado cada uno de sus parámetros para obtener los mejores resultados. Para la evaluación de los modelos se utiliza varias medidas propias de su metodología y para la comparación de modelos se implementa training test y validación cruzada con distintas semillas.

6.2.1 Árbol de Clasificación

Variable Objetivo: Consumo de Marihuana.

Se han construido varios árboles utilizando distintos criterios para la selección del punto de corte y de variables; gestionando los missing para el conjunto de datos “B”, aplicando también el ajuste del p-valor, entre otros aspectos con el fin de predecir de mejor manera las variables objetivos (ha consumido alguna vez droga).

En resumen se realiza una partición de datos, en algunos casos 70 entrenamiento y 30 test o 60 entrenamiento, 20 validación y 20 test; se inicia construyendo un árbol con las características mayor, para observar el error en el gráfico de evaluación de subárbol y poder identificar sus niveles para ver si es posible generar un árbol más pequeño donde se tenga un error similar, pero con menos hojas (menos complicado), luego se ha generado varios árboles especificando el número de hojas y realizando la poda para evitar el sobreajuste, cambiando los criterios que se mencionaron anteriormente. Finalmente se realiza un training test con 10 repeticiones utilizando en algunos casos validación cruzada para así poder evaluar el comportamiento de los modelos en sesgo y varianza.

A continuación, en la tabla 6.2.1.1 se resumen las características de los modelos de árboles de clasificación que han presentado mejores resultados.

Árbol - Conjunto de Datos	Características	Medidas de Evaluación	Medida de Evaluación
	Punto de Corte: ProbChisq Profundidad máx: 6 Tam, hoja: 5	**Missclasification Rate**	**Índice Roc**
Árbol 1 - A	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	0.1814	0.881
Árbol 9 - B	p-valor:0.20 Gestión missing: utilizar en búsqueda	0.1805	0.882
	Punto de Corte: ProbChisq Profundidad máx: 8 Tam. hoja: 10	**Missclasification Rate**	**Índice Roc**
Árbol 2 - A	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad - Si	**0.1794**	**0.892**
Árbol 10 - B	p-valor:0.15 Gestión missing: rama + correlada	**0.1784**	**0.892**
	Punto de Corte: Gini Profundidad máx: 6 Tam. hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 4 - A	Gestión missing: utilizar en búsqueda	0.1821	0.88
Árbol 11 - B		0.1810	0.88
	Punto de Corte: Entropía Profundidad máx: 10 Tam. hoja: 9	**Missclasification Rate**	**Índice Roc**
Árbol 5 - A	Gestión missing: utilizar en búsqueda	0.1799	0.892
Árbol 12 - B		0.1821	0.89
	Punto de Corte: ProbChisq Profundidad máx: 8 Tam, hoja: 7	**Missclasification Rate**	**Índice Roc**
Árbol 6 - A	Ajuste p-valor: Bonferroni - Después Ajuste p-valor Profundidad: Si	**0.1780**	**0.893**
Árbol 13 - B	p-valor:0.20 Gestión missing: utilizar en búsqueda	**0.1787**	0.892
	Punto de Corte: ProbChisq Profundidad máx: 14 Tam, hoja: 12	**Missclasification Rate**	**Índice Roc**
Árbol 7 - A	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	**0.1792**	**0.893**
Árbol 14 - B	p-valor:0.1 Gestión missing: rama + grande	**0.1799**	**0.892**
	Punto de Corte: ProbChisq Profundidad máx: 18 Tam, hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 8 - A	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	**0.1790**	**0.892**
Árbol 15 - B	p-valor:0.25 Gestión missing: utilizar en búsqueda	**0.1790**	**0.891**

Tabla 6.2.1.1. Consumo de Marihuana: Mejores modelos Árboles de Clasificación.

Se han seleccionado los mejores árboles, considerando los criterios de evaluación de Missclasification Rate (menor) y Índice Roc (mayor), este último indica que el modelo tiene un alto poder predictivo y como se observa en la figura 6.2.1.1, no presenta mayor diferencia entre los modelos analizados; se ha tomado particularmente los resultados de Missclasification Rate, que a partir de los diferentes conjuntos de datos y variables seleccionadas han dado lugar como mejor modelo al ÁRBOL 6 con características de punto de corte ProbChisq, profundidad 8, tamaño de hoja 7 y p-valor 0.20. Otros buenos resultados se presentan en los ÁRBOLES 2-7-10-13-14-15, se ha de considerar el principio de parsimonia es decir un árbol con menos hojas, para su mejor interpretación.

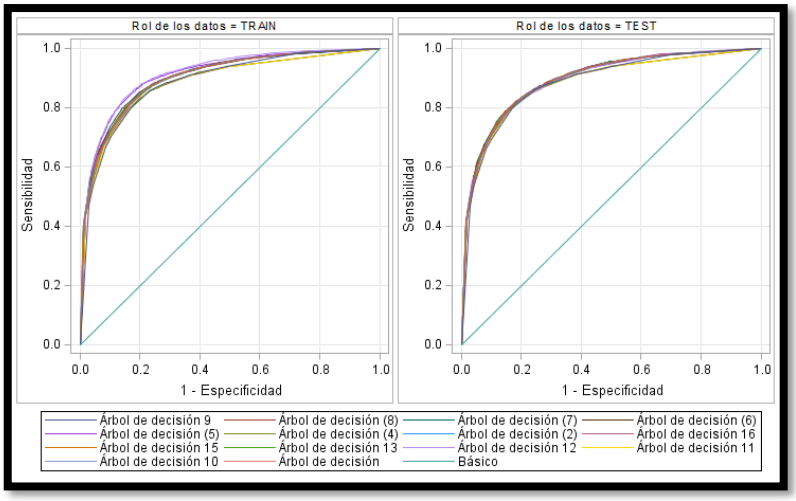


Figura 6.2.1.1. Consumo de Marihuana: Curva Roc modelos Árboles de Clasificación.

Se realiza un Training-Test con distintas semillas para la partición aleatoria, considerando también validación cruzada y así observar si existen diferencias en el comportamiento de los mejores modelos descritos anteriormente, tanto desde el punto de vista de sesgo, como de varianza. Se realiza un training-test con diez repeticiones y validación cruzada con diez subconjuntos y diez repeticiones.

En la figura 6.2.1.2, en un diagrama de cajas se resume el comportamiento de los modelos, donde se observa la magnitud de los errores de la tasa de clasificación errónea y la variabilidad de estos.

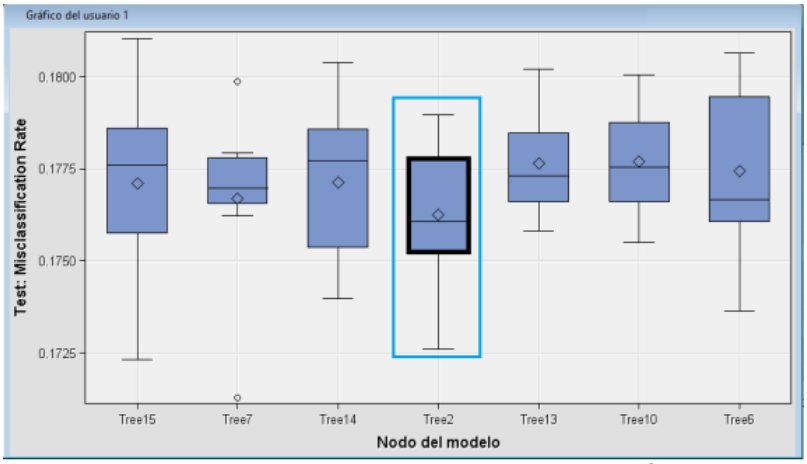


Figura 6.2.1.2. Consumo de Marihuana: Comparación de modelos Árboles de Clasificación.

Observando los resultados, el mejor modelo viene dado por el **Árbol 2** que ha resultado ganador en 7 de las 10 iteraciones, aunque presenta más variabilidad que otros modelos, se observa que es el que menor error promedio presenta **0.1761**, sus características son conjunto de datos "A", punto de corte ProbChisq, profundidad 8, tamaño de hoja 10, con p-valor 0.15 y gestión de missing rama más correlada.

A continuación se presentan algunas características del mejor modelo seleccionado, más adelante estos resultados serán analizados en la sección de análisis de resultados, donde se describirán aspectos importantes para conocer y comprender posibles

escenarios de una serie de aspectos relacionados con la población escolar de Chile en el haber consumido marihuana.

Las variables más importantes se muestran a continuación en la tabla 6.2.1.2.

Nombre de la variable	Número de reglas de división	Importancia
P7	1.0	1.0
P95_a	2.0	0.658
P27	3.0	0.316
P11	8.0	0.240
P23_a	6.0	0.224
P99	7.0	0.223
P84	6.0	0.176
P96	7.0	0.174
P79	5.0	0.100
P82_d	4.0	0.088
P82_c	3.0	0.086
P85	3.0	0.079
P97	2.0	0.064
P86	1.0	0.056
P9	1.0	0.053
P95_c	2.0	0.048
P26	1.0	0.048
P8_b	1.0	0.047
P25	2.0	0.044
P12	1.0	0.040
P95_d	1.0	0.039
P23_c	2.0	0.039
P95_e	1.0	0.039
P15	2.0	0.038

Tabla 6.2.1.2. Consumo de Marihuana: Variables más importantes.

Se muestra también en la figura 6.2.1.3, **la matriz de confusión** que contiene las observaciones que han sido bien y mal clasificadas, además se muestran algunas medidas de clasificación que se pueden considerar interesantes en el análisis de resultados.

Tabla de REP_P35 por U_REP_P35

REP_P35(Replacement: P35)
U_REP_P35(Unnormalized Into: REP_P35)

Frecuencia|	0|	1|	Total
0	18167	2900	21067
1	3022	8919	11941
Total	21189	11819	33008

Tasa de acierto	82.06%
Tasa de Fallo	17.94%
Sensibilidad o Tasa de verdaderos positivos	86.23%
Especificidad o Tasa de verdades negativos	74.69%

Figura 6.2.1.3. Consumo Marihuana: Matriz de Confusión y medidas de clasificación.

Variable Objetivo: Consumo de Cocaína o Pasta Base.

Del mismo modo que el estudio anterior, con los mismos conjuntos de Datos, refiriéndose a estos con la utilización de datos missing y emputados pero con otros set de variables y diferentes número de observaciones (Conjunto de datos A' y B'), se han construido varios modelos de árboles de clasificación, especificando y variando sus características. Se ha utilizado una partición estratificada con respecto a las variables objetivo con 60 en entrenamiento, 20 validación y 20 test para la construcción y evaluación.

A continuación, en la tabla 6.2.1.3 se resumen las características de los modelos de árboles de clasificación que han presentado mejores resultados para la variable dependiente Consumo de Cocaína y Pasta Base.

Árbol - Conjunto de Datos	Características	Medidas de Evaluación	Medida de Evaluación
	Punto de Corte: ProbChisq Profundidad máx: 6 Tam, hoja: 5	**Missclasification Rate**	**Índice Roc**
Árbol 29 - A'	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	0.0721	0.847
Árbol 16 - B'	p-valor:0.20 Gestión missing: utilizar en búsqueda	0.0696	0.847
	Punto de Corte: ProbChisq Profundidad máx: 8 Tam. hoja: 10	**Missclasification Rate**	Índice Roc
Árbol 23 - A'	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad - Si	0.0724	0.847
Árbol 17 - B'	p-valor:0.15 Gestión missing: rama + correlada	0.0724	0.854
	Punto de Corte: Gini Profundidad máx: 6 Tam. hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 24 - A'	Gestión missing: utilizar en búsqueda	0.0726	0.85
Árbol 18 - B'		0.0725	0.849
	Punto de Corte: Entropía Profundidad máx: 10 Tam. hoja: 9	**Missclasification Rate**	**Índice Roc**
Árbol 25 - A'	Gestión missing: utilizar en búsqueda	**0.0777**	**0.863**
Árbol 19 - B'		**0.0763**	**0.864**
	Punto de Corte: ProbChisq Profundidad máx: 8 Tam, hoja: 7	**Missclasification Rate**	**Índice Roc**
Árbol 26 - A'	Ajuste p-valor: Bonferroni - Después Ajuste p-valor Profundidad: Si	**0.0731**	**0.853**
Árbol 20 - B'	p-valor:0.20 Gestión missing: utilizar en búsqueda	**0.0724**	**0.854**
	Punto de Corte: ProbChisq Profundidad máx: 14 Tam, hoja: 12	**Missclasification Rate**	**Índice Roc**
Árbol 27 - A'	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	**0.0727**	**0.845**
Árbol 21 - B'	p-valor:0.1 Gestión missing: rama + grande	**0.0725**	**0.847**
	Punto de Corte: ProbChisq Profundidad máx: 18 Tam, hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 28 - A'	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	**0.0729**	**0.854**
Árbol 22 - B'	p-valor:0.25	**0.0732**	**0.847**

	Gestión missing: utilizar en búsqueda		

Tabla 6.2.1.3. Consumo de Cocaína o Pasta Base: Mejores modelos Árboles de Clasificación

Se han identificado los mejores modelos (resaltados en verde) según los criterios de evaluación de Missclasification Rate (menor) y Índice Roc (mayor). Para una mejor interpretación en cuanto a la magnitud del error y su variabilidad de los modelos seleccionados, se utiliza validación cruzada repetida con diez subconjuntos y diez repeticiones. En la figura 6.2.1.4, en un diagrama de cajas con el objetivo de seleccionar un único y mejor modelo, se muestra su representación.

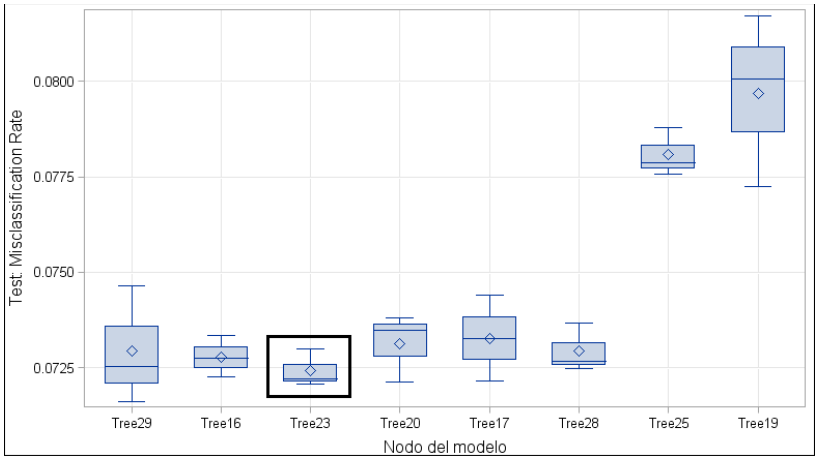


Figura 6.2.1.4. Consumo de Cocaína y Pasta Base: Comparación de modelos Árboles de Clasif.

Observando los resultados, el mejor modelo viene dado por el **Árbol 23**, que ha resultado ganador en 7 de las 10 iteraciones, aunque es peor en media que otros modelos, se observa que es el que menor error promedio presenta **0.0722**, sus características son conjunto de datos " A' " con missing, punto de corte ProbChisq, profundidad 8, tamaño de hoja 10, con p-valor 0.15 y gestión de missing rama más correlada.

A continuación se presentan algunas características del mejor modelo seleccionado, más adelante del mismo modo que el anterior análisis los resultados serán analizados en la sección de análisis de resultados, donde se describirán aspectos importantes para conocer y comprender posibles escenarios de una serie de aspectos relacionados con la población escolar de Chile en el haber consumido cocaína o pasta base.

Las variables más importantes se muestran a continuación, en la tabla 6.2.1.4.

Nombre de la variable	Número de reglas de división	Importancia
P95_b	2.0	1.0
P35	2.0	0.604
P95_c	1.0	0.371
P27	1.0	0.356
P23_c	2.0	0.332
P85	1.0	0.210
P21	1.0	0.207
P20_d	3.0	0.192

P97	1.0	0.187
P74_b	1.0	0.159
P11	1.0	0.150
P82_d	1.0	0.141
P96	1.0	0.139
P108	1.0	0.111
P79	1.0	0.107
P8_a	1.0	0.106
P16	1.0	0.104
P22	1.0	0.080
P20_b	1.0	0.077
P18	1.0	0.076
P20_a	1.0	0.075
P3	1.0	0.072
P94	1.0	0.072

Tabla 6.2.1.4. Consumo de Cocaína o Pasta Base: Variables más importantes.

Se muestra también para la variable objetivo Consumo de Cocaína en la figura 6.2.1.5, **la matriz de confusión,** que contiene las observaciones que han sido bien y mal clasificadas; además, se muestran también algunas medidas de clasificación de interés para la sección de análisis de resultados.

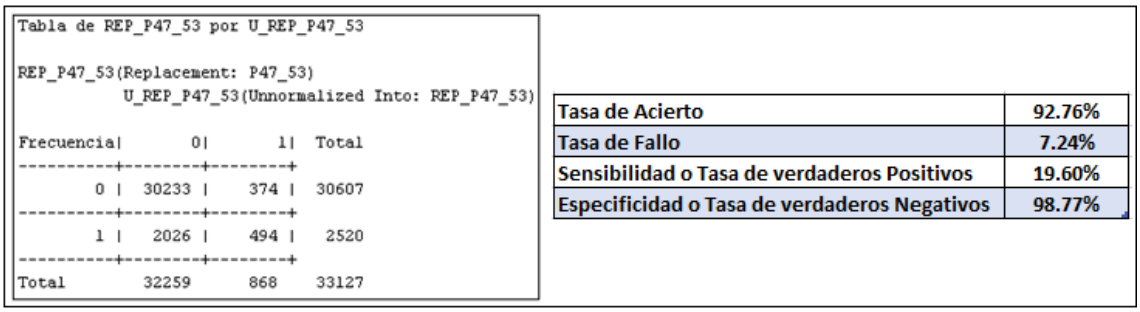


Tabla de REP_P47_53 por U_REP_P47_53

REP_P47_53(Replacement: P47_53)
U_REP_P47_53(Unnormalized Into: REP_P47_53)

Frecuencia|	0|	1|	Total
0	30233	374	30607
1	2026	494	2520
Total	32259	868	33127

Tasa de Acierto	92.76%
Tasa de Fallo	7.24%
Sensibilidad o Tasa de verdaderos Positivos	19.60%
Especificidad o Tasa de verdaderos Negativos	98.77%

Figura 6.2.1. 5. Consumo Cocaína o Pasta Base: Matriz de Confusión y medidas de clasificación.

Variables Objetivo: Consumo Otras Drogas.

De igual manera, para la variable objetivo consumo de otras drogas, utilizando los conjuntos de datos A" y B" y las misma partición de datos del análisis anterior, se construyen varios modelos de árboles de clasificación, cuyas características y medidas de evaluación, se encuentran parametrizadas en la tabla 6.2.1.5 que se muestra a continuación.

Árbol - Conjunto de Datos	Características	Medidas de Evaluación	Medida de Evaluación
	Punto de Corte: ProbChisq Profundidad máx: 6 Tam, hoja: 5	**Missclasification Rate**	**Índice Roc**
Árbol 43 - A"	Ajuste p-valor: Bonferroni - Antes Ajuste p-valor Profundidad: Si	**0.0684**	**0.773**
Árbol 30 - B"	p-valor:0.20 Gestión missing: utilizar en búsqueda	**0.0674**	**0.851**
	Punto de Corte: ProbChisq Profundidad máx: 8	**Missclasification Rate**	**Índice Roc**

Árbol 37 - A"	Tam. hoja: 10 Ajuste p-valor: Bonferroni - Antes	**0.0686**	**0.775**
Árbol 31 - B"	Ajuste p-valor Profundidad - Si p-valor:0.15 Gestión missing: rama + correlada	**0.0688**	**0.859**
	Punto de Corte: Gini Profundidad máx: 6 Tam. hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 38 - A"	Gestión missing: utilizar en búsqueda	**0.0681**	**0.843**
Árbol 32 - B"		**0.0679**	**0.781**
	Punto de Corte: Entropía Profundidad máx: 10 Tam. hoja: 9	**Missclasification Rate**	**Índice Roc**
Árbol 39 - A"	Gestión missing: utilizar en búsqueda	**0.0717**	**0.873**
Árbol 33 - B"		**0.0726**	**0.873**
	Punto de Corte: ProbChisq Profundidad máx: 8 Tam, hoja: 7	**Missclasification Rate**	**Índice Roc**
Árbol 40 - A"	Ajuste p-valor: Bonferroni - Después	**0.0696**	**0.856**
Árbol 34 - B"	Ajuste p-valor Profundidad: Si p-valor:0.20 Gestión missing: utilizar en búsqueda	**0.0689**	**0.859**
	Punto de Corte: ProbChisq Profundidad máx: 14 Tam, hoja: 12	**Missclasification Rate**	**Índice Roc**
Árbol 41 - A"	Ajuste p-valor: Bonferroni - Antes	**0.068**	**0.844**
Árbol 35 - B"	Ajuste p-valor Profundidad: Si p-valor:0.1 Gestión missing: rama + grande	**0.0691**	**0.859**
	Punto de Corte: ProbChisq Profundidad máx: 18 Tam, hoja: 15	**Missclasification Rate**	**Índice Roc**
Árbol 42 - A"	Ajuste p-valor: Bonferroni - Antes	**0.0696**	**0.855**
Árbol 36 - B"	Ajuste p-valor Profundidad: Si p-valor:0.25 Gestión missing: utilizar en búsqueda	**0.0689**	**0.858**

Tabla 6.2.1.5. Consumo Otras Drogas: Mejores modelos Árboles de Clasificación.

De los mejores modelos resaltados en la tabla anterior, se muestra en la figura 6.2.1.6, la representación en un diagrama de cajas de su error y variabilidad, utilizando también validación cruza de diez subconjuntos y diez repeticiones.

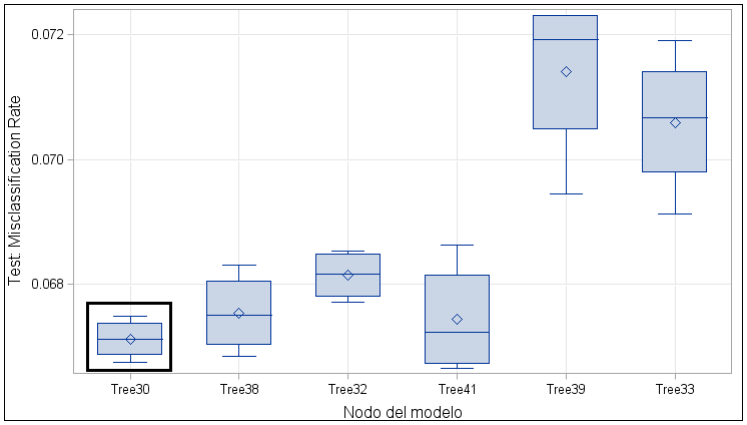


Figura 6.2.1.6. Consumo Otras Drogas: Comparación de modelos Árboles de Clasificación.

Observando los resultados, el mejor modelo viene dado por el **Árbol 30**, que ha resultado ganador en 6 de las 10 iteraciones, es el que menor error promedio presenta

0.068 y baja variabilidad, en cuanto a su capacidad predictiva tiene un valor de ROC de 0.85; sus características son conjunto de datos B" imputados, punto de corte ProbChisq, profundidad 6, tamaño de hoja 5, con un p-valor 0.20.

A continuación, en la tabla 6.2.1.6 y figura 6.2.1.7 se presentan las variables importantes del análisis y su matriz de confusión respectivamente, está última con varias medidas de clasificación del mejor modelo seleccionado.

Nombre de la variable	Número de reglas de división	Importancia
P47_53	1.0	1.0
P95_d	2.0	0.576
P23_c	1.0	0.298
P91_b	2.0	0.248
P21	3.0	0.176
P95_c	2.0	0.173
P90_e	1.0	0.141
P85	1.0	0.134
P75	2.0	0.120
P95_e	1.0	0.103
P73	1.0	0.096
P20_b	1.0	0.091
P16	1.0	0.076
P95_a	1.0	0.070
P18	1.0	0.067
P27	1.0	0.061
P82_d	1.0	0.058
P82_c	1.0	0.055
P91_c	1.0	0.053

Tabla 6.2.1.6. Consumo Otras Drogas: Variables más importantes

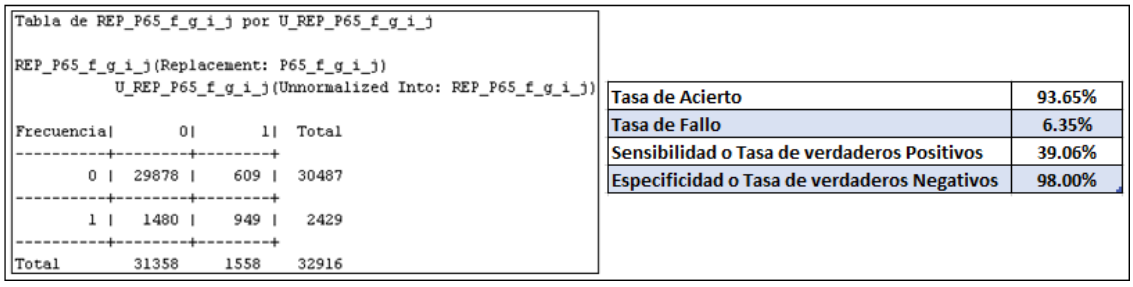


Tabla de REP_P65_f_g_i_j por U_REP_P65_f_g_i_j

REP_P65_f_g_i_j(Replacement: P65_f_g_i_j)
U_REP_P65_f_g_i_j(Unnormalized Into: REP_P65_f_g_i_j)

Frecuencia	0	1	Total
0	29878	609	30487
1	1480	949	2429
Total	31358	1558	32916

Tasa de Acierto	93.65%
Tasa de Fallo	6.35%
Sensibilidad o Tasa de verdaderos Positivos	39.06%
Especificidad o Tasa de verdaderos Negativos	98.00%

Figura 6.2.1.7. Consumo Otras Drogas: Matriz de Confusión y medidas de clasificación.

6.2.2 Regresión Logística

Luego del proceso de depuración, así como del estudio para una primer instancia de selección de las variables de intervalo y de clase más influyentes con respecto a las variables objetivos (esto se realizó previo a no tener variables que no aportan), se han construido varios modelos de Regresión Logística, considerando los datos imputados conjunto "B", además teniendo en cuenta en algunos modelos únicamente el set de variables más importantes del modelo ganador de árboles de decisión, se lo identificara

como conjunto " C' ", este último para observar si se tiene un error semejante con un modelo mucho más sencillo.

Se construyen varios modelos de Regresión Logística, realizando otros procesos de selección de variables que ofrece el modelo (Forward, Backward y StepWise), esto con el fin de rechazar variables que realmente no aportan nada y predecir de mejor manera la variable objetivo.

En resumen, se realiza una partición de datos de 70 entrenamiento y 30 Test; se ha empezado construyendo modelos de regresión sin ninguna característica de selección de variables, luego construyendo modelos aplicando las características de selección de variables mencionadas; para luego comparar los modelos a través de los estadísticos de medidas de ajuste y comparación de modelos. Se ha realizado un training test para evaluar el comportamiento de los modelos en sesgo y varianza.

A continuación, en la tabla 6.2.2.1 se han resumido las características de los distintos modelos de regresión empleados, indicando cual es el mejor, y que será evaluado en el análisis de resultados.

Modelo – Conjunto de Datos	Tasa de Clasif. Errónea	Medidas de Evaluación AIC	Medidas de Evaluación SBC	Medidas de Evaluación ROC	Número de Parámetros DFM
Regresión 9 normal – "B"	0.1552	54265.65	55977.22	0.919	185
Regresión 1 normal – "C' "	0.1601	56070.57	56625.68	0.914	60
Regresión 11 Backward – "B"	0.1562	54235.12	55632.13	0.919	151
Regresión 3 Backward – "C' "	0.1603	56067.28	56603.88	0.914	58
Regresión 13 Fordward – "B"	0.1556	54239.7	55581.2	0.919	145
Regresión 5 Fordward – "C' "	0.1603	56067.28	56603.88	0.914	58
Regresión 15 StepWise – "B"	0.1556	54239.7	55581.2	0.919	145
Regresión 7 StepWise – "C' "	0.1603	56067.28	56603.88	0.914	58
Regresión LARS LASSO - "B"	0.1564			0.918	
Regresión LARS2 LASSO - "C' "	0.1608			0.913	

Tabla 6.2.2.1. Consumo de Marihuana: Mejores Modelos Regresión Logística.

Se identifican los mejores resultados en los diferentes criterios de los modelos, el color verde claro es el mejor en su criterio de evaluación, mientras que el amarillo es el segundo mejor.

Los modelos que recogían el set de variables más importantes del modelo ganador en árboles de decisión denominado Conjunto "C' ", no ha resultado ganadores en los criterios de evaluación (no considerar DFM) pero en si se observa que minimizando muchas variables; no difiere en gran cantidad el error de tasa de clasificación errónea, por ello no se perdería mucho valor predictivo y considerando que tiene un menor

número de parámetros (58) se tendría un modelo más sencillo en el que se podría realizar mejor la interpretación de los resultados.

Según los modelos evaluados con el conjunto de datos “B”, Regresión Fordward y StepWise se comportan similarmente, ganan en dos criterios y dos en segundo lugar; Regresión Backward resulta también ganador en dos criterios y uno en segundo lugar; el modelo de regresión sin selección de variable tiene el menor error en la tasa de clasificación errónea, pero tiene un gran número de parámetros; consideraremos estos modelos incluyendo uno del conjunto de Datos “C”, así como el mejor de regresión Lasso, para realizar un training-test con diez repeticiones con distintas semillas para la partición aleatoria y observar si existen diferencias en sus criterios de evaluación de modelo, tanto desde el punto de vista del sesgo y varianza.

En la figura 6.2.2.1 en un diagrama de cajas, se resume el comportamiento de los modelos, donde se observa la magnitud de los errores de la tasa de clasificación errónea y la variabilidad de estos.

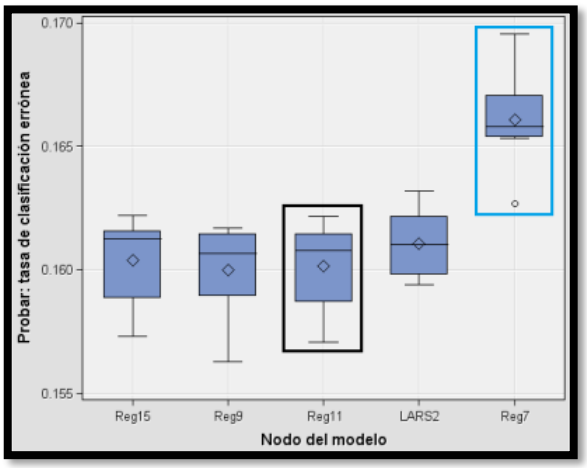


Figura 6.2.2.1. Consumo de Marihuana: Comparación de modelos Regresión Logística.

Observando los resultados, se considera que el mejor modelo (resaltado en negro) con un error promedio de **0.1601** viene dado por **Regresión11 con selección de variables Backward,** conjunto de datos “B” (ganador en 3 de las 10 iteraciones con 151 parámetros), si bien Regresión9 presenta mejores resultados (ganador en 5 de las 10 iteraciones con 185 parámetros), se considera un modelo con menor número de parámetros; el modelo de Regresión7 con selección de variables StepWise presenta el modelo con menor número de parámetros 58 (resaltado en azul), pero su error es más alto

A continuación se presentan algunas características del mejor modelo seleccionado, más adelante estos resultados serán analizados en la sección de análisis de resultados, donde se describirá la influencia de los aspectos considerados de la población escolar (variables input), en relación con el consumo de marihuana.

En la tabla 6.2.2.3, se encuentra por orden las **variables más importantes.**

Orden	Variable	DF	Wald Chi-Square	Pr > ChiSq	Orden	Variable	DF	Wald Chi-Square	Pr > ChiSq
1	P95_a	2	2251.76	<.0001	**20**	P82_c	3	84.1432	<.0001
2	P99	4	985.83	<.0001	**21**	P95_c	2	74.4925	<.0001
3	P96	3	797.39	<.0001	**22**	P11	1	72.1609	<.0001
4	P84	2	657.86	<.0001	**23**	P108	1	71.6569	<.0001
5	P79	1	460.51	<.0001	**24**	P23_c	1	63.4362	<.0001
6	P98	4	329.45	<.0001	**25**	P25	4	54.4572	<.0001
7	P23_a	2	318.23	<.0001	**26**	P106	7	53.1856	<.0001
8	P97	3	267.10	<.0001	**27**	P8_b	3	51.0282	<.0001
9	P10	3	257.10	<.0001	**28**	P2	1	46.9789	<.0001
10	P95_e	2	224.13	<.0001	**29**	REP_P26	4	41.4997	<.0001
11	P85	5	201.31	<.0001	**30**	P105	5	40.3236	<.0001
12	P27	4	188.53	<.0001	**31**	P18	3	40.0586	<.0001
13	P80	4	181.57	<.0001	**32**	P75	7	39.3598	<.0001
14	P19	3	166.64	<.0001	**33**	P81_a	5	36.1297	<.0001
15	P82_d	3	151.80	<.0001	**34**	P8_a	3	25.9784	<.0001
16	P20_d	2	129.33	<.0001	**35**	P16	3	24.1146	<.0001
17	P7	1	107.95	<.0001	**36**	P82_a	3	24.0225	<.0001
18	P9	3	100.71	<.0001	**37**	P86	2	23.7910	<.0001
19	P81_b	5	95.48	<.0001					

Tabla 6.2.2.2. Consumo de Marihuana: Mejores Variables.

Se considera, en la figura 6.2.2.2, una porción del análisis de la estimación de máxima verosimilitud que ayudara en el análisis de resultados para ejemplos de la estimación de parámetros.

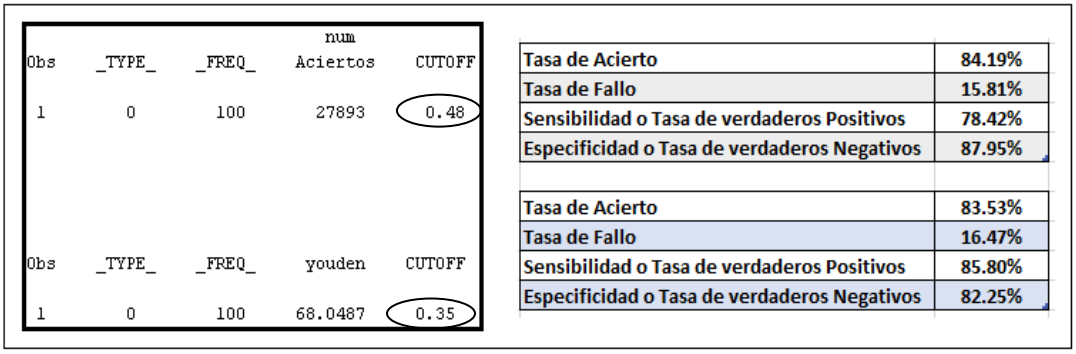
Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Estimador estandarizado	Exp(Est)
Intercept		1	-1.5185	0.1733	76.80	<.0001		0.219
IMP_P10	1	1	0.3753	0.0271	191.63	<.0001		1.455
IMP_P10	2	1	0.1591	0.0267	35.47	<.0001		1.172
IMP_P10	3	1	-0.1657	0.0273	36.78	<.0001		0.847
IMP_P105	1	1	0.0334	0.0842	0.16	0.6914		1.034
IMP_P105	2	1	0.0965	0.0861	1.26	0.2620		1.101
IMP_P105	3	1	0.0516	0.0893	0.33	0.5633		1.053
IMP_P105	4	1	0.2143	0.0854	6.30	0.0121		1.239
IMP_P105	8	1	-0.5122	0.4109	1.55	0.2126		0.599
IMP_P106	0	1	0.2198	0.0343	41.18	<.0001		1.246
IMP_P106	1	1	0.0269	0.0281	0.92	0.3388		1.027
IMP_P106	2	1	-0.0333	0.0232	2.07	0.1501		0.967
IMP_P106	3	1	-0.0882	0.0254	12.09	0.0005		0.916
IMP_P106	4	1	-0.0696	0.0292	5.69	0.0170		0.933
IMP_P106	5	1	-0.0120	0.0352	0.12	0.7344		0.988
IMP_P106	6	1	-0.0301	0.0404	0.55	0.4566		0.970
IMP_P108	1	1	0.1183	0.0140	71.66	<.0001		1.126
IMP_P109	1	1	-0.0135	0.0265	0.26	0.6112		0.987
IMP_P109	2	1	0.0694	0.0318	4.77	0.0289		1.072
IMP_P109	3	1	-0.0398	0.0275	2.10	0.1477		0.961
IMP_P109	4	1	-0.0951	0.0630	2.28	0.1311		0.909
IMP_P109	5	1	0.0852	0.0403	4.47	0.0344		1.089
IMP_P11		1	0.0172	0.00202	72.16	<.0001	0.0706	1.017

Figura 6.2.2.2. Análisis de estimaciones de máxima Verosimilitud.

Se muestra también en la figura 6.2.2.3, dos puntos de corte, el primer punto de corte en cara a maximizar la tasa de aciertos y el otro que maximiza el índice de youden el cual establece una solución para que optimice el valor de ambos estadígrafos sensibilidad y especificidad; es decir que este último considera igualmente graves los FP (falsos positivos) y FN (falsos negativos). Junto a estos valores se pueden observar

sus medidas de clasificación correspondiente, que contiene las observaciones que han sido bien y mal clasificadas, además se muestran algunas medidas de clasificación que se consideran interesantes en el análisis de resultados.



Obs	_TYPE_	_FREQ_	num Aciertos	CUTOFF
1	0	100	27893	0.48

Tasa de Acierto	84.19%
Tasa de Fallo	15.81%
Sensibilidad o Tasa de verdaderos Positivos	78.42%
Especificidad o Tasa de verdaderos Negativos	87.95%

Obs	_TYPE_	_FREQ_	youden	CUTOFF
1	0	100	68.0487	0.35

Tasa de Acierto	83.53%
Tasa de Fallo	16.47%
Sensibilidad o Tasa de verdaderos Positivos	85.80%
Especificidad o Tasa de verdaderos Negativos	82.25%

Figura 6.2.2.3. Consumo Marihuana: Punto de corte y medidas de clasif. Regresión Logística.

6.2.3 Redes Neuronales

Teniendo en cuenta que existen varias variables que no se encuentran relacionadas linealmente, además la complejidad de los datos y considerando el número de observaciones y variables, se han construido varios modelos de redes Neuronales para los conjuntos de datos imputados "C" y "D" con sus diferentes set de variables.

Con el objetivo de encontrar buenos modelos predictivos con la técnica de redes neuronales, en su construcción, se han variado algunas características, el número de nodos permitidos de la capa oculta, la función de activación y algoritmo de optimización para encontrar los pesos que minimicen el error; además se ha iniciado estudiando Early Stoping para identificar el número de iteraciones adecuados de los modelos y evitar el sobreajuste, como se observa en la tabla 6.2.3.1 no es necesario muchas iteraciones en los conjuntos de datos y los mejores resultados que minimizan el error de la variable objetivo, se tienen con la función de activación TanH (Tangente Hiperbólica) con algoritmo de optimización quanew, fueron los primeros candidatos a emplear en los modelos de redes construidos, aunque no los mejores como se evidencia más adelante.

Conjunto de Datos	Semilla	# Num. Nodos	Función Activación	Algoritmo de Optimización	Iteraciones recomendadas Early Stoping	_VOBJ_
"D"	442711	3	TanH	LEVMAR	9	0.789
"D"	123456	10	Softmax	QUANEW	21	0.758
"D"	123456	3	Softmax	LEVMAR	10	0.7418
"D"	256878	5	Log	LEVMAR	12	0.719
"D"	123467	3	TanH	QUANEW	44	0.711
"C"	124562	4	TanH	LEVMAR	12	0.778
"C"	471145	15	Softmax	QUANEW	24	0.762
"C"	471145	15	Softmax	LEVMAR	10	0.773
"C"	442711	5	Log	LEVMAR	12	0.8035
"C"	123467	3	TanH	QUANEW	44	0.729

Tabla 6.2.3.1. Consumo Marihuana: Early Stoping Redes Neuronales.

A continuación en la tabla 6.2.3.2, se resumen los distintos modelos de Redes Neuronales con sus diferentes características; los modelos de redes han trabajo mejor con un menor número de capas, con función de activación Softmax y algoritmo con

segunda derivada Levmar, se evidencia a través de las medidas de evaluación tasa de clasificación errónea y el área bajo la curva (ROC).

Conjunto de Datos	Modelo de Red	# Num. Nodos	Función Activación	Algoritmo de Optimización	Tasa de Clasif. Errónea	Medidas de Evaluación ROC
"D"	RED 1	3	Log	LEVMAR	0.2867	0.822
"D"	RED 2	3	TanH	QUANEW	0.3619	0.403
"D"	RED 3	10	Log	LEVMAR	0.3316	0.685
"D"	RED 4	5	TanH	QUANEW	0.2344	0.758
"D"	RED 5	10	TanH	TRUST-REGION	0.3606	0.607
"D"	RED 6	3	Softmax	LEVMAR	0.1656	0.909
"C"	RED 7	3	Log	Back Prop (mom = 0.01)	0.3619	0.403
"C"	RED 8	3	Log	LEVMAR	0.2686	0.762
"C"	RED 9	3	TanH	QUANEW	0.2831	0.647
"C"	RED 10	10	Log	LEVMAR	0.3619	0.358
"C"	RED 11	5	TanH	QUANEW	0.2777	0.567
"C"	RED 12	10	TanH	TRUST-REGION	0.3352	0.548
"C"	RED 13	3	Softmax	LEVMAR	0.1720	0.905
"C"	RED 14	3	Log	Back Prop (mom = 0.01)	0.2831	0.647

Tabla 6.2.3.2. Consumo Marihuana: Modelos de Redes Neuronales.

Se realiza la comparación con remuestreo de los modelos que mejor resultados han presentado, utilizando validación cruzada repetida 4 grupos y 21 repeticiones, para evaluar nuestros modelos de Redes Neuronales en Sesgo y varianza; considerando que se obtuvieron dos buenos modelos (resaltados en verde), a partir de ellos se han construido otros variando el número de capas, además se han incluido también, pocos modelos con características diferentes de la red (función de activación y algoritmo) para comprobar los resultados; en la tabla 6.2.3.3 se encuentran dichos modelos . En la figura 6.2.3.1 en un diagrama de cajas se resume el comportamiento de los modelos, para identificar fácilmente el mejor modelo predictivo de redes neuronales.

Conjunto de Datos	Modelo de Red	# Num. Nodos	Función Activación	Algoritmo de Optimización	Tasa de Clasif. Errónea	Medidas de Evaluación ROC
"D"	RED 6	3	Softmax	LEVMAR	0.1692	0.907
"D"	RED 15	4	Softmax	LEVMAR	0.1707	0.905
"D"	RED 16	2	Softmax	LEVMAR	0.1612	0.913
"C"	RED 13	3	Softmax	LEVMAR	0.1690	0.906
"C"	RED 17	7	Softmax	LEVMAR	0.1729	0.902
"C"	RED 18	1	Softmax	LEVMAR	0.1671	0.911
"D"	RED 19	10	TanH	QUANEW	0.3612	0.658
"C"	RED 20	15	TanH	QUANEW	0.3615	0.567

Tabla 6.2.3.3. Consumo Marihuana: Remuestreo mejores modelos de Redes Neuronales.

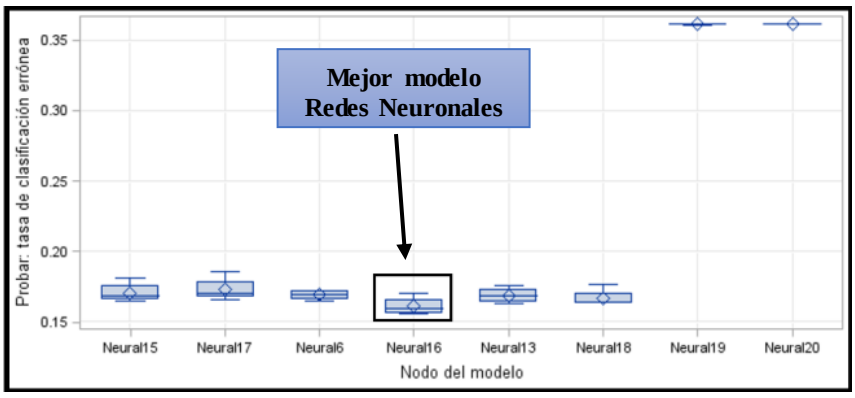


Figura 6.2.3. 1. Consumo de Marihuana: Comparación de Modelos Redes Neuronales.

En la tabla y gráfico anterior de remuestreo y comparación de modelos, se encuentra identificado el mejor modelo de Redes Neuronales, el cual ha resultado ganador en los dos criterios de evaluación, es decir menor error en tasa de clasificación errónea con valor de 0.1612 y índice ROC de 0.913, cuyo valor indica gran valor predictivo, en cuanto a sus característica se han obtenido del conjunto de Datos "D", con función de activación Softmax, con dos números de nodos en la capa oculta y algoritmo de optimización Levmar, sin necesidad de Early Stoping.

En la figura 6.2.3.2 se pueden identificar las medidas de clasificación del modelo. En el anexo V en la sección de Redes Neuronales, se encuentra el código Sas, gráfica de Early Stoping y comparación de remuestreo de los mejores modelos.

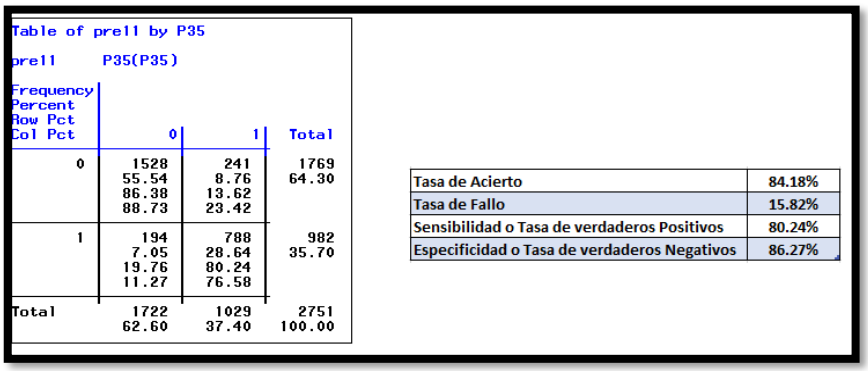


Table of prell by P35

prell / P35(P35)	0	1	Total
0	1528 55.54 86.38 88.73	241 8.76 13.62 23.42	1769 64.30
1	194 7.05 19.76 11.27	788 28.64 80.24 76.58	982 35.70
Total	1722 62.60	1029 37.40	2751 100.00

Tasa de Acierto	84.18%
Tasa de Fallo	15.82%
Sensibilidad o Tasa de verdaderos Positivos	80.24%
Especificidad o Tasa de verdaderos Negativos	86.27%

Figura 6.2.3. 2. Consumo de Marihuana: Medidas de Clasificación Redes Neuronales.

6.2.4 Random Forest

Se han implementado algunos modelos utilizando esta técnica que está basado en árboles de decisión, pero solucionan las desventajas de estos combinando los resultados de varios árboles; la idea es construir varios modelos con diferentes submuestras, para luego proceder a promediarlos, en donde se podrá obtener las probabilidades estimadas y obtener una clasificación a partir de un punto de corte o por majority voting; así se reducirá la dependencia de la estructura de los datos y podremos mejorar la varianza y sesgo de los modelos; tratando de encontrar un mejor modelo predictivo para detectar si un estudiante ha consumido marihuana.

A continuación, se construyen varios modelos a partir de los Conjuntos de Datos "C" y "D" con missing (se han incorporado a los dos conjuntos de datos dos variables continuas P2 y P12), en los primeros modelos, se utilizan todas las variables (Bagging caso particular de Random Forest), en los demás, se va incorporando aleatoriedad en las variables para segmentar cada nodo del árbol, estos últimos ayudan a la selección evitando decidir por un set de variables; es decir, se tendrá dos fuentes de variabilidad el remuestreo de observaciones y de variables, así se podrá reducir el sobreajuste. Se han parametrizado algunas características bastante influyentes con el objetivo de encontrar un buen modelo predictivo con la técnica de Random Forest, entre ellas están el número de árboles a promediar (iteraciones), el número de variables, el tamaño mínimo de nodos terminales, el número de significación p-valor que se utiliza por defecto 0.05. Una vez se identifique el modelo que minimice la tasa de fallos en cada iteración, se identificará la proporción muestral adecuada a sortear en cada iteración y se aumentará el número de árboles.

En la tabla 6.2.4.1 se observan los modelos construidos, se implementa validación cruzada de cuatro grupos y cinco repeticiones.

Modelo identificador (Random Forest)	Conjunto de Datos	# Num. Iteraciones	Nodos terminales (nodesize)	Número de variables	Promedio Tasa de fallos	Promedio de AUC (ROC)
bagging1	**"D"**	**200**	**10**	**25**	**0.1630**	**0.9107**
bagging2	**"D"**	**200**	**15**	**25**	**0.1633**	**0.9112**
bagging3	**"C"**	**200**	**10**	**15**	**0.1667**	**0.9060**
bagging4	**"C"**	**200**	**15**	**15**	**0.1656**	**0.9062**
rf1	**"D"**	**200**	**10**	**15**	**0.1623**	**0.9114**
rf2	**"D"**	**200**	**10**	**10**	**0.1621**	**0.9116**
rf3	**"D"**	**200**	**12**	**8**	**0.1618**	**0.9114**
rf4	**"D"**	**200**	**12**	**12**	**0.1620**	**0.9115**
rf5	**"D"**	**200**	**15**	**17**	**0.1614**	**0.9113**
rf6	**"C"**	**200**	**10**	**11**	**0.1654**	**0.9064**
rf7	**"C"**	**200**	**14**	**10**	**0.1662**	**0.9070**
rf8	**"C"**	**200**	**11**	**8**	**0.1641**	**0.9071**
rf9	**"C"**	**200**	**15**	**12**	**0.1652**	**0.9071**
rf10	**"C"**	**200**	**8**	**7**	**0.1655**	**0.9073**

Tabla 6.2.4.1 Consumo Marihuana: Mejores Modelos Random Forest.

El mejor modelo ha resultado del identificador rf5, el cual está formado por el conjunto de Datos "D", con 200 iteraciones, con 15 nodos terminales y 17 variables para segmentar cada nodo del árbol. En las figuras 6.2.4.1 y 6.2.4.2 se observan en un diagrama de cajas la tasa de fallos y el valor ROC de los modelos construidos, donde se evidencia y se encuentra identificado el mejor modelo. En cuanto a la tasa de fallos se tiene menor error y distribución en modelo rf5, observan el valor ROC, se tienen buenos resultados en varios modelos.

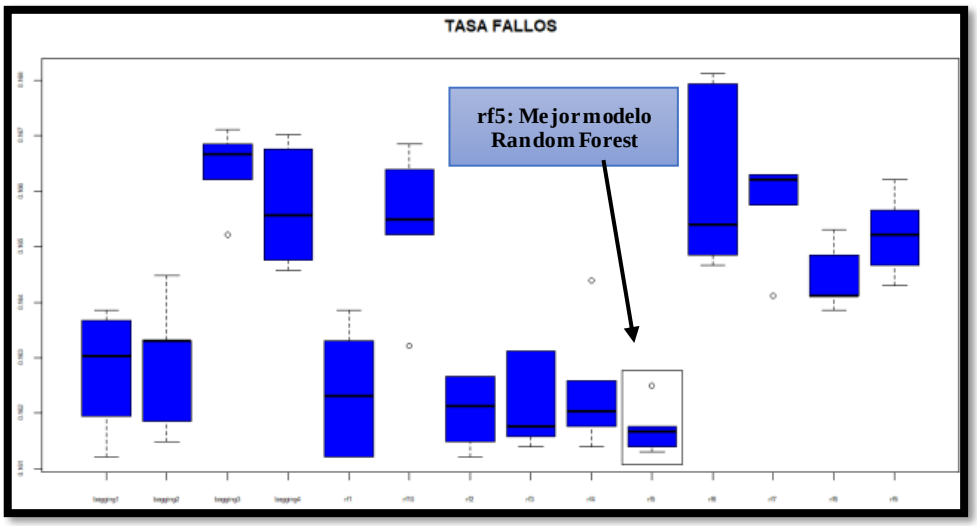


Figura 6.2.4.1. Consumo de Marihuana: Comparación de Modelos Random Forest Tasa de Fallos.

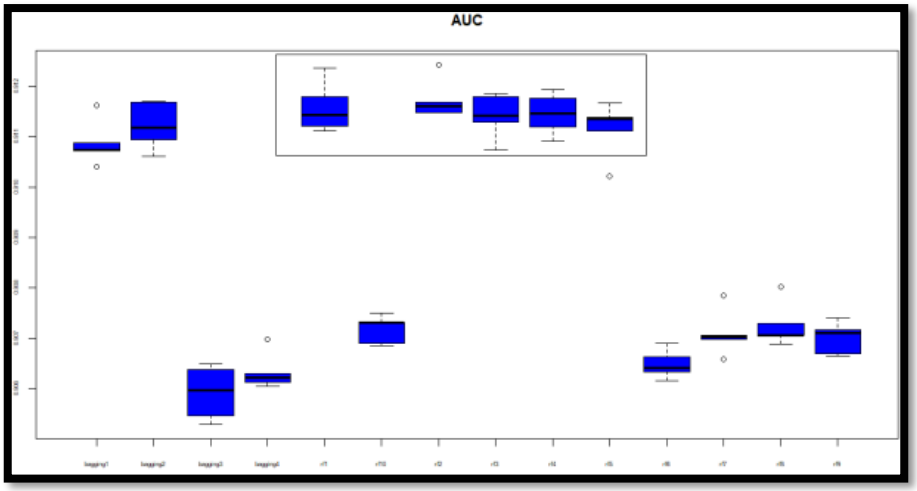


Figura 6.2.4.2. Consumo de Marihuana: Comparación de Modelos Random Forest AUC.

En el anexo VI, en la sección de Random Forest, se encuentra el código R del mejor modelo, así como el gráfico de importancia de las variables.

6.2.5 Gradient Boosting

Se utiliza la técnica de gradient boosting, para la construcción de modelos predictivos, para detectar en nuevas observaciones de la población escolar el haber consumido marihuana; en esta técnica muy popular y considerada de gran eficacia predictiva, se construyen varios modelos de árboles de clasificación (en el caso del presente estudio), en donde se modifican o se irán ajustando ligeramente las predicciones iniciales, intentando minimizar los residuos en la dirección de decrecimiento, con lo que las predicciones se irán ajustando cada vez más a los datos logrando fortalecer el modelo en respecto a la construcción de un solo árbol.

Al igual que en la análisis anterior, existen varios parámetros que se pueden monitorizar, entre los cuales se han variado el número de iteraciones, el tamaño máximo de nodos finales (minobsnnode), incorporando en esta técnica la constante de regulación shrink (v), variando los valores entre (0.001 y 0.1); que mide la velocidad de ajuste y que se ve influenciado por el número de iteraciones, así a menor v, más lento ira convergiendo al valor real, siendo necesario más iteraciones.

Utilizando el conjunto de Datos “C” y “D” missing (incorporado también las variables P2 y P12), se inicia estudiando Early Stoping para determinar el número de iteraciones, en donde se ha variado la constante de regulación. Para los dos conjuntos de datos no ha sido necesario Early Stoping; también como se observa en la figura 6.2.5.1, se realiza un tuneado, para determinar la configuración de modelos para obtener una mayor tasa de exactitud de predicción, se han ido fijando algunos parámetros para ver cómo evoluciona la constante de regularización, en función de las iteraciones, utilizando y variando el tamaño máximo de nodos finales. En la tabla 6.2.5.1 considerando los resultados sin una gran diferencia con respecto al número de iteraciones, se han identificado en orden los mejores resultados según las características descritas.

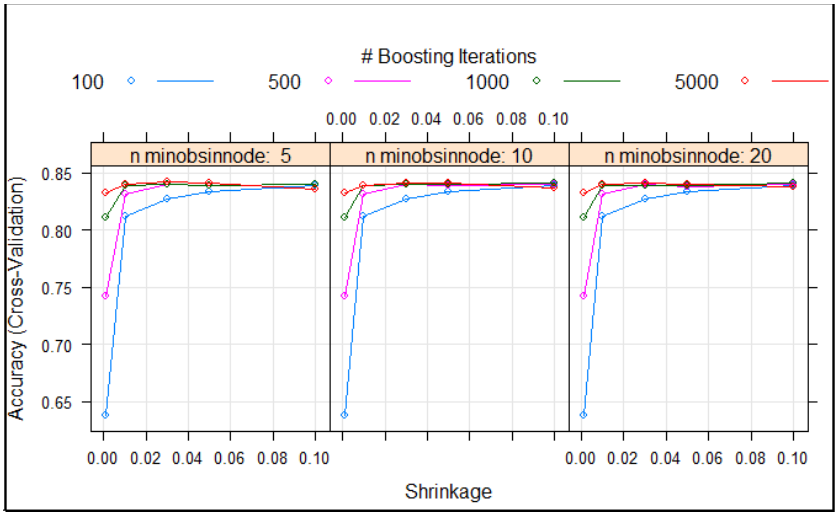


Figura 6.2.5.1. Tuneado Gradient Boosting.

shrinkage	n.minobsinnode	n.trees	Accuracy
0.100	5	500	0.8402410
0.100	20	500	0.8400591
0.030	10	500	0.8398775
0.030	20	500	0.8398775
0.030	5	500	0.8396957
0.010	20	5000	0.8396051
0.050	10	1000	0.8396051
0.100	10	500	0.8396046
0.010	10	5000	0.8395142
0.050	5	500	0.8391504
0.050	20	1000	0.8390597
0.010	5	1000	0.8388779
0.050	10	500	0.8386961
0.001	5	5000	0.8321529
0.001	10	5000	0.8321529
0.001	20	5000	0.8321529

Tabla 6.2.5.1. Tuneado Gradient Boosting mejores resultados.

Tomando en cuenta las características de los mejores resultados de la tabla anterior, utilizando los conjuntos de datos mencionados anteriormente, se han construido varios modelos de Gradient Boosting. Partiendo de ellos, se han propuesto otros modelos variando el tamaño del número de nodos para mejorar la varianza del error.

En resumen, para los modelos que se identifican en la tabla 6.2.5.2 con sus diferentes características, se ha utilizado validación cruzada repetida con 4 grupos y 5 repeticiones con el objetivo de observar, si existen diferencias en sus criterios de evaluación (tasa de fallos y el AUC) tanto desde el punto de vista de sesgo y varianza.

Modelo identificador (Gradient Boosting)	Conjunto de Datos	# Num. Iteraciones	Constante de regularización (v) Shrink	minobsinnode	Promedio Tasa de fallos	AUC

gbm1	“D”	500	0.1	5	0.1605	0.9168
gbm2	“D”	500	0.1	20	0.1597	0.9167
gbm3	“D”	500	0.03	10	0.1605	0.9152
gbm4	“D”	1000	0.01	15	0.1620	0.9132
gbm5	“D”	1000	0.05	15	0.1604	0.9174
gbm6	“D”	300	0.05	10	0.1600	0.9152
gbm7	“D”	1500	0.009	15	0.1608	0.9146
gbm8	“D”	700	0.3	15	0.1667	0.9066
gbm9	“D”	1000	0.2	20	0.1662	0.9093
gbm10	“C”	500	0.1	5	0.1635	0.9110
gbm11	“C”	500	0.1	20	0.1641	0.9108
gbm12	“C”	500	0.03	10	0.1638	0.9116
gbm13	“C”	1000	0.01	15	0.1649	0.9108
gbm14	“C”	1000	0.05	15	0.1634	0.9116
gbm15	“C”	300	0.05	10	0.1636	0.9116
gbm16	“C”	1500	0.009	15	0.1634	0.9116
gbm17	“C”	700	0.3	15	0.1738	0.9026
gbm18	“C”	1000	0.2	20	0.1722	0.9043

Tabla 6.2.5. 2. Mejores Modelos Gradient Boosting.

En la tabla anterior, se identifican los cuatro mejores modelos (sombreados en verde), los cuales han resultado del conjunto de Datos “D”, con constantes de regularización 0.1 y 0.05, un mínimo de 500 y máximo 1000 iteraciones, en cuanto a los nodos finales, estos varían entre todos los utilizados. Para seleccionar el mejor modelo en la figura 6.2.5.2 y 6.5.5.3, se muestran en diagramas de cajas la representación de la tasa de fallos y AUC respectivamente acompañados de su varianza.

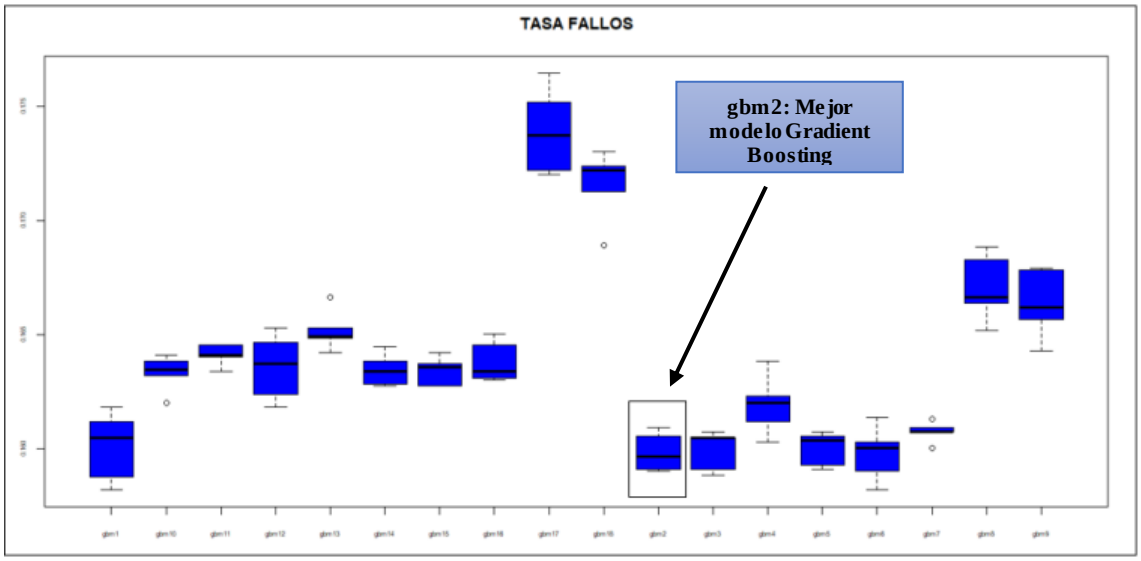


Figura 6.2.5.2. Comparación de Modelos Gradient Boosting Tasa de Fallos.

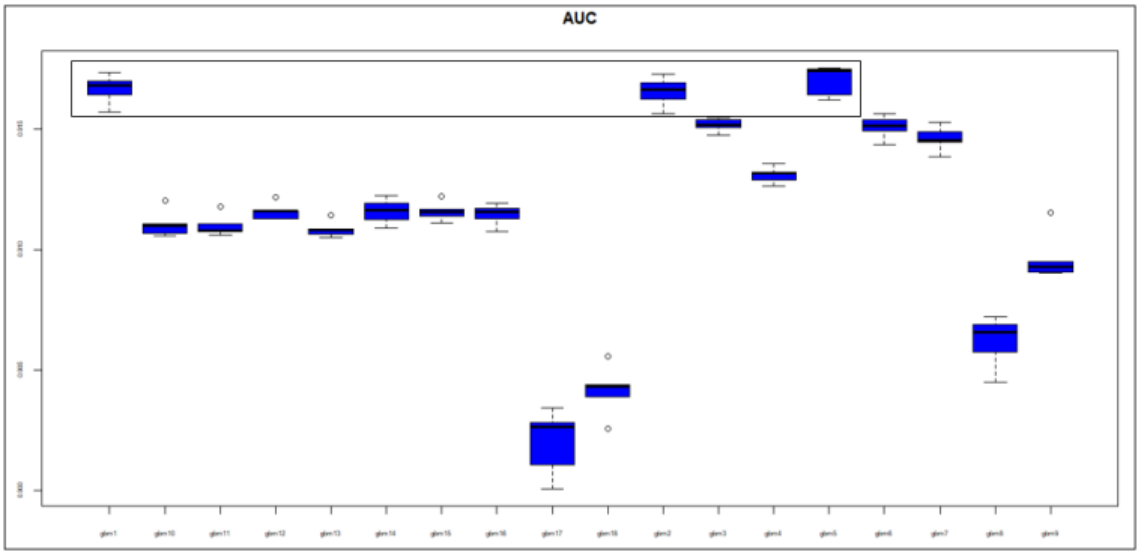


Figura 6.2.5. 3. Comparación de Modelos Gradient Boosting AUC.

En la figura anterior de la tasa de fallos, se encuentra identificado el mejor modelo gbm2, con una tasa media de fallos de 0.1597 y AUC 0.9167; que ha resultado del conjunto de datos "D", con constante de regularización 0.1, 500 iteraciones y un máximo de 20 nodos finales. En el anexo VI, en la sección de Gradient boosting, se encuentra el código R, así como un gráfico de Early Stoping y la importancia de variables.

6.2.6 Ensamblado y Comparación de Modelos

Para el estudio de métodos de ensamblado, se ha utilizado la técnica Stacking, es decir una combinación de modelos, que permite obtener el promedio de las probabilidades. Los métodos de ensamblados se realizan, a partir de los mejores modelos que se obtuvieron en cada técnica.

En lo que respecta a los mejores modelos de Redes Neuronales, Random Forest y Gradient Boosting, se han obtenido del conjunto de datos "D" imputados y con missing; para los mejores modelos de Árboles de clasificación y Regresión logística se han utilizado los conjuntos de Datos "A" y "B" con un mayor número de set de variables; se han replicado en el lenguaje R algunas características de estas dos últimas técnicas con el conjunto de Datos "D", utilizando datos con missing en el caso de árboles e imputados para regresión; no se han tenido buenos resultados en árboles de clasificación, por ello y dado que en el análisis se utilizan otras técnicas más sofisticadas basadas en árboles, no han sido incluidos en las técnicas de ensamblado. A todas las técnicas descritas anteriormente y que participarán en los métodos de ensamblado, se han vuelto ejecutar utilizando distintas semillas, validación cruzada de cuatro grupos y diez repeticiones.

Como se observa en la tabla 6.2.6.1, se han construido varios ensamblados combinando las predicciones de 2, 3 y 4 modelos, correspondientes a las técnicas con sus identificadores Regresión Logística (Log), Redes Neuronales (Red), Random Forest (RanForest) y Gradient Boosting (GradBoost), se han agregado también en la tabla, los resultados de las mejores modelos encontradas con cada técnica para su comparación con la medida de evaluación de la Tasa de Fallos, en cuanto a su valor AUC no se ha considerado la mayoría de los modelos han tenido gran capacidad predictiva.

Modelo o Método identificador	Características (mejores modelos)	Medida de Evaluación Tasa de Fallos
Logi	Mejor modelo Regresión Logística (Log)	**0.1587**
red	Mejor modelo Red Neuronal (Red)	**0.1592**
rf	Mejor modelo Random Forest (RanForest)	**0.1677**
gbm	Mejor modelo Gradient Boosting (GradBoost)	**0.1591**
ensamb1	**Formado:** Log + Red	**0.1578**
ensamb2	**Formado:** Log + RanForest	**0.1576**
ensamb3	**Formado:** Log + GradBoost	**0.1583**
ensamb4	**Formado:** Red + RanForest	**0.1583**
ensamb5	**Formado:** Red + GradBoost	**0.1584**
ensamb6	**Formado:** RanForest + GradBoost	**0.1589**
ensamb7	**Formado:** Log + Red + RanForest	**0.1577**
ensamb8	**Formado:** Log + Red + GradBoost	**0.1577**
ensamb9	**Formado:** Log + RanForest + GradBoost	**0.1577**
ensamb10	**Formado:** Red + RanForest + GradBoost	**0.1578**
ensamb11	**Formado:** Log+ Red + RanForest + GradBoost	**0.1574**

Tabla 6.2.6.1. Mejores Modelos incluidos Ensamblado.

Se puede observar que los métodos de ensamblado han mejorado un poco la tasa de fallos con respecto a las técnicas de modelización generadas individualmente, los mejores ensamblados se encuentran resaltados en verde. A continuación en la figura 6.2.6.1, se muestra en una diagrama de cajas, la representación de la tasa de fallos, ordenada de menor a mayor para una mejor interpretación y poder seleccionar el mejor modelo de predicción que se ha encontrado durante el estudio.

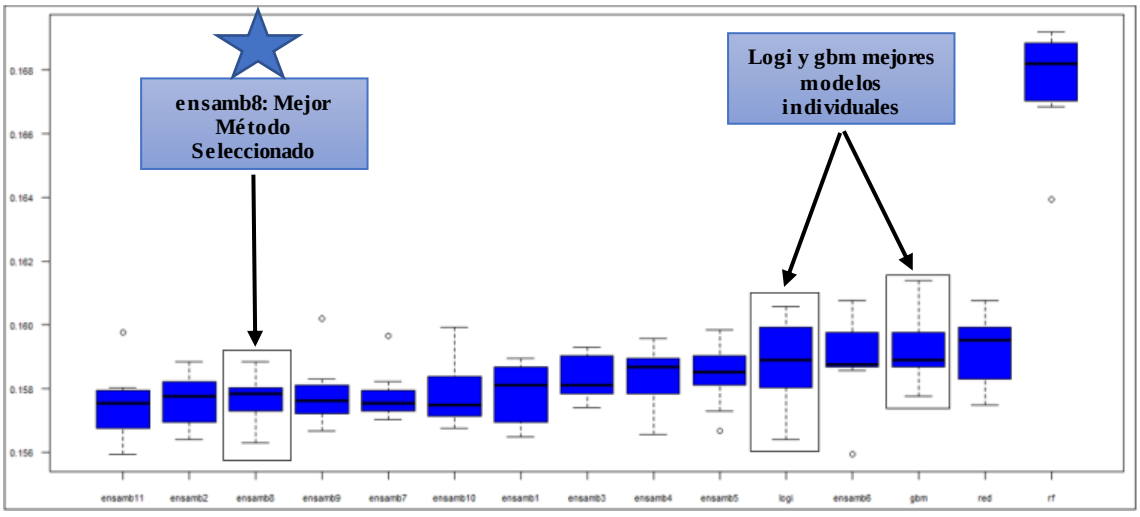


Figura 6.2.6.1. Comparación de Modelos: Mejor Modelo Predictivo Seleccionado.

En general no han discrepado mucho los valores en la tasa de fallos, varios modelo ofrecen resultados semejantes. Se ha considerado el mejor modelo de predicción la técnica de ensamblado con identificador **ensamb8**, el cual ha resultado de la combinación de los mejores modelos de Logística (Log) + Red Neuronal (Red) y Gradient Boosting (GradBoost), ya que no arroja valores atípicos, tiene poca variabilidad y un valor (considerado bajo) en la tasa de fallos **0.1577**. En lo que respecta a las técnicas individuales, los modelos de Regresión Logística y **Gradient Boosting** tienen los mejores resultados, le sigue el modelo de Red Neuronal, quedando en último lugar el modelo de Random Forest.

En el análisis de resultados, se analizan algunas medidas de clasificación del mejor método de ensamblado identificado y del modelo de Gradient boosting, el cual se ha considerado el mejor modelo con respecto a las técnicas individuales, por su menor variabilidad con respecto a Regresión Logística (el cual será analizado con otro set de variables, el conjunto de datos "B", evidenciando buenos resultados). El mejor modelo de gradient boosting a resultado del conjunto de datos "D", con constante de regularización 0.1, 500 iteraciones y un máximo de 20 nodos finales; presentando una valor de la tasa de fallos de **0.1591**.

7. ANÁLISIS DE RESULTADOS

7.1 Árboles de Clasificación variable objetivo: Consumo de Marihuana

En cuanto al análisis predictivo en la Figura 7.1.1, se tiene un total del 74.69% de observaciones bien clasificadas (Tasa de verdaderos positivos) cuando el nivel de la variable objetivo es uno, es decir si ha consumido marihuana, por ende un 25.31% de observaciones mal clasificadas para este nivel.

Se tiene un total del 86.23% (Tasa de verdaderos negativos) de observaciones bien clasificadas cuando el nivel de la variable objetivo es cero, es decir no ha consumido marihuana, por ende un 13.77% de observaciones mal clasificadas para este nivel. Lo que equivaldría a un total del 82.06% de observaciones de prueba bien clasificados utilizando el modelo de árboles de clasificación que es la tasa de aciertos.

En cuanto a los valores predictivos positivos (probabilidad de haber consumido marihuana si el resultado de la prueba es positivo.) y negativo (probabilidad de no haber consumido marihuana si el resultado de la prueba es negativo) se tiene un valor del 75.46% y 85.73% respectivamente.

Si se compara el modelo de árboles de clasificación con el no modelo, se observa que este ha mejorado significativamente, ha disminuido el error de un 36% a un 17%, además, si se lo compara con la media que tiene una probabilidad de acierto del 62.82% (no modelo), ahora con el modelo ganador se tiene una probabilidad de acierto del 82.06% (mejor modelo árbol).

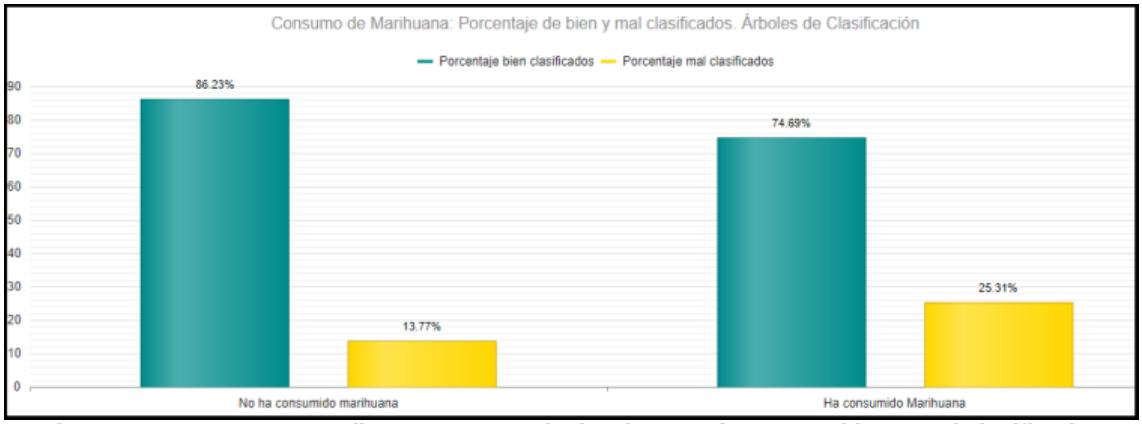


Figura 7.1.1. Consumo Marihuana: Porcentaje de Observaciones Test bien y mal clasificadas Árboles de Clasif.

En cuanto a la relación e interpretación, en primer lugar se analizará las variables más importantes e influentes en el estudio, en este sentido hay que recordar que mientras más arriba estén las variables de entrada en el árbol, más importantes resultan en la clasificación de salida. En resumen las dos primeras variables en el proceso de construcción han variado su orden de importancia, estas son el consumo de cigarrillos y la permanencia de estar cerca de un grupo que consume marihuana, entre otras variables que se encuentran en la tabla 7.1.1.

P7	**¿Has fumado cigarrillos alguna vez en la vida? ① Sí ② No**
P95_a	**Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo marihuana con el evidente**

	propósito de volarse, drogarse? **①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido**
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ①: Nunca he tomado alcohol ②: Ninguno ③: Uno o Menos de uno ④: Entre 2 y 5 ⑤: Entre 6 o más.
P11	¿Cuántos días has fumado cigarrillos en los últimos 30 días? N° de días: Marca "0" en la hoja de respuestas si no has fumado en los últimos 30 días
P23_a	¿Cuántas veces en tú vida te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪: Nunca ①:1-2 veces ②3-5 veces ③:6 veces o más.
P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana? Digamos, todos los fines de semana o más seguido ①:Ninguno ②:Menos de la mitad ③:Como la mitad ④:Más de la mitad ⑤:Todos o casi todos
P84	Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste al colegio una parte importante de la jornada o en toda la jornada ①:Nunca o Casi nunca ②:Pocas veces ③:Varias o Muchas veces
P96	Si en tu grupo de amigas yamigos cercanos supieran que fumas marihuana ¿tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven? (no consideres alcohol, cigarrillos o tranquilizantes) ①:Sí ②:No
P82_d	Si tu mamá descubriera que fumas marihuana: ①:Extremadamente o Bastante molesta ②:Algo o poco molesta ③:Indiferente ⑥:No aplica
P82_c	Si tu papá descubriera que fumas marihuana: ①:Extremadamente o Bastante molesto ②:Algo o poco molesto ③:Indiferente ⑥:No aplica
P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ①:Menos de 4,5 ②:Entre 4,5 y 4,9 ③:Entre 5,0 y 5,4 ④:Entre 5,5 y 5,9 ⑤:Entre 6,0 y 6,4 ⑥:Entre 6,5 y 7,0
P97	¿Si en tu grupo de amigas yamigos más cercanos supieran que has probado una droga distinta a la marihuana como cocaína, pasta base, éxtasis, ácidos o cosas parecidas, tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P86	¿Cuántos cursos has repetido en tu vida escolar? ①:Ninguno ②:Uno ③:Dos o más
P9	¿Cuándo fue la primera vez que fumaste cigarrillos? ①:Durante los últimos 30 días ②: Hace más de un mes, pero menos de un año ③:Hace más de un año ④:Nunca he probado
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P26	Qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol (99) No ha consumido ① Poco o casi nada ② Medio tomado ③ Bien tomado (88) No responde
P8_b	¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días? Marca"0" en la hoja de respuestas si no has fumado todos o casi todos los días . **Consideren los siguientes rangos** ①:(5 a 12 años) , ②:(13 a 17 años), ③:(18 a 21 años)
P25	Pensando en el último día que consumiste alcohol ¿cuál de las siguientes bebidas alcohólicas fue la que más tomaste ese día? Marca aquella bebida (o tipo de alcohol) que más consumiste ①:Cerveza ②:Vino ③:Espumantes (champaña, Manquehuito, vinos con sabores u otros) ④:Tragos fuertes solos o combinados (piscola, roncola, vodka naranja u otro) ⑤:No consumo alcohol
P12	Considerando sólo los días que fumaste en el último mes. ¿Aproximadamente, cuántos cigarrillos fumaste al día? N° de cigarrillos: Marca"0" en la hoja de respuestas si no has fumado en los últimos 30 días
P95_d	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo Inhalables con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪: Nunca ①:una o más de una vez

P95_e	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alcohol con el evidente embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P15	¿Has probado alcohol alguna vez en la vida (cerveza, vinos o tragos fuertes)?

Tabla 7.1.1. Consumo de marihuana: Descripción de variables importantes árbol de clasificación.

Tomando en cuenta la gran cantidad de variables y conociendo que los árboles de clasificación permiten una representación gráfica de una serie de reglas sobre las decisiones tomadas (aspectos de la encuesta) para asignar un valor de salida a una determinada entrada, se describirán únicamente los aspectos más importantes.

En resumen, visualizando el árbol e identificando el nivel de importancia de las variables independientes (estas contienen aspectos relacionados a la población escolar como son ámbitos personales, entorno familiar y escolar, percepción del estudiante en el consumo de alcohol y tabaco, y alguna información social de su entorno) las variables más útiles y que mejor discriminan la variable objetivo (consumo de marihuana) tiene que ver con:

- Las amistades y la relación que mantiene el estudiante con estas, siendo la más importante P95_a: **Si le ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo marihuana con el evidente propósito de volarse o drogarse**, seguido de P99: **La cantidad de amigos que fuman marihuana**, entre otros aspectos relacionados a sus amistades y el consumo de sustancias (pasta base e Inhalables).

- El consumo de alcohol y cigarrillo, siendo la variable más importante en todo el estudio P7: **haber consumido cigarrillo,** seguido de P27: la cantidad de alcohol que toma el estudiante en una salida por la noche, entre varias aspectos sobre el consumo de estas sustancias que puede observar en las variables importantes.

- La relación de cómo se siente o comporta en el colegio y con quien vive, también mantiene una influencia menor a los otros aspectos, entre las más importantes se encuentran P84: cantidad de veces que se ha escapado del colegio, P79: Si considera que sus padres o apoderado han consumido drogas cuando eran jóvenes.

El árbol de clasificación obtenido permite predecir con un error del 0.17 si un estudiante ha consumido marihuana o no a partir de una serie de aspectos considerados en el estudio, así se observa en la figura 7.1.4 del resumen del árbol que:

Considerando el aspecto más importante que es el consumo de cigarrillos, un estudiante que no haya consumido cigarrillo con una probabilidad del 91.66% será clasificado como que no ha consumido marihuana.

Finalmente se describirá el “mejor” de los casos del camino y reglas del árbol, que deba tener un estudiante sobre los aspectos de la encuesta, para ser clasificado con más alta probabilidad de que haya consumido marihuana o no haya consumido.

Para considerar que haya consumido marihuana con el 89.51% de probabilidad debe haber consumido cigarrillos (57.8%) + haber estado de vez en cuando o muy seguido

cerca de alguien o con un grupo que consume marihuana (durante el último año) (72.56%) + Tomar alcohol entre 2 o más vasos en una salida por la noche (81.94%) + Tener la mitad o más de amigos(as) que fumen regularmente marihuana (89.51%). Esta representación se puede visualizar en el gráfico del árbol y en la figura 7.1.2.

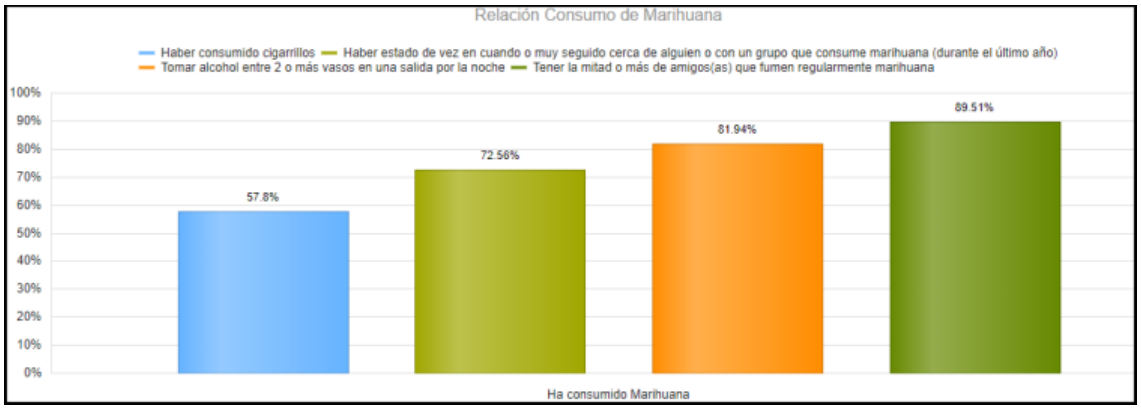


Figura 7.1.2. Características para el Consumo de Marihuana.

Para considerar que no haya consumido marihuana con el 98.18% de probabilidad debe no haber consumido cigarrillos (91.66) + no haber estado nunca cerca de alguien o con un grupo de que consume marihuana (durante el último año) (96.61%) + no haberse emborrachado o intoxicado nunca tomando alcohol (97.46%) + Tener un promedio de notas del año pasado entre 4.5 o más (97.83%) + Tener una percepción de que el padre estaría bastante molesto, si descubriera que fuma marihuana o haber contestado no aplica (98.18%). Se observa esta representación en la figura 7.1.3.

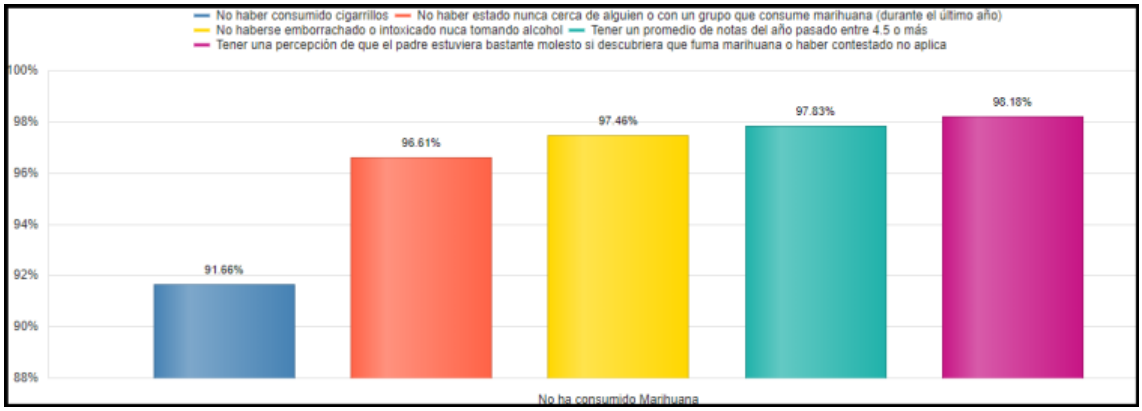


Figura 7.1.3. Características para el no Consumo de Marihuana.

Así se podrá analizar en el resumen del diagrama del árbol representada en la figura 7.1.4 una gran cantidad de relaciones y escenarios. Para observar todo el diagrama del árbol consultar en el Anexo VII.

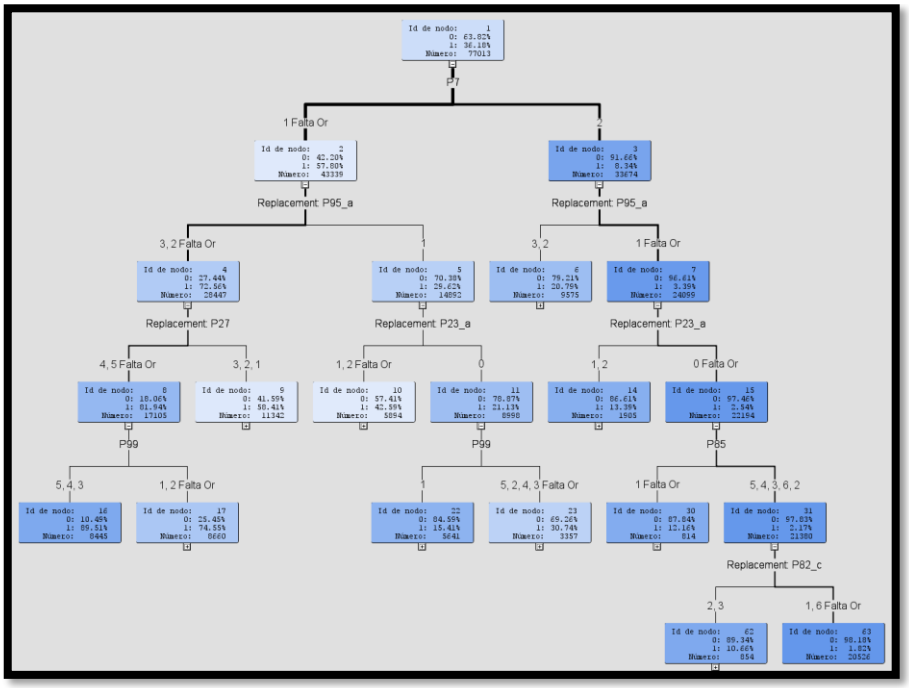


Figura 7.1.4. Consumo de Marihuana: Resumen árbol de clasificación.

7.2 Árboles de Clasificación variable objetivo: Consumo de Cocaína o Pasta Base

En cuanto al análisis predictivo en la Figura 7.2.1, se tiene un total del 19.60% de observaciones bien clasificadas (Tasa de verdaderos positivos) cuando el nivel de la variable objetivo es uno, es decir si ha consumido cocaína o pasta base, por ende un 80.40 % de observaciones mal clasificadas para este nivel.

Se tiene un total del 98.77% (Tasa de verdaderos negativos) de observaciones bien clasificadas cuando el nivel de la variable objetivo es cero, es decir no ha consumido cocaína o pasta base, por ende un 1.23% de observaciones mal clasificadas para este nivel. Lo que equivaldría a un total del 92.76% de observaciones de prueba bien clasificados (tasa de aciertos) utilizando el modelo de árboles de clasificación.

En cuanto a los valores predictivos positivos (probabilidad de haber consumido cocaína o pasta base si el resultado de la prueba es positivo.) y negativo (probabilidad de no haber consumido cocaína o pasta base si el resultado de la prueba es negativo) se tiene un valor del 56.91% y 93.71% respectivamente.

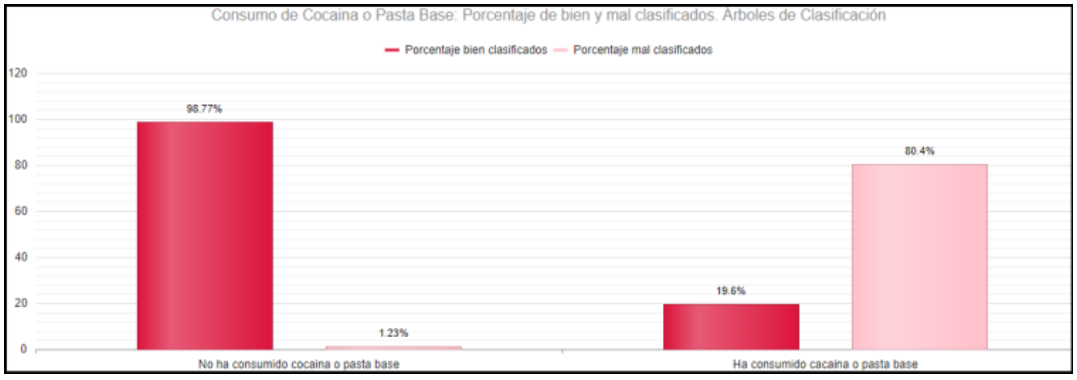


Figura 7.2.1. Consumo de Cocaína o Pasta Base: Observaciones Test bien y mal clasificadas Árboles de Clasif.

En cuanto a la relación e interpretación, se muestran en la tabla 7.2.1 las variables más importantes e influentes en el estudio, se observa aquí también **que el estar cerca de alguien o alrededor de un grupo que consuma las drogas analizadas influye en el estudiante para que haya probado las sustancias de cocaína o pasta base**, estas corresponden a (P95_b y P95_c); otras variables influyentes, **es el haber consumido marihuana alguna vez** (P35), **la cantidad de alcohol que toma en una** noche (P27), así como el **número de veces que se ha emborrachado o intoxicado últimamente** (P23_c).

P95_b	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo cocaína con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P35	¿Has consumido marihuana alguna vez en la vida? ① Sí ⓪ No
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ①: Nunca he tomado alcohol ②: Ninguno ③: Uno o Menos de uno ④: Entre 2 y 5 ⑤: Entre 6 o más.
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪: Nunca ①:una o más de una vez
P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ①:Menos de 4,5 ②:Entre 4,5 y 4,9 ③:Entre 5,0 y 5,4 ④:Entre 5,5 y 5,9 ⑤:Entre 6,0 y 6,4 ⑥:Entre 6,5 y 7,0
P21	Piensa en los últimos 30 días ¿Cuántos días has consumido algún tipo de alcohol? N° de días: Marca"0" en la hoja de respuestas si no has consumido
P20_d	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Tener relaciones sexuales sin condón ①:Sí ②:No
P97	¿Si en tu grupo de amigas yamigos más cercanos supieran que has probado una droga distinta a la marihuana como cocaína, pasta base, éxtasis, ácidos o cosas parecidas, tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P74_b	¿Qué educación alcanzo tu madre? ①:Básica incompleta ②:Básica completa ③:Media incompleta ④:Media completa ⑤:Técnica superior incompleta o Universitaria incompleta ⑥:Técnica superior completa ⑦:Universitaria completa ⑨No sé o N/A ⑨
P11	¿Cuántos días has fumado cigarrillos en los últimos 30 días? N° de días: Marca "0" en la hoja de respuestas si no has fumado en los últimos 30 días
P82_d	Si tu mamá descubriera que fumas marihuana como crees que estuviera:①:Extremadamente o Bastante molesta ②:Algo o poco molesta ③:Indiferente ⑥:No aplica
P96	Si en tu grupo de amigas yamigos cercanos supieran que fumas marihuana ¿tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P108	¿Trabajas regularmente además de estudiar? ①:Sí ②:No
P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven? (no consideres alcohol, cigarrillos o tranquilizantes) ①:Sí ②:No
P8_a	¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez? Se consideran los siguientes rangos (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
P16	¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? Se consideran los siguientes rangos (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
P22	¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol? ① 1 a 2 tragos ② 3 a 4 tragos ③ 5 o mas
P20_b	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Consumir alcohol estando solo o sola ① Sí ② No
P18	¿Cuándo fue la última vez que tomaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado

P20_a	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Amigos, amigas o familiares te han sugerido o mencionado que disminuyas el consumo de alcohol ① Sí ② No
P3	Después de que sales del colegio o durante los fines de semana, ¿cuántas veces ocurre que tu madre, padre, apoderada o apoderado no saben dónde estás? Ya sea por un período de una hora o más. ① Nunca o casi nunca saben dónde estoy ② A veces no saben ③ Siempre o casi siempre saben dónde estoy
P94	¿Cuán probable es que sigas estudiando después del colegio? (en la Universidad, Instituto Profesional, Centro de Formación técnica u otro) ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable o Imposible

Tabla 7.2.1. Consumo de Cocaína o Pasta Base: Descripción de Variables Importantes

Dado el gran número de variables, se describen únicamente los aspectos más importantes. En resumen, visualizando el árbol e identificando el nivel de importancia de las variables independientes, las variables más útiles y que mejor discriminan la variable objetivo (consumo de cocaína y pasta base) tiene que ver con:

- El consumo de alcohol y cigarrillo, siendo el aspecto del consumo de alcohol el que mayor número de variables importantes posee, se describió anteriormente las más importantes, correspondientes a P27 y P23_c, con respecto al consumo de cigarrillos, el número de días que ha consumido últimamente es el más importante (P11).

- Los aspectos de la relación de amistades y de las personas con las que vives también influyen en los resultados encontrados, siendo más importante el aspecto de relación que tiene el estudiante con sus amistades; en él, se identifica las variables más importantes del estudio descritas anteriormente P95_b y P95_c; existen otras variables dentro de este aspecto, como la percepción que tendría el estudiante si los amigos supieran que ha experimentado con el consumo de cocaína o pasta base u otras drogas (P97-P96); con respecto al factor de convivencia, el grado de estudios que alcanzo la madre (P74_b), así como la percepción del grado de molestia que la madre tendría si descubriera que ha fumado marihuana (P82_d) son las variables más importantes.

- La relación de cómo se siente o comporta en el colegio también mantiene una influencia menor comparado con los otros aspectos que se han comentado, siendo en este factor la variable más importante el promedio de notas con la que termino el estudiante el año pasado (P85).

- Considerando el aspecto personal, el consumo de marihuana mantiene una relación en el consumo de cocaína y pasta base; es una de las variables más importantes en este estudio.

El árbol de clasificación obtenido permite clasificar de mejor manera a los estudiantes que no han consumido cocaína o pasta base, siendo la proporción de aciertos para este nivel del 93.71%. En la figura 7.2.1, se observa un resumen de la representación gráfica de una serie de reglas sobre las decisiones tomadas con respecto a las variables importantes para clasificar a un estudiante con una probabilidad de acierto como que ha consumido o no la sustancia. Se describe a continuación las características que debe tener un estudiante para ser considerado con mayor probabilidad que haya o no consumido las drogas de cocaína o pasta base.

Para considerar que no haya consumido cocaína o pasta base con el 98.74% de probabilidad, debe no haber estado durante el último año, cerca de alguien o alrededor de un grupo que hayan consumido estas sustancias (96.84%) + no haber consumido marihuana (98.74%). En la figura 7.2.2 se puede observar su representación.

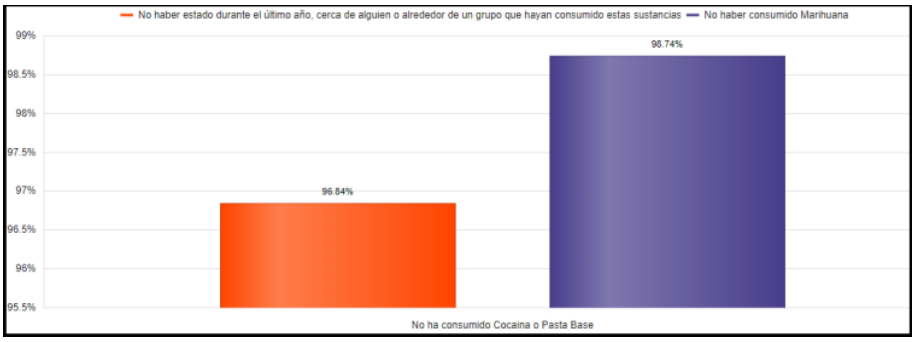


Figura 7.2.2. Características para el no Consumo de Cocaína o Pasta Base.

Para considerar que haya consumido cocaína o pasta base con 75.28% de probabilidad debe haber estado de vez en cuando o muy seguido con un grupo que haya estado consumiendo cocaína (26.74%) + haber consumido marihuana (36.13%) + no haber tomado alcohol nunca o llegar a tomar en una salida por la noche entre 6 o más vasos de licor (48.82%) + consumir alcohol más de 13 días en un mes (66.35%) + haber probado una bebida alcohólica en las edades de 5 a 12 años (75.28%). En la figura 7.2.3 se puede observar su representación.

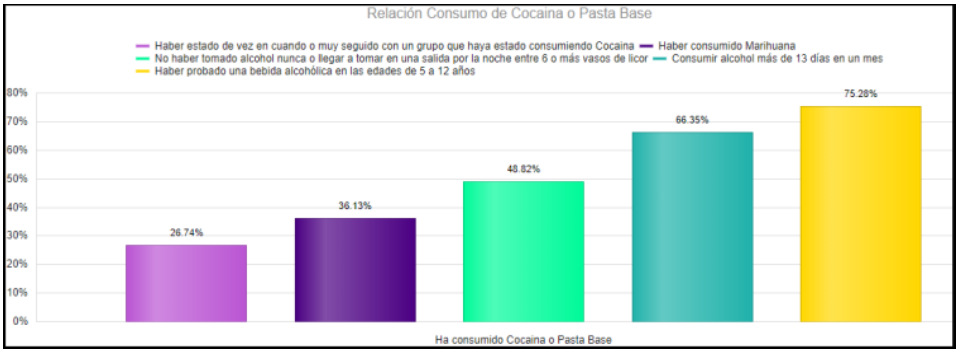


Figura 7.2.3. Características para el Consumo de Cocaína o Pasta Base.

Así se podrá analizar en el resumen del diagrama del árbol de la figura 7.2.4 algunas relaciones y escenarios. El diagrama completo se puede consultar en el Anexo VII.

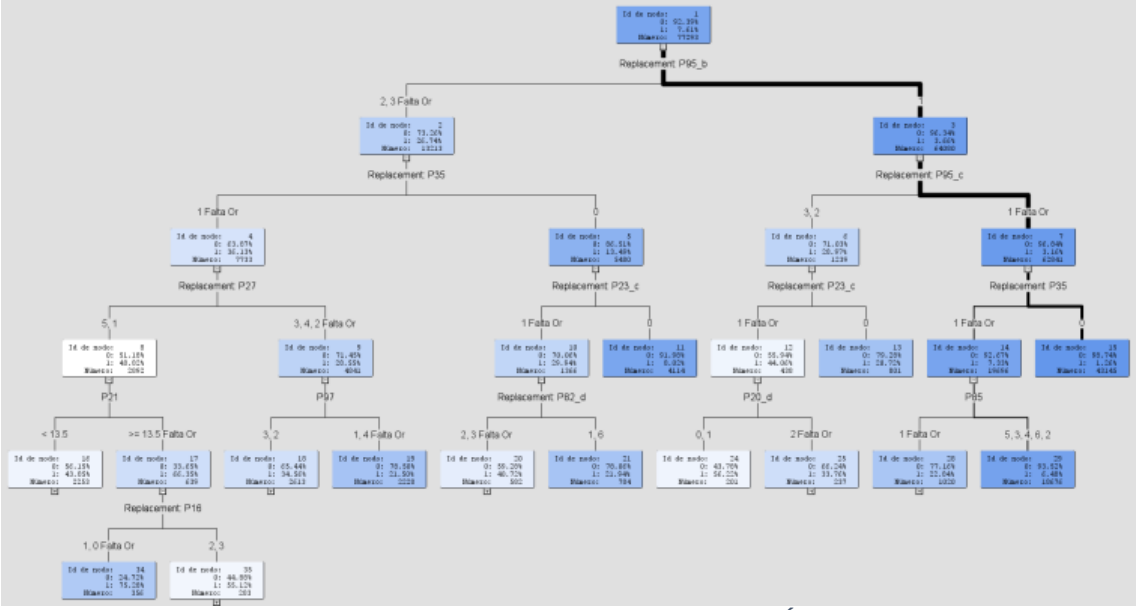


Figura 7.2.4. Consumo de Cocaína o Pasta Base: Resumen Árbol de Clasificación.

7.3 Árboles de Clasificación variable objetivo: Consumo Otras Drogas

En cuanto al análisis predictivo en la Figura 7.3.1 se tiene un total del 39.06% de observaciones bien clasificadas (Tasa de verdaderos positivos) cuando el nivel de la variable objetivo es uno, es decir si ha consumido otras drogas, por ende un 60.94 % de observaciones mal clasificadas para este nivel.

Se tiene un total del 98.00 % (Tasa de verdaderos negativos) de observaciones bien clasificadas cuando el nivel de la variable objetivo es cero, es decir no ha consumido otras drogas, por ende un 12.00% de observaciones mal clasificadas para este nivel.

Lo que equivaldría a un total del 93.65% de observaciones de prueba bien clasificados utilizando el modelo de árboles de clasificación que es la tasa de aciertos.

En cuanto a los valores predictivos positivos (probabilidad de haber consumido otras drogas si el resultado de la prueba es positivo.) y negativo (probabilidad de no haber consumido otras drogas si el resultado de la prueba es negativo) se tiene un valor del 60.91% y 95.28% respectivamente.

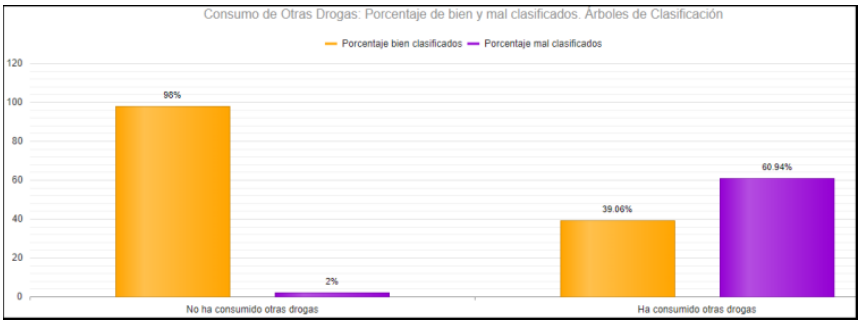


Figura 7.3.1. Consumo Otras Drogas: Porcentaje de observaciones Test bien y mal clasificadas Árbol de Clasif.

En cuanto a la relación e interpretación, se identifican las variables más importantes, que tienen que ver para el consumo de otras drogas, **el haber consumido cocaína o**

pasta base (P47_53) es la que más influye en el estudio, **seguido del estar cerca de alguien o un grupo que consume otras drogas como inhalables** (P95_d) y del **número de veces que el estudiante se haya embriagado o intoxicado tomado alcohol durante el último mes** (P23_c); así en la tabla 7.3.1 se pueden observar en orden de importancia la descripción de las otras variables influyentes del estudio.

P47_53	¿Has consumido cocaína o pasta base alguna vez en la vida? ①Sí ⓪No
P95_d	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo Inhalables con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:Casi nunca o de vez en cuando ③:Bastante seguido o muy seguido
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪: Nunca ①:una o más de una vez
P91_b	Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio. ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
P21	Piensa en los últimos 30 días ¿Cuántos días has consumido algún tipo de alcohol? N° de días: Marca"0" en la hoja de respuestas si no has consumido
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P90_e	Has robado algo a alguien en el colegio: ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ①:Menos de 4,5 ②:Entre 4,5 y 4,9 ③:Entre 5,0 y 5,4 ④:Entre 5,5 y 5,9 ⑤:Entre 6,0 y 6,4 ⑥:Entre 6,5 y 7,0
P75	¿Con qué personas vives actualmente? ①:Padre y madre ②:Padre y su pareja ③:Madre y su pareja ④:Sólo con el padre ⑤:Sólo con la madre ⑥:Sólo con Hermana(s) o hermano(s) ⑦: Sólo con Abuelo(s) o Abuela(s) ⑧:Otro adulto responsable
P95_e	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alcohol con el evidente embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P73	¿Quién es el jefe de tu hogar? Jefe de hogar se define como la persona, hombre o mujer, reconocida como tal por los integrantes del hogar: ① Padre ② Madre ③ Abuela o Abuelo ④ Otro
P20_b	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Consumir alcohol estando solo o sola ① Sí ② No
P16	¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? Se consideran los siguientes rangos (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
P95_a	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo marihuana con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P18	¿Cuándo fue la última vez que tomaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ①: Nunca he tomado alcohol ②: Ninguno ③: Uno o Menos de uno ④: Entre 2 y 5 ⑤: Entre 6 o más.
P82_d	Si tu mamá descubriera que fumas marihuana como crees que estuviera:①:Extremadamente o Bastante molesta ②:Algo o poco molesta ③:Indiferente ⑥:No aplica
P82_c	Si tu papá descubriera que fumas marihuana como crees que estuviera: ①:Extremadamente o Bastante molesta ②:Algo o poco molesta ③:Indiferente ⑥:No aplica
P91_c	Has estado en un grupo que ha sido atacado por otro grupo ① Nunca ② Una vez ③Dos veces ④ 3 o más veces

Tabla 7.3.1. Consumo de Otras Drogas: Descripción de variables importantes.

Se describen también los aspectos más influyentes, considerando la importancia y la cantidad de variables de los diferentes aspectos relacionados a la encuesta. Las variables más útiles y que mejor discriminan la variable objetivo (Consumo de Otras drogas) tiene que ver con:

- En cuanto al aspecto personal, el haber consumido cocaína o pasta base es la variable más importante para discriminar en el árbol de decisión si el estudiante

ha consumido o no otras drogas (crack, éxtasis, heroína, alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos).

- El aspecto del consumo de alcohol es el que mayor número de la variables importantes relacionadas posee; la más importante se describió anteriormente P23_c, que también era la variable más influyente dentro de este aspecto en el consumo de cocaína y pasta base, le sigue el número de días que ha consumido alcohol durante el último mes (P21).

- Los aspectos de la relación de amistades, de cómo se siente en el colegio y de las personas con las que vive también influyen en los resultados encontrados; con respecto a la relación de amistad, el haber estado cerca de alguien o de un grupo que ha estado consumiendo inhalables o pasta base son las más importantes (P95_c – P95_d); las variables referentes al aspecto de cómo se siente en el colegio en el análisis de este estudio tiene una mayor participación, las más importantes son si el estudiante ha sido agredido por alguien el colegio (P91_b), si ha robado en la institución (P90_e) y el promedio de notas del años pasado (P85); y con respecto al último aspecto mencionado es decir de las personas con quien vives, las variables de con quien vives actualmente (P75) y quien es el jefe de hogar (P73) son las que influyen con una menor grado de importancia con respecto a los otros aspectos.

El modelo obtenido de árbol de clasificación en este análisis como en el anterior, permite clasificar de mejor manera a los estudiantes que no han consumido otras drogas, siendo la proporción de aciertos para este nivel del 95.28%. En la figura 7.3.2, se observa un resumen de la representación gráfica de una serie de reglas sobre las decisiones tomadas con respecto a las variables importantes para clasificar a un estudiante con una probabilidad de acierto como que ha consumido o no la sustancia. Se describe a continuación las características que debe tener un estudiante para ser considerado con mayor probabilidad que haya o no consumido otras drogas.

Para considerar que no haya consumido otras drogas con el 97.71% de probabilidad, debe no haber consumido cocaína o pasta base (95.86%) + no haber estado durante el último año cerca de alguien o algún grupo que ha estado consumiendo inhalables (97.71%). Se visualiza esta representación en la figura 7.3.2.

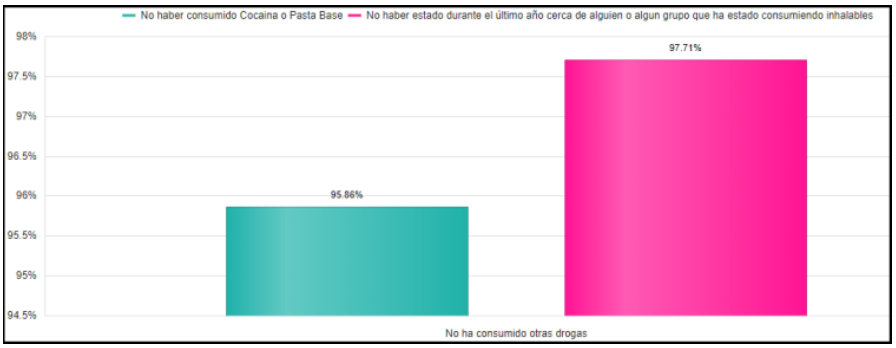


Figura 7.3.2. Características para el no Consumo de Otras Drogas.

Para considerar que haya consumido otras drogas con 70.67% de probabilidad debe haber consumido cocaína o pasta base (47.40%) + haber estado de vez en cuando o

muy seguido con un grupo que haya estado consumiendo otra droga como inhalables (63.46%) + haber robado a alguien en el colegio (70.67%). Se observa esta representación en la figura 7.3.3.

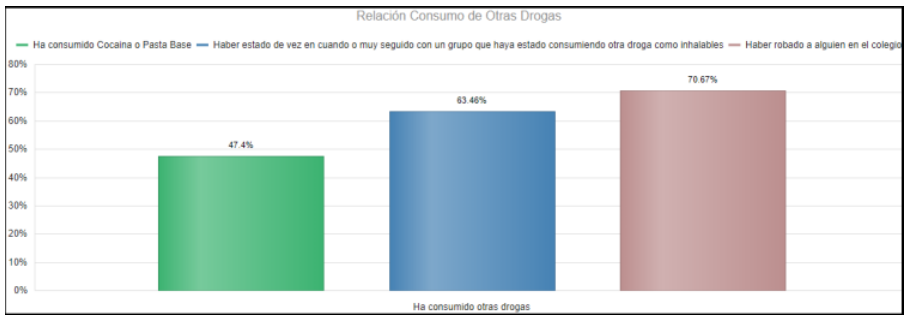


Figura 7.3.3.Características para el Consumo de Otras Drogas.

Otro caso para considerar que haya consumido otras drogas con el 70.49%, partiendo de las características anteriores, pero con la diferencia de tener la decisión de que el estudiante no ha robado en el colegio (54.06%), debe haber sido agredido físicamente estando solo o sola en el colegio (70.49%).

Analizando un último caso, para considerar que el estudiante ha consumido otras drogas con el 73.81%, partiendo así mismo de las condiciones anteriores, pero considerando que no ha robado y que no ha sido agredido físicamente en el colegio (49.30%), debe haber consumido alcohol más de once días durante el último mes (66.38%) + haber estado cerca de alguien o de un grupo que este consumiendo pasta base (73.81%). Estos resultados, así como los demás mostrados de las drogas analizadas, se pueden encontrar en formato HTML en el Anexo VII.

Así se podrán analizar otros escenarios en la figura 7.3.4 del resumen del árbol. El diagrama completo se puede consultar en el Anexo VII.

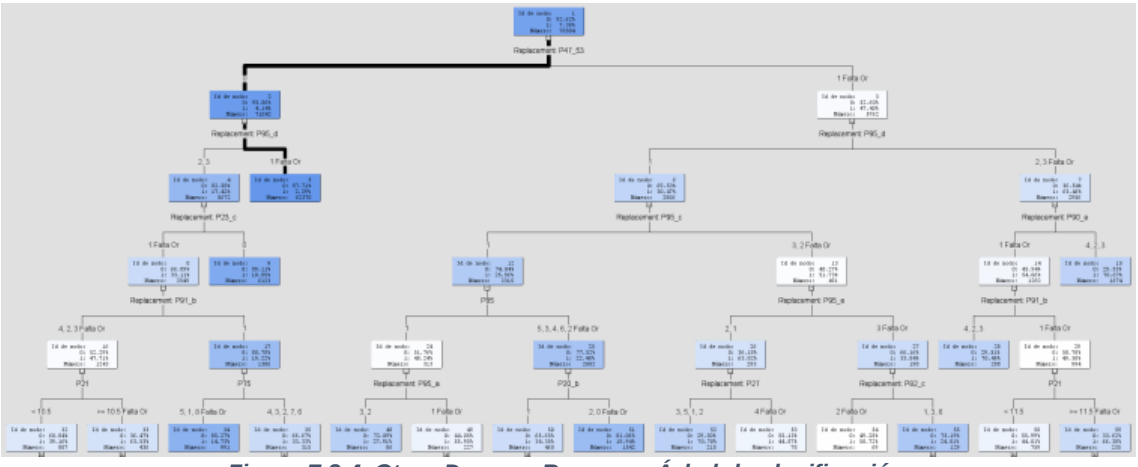


Figura 7.3.4. Otras Drogas: Resumen Árbol de clasificación.

7.4 Regresión Logística variable objetivo: Consumo de Marihuana

En cuanto al análisis predictivo considerando el índice de youden el modelo ganador de regresión logística muestra mejores resultados estadísticos, que el mejor modelo de árbol de clasificación, en la Figura 7.4.1, se tiene un total del 85.80% de observaciones

bien clasificadas (Tasa de verdaderos positivos) cuando el nivel de la variable objetivo es uno, es decir si ha consumido marihuana, por ende un 14.20% de observaciones mal clasificadas para este nivel.

Se tiene un total del 82.25% (Tasa de verdaderos negativos) de observaciones bien clasificadas cuando el nivel de la variable objetivo es cero, es decir no ha consumido marihuana, por ende un 17.75 % de observaciones mal clasificadas para este nivel. Lo que equivaldría a un total del 84.19% de observaciones de prueba bien clasificadas (tasa de aciertos) utilizando el modelo de regresión logística.

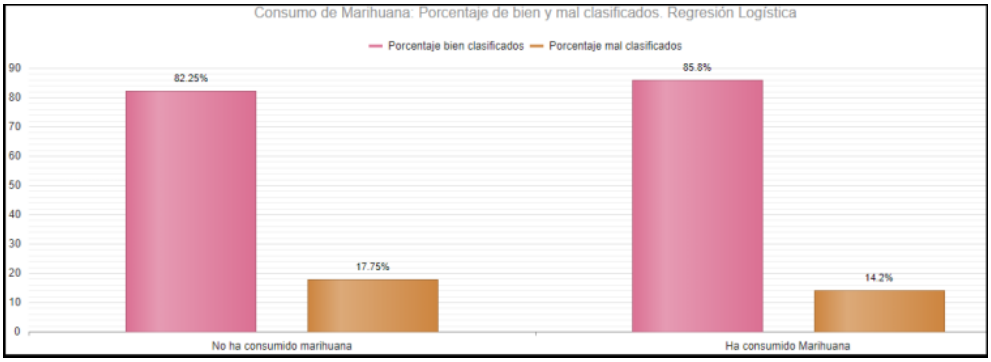


Figura 7.4.1. Consumo Marihuana: Porcentaje de Observaciones Test bien y mal clasificadas Regresión Logística.

En cuanto a los valores predictivos positivos (probabilidad de haber consumido marihuana si el resultado de la prueba es positivo.) y negativo (probabilidad de no haber consumido marihuana si el resultado de la prueba es negativo) se tiene un valor del 73.26% y 91.08% respectivamente, el valor predictivo ha bajado un poco con respecto al valor predictivo positivo del modelo de árboles, pero el valor predictivo negativo ha aumentado considerablemente comparado con él mismo.

Si se compara el modelo de regresión Logística con el no modelo, se observa que este ha mejorado significativamente, ha disminuido el error de un 36% a un 16%; además, si se lo compara con la media que tiene una probabilidad de acierto del 62.82% (no modelo), ahora con el modelo ganador se tiene una probabilidad de acierto del 84.19% (mejor modelo regresión).

Considerando remuestreo, entre las técnicas de árboles y regresión, se observa en el figura 7.4.2 que el modelo de regresión presenta un menor error de tasa de clasificación errónea.

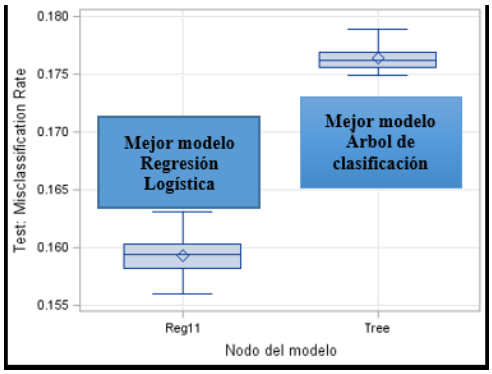


Figura 7.4.2. Comparación mejor Modelos Árbol de Clasif. Vs Regresión Logística.

En cuanto a la relación e interpretación, se analizará del mismo modo que en el modelo de árboles de clasificación las variables más importantes e influentes en el estudio, en este sentido se observa el valor de la prueba de Chi cuadrado de Wald, que indicara si las variables en el modelo son significativas, en este caso mientras más alto su valor, más importante se consideran.

En el análisis de regresión la variable P95_a: permanencia de estar en un grupo que consume marihuana es la variable más influyente, fue considerada una de las más importantes en el modelo árbol de clasificación, la segunda variable más importante o influyente encontrada también en las dos técnicas es P99: la cantidad de amigos que fuman marihuana que tiene el estudiante. La mayoría de las variables consideras importantes en el análisis de árboles de clasificación, también son consideradas en el modelo de regresión, es así que de las 24 primeras variables, 17 coinciden en los dos análisis (resaltadas en verde). En la tabla 7.4.1, se pueden observar las variables más importantes.

P95_a	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo marihuana con el evidente propósito de volarse, drogarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana? Digamos, todos los fines de semana o más seguido ①:Ninguno ②:Menos de la mitad ③:Como la mitad ④:Más de la mitad ⑤:Todos o casi todos
P96	Si en tu grupo de amigas yamigos cercanos supieran que fumas marihuana ¿tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P84	Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste al colegio una parte importante de la jornada o en toda la jornada ①:Nunca o Casi nunca ②:Pocas veces ③:Varias o Muchas veces
P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven? (no consideres alcohol, cigarrillos o tranquilizantes) ①:Sí ②:No
P98	¿Cuántos de tus amigas y amigos toman regularmente alcohol? Digamos, todos los fines de semana o más seguido ①:Ninguno ②:Menos de la mitad ③:Como la mitad ④:Más de la mitad ⑤:Todos o casi todos
P23_a	¿Cuántas veces en tú vida te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪:Nunca ①:1-2 veces ②:más de 3 veces
P97	¿Si en tu grupo de amigas yamigos más cercanos supieran que has probado una droga distinta a la marihuana como cocaína, pasta base, éxtasis, ácidos o cosas parecidas, tú crees que: ①:Te harían algún reproche o te dirían algo para que no lo hicieras ②:Algunos te harían reproches y otro no ③:No te harían ningún problema ④:Te alentarían a que lo siguieras haciendo
P10	¿Cuándo fue la última vez que fumaste un cigarrillo? ①:Durante los últimos 30 días ②:Hace más de un mes, pero menos de un año ③:Hace más de un año ④:Nunca he probado
P95_e	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alcohol con el evidente embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ①:Menos de 4,5 ②:Entre 4,5 y 4,9 ③:Entre 5,0 y 5,4 ④:Entre 5,5 y 5,9 ⑤:Entre 6,0 y 6,4 ⑥:Entre 6,5 y 7,0
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ①: Nunca he tomado alcohol ②: Ninguno ③: Uno o Menos de uno ④: Entre 2 y 5 ⑤: Entre 6 o más.
P80	Hasta donde tú conoces ¿alguno de tus hermanos o hermanas consume alguna droga ilícita (ilegal)? ①:Estoy seguro de que no lo ha(n) hecho ②:Creo que no lo ha(n) hecho ③:Creo que lo hace(n) ④:Estoy seguro de que lo hace(n) ⑤:No tengo hermanos o hermanas

P19	¿Cuán difícil te sería comprar alguna bebida alcohólica, si quisieras hacerlo? ①:Me sería muy fácil ②:Me sería fácil ③:Me sería difícil ④:Me sería muy difícil ⑤:No podría comprarla ⑨:No sé
P82_d	Si tu mamá descubriera que fumas marihuana como crees que estuviera:①:Extremadamente o Bastante molesta ②:Algo o poco molesta ③:Indiferente ⑥:No aplica
P20_d	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Tener relaciones sexuales sin condón ①:Sí ②:No
P7	¿Has fumado cigarrillos alguna vez en la vida? ①:Sí ②:No
P9	¿Cuándo fue la primera vez que fumaste cigarrillos? ①:Durante los últimos 30 días ②: Hace más de un mes, pero menos de un año ③:Hace más de un año ④:Nunca he probado
P81_b	¿Cómo describirías el hábito que tiene tu madre respecto al alcohol (vino, cerveza, licor)? ①:Nunca toma alcohol, ②:Solo en ocasiones especiales, ③:Solo en fines de semana, pero nunca en días de semana, ④:Toma alcohol diariamente, uno o dos tragos, ⑤:Toma alcohol diariamente, más de dos tragos, ⑨:No aplica, no tiene padre o madre vivo, no lo ve nunca
P82_c	Si tu papá descubriera que fumas marihuana: ①:Extremadamente o Bastante molesto ②:Algo o poco molesto ③:Indiferente ⑥:No aplica
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse? ①:Nunca ②:De vez en cuando o casi nunca ③:Bastante o muy seguido
P11	¿Cuántos días has fumado cigarrillos en los últimos 30 días? N° de días: Marca "0" en la hoja de respuestas si no has fumado en los últimos 30 días
P108	¿Trabajas regularmente además de estudiar? ①:Sí ②:No
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? ⓪: Nunca ①:una o más de una vez

Tabla 7.4.1. Consumo Marihuana: Descripción de variables importantes Regresión Logística.

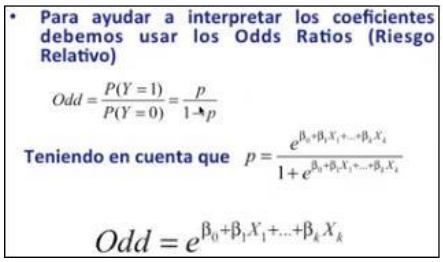
Analizando el número de variables, así como la importancia que mantienen en el modelo; entendiendo que las variables tienen que ver con ciertos aspectos de los estudiantes, se consideran los más importantes en este orden.

- Las amistades y la relación que mantienen los estudiantes con estas, como se observa en la tabla anterior existen varias variables que tiene relación con este aspecto, además son consideradas en el orden entre las más importantes; estas coinciden con las encontradas en el modelo de árbol de clasificación, identificándolas son las dos más importantes que se describieron anteriormente P95_a y P99, existen otras variables dentro de este aspecto como el pensar que dirían tus amistades si supieran que consumes marihuana u otra droga identificadas por P96 y P97.

- El consumo de alcohol y cigarrillo; el mayor número de variables influyentes, tienen que ver con este aspecto, pese a que el orden de importancia es algo menor con respecto al primer aspecto analizado; siendo las variables más importantes en este aspecto P23_a: la cantidad de veces que te has emborrachado y P10: el tiempo de la última vez que consumiste cigarrillo, nótese que en el estudio de árboles de clasificación otra variable considerada la más importante tenía que ver con el consumo de cigarrillo.

- La relación de con quien vive y como se siente o comporta en el colegio también se ve reflejado en el estudio, entre las variables más importantes se encuentra P84: cantidad de veces que se ha escapado del colegio y P79: Si considera que sus padres o apoderado han consumido drogas cuando eran jóvenes; en el estudio analizado anteriormente se consideraron también estas variables.

Una vez encontrado el mejor modelo de regresión, se realiza una interpretación de sus variables y parámetros del modelo con el fin de observar en cierto modo la relación e influencia de las variables que engloban varios aspectos de la población escolar en el consumo de drogas, que es el objetivo principal del estudio.



Teniendo en cuenta la formula, dos cuestiones son consideradas importantes sobre los coeficientes βi, el signo y el tamaño; si es positivo incrementos en Xi e incrementan la probabilidad de pertenecer al grupo de interés, es decir la probabilidad de que un estudiante haya consumido drogas (marihuana para el estudio descrito), signo negativo lo contrario incrementos en Xi disminuyen las probabilidades. Se han considerado en la tabla 7.4.2 de Análisis de estimaciones de máxima verosimilitud las variables más importantes para su interpretación.

Analysis of Maximum Likelihood Estimates							
Parameter	**Category**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Exp(Est)**
P10	**1**	1	0.3753	0.0271	191.63	<.0001	1.455
P10	**2**	1	0.1591	0.0267	35.47	<.0001	1.172
P10	**3**	1	-0.1657	0.0273	36.78	<.0001	0.847
P7	**1**	1	0.3322	0.0320	107.95	<.0001	1.394
P99	**1**	1	-0.8070	0.0262	946.09	<.0001	0.446
P99	**2**	1	-0.2375	0.0215	122.50	<.0001	0.789
P99	**3**	1	0.2093	0.0270	60.09	<.0001	1.233
P99	**4**	1	0.3548	0.0347	104.51	<.0001	1.426
P23_a	**0**	1	-0.3594	0.0202	318.01	<.0001	0.698
P23_a	**1**	1	0.0900	0.0174	26.67	<.0001	1.094
P84	**1**	1	0.2732	0.0226	146.23	<.0001	1.314
P84	**2**	1	-0.3327	0.0220	228.87	<.0001	0.717
P95_a	**1**	1	-0.9295	0.0204	081.50	<.0001	0.395
P95_a	**2**	1	0.3478	0.0166	436.65	<.0001	1.416
P79	1	1	0.2432	0.0113	460.51	<.0001	1.275

Tabla 7.4.2. Análisis de Estimación de máxima verosimilitud Regresión Logística.

A continuación, se muestra la interpretación considerando algunas categorías de las variables importantes.

Variable P95_a

Si un estudiante en el último año, no se ha encontrado nunca cerca de un grupo que estado consumiendo droga, tiene 2.5 veces menos posibilidad de haber consumido

marihuana alguna vez, en comparación de un estudiante que ha estado bastante seguido cerca de un grupo que ha consumido droga.

Si un estudiante en el último año se ha encontrado casi nunca o de vez en cuando cerca de un grupo que estado consumiendo droga, tiene 1.4 veces más posibilidad de haber consumido marihuana alguna vez, en comparación de un estudiante que ha estado bastante seguido cerca de un grupo que ha consumido droga.

Variable P99
Si un estudiante no tiene ningún amigo que fume regularmente marihuana tienen 2.2 veces menos posibilidad de haber consumido marihuana alguna vez, en comparación de un estudiante en donde todos o casi todos sus amigos fuman marihuana regularmente.

Variable P84
Si un estudiante se ha escapado del colegio pocas veces tiene 1.3 veces más posibilidad de no haber consumido marihuana en comparación de un estudiante que se ha escapado varias o muchas veces del colegio

Variable P23_a
Si un estudiante nunca se ha emborrachado en su vida tiene 1.4 veces más posibilidad de no haber consumido marihuana en comparación de un estudiante que se han emborrachado en su vida más de tres veces.

Variable P10
Si un estudiante ha consumido cigarrillos durante los últimos treinta días tiene 1.4 veces más posibilidad de consumir marihuana en comparación de un estudiante que no ha probado cigarrillo. Si un estudiante ha consumido cigarrillo hace más de un año tiene 1.18 veces menos posibilidad de consumir marihuana que un estudiante que no ha consumido cigarrillo.

Variable P7
Si un estudiante ha consumido cigarrillo tiene 1.3 veces más posibilidad de consumir marihuana que un estudiante que no ha consumido cigarrillo.

Variable P79
Si un estudiante piensa que el padre, madre o apoderado ha consumido alguna droga tiene 1.2 veces más de posibilidad de consumir que un estudiante que piensa que no han consumido drogas

En general en el modelo propuesto para el consumo de marihuana se observa considerando el factor más importante de relación de amistad, que el estar más cerca de alguien, así como tener varias amigos que consumen marihuana incrementa de manera significativa la Odds del consumo de marihuana en los estudiantes. Se observó en la interpretación de categorías de algunas variables analizadas, por ejemplo el estar poco o de vez en cuando en grupo o tener más de la mitad de los amigos que estén consumiendo marihuana (considerarlo como término medio) en comparación con el estar bastante seguido de alguien, así como tener casi todos los amigos que fumen esta

sustancia (considerarlo como termino extremo) se tiene más posibilidad que los estudiantes que estén en el término medio consuman marihuana que los que están en el término extremo; esto puede deberse y tomando en cuenta las variables analizadas que para los estudiantes que están en este término extremo y no hayan consumido sea ya más fácil sentirse parte del grupo o resistir la presión (menos vulnerables), pero ya es un tema de interpretación y discusión. En el modelo se pueden observar varias variables con el llamado término medio que analizadas individualmente tienen más posibilidades aunque con bajo valor de consumir marihuana que el termino extremo pero esto no debe sesgar la interpretación. Para el aspecto del consumo de alcohol y cigarrillos se observa en el modelo que a medida que los estudiante que no hayan consumido tabaco o no se hayan excedido nunca con el consumo de alcohol disminuye la odds del consumo de marihuana.

7.5 Mejores Modelos Predictivo variable objetivo: Consumo de Marihuana

El método de ensamblado **ensamb8** que ha resultado de la combinación de los mejores modelos de Regresión Logística, Red Neuronal y Gradient Boosting, se ha considerado la mejor técnica predictiva para que detecte en nuevos estudiantes de la población escolar de Chile el consumo de marihuana, quedando en segundo lugar el modelo de Gradient Boosting (**GradBoost**), el cual ofrece una capacidad predictiva similar.

Los modelos encontrados, ayudarán en la decisión de conocer si una estudiante ha consumido o no marihuana conociendo de ellos previamente cierta información de diversos aspectos, que durante el estudio han resultado ser de gran importancia, entre esta información se encuentran ámbitos personales, relación de amistades, entorno familiar y escolar, además de percepción del estudiante en el consumo de alcohol y tabaco.

Analizando la matriz de confusión del método de ensamblado ensamb8 y el modelo de Gradient Boosting como se observa en la figura 7.5.1, se han obtenido algunas medidas de clasificación que se describen a continuación.

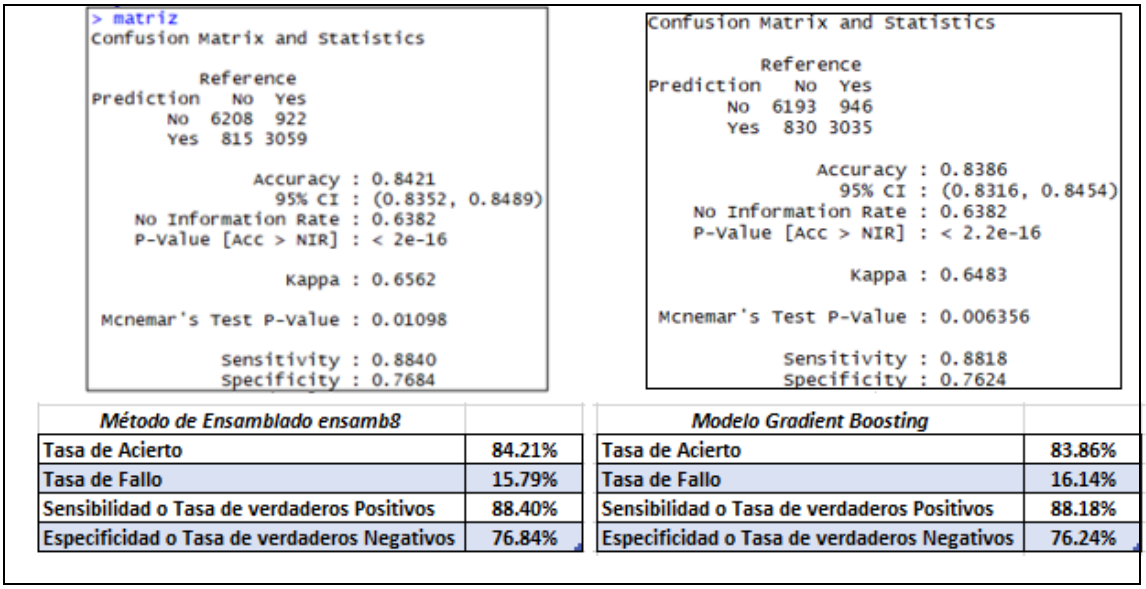
Se tiene un total del 88.40% (ensamb8) y 88.18% (GradBoost) de observaciones bien clasificadas (Tasa de verdaderos positivos) cuando el nivel de la variable objetivo es uno, es decir si ha consumido marihuana, por ende un 11.60% (ensamb8) y 11.82% (GradBoost) de observaciones mal clasificadas para este nivel.

Se tiene un total del 76.84% (ensamb8) y 76.24% (GradBoost) de observaciones bien clasificadas (Tasa de verdaderos negativos) cuando el nivel de la variable objetivo es cero, es decir no ha consumido marihuana, por ende un 23.16% (ensamb8) y 23.76% (GradBoost) de observaciones mal clasificadas para este nivel. Lo que equivaldría a un total del 84.21% de observaciones de prueba bien clasificadas utilizando la mejor técnica predictiva ensamb8 encontrada, que es la tasa de aciertos; en el caso de GradBoost, se tiene una tasa de acierto del 83.86%.

En cuanto a los valores predictivos positivos (probabilidad de haber consumido marihuana si el resultado de la prueba es positivo.) y negativo (probabilidad de no haber consumido marihuana si el resultado de la prueba es negativo) para el método de

ensamb8 se tiene un valor del 76.84% y 88.39% respectivamente; en el modelo de GradBoost estas medidas son del 76.23% y 88.18%.

Si se compara las mejores técnica de predicción encontradas con el no modelo, se observa que este ha mejorado significativamente, ha disminuido el error de un 36% a un 15%; además, si se lo compara con la media que tiene una probabilidad de acierto del 62% (no modelo), ahora con el modelo ganador se tiene una probabilidad de acierto alrededor del 84% (mejores modelos estudio de predicción).



Método de Ensamblado ensamb8	
Tasa de Acierto	84.21%
Tasa de Fallo	15.79%
Sensibilidad o Tasa de verdaderos Positivos	88.40%
Especificidad o Tasa de verdaderos Negativos	76.84%

Modelo Gradient Boosting	
Tasa de Acierto	83.86%
Tasa de Fallo	16.14%
Sensibilidad o Tasa de verdaderos Positivos	88.18%
Especificidad o Tasa de verdaderos Negativos	76.24%

Figura 7.5.1. Consumo de Marihuana: Medidas de Clasificación mejor modelo predictivo.

A continuación, en la tabla 7.5.1 se resumen los mejores modelos de predicción construidos, para detectar el consumo de marihuana en nuevos estudiantes de la población escolar de Chile; se muestran sus diferentes medidas de clasificación con el objetivo de evaluar la capacidad predictiva de los mejores modelos de machine Learning que han resultado durante el estudio; se encuentra resaltadas en verde las mejores medidas encontradas que corresponde a técnicas diferentes.

Modelos	Tasa de Acierto	Tasa de Fallos	Sensibilidad	Especificidad	Valor predictivo Positivo	Valor predictivo Negativo
Árbol de Decisión	*82.06%*	*17.94%*	*74.69%*	*86.23%*	*75.46%*	*85.73%*
Regresión Logística	*84.19%*	*15.81%*	*85.80%*	*82.25%*	*73.26%*	*91.08%*
Gradient Boosting	*83.86%*	*16.14%*	*88.18%*	*76.24%*	*76.23%*	*88.18%*
Ensamblado	*84.21%*	*15.79%*	*88.40%*	*76.84%*	*76.84%*	*88.39%*

Tabla 7.5.1. Medidas de clasificación, capacidad predictiva de los mejores modelos construidos.

8. DISCUSIÓN Y CONCLUSIONES

El estudio que se llevó a cabo tenía como objetivo principal analizar a través de técnicas de Machine Learning como influye la información personal y social de la población escolar de Chile en el consumo de drogas. Para ello se han considerado aspectos tales como monitoreo parental, relación de amistades, personas con las que habita, entorno escolar y familiar, percepción y patrones de consumo de algunas drogas licitas e ilícitas, entre otra información adicional como sociodemográfica. Utilizando esta información se han implementado modelos de predicción de árboles de clasificación y regresión logística que proporcionan modelos de predicción y permiten analizar y evaluar el poder predictivo que tienen los aspectos mencionados en el consumo de drogas, tratando de encontrar relaciones de interés, donde se pueda realizar de cierto modo una interpretación de la información de los aspectos analizados; además se han implementado otras técnicas de predicción como Redes Neuronales, Random Forest, Gradient Boosting y métodos de Ensamblado con el objetivo de encontrar el mejor modelo predictivo para detectar si un nuevo estudiante ha experimentado el consumo de marihuana.

Lo resultados del presente estudio, indican en primer lugar que la información personal y social englobadas en los diferentes aspectos mencionados, muestran una relación con el consumo de drogas y permiten predecir con una capacidad aceptable la experimentación inicial del consumo de estas sustancias. El estudio confirma la importancia del aspecto de la relación e influencia de amistades para el consumo de drogas analizadas (marihuana) (cocaína o pasta base) (otras drogas como: crack, éxtasis, heroína, alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos), donde se evidencia que aquellos estudiantes que durante el último tiempo han estado cerca de personas que consumen drogas o qué a medida tengan amigo(a)s que las consumen tienen más posibilidad de experimentar con las mismas sustancias; otro aspecto de gran importancia es el consumo de tabaco y alcohol, en los modelos explorados cierta información de esta última sustancia licita aparece o se ha contemplado de gran utilidad a la hora de discriminar o clasificar a un estudiante con más alta posibilidad de haber consumido o no drogas; donde se pone en manifiesto que a medida que aumenta las veces que un estudiante se ha embriagado o intoxicado tomando alcohol, así como la cantidad de licor que toma en una salida por la noche aumenta la posibilidad de que consuma drogas.

En relación con el consumo de marihuana, la muestra de estudio indica que los estudiantes de la población secundaria de Chile obtienen una tasa alta de consumo, tomando en consideración e importancia la información mencionada anteriormente sobre la relación de amistad y el consumo de sustancias ilícitas. Se observa además con el modelo de árbol de clasificación que el haber experimentado el consumo de tabaco es la información más relevante junto con el número de días que ha consumido últimamente. El consumo de cigarrillo puede pronosticar el uso de marihuana, es así que a un estudiante que no ha experimentado el consumo de cigarrillo, teniendo en cuenta solo esta información, lo clasifica con alrededor del 91.5% como que no ha consumido marihuana. En lo que respecta a la interpretación con el modelo de regresión de esta variable, un estudiante que ha experimentado el consumo de cigarrillo tiene una probabilidad 30% mayor de consumir marihuana que un estudiante que no haya

consumido cigarrillo. Existen otros aspectos de menor influencia como información relacionada con quien vive y la vida escolar del estudiante; se evidencia que los estudiantes que no se han escapado nunca del colegio y que tengan una percepción de que sus padres o apoderado no haya consumido drogas tienen menos posibilidades de experimentar con el consumo de marihuana.

Se han encontrado varios modelos predictivos en cara a identificar si los estudiantes han consumido drogas. Evaluando el poder predictivo para cuando el objetivo es detectar el haber consumido marihuana, se tiene modelos aceptables para identificar ya sea si un estudiante ha consumido la sustancia o no lo ha hecho, así en el modelo de árboles de clasificación considerado el más explicativo o de mejor interpretación por el tipo de información que utiliza (casi el total de variables categóricas) se tiene un total 82.06% de observaciones de prueba bien clasificadas, siendo la tasa de verdaderos positivos del 74.69% y la de verdaderos negativos del 86.23%; además se puede aceptar la predicción de haber consumido o no marihuana cuando el modelo lo indica así con un 75.46% y 85.73% respectivamente. Estas medidas de clasificación que evalúan el poder predictivo de los modelos se han visto mejoradas en pequeña proporción con la utilización y comparación de otros modelos como regresión logística que también ha resultado útil en la interpretación de resultados, se han utilizado otras técnicas de gran capacidad predictiva como Random Forest y métodos de Ensamblado que han dado lugar a buenos resultados inclusive reduciendo la cantidad de información ingresada por su alto procesamiento, destacando la técnica de ensamblado como la mejor.

En relación con el consumo de cocaína o pasta base, la información sobre la relación de amistad y el consumo de sustancias ilícitas especialmente el alcohol toma gran importancia como se comentó anteriormente de forma general. Pero en este estudio se identifica además, que el consumo de marihuana guarda una relación importante al momento de evaluar el poder predictivo en los modelos para clasificar a un estudiante como que ha experimentado o no con estas sustancias ilícitas, siendo esta información de carácter personal del estudiante un factor muy importante junto al factor relacionado con la amistad (información más importante) e información referente al consumo de alcohol (gran cantidad de información relacionada); así se pone en manifiesto que el consumo de marihuana junto a los factores mencionados guardan un relación y permiten clasificar a un estudiante como consumidor o no consumidor de cocaína o pasta base.

En relación con el consumo de otras drogas, además de los aspectos descritos anteriormente sobre los factores de la relación de amistad y el consumo de alcohol, el estudio confirma que tiene aún mayor efecto el haber consumido cocaína o pasta base para clasificar a un estudiante con más alta posibilidad de haber experimentado o no el consumo de otras drogas. Dada la muestra de estudio con más proporción de estudiantes que no han consumido esta sustancia; el modelo clasifica de mejor manera a los estudiantes que no lo han consumido, es decir la tasa de verdaderos negativos que tiene un total del 98% de observaciones de prueba bien clasificadas. Para este estudio a diferencia del consumo de marihuana y cocaína o pasta base, se observa mayor participación de características que tienen que ver con experiencias sucedidas en el colegio como si el estudiante ha sido agredido por alguien o si ha robado en la institución educativa, en el caso de sean afirmativas el modelo tendrá menos

posibilidades de clasificar a un estudiante como no haber experimentado con las sustancias contempladas en otras drogas.

Los modelos encontrados en estos dos últimos análisis, dada la proporción de la muestra con más sucesos de estudiantes que no han consumido estas sustancia, clasifica de mejor manera la no experimentación de estas drogas en nuevos estudiantes, es así que para cuando el objetivo es el consumo de cocaína o pasta base, la probabilidad de no haber consumido estas sustancias cuando así el modelo lo indique tiene un valor 93.71%, en cambio en valores predictivos positivos es decir probabilidad de haber consumido estas sustancias cuando el resultado lo indique así tiene un valor del 56.91%; en el caso del consumo de otras drogas estos mismos valores son del 60.91% y 95.28%.

En último lugar, mencionar que existen varias variables (información) que no han sido de utilidad en los modelos, entre ellas los aspectos sociodemográficos del sexo y edad, que en ningún caso han resultado factores discriminatorios en el consumo de drogas cuando se han introducido como variables independientes; esto en comparación con las otras variables de los factores utilizados.

Los árboles de decisión han resultado de gran utilidad, pudiendo aumentar información que resulte de interés para mejorar su interpretación y aumentar la capacidad predictiva a la hora de experimentar con estas sustancias (drogas).

Dentro del marco de actuación, tener en cuenta los resultados y aspectos encontrados para tomar acciones sobre el consumo de drogas en la población escolar.

En términos generales se especifican algunas Conclusiones importantes:

- Los resultados confirman la relación de influencia de la información personal y social en la experimentación del consumo drogas.

- El factor más determinante en los estudios es la relación con el grupo de consumidores de drogas, es decir grupos de personas o amigos con comportamientos y consumo favorables a las mismas.

- Otro aspecto que discrimina y aumenta la posibilidad de consumo de drogas es la cantidad de alcohol que toma, así como el número de veces que se ha intoxicado por el exceso de esta sustancia.

- El consumo de cigarrillo tiene una relación alta para detectar si un estudiantes ha experimentado o no con el consumo de marihuana.

- Analizando el consumo de marihuana es un factor influyente para que un estudiante experimente con el consumo de cocaína o pasta base.

- El consumo de cocaína o pasta fase tiene un mayor efecto para que un estudiante haya consumido otras drogas.

- Los resultados obtenidos con respecto al factor escolar muestra que los estudiantes con un promedio escolar bajo, que han sido agredidos en el colegio o que han robado tienen mayor probabilidad de iniciarse en el consumo de drogas, en especial en el estudio que contempla las sustancias de otras drogas.

- Los aspectos de la relación de amistad, así como el consumo de alcohol, tabaco y drogas tienen un mayor efecto en las experimentación de sustancias ilícitas en comparación con aspectos como el monitoreo parental, entorno familiar, escolar y de convivencia.

- En el estudio de este análisis de predicción, con los construcción y comparación de modelos se puede predecir con una capacidad aceptable la experimentación inicial del consumo de drogas.

Trabajo Futuro

Se deben profundizar en los factores que puedan influenciar en el inicio de consumo de drogas en escolares de Chile y a nivel general, así se pueden ir incorporando más preguntas para abordar y comprender de mejor manera este fenómeno social. Se pueden agregar factores mencionados en otros estudios como aspectos económicos permisividad, impulsividad y búsqueda de sensaciones en los estudiantes.

Con respecto a los análisis de predicción, para obtener mejores modelos se deben incorporar todas la variables que aporten alguna información y trabajar con técnicas de gran capacidad predictiva. Además con el objetivo de evaluar la capacidad predictiva de los aspectos de la población escolar en el consumo de drogas se debe indagar de mejor manera las técnica de árboles, regresión y otras que resulten de utilidad e interés.

9. BIBLIOGRAFÍA

1. Valencia, R. (2015). Boletín 19 Involucramiento parental y consumo de drogas en escolares de Chile.pdf. (s. f.).
2. Fernández, H. M. (2012). El costo socioeconómico del consumo de drogas ilícitas en Chile. Revista Cepal (107), 93-114.
3. Montequín *et al. (2003). - METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE D.pdf*. (s. f.). Recuperado de https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf
4. Hastie, Tibshirani. (2009). Elements of Statistical Learning: Data mining, inference, and prediction. 2nd Edition. (s. f.). Recuperado de https://web.stanford.edu/~hastie/ElemStatLearn//
5. ENPE. *(2017)*. (s. f.). Recuperado de http://www.senda.gob.cl/wp-content/uploads/2019/01/ENPE-2017.pdf
6. García, E. G., & Pol, A. L. P. (2009). Predicción del consumo de cocaína en adolescentes mediante árboles de decisión. *Revista de investigación en educación*, *6*(1), 7-13.
7. Alfonso, J. P., Huedo-Medina, T. B., & Espada, J. P. (2009). Factores de riesgo predictores del patrón de consumo de drogas durante la adolescencia. *Anales de Psicología / Annals of Psychology*, *25*(2), 330-338.
8. *Victoria De Girón*, V. (2014). *Comportamiento adictivo de la familia como factor de riesgo de consumo de drogas en jóvenes y adolescentes adictos.* Recuperado de https://www.medigraphic.com/pdfs/revcubinvbio/cib-2014/cib144h.pdf.
9. *Franco*, A. J.-M., Agustín, A. B. S., Baile, A. M., Valero, P. G., & Puerta, I. N. de la. (2009). Consumo de drogas en estudiantes universitarios de primer curso. *Adicciones*, *21*(1), 21-28. Recuperdado de https://doi.org/10.20882/adicciones.248.
10. *García*, E. G., & Pol, A. L. P. (2009). Predicción del consumo de cocaína en adolescentes mediante árboles de decisión. *Revista de investigación en educación*, *6*(1), 7-13.

11. Rodríguez; De la Villa; & Sirvent. (2006). *Factores relacionados con las actitudes juveniles .pdf*. (s. f.).

12. *Delgado; & Martínez. (2016). Características Psicosociales Asociadas al Consumo de Alcohol, Tabaco y Drogas en Adolescentes de Chiapas*. (s. f.).

13. Saravia, J. C., Gutiérrez, C., & Frech, H. (2014). *Factores asociados al inicio de consumo de drogas ilícitas en adolescentes de educación secundaria. 18*(1), 8.

14. Informe *Europeo sobre Drogas: Tendencias y novedades. (2019)*. (s. f.). 100.

15. Martin*, J.* (s. f.). Recuperado de https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartin-slides.pdf

16. D. G. de S. P; Real, M. (2016). Encuesta sobre drogas a la población escolar. Recuperado de https://saludcantabria.es/uploads/pdf/profesionales/drogodependencias/Escolar%202016.pdf

17. Calviño*, A. (2019)*. Apuntes Machine Learning. Facultad de Estudios Estadísticos, Universidad Complutense de Madrid.

18. Rodriguez*, J. M. (2018)*. *Estudio comparativo de modelos de machine learning para la detección de dianas microARN*.

19. SampSize function | R Documentation. (s. f.). Recuperado de https://www.rdocumentation.org/packages/DoseFinding/versions/0.5-2/topics/sampSize

20. Tibshirani, S., & Friedman, H. (s. f.). *Valerie and Patrick Hastie*. 764.

21. Portela, J. (2019). Apuntes Machine Learning. Facultad de Estudios Estadísticos, Universidad Complutense de Madrid.

10. ANEXOS

ANEXO I: Depuración de los Conjuntos de Datos

Análisis Descriptivo

Análisis Descriptivo variables de Intervalo Consumo Cocaína

Variables de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
P11	P11	1389	109031	0	30	3.133714	7.425469	2.660881	6.000865
P12	P12	889	109531	0	40	1.864769	5.62102	4.891265	26.24662
P16	P16	904	109516	0	21	10.26497	6.097957	-0.91364	-0.8474
P2	P2	1499	108921	10	25	15.59727	1.588484	0.225306	-0.01085
P21	P21	1431	108989	0	30	2.134142	4.651659	3.808441	16.81211
P8_a	P8_a	1898	108522	0	21	7.211948	6.923263	-0.00221	-1.86116
P8_b	P8_b	1942	108478	0	21	3.257103	6.147528	1.41123	0.098144

Análisis Descriptivo variables de Clase Consumo Cocaína

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P1	P1	N	2	712
P10	P10	N	4	2091
P105	P105	N	6	2387
P106	P106	N	8	2549
P108	P108	N	2	2151
P109	P109	N	6	5357
P110	P110	N	8	1872
P15	P15	N	2	776
P17	P17	N	4	948
P18	P18	N	4	1026
P19	P19	N	6	677
P20_a	P20_a	N	3	931
P20_b	P20_b	N	3	980
P20_c	P20_c	N	3	2576
P20_d	P20_d	N	3	2731
P22	P22	N	6	967
P23_a	P23_a	N	7	647
P23_b	P23_b	N	7	703
P23_c	P23_c	N	7	671
P25	P25	N	5	1533
P26	P26	N	11	16592
P27	P27	N	7	414
P3	P3	N	3	332
P35	P35	N	2	548
P4	P4	N	4	383
P47_53	P47_53	N	2	0
P5	P5	N	3	673
P6_a	P6_a	N	5	903
P6_b	P6_b	N	5	819

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P6_b	P6_b	N	5	819
P6_c	P6_c	N	5	862
P6_d	P6_d	N	5	1128
P6_e	P6_e	N	5	971
P6_f	P6_f	N	5	979
P6_g	P6_g	N	5	1300
P7	P7	N	2	511
P73	P73	N	4	2415
P74_a	P74_a	N	9	1421
P74_b	P74_b	N	9	1022
P74_c	P74_c	N	9	2262
P75	P75	N	8	2285
P76	P76	N	4	1062
P77	P77	N	2	651
P78	P78	N	3	711
P79	P79	N	2	1038
P80	P80	N	5	944
P81_a	P81_a	N	6	1124
P81_b	P81_b	N	6	731
P82_a	P82_a	N	6	1042
P82_b	P82_b	N	6	797
P82_c	P82_c	N	6	1104
P82_d	P82_d	N	6	929
P83	P83	N	3	762
P84	P84	N	5	676
P85	P85	N	6	869
P86	P86	N	3	721
P87	P87	N	2	772
P88	P88	N	2	813

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P89	P89	N	2	1575
P9	P9	N	4	1345
P90_a	P90_a	N	5	963
P90_b	P90_b	N	5	826
P90_c	P90_c	N	5	958
P90_d	P90_d	N	5	1073
P90_e	P90_e	N	5	1015
P91_a	P91_a	N	5	1128
P91_b	P91_b	N	5	1189
P91_c	P91_c	N	5	1347
P91_d	P91_d	N	5	1589
P91_e	P91_e	N	5	1666
P92	P92	N	5	869
P93	P93	N	5	963
P94	P94	N	5	1221
P95_a	P95_a	N	5	1801
P95_b	P95_b	N	5	1321
P95_c	P95_c	N	5	1323
P95_d	P95_d	N	5	1545
P95_e	P95_e	N	5	1515
P96	P96	N	4	1571
P97	P97	N	4	1683
P98	P98	N	5	1737
P99	P99	N	5	2203

Análisis Descriptivo variables de Intervalo Consumo Otras Drogas

Variables de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
P11	P11	1352	108368	0	30	3.129632	7.421833	2.662505	6.010383
P12	P12	870	108850	0	40	1.856206	5.604366	4.905457	26.41137
P16	P16	887	108833	0	21	10.27234	6.095091	-0.91697	-0.8418
P2	P2	1464	108256	10	25	15.5973	1.587458	0.223884	-0.0202
P21	P21	1388	108332	0	30	2.130183	4.643949	3.813582	16.8716
P8_a	P8_a	1852	107868	0	21	7.211842	6.923533	-0.00258	-1.8619
P8_b	P8_b	1893	107827	0	21	3.249279	6.143372	1.414702	0.106927

Análisis Descriptivo variables de Clase Consumo Otras Drogas

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P1	P1	N	2	695
P10	P10	N	4	2070
P105	P105	N	6	2340
P106	P106	N	8	2531
P108	P108	N	2	2114
P109	P109	N	6	5298
P110	P110	N	8	1852
P15	P15	N	2	766
P17	P17	N	4	936
P18	P18	N	4	1015
P19	P19	N	6	670
P20_a	P20_a	N	3	912
P20_b	P20_b	N	3	950
P20_c	P20_c	N	3	2525
P20_d	P20_d	N	3	2679
P22	P22	N	6	949
P23_a	P23_a	N	7	643
P23_b	P23_b	N	7	672
P23_c	P23_c	N	7	635
P25	P25	N	5	1499
P26	P26	N	11	16520
P27	P27	N	7	409
P3	P3	N	3	322
P35	P35	N	2	549
P4	P4	N	4	374
P5	P5	N	3	652
P65_f_g_i_j	P65_f_g_i_j	N	2	0
P6_a	P6_a	N	5	882
P6_b	P6_b	N	5	769

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P7	P7	N	2	498
P73	P73	N	4	2377
P74_a	P74_a	N	9	1376
P74_b	P74_b	N	9	966
P74_c	P74_c	N	9	2182
P75	P75	N	8	2245
P76	P76	N	4	1031
P77	P77	N	2	624
P78	P78	N	3	682
P79	P79	N	2	1013
P80	P80	N	5	911
P81_a	P81_a	N	6	1086
P81_b	P81_b	N	6	684
P82_a	P82_a	N	6	1003
P82_b	P82_b	N	6	749
P82_c	P82_c	N	6	1044
P82_d	P82_d	N	6	869
P83	P83	N	3	747
P84	P84	N	5	653
P85	P85	N	6	853
P86	P86	N	3	692
P87	P87	N	2	740
P88	P88	N	2	786
P89	P89	N	2	1536
P9	P9	N	4	1333
P90_a	P90_a	N	5	939
P90_b	P90_b	N	5	772
P90_c	P90_c	N	5	895
P90_d	P90_d	N	5	1014

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
P90_e	P90_e	N	5	954
P91_a	P91_a	N	5	1090
P91_b	P91_b	N	5	1126
P91_c	P91_c	N	5	1272
P91_d	P91_d	N	5	1514
P91_e	P91_e	N	5	1594
P92	P92	N	5	847
P93	P93	N	5	930
P94	P94	N	5	1196
P95_a	P95_a	N	5	1763
P95_b	P95_b	N	5	1256
P95_c	P95_c	N	5	1252
P95_d	P95_d	N	5	1474
P95_e	P95_e	N	5	1442
P96	P96	N	4	1548
P97	P97	N	4	1648
P98	P98	N	5	1700
P99	P99	N	5	2150

Corrección de Errores

AGRUPACIÓN DE CATEGORIAS
- **P4 ¿Cuán atentos están tu padre, madre, apoderado o apoderada (o alguno de ellos) respecto de lo que haces en el colegio? ① □Mucho ② □Bastante ③ □Poco ④ □Nada**
Se unen las categorías mucho y bastante, al tener significados semejantes, **(1) Y (2) se unen = (1)** → Mucho o bastante.

- **P6c. Fumar una o más cajetillas de cigarros al día : Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨**

Se unen la categoría de ningún riesgo y riesgo leve, **(1) Y (2) se unen = (1) → riesgo leve o ninguno**

- **P6f. Emborracharse con alcohol: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨**

Se unen la categoría ningún riesgo y riesgo leve, **(1) Y (2) se unen = (1) → riego leve o ninguno**

- **P6g: Tomar uno o dos tragos de alcohol todos o casi todos los días: Ningún riesgo ① Riesgo leve ② Riesgo moderado ③ Riesgo grande ④ No sé ⑨**

Se unen las categorías ningún riesgo y riego leve, **(1) Y (2) se unen = (1) → riego leve o ninguno**

- **P74_a: ¿Qué educación alcanzaron tu padre?**

Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta ⑤ Técnica superior completa ⑥ Universitaria incompleta ⑦ Universitaria completa ⑧ No sé o No aplica ⑨

Se unen la categoría **5 y 7 = 5 → (Técnica superior incompleta o universitaria Incompleta)**
todas bajan un número de identificador, 9 se mantiene

- **P76 ¿Quién es tu apoderado/apoderada? Apoderado/apoderada es quien se responsabiliza por ti ante las autoridades del colegio ① □Padre ② □Madre ③ □Abuela o Abuelo ④ □Otro**

Se unen la categoría **(3) y (4) = (3) → Abuelo(a) u otro familiar**

¿Cómo crees que estaría tu papá y tu mamá en estas situaciones?

- **P82_a: Si tu mamá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥**
- **P82_b: Si tu mamá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥**
- **P82_c: Si tu papá descubriera que fumas marihuana: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥**
- **P82_d: Si tu mamá descubriera que fumas marihuana: Extremadamente molesto(a) ① Bastante molesto(a) ② Algo molesto(a) ③ Poco molesto(a) ④ Indiferente ⑤ No aplica ⑥**

Se unirán las categorías **(1 y 2)=(1)→ extremada o baste molest**o(a) **y (3 y 4)=(2)→ algo o poco molesto(a)**
todas bajan 2 puestos en el número de identificador 6 se mantiene

- **P84: Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste al colegio una parte importante de la jornada o en toda la jornada ① □Nunca ② □Casi nunca ③ □Pocas veces ④ □Varias veces ⑤ □Muchas veces**

Se unirán las categorías **(2 y 3)=2 y (4 y 5)=3**

- **Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio?**

P90_a.	Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio? 90a. Participado en un grupo que molesta a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P90_b.	90b. Participado en un grupo que ha agredido físicamente a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P90_c.	90c. Participado en un grupo que ha comenzado una pelea con otro grupo : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P90_d.	90d. Comenzado una pelea solo con otro/a compañero/a : ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P90_e.	90e. Has robado algo a alguien en el colegio: ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_a.	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio 91a. Has sido molestado/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces

P91_b.	91b. Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_c.	91c. Has estado en un grupo que ha sido atacado por otro grupo ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_d.	91d. Alguien solo/sola ha iniciado una pelea contigo ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces
P91_e.	91e. Te han robado algo en el colegio ① Nunca ② Una vez ③Dos veces ④ 3 o 4 veces ⑤5 o más veces

Se unirán las categorías **(4 y 5)=(4)→3 o 4 veces con 5 o más veces**

- **P92 ¿Cuán probable es que pases de curso este año? ①□Es seguro ②□Muy probable ③□Más o menos probable ④□Poco probable ⑤□Imposible**
- **P93 ¿Cuán probable es que termines cuarto medio? ①□Es seguro ②□Muy probable ③□Más o menos probable ④□Poco probable ⑤□Imposible**
- **P94¿Cuán probable es que sigas estudiando después del colegio? (en la Universidad, Instituto Profesional, Centro de Formación técnica u otro) ①□Es seguro ②□Muy probable ③□Más o menos probable ④□Poco probable ⑤□Imposible**

Se unen las categorías **(1 y 2) = 1 Seguro o Muy probable**
Se unen las categorías **(4 y 5) = 3 Poco probable o Imposible**

Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alguna de estas sustancias con el evidente propósito de volarse, drogarse o embriagarse?

- **95a. Marihuana: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95b. Cocaína: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95c. Pasta base: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95d. Inhalables: Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**
- **95e. Alcohol : Nunca ① Casi nunca ② De vez en cuando ③ Bastante seguido ④ Muy seguido⑤**

Se unen las categorías **(2 y 3)=2 Casi Nunca o de vez en cuando y (4 y 5)=3 Bastante o muy seguido**

- **P8a: ¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez? No consideres si tus padres o algún adulto te dieron a probar siendo niño.**
- **P8b: ¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días?**

Edad en años: Marca"0" en la hoja de respuestas si no has fumado todos o casi todos los días

Se consideran los siguientes rangos **(5 a 12)=1 , (13 a 17)=2, (18 a 21)=3**

- **P16. ¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? No consideres si tu padre, madre o una persona adulta te dieron a probar siendo niño/niña.**

Edad en años: Marca"0"en la hoja de respuestas si no has probado

Se consideran los siguientes rangos **(5 a 12)=1 , (13 a 17)=2, (18 a 21)=3**

- **P19 ¿Cuán difícil te sería comprar alguna bebida alcohólica, si quisieras hacerlo? ① Me sería muy fácil ② Me sería fácil ③ Me sería difícil ④ Me sería muy difícil ⑤ No podría comprarla ⑨ No sé**

Se unen las categorías **(1 y 2)=1 → fácil y (3 y 4)=2 → Difícil**

- **P22. ¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol? Guíate por la siguiente tabla para saber cuántos tragos consumes**

1 trago (una botella o lata individual de cerveza (333 cc.); Un vaso de vino (140 cc.); Un trago de licor (40 cc. de pisco, ron, vodka o whisky, sólo o combinado) 1 trago y medio (medio litro de cerveza) 3 tragos (un litro de cerveza) 6 tragos (una botella de vino (750 cc.) 8 tragos (una caja de vino (1 litro) 18 tragos (una botella de licor (750 cc.)
① 1 a 2 tragos ② 3 a 4 tragos ③ 5 a 6 tragos ④ 7 a 8 tragos ⑤ 9 o más tragos ⑨ Nunca o casi nunca consumo alcohol

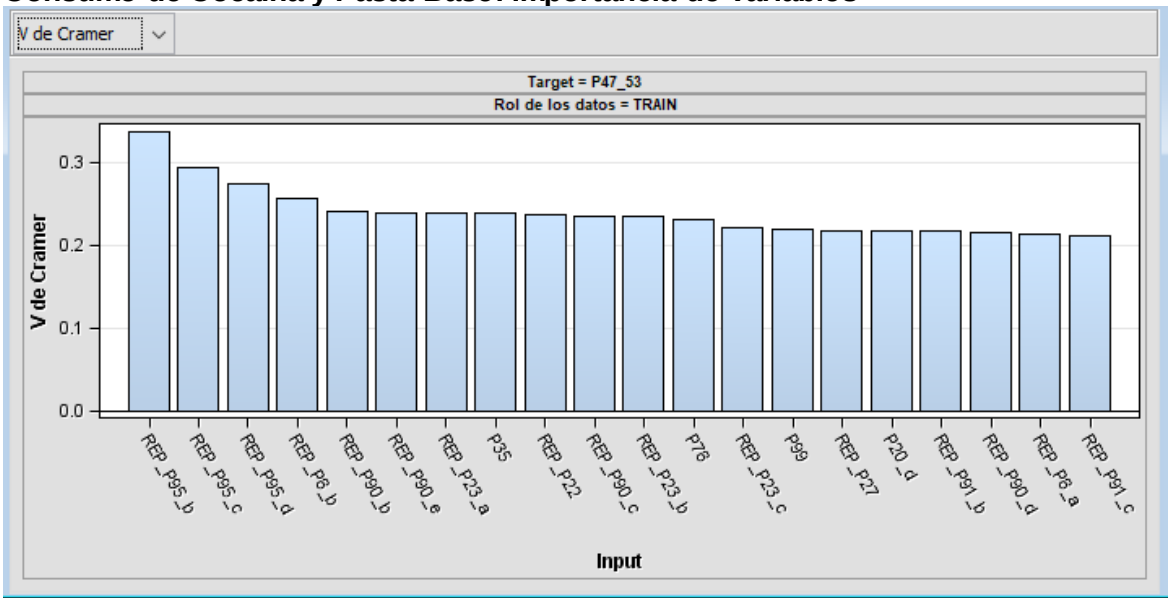
Se unen las categorías **(3-4-5)=3 → 5 o más tragos**

¿Cuántas veces te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?

Nunca ⓪ 1-2 veces ① 3-5 veces ② 6-9 veces ③ 10-19 veces ④ 20-39 veces ⑤ 40 o más veces ⑥ - **P23a: En tu vida** - **P23b: En los últimos 12 meses** - **P23c: En los últimos 30 días**
Se unen las categorías **(2,3,4,5,6)=2→ más de 3 veces** Se unen las categorías **(1 a 6)=1→ una o más de una vez**
- **P26: Indica, de 1 a 10, qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol, donde 1 equivale a "tomé alcohol pero no sentí ningún efecto" y 10 equivale a "estaba tan borracho que no me acuerdo de nada".** **No me hizo efecto 1 2 3 4 5 6 7 8 9 10 No me acuerdo de nada, Marca "99" en la hoja de respuestas si no has consumido**
Se unen las categorías **(1 a 3)=1 → Poco o casi nada** Se unen las categorías **(4 a 7)=2 → Medio tomado** Se unen las categorías **(8 a 10)=3 → Bien tomado** Se considera una nueva categoría **"no responde" (88), debido al gran número de observaciones.**
- **P27 Pensando en una SALIDA DE SÁBADO POR LA NOCHE ¿Cuántos vasos de cerveza, vino o licor llegas a tomar?** **① □Nunca he tomado alcohol ② □Ninguno ③ □Menos de 1 ④ □Uno ⑤ Entre 2 y 5 ⑥ Entre 6 y 10 ⑦ Más de 10**
Se unen las categorías **(3 y 4)=3 → Uno o menos de uno** Se unen las categorías **(6 y 7)=5 → más de 5 o 6 en adelante** El identificador baja (5 vale ahora 4).

ANEXO II: Importancia de Variables Otras variables objetivos

Consumo de Cocaína y Pasta Base: Importancia de Variables



Consumo de Cocaína y Pasta Base: Descripción de variables más importantes

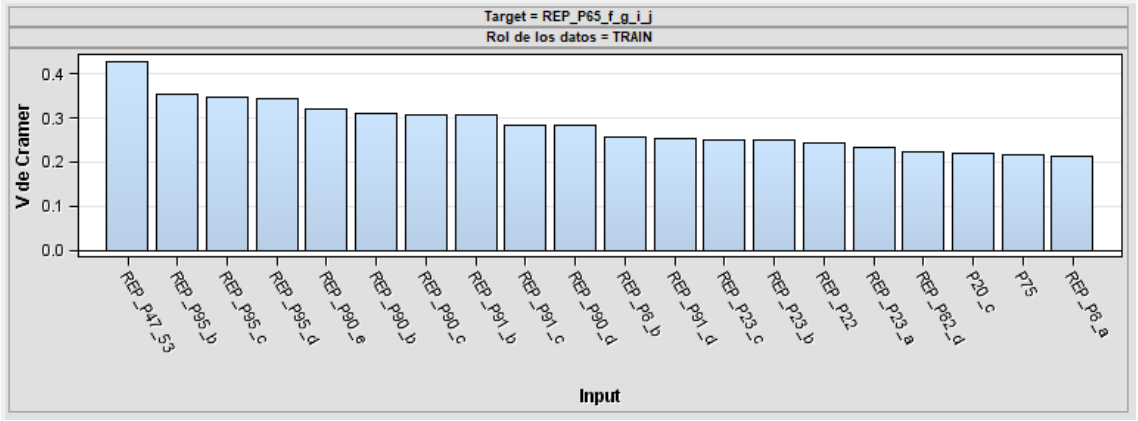
P95_b	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo cocaína con el evidente propósito de volarse, drogarse o embriagarse?
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse?
P95_d	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo inhalables con el evidente propósito de volarse, drogarse o embriagarse?
P8_b	¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días?

P90_b	Durante los últimos 12 meses, ¿cuán seguido has participado en un grupo que ha agredido físicamente a un compañero/a que está solo/a?
P90_e	Durante los últimos 12 meses, ¿cuán seguido has robado algo a alguien en el colegio?
P23_a	¿Cuántas veces en tú vida te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P35	**¿Has consumido marihuana alguna vez en la vida?**
P22	¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol?
P90_c	Durante los últimos 12 meses, ¿cuán seguido has participado en un grupo que ha comenzado una pelea con otro grupo?
P23_b	¿Cuántas veces en los últimos 12 meses te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P76	¿Quién es tu apoderado/apoderada?
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana?
P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar?
P20_d	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Tener relaciones sexuales sin condón
P91_b	Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio
P90_d	Durante los últimos 12 meses, Has Comenzado una pelea solo con otro/a compañero/a
P8_a	¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez?
P91_c	Has estado en un grupo que ha sido atacado por otro grupo

Consumo de Cocaína y Pasta Base: Variables Excluidas del Estudio

P88	En general, ¿consideras que en tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos dentro del colegio?
P77	¿Has conversado con tu padre, madre o apoderado/a acerca de las consecuencias del consumo de drogas?
P6_a	¿Cuál crees tú que es el riesgo que corre una persona que fuma cigarrillos de vez en cuando?
P6_d	¿Cuál crees tú que es el riesgo que corre una persona que toma bebidas alcohólicas de vez en cuando?
P81_a	¿Cómo describirías el hábito que tiene tu padre respecto al alcohol (vino, cerveza, licor)?
P89	¿consideras que en los alrededores de tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos en las afueras o cercanías del colegio?
P83	Durante este año, ¿Te ha tocado asistir o participar en el colegio en actividades específicamente destinadas a prevenir el consumo de drogas, como por ejemplo charlas o talleres?
P1	Sexo
P87	En general, ¿consideras que en tu colegio hay estudiantes que traen, toman o comparten alcohol dentro del colegio?
P15	¿Has probado alcohol alguna vez en la vida (cerveza, vinos o tragos fuertes)?
P5	En general, ¿cuánto crees que tu padre, madre, apoderado o apoderada (o alguno de ellos) conocen a tus amigos y amigas más cercanos/as?
P74_c	¿Qué educación alcanzo tu jefe/jefa de hogar?
P106	¿De cuánto dinero al mes dispones generalmente para tus gastos? Haz un cálculo mensual
P110	Pensando en los últimos 7 días, ¿cuántos días hiciste ejercicio o actividad física, fuera del horario de clases, durante al menos 20 minutos y que te haya hecho transpirar o respirar fuertemente?

Consumo Otras Drogas: Importancia de Variables



Consumo Otras Drogas: Descripción de variables más importantes (Se encuentran resaltadas las variables que coinciden con las variables importantes de la variable objetivo Consumo de Cocaína o Pasta Base**)**

P95_b	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo cocaína con el evidente propósito de volarse, drogarse o embriagarse?
P95_c	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo pasta base con el evidente propósito de volarse, drogarse o embriagarse?
P95_d	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo inhalables con el evidente propósito de volarse, drogarse o embriagarse?
P8_b	¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días?
P90_b	Durante los últimos 12 meses, ¿cuán seguido has participado en un grupo que ha agredido físicamente a un compañero/a que está solo/a?
P90_e	Durante los últimos 12 meses, ¿cuán seguido has robado algo a alguien en el colegio?
P23_a	¿Cuántas veces en tú vida te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P91_d	Durante los últimos 12 meses, Alguien solo/sola ha iniciado una pelea contigo
P22	¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol?
P90_c	Durante los últimos 12 meses, ¿cuán seguido has participado en un grupo que ha comenzado una pelea con otro grupo?
P23_b	¿Cuántas veces en los últimos 12 meses te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P82_d	¿Cómo crees que estaría tu mamá si descubriera que fumas marihuana
P23_c	¿Cuántas veces en los últimos 30 días te has emborrachado o intoxicado tomando alcohol, por ejemplo: tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió?
P20_c	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Peleas con golpes, empujones o patadas
P75	¿Con qué personas vives actualmente?
P20_d	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Tener relaciones sexuales sin condón
P91_b	Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio
P90_d	Durante los últimos 12 meses, Has Comenzado una pelea solo con otro/a compañero/a
P8_a	¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez?
P91_c	Has estado en un grupo que ha sido atacado por otro grupo

Consumo Otras Drogas: Variables Excluidas del Estudio (Semejantes a las variables excluidas de Consumo de Cocaína o Pasta Base**)**

P88	En general, ¿consideras que en tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos dentro del colegio?
P77	¿Has conversado con tu padre, madre o apoderado/a acerca de las consecuencias del consumo de drogas?

P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven
P6_a	¿Cuál crees tú que es el riesgo que corre una persona que fuma cigarrillos de vez en cuando?
P6_d	¿Cuál crees tú que es el riesgo que corre una persona que toma bebidas alcohólicas de vez en cuando?
P81_a	¿Cómo describirías el hábito que tiene tu padre respecto al alcohol (vino, cerveza, licor)?
P89	¿consideras que en los alrededores de tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos en las afueras o cercanías del colegio?
P83	Durante este año, ¿Te ha tocado asistir o participar en el colegio en actividades específicamente destinadas a prevenir el consumo de drogas, como por ejemplo charlas o talleres?
P1	Sexo
P87	En general, ¿consideras que en tu colegio hay estudiantes que traen, toman o comparten alcohol dentro del colegio?
P15	¿Has probado alcohol alguna vez en la vida (cerveza, vinos o tragos fuertes)?
P5	En general, ¿cuánto crees que tu padre, madre, apoderado o apoderada (o alguno de ellos) conocen a tus amigos y amigas más cercanos/as?
P74_a	¿Qué educación alcanzo tu padre?
P74_b	¿Qué educación alcanzo tu madre?
P74_c	¿Qué educación alcanzo tu apoderado/a?
P106	¿De cuánto dinero al mes dispones generalmente para tus gastos? Haz un cálculo mensual
P110	Pensando en los últimos 7 días, ¿cuántos días hiciste ejercicio o actividad física, fuera del horario de clases, durante al menos 20 minutos y que te haya hecho transpirar o respirar fuertemente?

ANEXO III: Selección de Variables para técnicas de predicción

IMPORTANCIA DE VARIABLES GRADIENT BOOSTING	IMPORTANCIA INICIAL DE VARIABLES	MEJORES VARIABLES ÁRBOLES	MEJORES VARIABLES REGRESION
P95_a	**P10**	**P7**	**P95_a**
P9	**P9**	**P95_a**	**P99**
P7	**P7**	**P27**	**P96**
P10	**P8_a**	**P11**	**P84**
P8_a	**P27**	**P23_a**	**P79**
P27	**P95_a**	**P99**	**P98**
P99	**P23_a**	**P84**	**P23_a**
P79	**P22**	**P96**	**P97**
P22	**P99**	**P79**	**P10**
P19	**P18**	**P82_d**	**P95_e**
P96	**P26**	**P82_c**	**P85**
P84	**P8_b**	**P85**	**P27**
P85	**P23_b**	**P97**	**P80**
P23_a	**P25**	**P86**	**P19**
P80	**P19**	**P9**	**P82_d**
P2	**P20_d**	**P95_c**	**P20_d**
P25	**P16**	**P26**	**P7**
P16	**P96**	**P8_b**	**P9**
P98	**P17**	**P25**	**P81_b**
P109	**P20_a**	**P12**	**P82_c**

ANEXO IV: Diccionario de Datos Corregido

DICCIONARIO DE DATOS En el siguiente cuadro se identifican el conjunto de variables de estudio con las categorías corregidas.		
Aspecto	Nombre	Descripción
	P1	Sexo
	P2	Edad
Monitoreo Parental	P3	Después de que sales del colegio o durante los fines de semana, ¿cuántas veces ocurre que tu madre, padre, apoderada o apoderado no saben dónde estás? Ya sea por un período de una hora o más. ① Nunca o casi nunca saben dónde estoy ② A veces no saben ③ Siempre o casi siempre saben dónde estoy
	P4	¿Cuán atentos están tu padre, madre, apoderado o apoderada (o alguno de ellos) respecto de lo que haces en el colegio? ① Mucho o Bastante ② Poco ③ Nada
	P5	5. En general, ¿cuánto crees que tu padre, madre, apoderado o apoderada (o alguno de ellos) conocen a tus amigos y amigas más cercanos/as? ① Bastante ② Más o menos ③ Poco
Percepción de riesgo tabaco y alcohol	P6a	¿Cuál crees tú que es el riesgo que corre una persona que hace alguna de estas cosas? Fumar cigarrillos de vez en cuando (ocasionalmente): Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6b	Fumar cigarrillos frecuentemente: Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6c.	Fumar una o más cajetillas de cigarros al día: Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6d.	Tomar bebidas alcohólicas de vez en cuando (ocasionalmente): Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6e.	Tomar alcohol frecuentemente: Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6f.	Emborracharse con alcohol: Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
	P6g.	Tomar uno o dos tragos de alcohol todos o casi todos los días: Riesgo leve o ninguno ① Riesgo moderado ② Riesgo grande ③ No sé ⑨
Las siguientes preguntas son acerca de tu madre y padre	P73	¿Quién es el jefe de tu hogar? Jefe de hogar se define como la persona, hombre o mujer, reconocida como tal por los integrantes del hogar: ① Padre ② Madre ③ Abuela o Abuelo ④ Otro
	P74_a	¿Qué educación alcanzaron tu padre? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta o Universitaria incompleta ⑤ Técnica superior completa ⑥ Universitaria completa ⑦ No sé o N/A ⑨
	P74_b	¿Qué educación alcanzo tu madre? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta o Universitaria incompleta ⑤ Técnica superior completa ⑥ Universitaria completa ⑦ No sé o N/A ⑨
	P74_c	¿Qué educación alcanzaron el jefe/a de hogar? Básica incompleta ① Básica completa ② Media incompleta ③ Media completa ④ Técnica superior incompleta o Universitaria incompleta ⑤ Técnica superior completa ⑥ Universitaria completa ⑦ No sé o N/A ⑨
	P75	¿Con qué personas vives actualmente? ① Padre y madre ② Padre y su pareja ③ Madre y su pareja ④ Sólo con el padre ⑤ Sólo con la madre ⑥ Sólo con Hermana(s) o hermano(s) ⑦ Sólo con Abuelo(s) o Abuela(s) ⑧ Otro adulto responsable

Las siguientes preguntas tienen relación con las personas con quienes vives	P76	¿Quién es tu apoderado/apoderada? Apoderado/apoderada es quien se responsabiliza por ti ante las autoridades del colegio ① Padre ② Madre ③ Abuela o Abuelo u Otro
	P77	¿Has conversado con tu padre, madre o apoderado/a acerca de las consecuencias del consumo de drogas? ① Sí ② No
	P78	¿Crees tú que tu padre, madre o apoderado/a sabe que has probado o consumido alguna droga? (no consideres alcohol, cigarrillos o tranquilizantes) ① Si ② No ③ Nunca he probado drogas
	P79	Pensando en tu padre, madre o apoderado/a, ¿crees que hayan consumido alguna droga cuando joven? (no consideres alcohol, cigarrillos o tranquilizantes) ① Sí ② No
	P80	Hasta donde tú conoces ¿alguno de tus hermanos o hermanas consume alguna droga ilícita (ilegal)? ① Estoy seguro que no lo ha(n) hecho ② Creo que no lo ha(n) hecho ③ Creo que lo hace(n) ④ Estoy seguro que lo hace(n) ⑤ No tengo hermanos o hermanas
	P81_a	¿Cómo describirías el hábito que tiene tu padre respecto al alcohol (vino, cerveza, licor)? Nunca toma alcohol ① Solo en ocasiones especiales ② Solo en fines de semana, pero nunca en días de semana ③ Toma alcohol diariamente, uno o dos tragos ④ Toma alcohol diariamente, más de dos tragos ⑤ No aplica, no tiene padre o madre vivo, no lo ve nunca ⑨
	P81_b	¿Cómo describirías el hábito que tiene tu madre respecto al alcohol (vino, cerveza, licor)? Nunca toma alcohol ① Solo en ocasiones especiales ② Solo en fines de semana, pero nunca en días de semana ③ Toma alcohol diariamente, uno o dos tragos ④ Toma alcohol diariamente, más de dos tragos ⑤ No aplica, no tiene padre o madre vivo, no lo ve nunca ⑨
	P82_a	¿Cómo crees que estaría tu papá y tu mamá en estas situaciones? Si tu papá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) o Bastante molesto(a) ① Algo molesto(a) o Poco molesto(a) ② Indiferente ③ No aplica ⑥
	P82_b	Si tu mamá te sorprende llegando a casa con unos tragos de más: Extremadamente molesto(a) o Bastante molesto(a) ① Algo molesto(a) o Poco molesto(a) ② Indiferente ③ No aplica ⑥
	P82_c	Si tu papá descubriera que fumas marihuana: Extremadamente molesto(a) o Bastante molesto(a) ① Algo molesto(a) o Poco molesto(a) ② Indiferente ③ No aplica ⑥
	P82_d	Si tu mamá descubriera que fumas marihuana: Extremadamente molesto(a) o Bastante molesto(a) ① Algo molesto(a) o Poco molesto(a) ② Indiferente ③ No aplica ⑥
Las siguientes preguntas tienen relación con cómo te sientes en el colegio en el que estás actualmente	P83	Durante este año, ¿Te ha tocado asistir o participar en el colegio en actividades específicamente destinadas a prevenir el consumo de drogas, como por ejemplo charlas o talleres? ① No ② Sí, una vez ③ Sí, más de una vez
	P84	Durante este año, ¿has hecho la cimarra o la chancha? Digamos no fuiste al colegio una parte importante de la jornada o en toda la jornada ① Nunca o Casi nunca ② Pocas veces ③ Varias o Muchas veces
	P85	¿Cuál es el promedio de notas con el que terminaste el año pasado? Descríbelo en estos rangos ① Menos de 4,5 ② Entre 4,5 y 4,9 ③ Entre 5,0 y 5,4 ④ Entre 5,5 y 5,9 ⑤ Entre 6,0 y 6,4 ⑥ Entre 6,5 y 7,0
	P86	¿Cuántos cursos has repetido en tu vida escolar? ① Ninguno ② Uno ③ Dos o más
	P87	En general, ¿consideras que en tu colegio hay estudiantes que traen, toman o comparten alcohol dentro del colegio? ① Sí ② No

	P88	En general, ¿consideras que en tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos dentro del colegio? ① Sí ② No
	P89	¿Y consideras que en los alrededores de tu colegio hay drogas, es decir, algunos estudiantes traen, prueban o se pasan droga entre ellos en las afueras o cercanías del colegio? ① Sí ② No
	P90_a	Durante los últimos 12 meses, ¿cuán seguido has hecho alguna de las siguientes cosas en el colegio? Participado en un grupo que molesta a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P90_b	Participado en un grupo que ha agredido físicamente a un compañero/a que está solo/a : ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P90_c	Participado en un grupo que ha comenzado una pelea con otro grupo : ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P90_d	Comenzado una pelea solo con otro/a compañero/a : ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P90_e	Has robado algo a alguien en el colegio: ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P91_a	Durante los últimos 12 meses, ¿cuán seguido te ha sucedido alguna de las siguientes cosas en el colegio Has sido molestado/a estando solo/sola, por un grupo del colegio ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P91_b	Has sido físicamente agredido/a estando solo/sola, por un grupo del colegio. ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P91_c	Has estado en un grupo que ha sido atacado por otro grupo ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P91_d	Alguien solo/sola ha iniciado una pelea contigo ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P91_e	Te han robado algo en el colegio ① Nunca ② Una vez ③Dos veces ④ 3 o más veces
	P92	¿Cuán probable es que pases de curso este año? ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable o Imposible
	P93	¿Cuán probable es que termines cuarto medio? ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable o Imposible
	P94	¿Cuán probable es que sigas estudiando después del colegio? (en la Universidad, Instituto Profesional, Centro de Formación técnica u otro) ① Es seguro ② Muy probable ③ Más o menos probable ④ Poco probable o Imposible
Las siguientes preguntas son acerca de tus amistades y la relación que mantienes con ellos y ella	P95_a	Durante los últimos 12 meses, ¿cuán seguido te ha tocado estar cerca de alguien o alrededor de un grupo que ha estado consumiendo alguna de estas sustancias con el evidente propósito de volarse, drogarse o embriagarse? 95a. Marihuana: Nunca ① Casi nunca o de vez en cuando ② Bastante seguido o muy seguido ③
	P95_b	95b. Cocaína: Nunca ① Casi nunca o de vez en cuando ② Bastante seguido o muy seguido ③
	P95_c	95c. Pasta base: Nunca ① Casi nunca o de vez en cuando ② Bastante seguido o muy seguido ③
	P95_d	95d. Inhalables: Nunca ① Casi nunca o de vez en cuando ② Bastante seguido o muy seguido ③
	P95_e	95e. Alcohol : Nunca ① Casi nunca o de vez en cuando ② Bastante seguido o muy seguido ③
	P96	Si en tu grupo de amigas y amigos cercanos supieran que fumas marihuana ¿tú crees que: ① Te harían algún reproche o te dirían algo para que no lo hicieras ② Algunos te harían reproches y otro no ③ No te harían ningún problema ④ Te alentarían a que lo siguieras haciendo
	P97	¿Si en tu grupo de amigas y amigos más cercanos supieran que has probado una droga distinta a la marihuana como cocaína, pasta base, éxtasis, ácidos o cosas parecidas, tú crees que: ① Te harían algún reproche o te dirían algo

		para que no lo hicieras ② Algunos te harían reproches y otro no ③ No te harían ningún problema ④ Te alentarían a que lo siguieras haciendo
	P98	¿Cuántos de tus amigas y amigos toman regularmente alcohol? Digamos, todos los fines de semana o más seguido ① Ninguno ② Menos de la mitad ③ Como la mitad ④ Más de la mitad ⑤ Todos o casi todos
	P99	¿Cuántos de tus amigas y amigos fuman regularmente marihuana? Digamos, todos los fines de semana o más seguido ① Ninguno ② Menos de la mitad ③ Como la mitad ④ Más de la mitad ⑤ Todos o casi todos
Las siguientes preguntas son acerca del consumo de tabaco	P7	¿Has fumado cigarrillos alguna vez en la vida? ① Sí ② No
	P8_a	8a. ¿Qué edad tenías cuando comenzaste a fumar cigarrillos por primera vez? No consideres si tus padres o algún adulto te dieron a probar siendo niño. Edad en años: Marca"0" en la hoja de respuestas si no has fumado. La variable de convierte en categórica: **Se consideran los siguientes rangos** (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
	P8_b	¿Qué edad tenías cuando comenzaste a fumar cigarrillos todos o casi todos los días? Edad en años: Marca"0" en la hoja de respuestas si no has fumado todos o casi todos los días La variable de convierte en categórica: **Se consideran los siguientes rangos** (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
	P9	¿Cuándo fue la primera vez que fumaste cigarrillos? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P10	¿Cuándo fue la última vez que fumaste un cigarrillo? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P11	¿Cuántos días has fumado cigarrillos en los últimos 30 días? N° de días: Marca "0" en la hoja de respuestas si no has fumado en los últimos 30 días
	P12	Considerando sólo los días que fumaste en el último mes. ¿Aproximadamente, cuántos cigarrillos fumaste al día? N° de cigarrillos: Marca"0" en la hoja de respuestas si no has fumado en los últimos 30 días
Las siguientes preguntas son acerca del consumo de bebidas alcohólicas	P15	¿Has probado alcohol alguna vez en la vida (cerveza, vinos o tragos fuertes)? ① Sí ② No
	P16	¿Qué edad tenías cuando probaste por primera vez alguna bebida alcohólica? No consideres si tu padre, madre o una persona adulta te dieron a probar siendo niño/niña. Edad en años: Marca"0"en la hoja de respuestas si no has probado La variable de convierte en categórica: **Se consideran los siguientes rangos** (5 a 12)= ① , (13 a 17)= ②, (18 a 21)= ③
	P17	¿Cuándo fue la primera vez que probaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P18	¿Cuándo fue la última vez que tomaste alcohol? ① Durante los últimos 30 días ② Hace más de un mes, pero menos de un año ③ Hace más de un año ④ Nunca he probado
	P19	¿Cuán difícil te sería comprar alguna bebida alcohólica, si quisieras hacerlo? ① Me sería fácil o muy fácil ② Me sería difícil o muy difícil ③ No podría comprarla ⑨ No sé
	P20_a	Piensa en los últimos 12 meses, ¿Te han ocurrido alguna de las siguientes cosas producto de tu consumo de ALCOHOL? Marca"0" en la hoja de respuestas si no has consumido alcohol en los últimos 12 meses 20a. Amigos, amigas o familiares te han sugerido o mencionado que disminuyas el consumo de alcohol ① Sí ② No
	P20_b	20b. Consumir alcohol estando solo o sola ① Sí ② No
	P20_c	20c. Peleas con golpes, empujones o patadas ① Sí ② No

	P20_d	20d. Tener relaciones sexuales sin condón ① Sí ② No
	P21	Piensa en los últimos 30 días ¿Cuántos días has consumido algún tipo de alcohol? N° de días: Marca"0" en la hoja de respuestas si no has consumido
	P22	¿Cuántos tragos sueles tomar en un día típico de consumo de alcohol? Guíate por la siguiente tabla para saber cuántos tragos consumes 1 trago (una botella o lata individual de cerveza (333 cc.); Un vaso de vino (140 cc.); Un trago de licor (40 cc. de pisco, ron, vodka o whisky, sólo o combinado) 1 trago y medio (medio litro de cerveza) 3 tragos (un litro de cerveza) 6 tragos (una botella de vino (750 cc.) 8 tragos (una caja de vino (1 litro) 18 tragos (una botella de licor (750 cc.) ① 1 a 2 tragos ② 3 a 4 tragos ③ 5 o mas
	P23_a	¿Cuántas veces te has emborrachado o intoxicado tomando alcohol, por ejemplo tambalearse al caminar, no ser capaz de hablar bien, vomitar o no recordar qué ocurrió? Nunca⓪ 1-2 veces① más de 3 veces② 23a. En tu vida
	P23_b	Nunca⓪ 1-2 veces① más de 3 veces② 23b. En los últimos 12 meses
	P23_c	Nunca⓪ una o más de una vez① 23c. En los últimos 30 días
	P25	Pensando en el último día que consumiste alcohol ¿cuál de las siguientes bebidas alcohólicas fue la que más tomaste ese día? Marca aquella bebida (o tipo de alcohol) que más consumiste ① Cerveza ② Vino ③ Espumantes (champaña, Manquehuito, vinos con sabores u otros) ④ Tragos fuertes solos o combinados (piscola, roncola, vodka naranja u otro) ⑤ No consumo
	P26	Qué tan borracho/borracha consideras que estuviste el último día que consumiste alcohol (99) No ha consumido ① Poco o casi nada ② Medio tomado ③ Bien tomado (88) No responde
	P27	Pensando en una salida de sábado por la noche ¿Cuántos vasos de cerveza, vino o licor llegas a tomar? ① Nunca he tomado alcohol ② Ninguno ③ Uno o menos de uno ④ Entre 2 y 5 ⑤ 5 o mas
sección aborda información adicional sobre ti.	P105	¿Con qué religión te identificas? ① Católica ② Evangélica/Protestante ③ Otra religión ④ Ninguna religión ⑨ No lo sé
	P106	¿De cuánto dinero al mes dispones generalmente para tus gastos? Haz un cálculo mensual ⓪ No dispongo de dinero para mis gastos ① Menos de $5.000 ② Entre $5.000 y $10.000 ③ Entre $10.001 y $20.000 ④ Entre $20.001 y $30.000 ⑤ Entre $30.001 y $40.000 ⑥ Entre $40.001 y $60.000 ⑦ Más de $60.000
	P108	¿Trabajas regularmente además de estudiar? ① Sí ② No
	P109	¿Cuál es el estado conyugal actual de tus padres? ① Casados ② Convivientes ③ Separados, anulados, divorciados o no viven juntos ④ Viudo o viuda ⑤ Soltero o soltera ⑥ Otra situación
	P110	Pensando en los últimos 7 días, ¿cuántos días hiciste ejercicio o actividad física, fuera del horario de clases, durante al menos 20 minutos y que te haya hecho transpirar o respirar fuertemente? N° de días: Marca "0" en la hoja de respuestas si no has realizado actividad física
Consumo Drogas	**P35**	¿Has consumido marihuana alguna vez en la vida? ① Sí ⓪ No → Es variable Objetivo
	P47_53	¿Has consumido cocaína o pasta base alguna vez en la vida? ①Sí ⓪No → Es variable Objetivo
	P66_f_g_i_j	¿Has consumido las siguientes sustancias alguna vez en la vida Crack, Éxtasis, Heroína, Alucinógenos sintéticos como LSD, PCP, polvo de ángel, u otros ácidos ① Si ⓪ No → Es variable Objetivo

ANEXO V: Código en SAS base

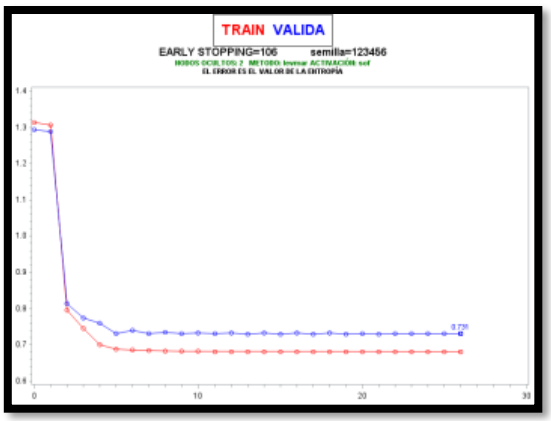
REDES NEURONALES

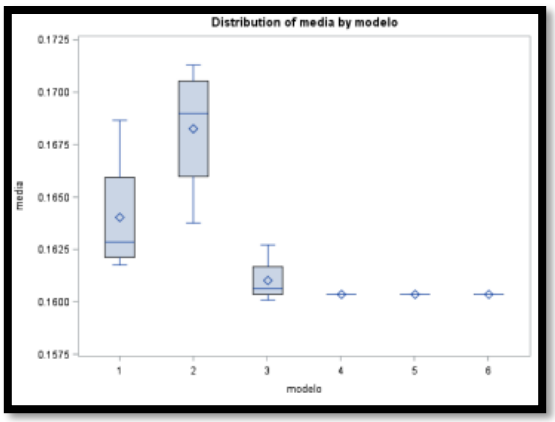
```
/* LA MACRO CRUZADABINARIANEURAL GENERA RESULTADOS POR CLASIFICACIÓN
BINARIA CON RED NEURONAL CON VARIAS SEMILLAS
Autor: J Portela (2019)
DESCRIPCIÓN DE PARÁMETROS:

archivo
vardepen        P35 variable con dos categorías excluyentes
conti    lista de variables continuas en el modelo
categor lista de variables categóricas en el modelo
ngrupos         grupos de validación cruzada
sinicio  semilla inicial de aleatorización
sifinal  semilla final de aleatorización
nodos           número de nodos red
algo            algoritmo
objetivo función objetivo para resumir en archivos y boxplot. Palabras clave:
         tasafallos (habitualmente se utilizará esta)
         porcenVN
         porcenFN
         porcenVP
         porcenFP
         sensi
         especif
         tasaciertos
         precision
*/
%macro
cruzadabinarianeural(archivo=,vardepen=,conti=,categor=,ngrupos=,sinicio=,sfinal=,nodos=,alg
o=,early=150,
acti=tanh,basura=c:\basura.txt,objetivo=tasafallos);
title ' ';
data final;run;
proc printto print=&basura;

/* Bucle semillas */
%do semilla=&sinicio %to &sfinal;
         data dos;set &archivo;u=ranuni(&semilla);
         proc sort data=dos;by u;run;
         data dos (drop=nume);
         retain grupo 1;
         set dos nobs=nume;
         if _n_>grupo*nume/&ngrupos  then grupo=grupo+1;
         run;
         data fantasma;run;
         %do exclu=1 %to &ngrupos;

         data trestr tresval;
set dos;if grupo ne &exclu then output trestr;else output tresval;
         PROC DMDB DATA=trestr dmdbcat=catatres;
         target &vardepen;
         var &conti;
         class &vardepen;
         %if &categor ne %then %do;class &categor &vardepen;%end;
         run;
         proc neural data=trestr dmdbcat=catatres random=789 ;
         input &conti;
         %if &categor ne %then %do;input &categor /level=nominal;%end;
```

```
        target &vardepen  /level=nominal;
        hidden &nodos /acti=&acti;

/* ESPECIFICACIONES  DE LA RED, SE PUEDEN CAMBIAR O AÑADIR COMO
PARÁMETROS */

/*nloptions maxiter=500*/;
netoptions randist=normal ranscale=0.15 random=15459;
/* Si se desea hacer early stopping se pone prelim 0 y se marca como comentario la línea
prelim 15...*/
/*prelim 0 */
prelim 15 preiter=10 pretech=&algo;
train maxiter=&early outest=mlpest technique=&algo;
score data=tresval role=valid out=sal ;
run;
data sal2;set sal;pro=1-%str(p_&vardepen)0;if pro>0.5 then pre11=1; else pre11=0;run;
proc freq data=sal2;tables pre11*&vardepen/out=sal3;run;

data estadisticos (drop=count percent pre11 &vardepen);
        retain vp vn fp fn suma 0;
        set sal3 nobs=nume;
        suma=suma+count;
        if pre11=0 and &vardepen=0  then vn=count;
        if pre11=0 and &vardepen=1  then fn=count;
        if pre11=1 and &vardepen=0  then fp=count;
        if pre11=1 and &vardepen=1  then vp=count;
        if _n_=nume then do;
        porcenVN=vn/suma;
        porcenFN=FN/suma;
        porcenVP=VP/suma;
        porcenFP=FP/suma;
        sensi=vp/(vp+fn);
        especif=vn/(vn+fp);
        tasafallos=1-(vp+vn)/suma;
        tasaciertos=1-tasafallos;
        precision=vp/(vp+fp);
        F_M=2*Sensi*Precision/(Sensi+Precision);
        output;
        end;
        run;
data fantasma;set fantasma estadisticos;run;
        %end;
        proc means data=fantasma sum noprint;var &objetivo;
        output out=sumaresi sum=suma mean=media;
        run;
        data sumaresi;set sumaresi;semilla=&semilla;
        data final (keep=suma media semilla);set final sumaresi;if suma=. then delete;run;
%end;
proc printto ;
proc print data=final;run;
%mend;

/*Lectura de Datos*/
data consumo2; set 'C:\Users\krlos\Documents\datosSas\datosD.sas7bdat';  run;


/* Probando early stopping con el conjunto de datos D*/
%redneuronalbinaria(archivo=consumo2,listclass=P95_a  P7 P99 P9 P27 P23_a P96 P79 P10
P84 P85 P25 P19 P8_a P8_b P22 P98 P97 P82_c P82_d  P80 P26 P16 P20_d,
listconti=P2,
```

```
vardep=P35,porcen=0.80,semilla=123456,ocultos=2,meto=levmar,acti=sof);

/*Ejecución Modelo de Red Neuronal Utilizando el Conjunto de datos "D" con 2 Nodos,
función de activación Softmax,
algoritmo opt levmar */

%cruzadabinarianeural(archivo=consumo2,vardepen=P35,
conti=P2,
categor=P95_a P7 P99 P9 P27 P23_a P96 P79 P10 P84 P85 P25 P19 P8_a P8_b P22 P98
P97 P82_c P82_d P80 P26 P16 P20_d,
ngrupos=4,sinicio=12345,sfinal=12348,nodos=2,algo=levmar,acti=sof);
data final3;set final;modelo=4;

/*COMPARACIÓN Y REPRESENTACIÓN GRÁFICA DE MODELOS REDES NEURONALES*/
data union;set final1 final2 final3 final4 final5 final6;/* final7 final8;*/
proc boxplot data=union;plot media*modelo;run;
```

Gráfica de Early Stoping y comparación de modelos





ANEXO VI: Código en R-Studio

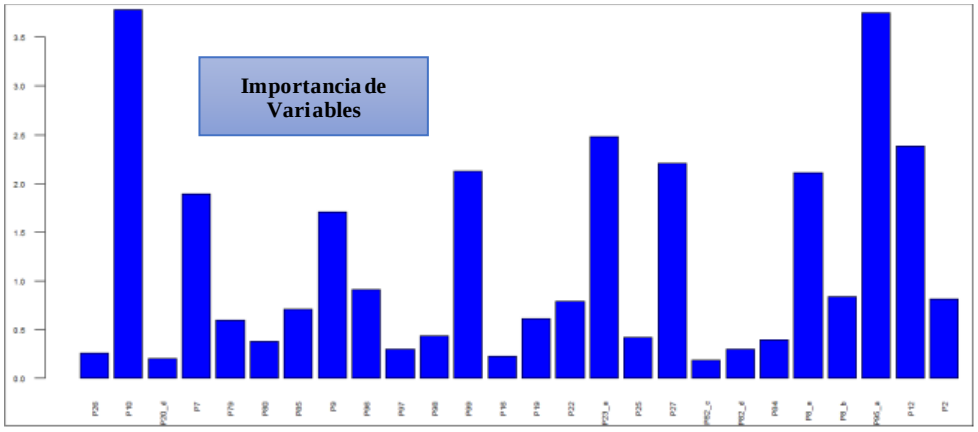
RANDOM FOREST

```
/*# FUNCIÓN PARA RANDOM FOREST (Bagging) CON R*/
cruzadarfbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,nodesize=20,
  mtry=2,ntree=50,replace=TRUE)
 {
  /*# Preparación del archivo
  # pasar las categóricas a dummies*/
  if (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis, listclass, sep = ".")
  } else {
   databis<-data[,c(vardep,listconti)]
  }

  /*#estandarizar las variables continuas
  # Calculo medias y dtipica de datos y estandarizo (solo las continuas)*/
  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)

  /*# Estandarización de variables continuas y se une la variables categoricas*/
```

```
 datacon<-scale(databis[,listconti],  center = means, scale = sds)
 numerocont<-which(colnames(databis)%in%listconti)
 databis<-cbind(datacon,databis[,-numerocont,drop=FALSE  ])
 databis[,vardep]<-as.factor(databis[,vardep])
 formu<-formula(paste("factor(",vardep,")~.",sep=""))
 /*# Preparo  caret  */
 set.seed(sinicio)
 control<-trainControl(method  = "repeatedcv",number=grupos,repeats=repe,
  savePredictions  = "all",classProbs=TRUE)

/*# Aplico caret y construyo modelo*/
 rfgrid  <-expand.grid(mtry=mtry)
 rf<- train(formu,data=databis,
  method="rf",trControl=control,
  tuneGrid=rfgrid,nodesize=nodesize,replace=replace,
  ntree=ntree)

 print(rf$results)
 preditest<-rf$pred
 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold  <- sapply(preditest$prueba,  "[", 1)
 preditest$Rep  <- sapply(preditest$prueba,  "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y)  {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }
 /*# Aplicamos función sobre cada Repetición*/
 medias<-preditest  %>%
  group_by(Rep)  %>%
  summarize(tasa=1-tasafallos(pred,obs))
 /*# CalculamoS  AUC  por cada Repetición  de cv
 # Definimnos  función*/
 auc<-function(x,y)  {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }
 /*# Aplicamos función sobre cada Repetición*/
 mediasbis<-preditest  %>%
  group_by(Rep)  %>%
  summarize(auc=auc(obs,Yes))
 /*# Unimos  la info  de auc y de tasafallos*/
 medias$auc<-mediasbis$auc
 return(medias)
}


/*# Ejecución de la función descrita anteriormente de Random Forest
medias9<-cruzadarfbin(data=datosD,
                      vardep="P35",listconti=c("P2","P12"),

listclass=c("P95_a","P9","P7","P99","P27","P23_a","P96","P79","P10","P84","P85","P25","P19","
P8_a","P8_b","P22","P98","P97","P82_c","P82_d","P80","P26","P16","P20_d"),
                      grupos=4,sinicio=1234,repe=5,nodesize=15,
                      mtry=17,ntree=200,replace=TRUE)

medias9$modelo="rf5"
```

```
summary(medias9)

/*# Ejecución del mejor modelo, utilizando la librería random forest y caret para
validación cruzada,
# se comprueba también el tamaños muestral*/
library(randomForest)
set.seed(12345)
rfbis<-randomForest(factor(P35)~.,data=datosD,
mtry=17,ntree=1000,sampsize=300,nodesize=15,replace=TRUE)
plot(rfbis$err.rate[,1])

for (muestra in seq(100,200,300))
{
 /* # controlamos la semilla*/
  set.seed(12345)
  rfbis<-randomForest(factor(P35)~.,data=datosD,
mtry=17,ntree=1000,sampsize=muestra,nodesize=15,replace=TRUE)
  plot(rfbis$err.rate[,1],main=muestra,ylim=c(0.25,0.5))
}
/*# Ahora se comprueba con validación cruzada con caret*/
rfgrid<-expand.grid(mtry=c(17))
rf<- train(factor(P35)~.,data=datosD,
          method="rf",trControl=control,tuneGrid=rfgrid,
          linout = FALSE,ntree=1000,sampsize=100,nodesize=15,replace=TRUE)
rf
summary(rf)
final<-rf$finalModel
/*# Importancia de Variables*/
tabla<-as.data.frame(importance(final))
tabla
barplot(tabla$MeanDecreaseGini,names.arg=rownames(tabla))
```

Importancia de variables



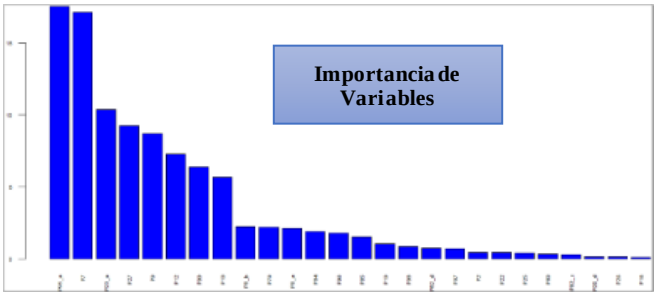
GRADIENT BOOSTING

```
/*# TUNEADO DE MODELOS UTILIZANDO LA TÉCNICA GRADIENT BOOSTING

# Caret permite tunear los siguientes parámetros básicos:
          # shrinkage:(parámetro v de regularización, mide la velocidad de ajuste, a menor v,
          más lento y necesita más iteraciones, pero es más fino en el ajuste)
          # n.minobsinnode: tamaño máximo de nodos finales (el principal parámetro que mide
          la complejidad)
          # n.trees= el número de iteraciones (árboles)
          # interaction.depth: (2 para árboles binarios) */
```

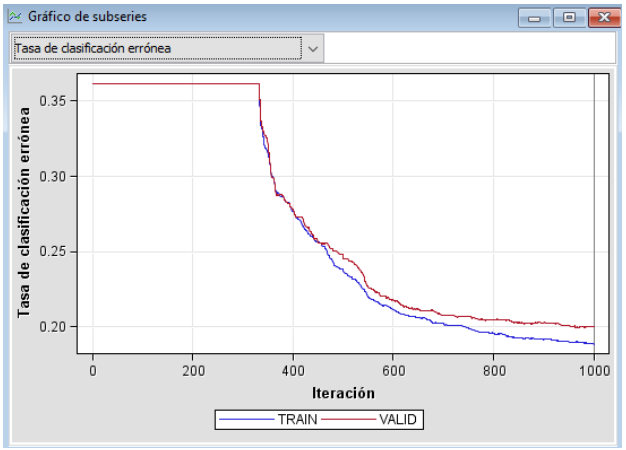
```
library(caret)
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.1,0.05,0.03,0.01,0.001),
                     n.minobsinnode=c(5,10,20),
                     n.trees=c(100,500,1000,5000),
                     interaction.depth=c(2))

control<-trainControl(method  = "cv",number=4,savePredictions  = "all",
                      classProbs=TRUE)
gbm<- train(factor(P35)~.,data=datosD,
            method="gbm",trControl=control,tuneGrid=gbmgrid,
            distribution="bernoulli",  bag.fraction=1,verbose=FALSE)
gbm
plot(gbm)

/*# FUNCIÓN PARA GRADIENT BOOSTING CON R*/
cruzadagbmbin<-
 function(data=data,vardep="vardep",
  listconti="listconti",listclass="listclass",
  grupos=4,sinicio=1234,repe=5,
  n.minobsinnode=20,shrinkage=0.1,n.trees=100,interaction.depth=2)
 {
  /*# Preparación  del archivo
  # pasar las categóricas a dummies*/
  if  (listclass!=c(""))
  {
   databis<-data[,c(vardep,listconti,listclass)]
   databis<- dummy.data.frame(databis,  listclass, sep = ".")
  }  else   {
   databis<-data[,c(vardep,listconti)]
  }
 /* # estandarizar  las variables  continuas
  # Calculo medias y dtipica de datos y estandarizo  (solo las continuas)*/
  means <-apply(databis[,listconti],2,mean)
  sds<-sapply(databis[,listconti],sd)

 /* # Estandarizo  solo las continuas y uno con las categoricas*/
  datacon<-scale(databis[,listconti],  center = means, scale = sds)
  numerocont<-which(colnames(databis)%in%listconti)
  databis<-cbind(datacon,databis[,-numerocont,drop=FALSE  ])

  databis[,vardep]<-as.factor(databis[,vardep])
  formu<-formula(paste("factor(",vardep,")~.",sep=""))

 /* # Preparo  caret */
  set.seed(sinicio)
  control<-trainControl(method  = "repeatedcv",number=grupos,repeats=repe,
   savePredictions  = "all",classProbs=TRUE)

 /* # Aplico caret y construyo modelo*/
   gbmgrid <-expand.grid(n.minobsinnode=n.minobsinnode,
    shrinkage=shrinkage,n.trees=n.trees,
    interaction.depth=interaction.depth)

  gbm<- train(formu,data=databis,
   method="gbm",trControl=control,
   tuneGrid=gbmgrid,distribution="bernoulli",verbose=FALSE)

  print(gbm$results)
```

```
 preditest<-gbm$pred

 preditest$prueba<-strsplit(preditest$Resample,"[.]")
 preditest$Fold  <- sapply(preditest$prueba,  "[", 1)
 preditest$Rep  <- sapply(preditest$prueba,  "[", 2)
 preditest$prueba<-NULL

 tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
 }

/* # Aplicamos función sobre cada Repetición*/
 medias<-preditest  %>%
  group_by(Rep)  %>%
  summarize(tasa=1-tasafallos(pred,obs))

/* # CalculamoS  AUC  por cada Repetición  de cv
 # Definimnos  función*/
 auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

 /*# Aplicamos función sobre cada Repetición*/
 mediasbis<-preditest  %>%
  group_by(Rep)  %>%
  summarize(auc=auc(obs,Yes))

/* # Unimos  la info  de auc y de tasafallos*/
 medias$auc<-mediasbis$auc

 return(medias)
}

/*# Ejecución de la función descrita anteriormente de Gradient Boosting
medias11<-cruzadagbmbin(data=datosD,vardep="P35",listconti=c("P2","P12"),

listclass=c("P95_a","P9","P7","P99","P27","P23_a","P96","P79","P10","P84","P85","P25","P19","
P8_a","P8_b","P22","P98","P97","P82_c","P82_d","P80","P26","P16","P20_d"),grupos=4,sinicio
=12579,repe=5,  N.minobsinnode=20,shrinkage= 0.1,n.trees=500,interaction.depth=2)

medias11$modelo="gbm2"
summary(medias11)

/*# Representación Diagrama de Cajas modelos Gradient Boosting*/
union4<-
rbind(medias10,medias11,medias12,medias13,medias14,medias15,medias16,medias17,media
s18,medias19,medias20,medias21,medias22,medias23,medias24,medias25,medias26,medias
27)
par(cex.axis=0.5, cex.lab=0.1)
boxplot(data=union4,tasa~modelo,main="TASA  FALLOS",col="blue")
boxplot(data=union4,auc~modelo,main="AUC",col="blue")

/*# CONSTRUCCIÓN DEL MEJOR MODELOS GRADIENT BOOSTING*/
library(caret)
set.seed(12345)
```

```
gbmgrid<-
expand.grid(shrinkage=c(0.1),n.minobsinnode=c(20),n.trees=c(500),inter
action.depth=c(2))
control<-trainControl(method = "cv",number=4,savePredictions = "all",
                      classProbs=TRUE)
gbm<- train(factor(P35)~.,data=datosD,
            method="gbm",trControl=control,tuneGrid=gbmgrid,
            distribution="bernoulli", bag.fraction=1,verbose=FALSE)
gbm
/*# IMPORTANCIA DE VARIABLES*/
par(cex=1.3)
summary(gbm)
tabla1<-summary(gbm)
par(cex=1.3,las=2.5)
barplot(tabla1$rel.inf,names.arg=row.names(tabla1),col="blue")
```




```
/*# Early Stoping con constante de regularización 0.001 y 1000
iteraciones*/
```



ANEXO VII: Accesos a gráficas, estadísticas y resultados en HTML.

En el siguiente link, se puede acceder a la representación del diagrama completo de árboles de clasificación, así como de las estadísticas de las muestras de estudio y los resultados de los análisis en formato HTML.

https://drive.google.com/open?id=1POEUPMUTjASOXM2WVhDaNXGQThSZku_U