Universidad del Desarrollo
Facultad de Ingeniería

# MODELO DE DESERCIÓN ESTUDIANTIL

POR: Rodrigo Manríquez Pacheco

Capstone proyect presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:
Dr. Cristian Esteban Candia Vallejos, Srta. Melanie Alejandra Oyarzun Wolf

Diciembre 2022
SANTIAGO

## TABLA DE CONTENIDO

## Resumen

A comienzos del 2022 la matrícula total de pregrado en la educación superior llegó a 1.211.797 alumnos en Universidades, Institutos Profesionales (IP) y Centros de formación técnica (CFT). Específicamente los CFT-IP suman más del 43,7% de la matrícula total en Chile, es decir 529.044 alumnos.

La deserción estudiantil en la Educación Técnico Profesional constituye una de las principales problemáticas para las Instituciones y sus alumnos, considerando además que, durante la pandemia, la deserción ha aumentado considerablemente en los principales planteles estudiantiles del País. El 45,1% de los alumnos de la cohorte 2016 llegaron hasta último año (permanencia), mientras que el 54,9% restante desertó de su carrera.

El presente trabajo tiene como objetivo construir modelos predictivos para la detección de la deserción en los IP en Chile, durante 2016 y el 2022. Lo anterior, se realizará mediante un proceso de identificación de las principales variables predictoras, mediante modelos de *Random Forest* y Regresiones Logísticas. El resultado de los modelos arrojo una precisión de predicción de un 84% para el modelo Logit y un 81% para el modelo *Random Forest*, lo cual se contrastó con el análisis de la curva de ROC.

# 1. Introducción

Considerando las complejidades propias de las Instituciones, sus familias, el nivel de educación de los padres y su ingreso familiar, la deserción estudiantil en la Educación Técnico Profesional constituye una de las principales problemáticas para las Instituciones y los alumnos. Por otra parte, durante la pandemia la deserción ha aumentado razonablemente en los principales planteles estudiantiles del País.

El presente trabajo tiene como objetivo construir modelos predictivos para la detección de la deserción en carreras técnico profesionales en Chile, durante 2016 y el 2022. Conectando la partida del estudio con el inicio de la gratuidad en los CFT-IP en Chile.

Lo anterior, se realizará mediante un proceso de identificación de las principales variables predictoras, aplicando técnicas econométricas y de *Machine Learning* para generar modelos que permitan predecir la deserción anual, estudiantil de alumnos de cuarto año, al interior del subsegmento de los Institutos Profesionales. Para ser más precisos, se utilizarán modelos de *Random Forest* y Regresiones Logísticas.

La comparación de los modelos se realizó en una primera instancia mediante el cálculo del F1 Score, mientras que para una segunda instancia se utilizó el análisis de la característica operativa del receptor (ROC) más para entender qué modelo se ajusta mejor a nuestras necesidades.

## 1.1. Historia de la Educación Técnico-Profesional

Uno de los primeros registros que se tiene de los IP-CFT en Chile es la creación de la Real Academia San Luis en 1797, la cual tenía como finalidad educar jóvenes en la doctrina de la Ilustración.

Un segundo registro aparece en siglo XIX, con la creación de la Escuela de Artes y Oficios (1842), y un tercer registro con la Creación del Consejo de la Educación Técnica (1886).

En 1947 se funda la Universidad Técnica del Estado (UTE), la cual fusionó las 4 principales escuelas politécnicas de país, impulsando el progreso de la enseñanza Técnico Profesional a nivel regional.

En 1965, reestructura el Sistema de Educación, por medio de 3 ejes: igualación de oportunidades educativas, cambios educacionales en función del desarrollo (económico, social y político), y un nuevo modelo pedagógico. Los cambios lograron ampliar la cobertura escolar y desarrollar una educación media que tenía 2 modalidades; científico-humanista y técnico-profesional.

A partir de 1965 se comienzan a crear Centros de Formación Técnico Profesional, desplegándose por todas las regiones, contribuyendo al desarrollo del país.

Hoy en Chile existen distintos tipos de Instituciones de Educación Superior las que se encuentran facultadas por Ley para impartir carreras profesionales, técnicas y de carácter militar y/o policial. Son las Universidades, Institutos Profesionales (IP) y Centros de Formación Técnica (CFT) los que imparten carreras. Cada carrera tiene un tiempo de duración determinado y dependiendo del tipo es el grado académico que pueden entregar.

Las Universidades imparten carreras de carácter profesional, así como también carreras técnicas, otorgando diferentes tipos de grados académicos. Por otra parte, los Institutos Profesionales ofrecen carreras de carácter profesional, a diferencia de las Universidades estas no poseen el grado de licenciatura, también imparten carreras técnicas. Finalmente, también existen los Centros de Formación Técnica (CFT) los cuales ofrecen carreras técnicas.

# 2. Trabajo Relacionado

La deserción es un fenómeno que se encuentra en los sistemas educativos (Eckert y Suénaga, 2015; Díaz, 2008). Es considerado un indicador relevante para la movilidad social, reflejando indirectamente la contribución social que realizan las Instituciones de Educación Superior (Grandón y Vargas, 2012).

Se revisó el trabajo realizado por Pyke y Sheridan (1993), quienes mediante regresiones logísticas obtuvieron resultados de permanencia y encontraron variables que influyen positivamente a la retención. También analizamos los modelos de deserción que incorporan árboles de decisión y *Random Forest* (Amat 2020).

Eckert y Suénaga (2015) observaron los principales los elementos que influyen en la deserción utilizando algoritmos de clasificación específicamente árboles de decisión, redes bayesianas.

Los autores Hernandez Gonzalez (et al., 2016), desarrollaron un modelo predictivo que entrega la probabilidad que tienen los alumnos de desertar. Donde también realizaron un estudio comparativo de algoritmos para predecir la deserción.

También se considerarán las investigaciones del Centro de Estudios del Ministerio de Educación, específicamente Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior.

# 3. Hipótesis y Objetivos

## 3.1. Objetivo General

El presente trabajo tiene como objetivo construir modelos predictivos para la detección de la deserción (o su defecto mediante la permanencia) en Chile para alumnos de la cohorte 2016. Conectando la partida del estudio con el inicio de la gratuidad en los CFT-IP en Chile.

## 3.2. Objetivos Específicos

- Identificar la deserción promedio de los planteles del subsegmento IP.
- Medir la deserción Técnico Profesional en carreras de 8 semestres de duración.
- Identificar las principales variables predictoras, aplicando técnicas econométricas y de *Machine Learning* para generar un modelo que permita predecir la deserción anual.
- Desarrollar modelos de árbol de decisión y regresiones logísticas.
- Comparar resultados de los modelos mediante el F1 Score y el análisis de ROC.

## 3.3. Hipótesis de Estudio

La deserción en los IP al 4to año es mayor en hombres que en mujeres.

# 4. Datos y Metodología

## 4.1. Datos

Antes de comenzar con el análisis de la información definiremos a la retención como la acción y efecto de retener[1]. La retención en las Instituciones de Educación Superior son medidas estrictamente año a año y se vinculan directamente con los indicadores Institucionales, donde la retención corresponde a la porción de los alumnos de una cohorte determinada que se matriculan en el segundo año (para el caso de la retención de segundo año). En palabras simples son los alumnos que comenzaron a estudiar una carrera determinada, en una institución en un periodo x y continúan estudiando en el periodo x+1.

Por otra parte, el Mineduc define a la retención como "La capacidad que tiene el sistema educativo para lograr la permanencia de los estudiantes en las aulas garantizando la terminación de ciclos y niveles en los tiempos previstos y asegurando el dominio de las competencias y conocimientos correspondientes".

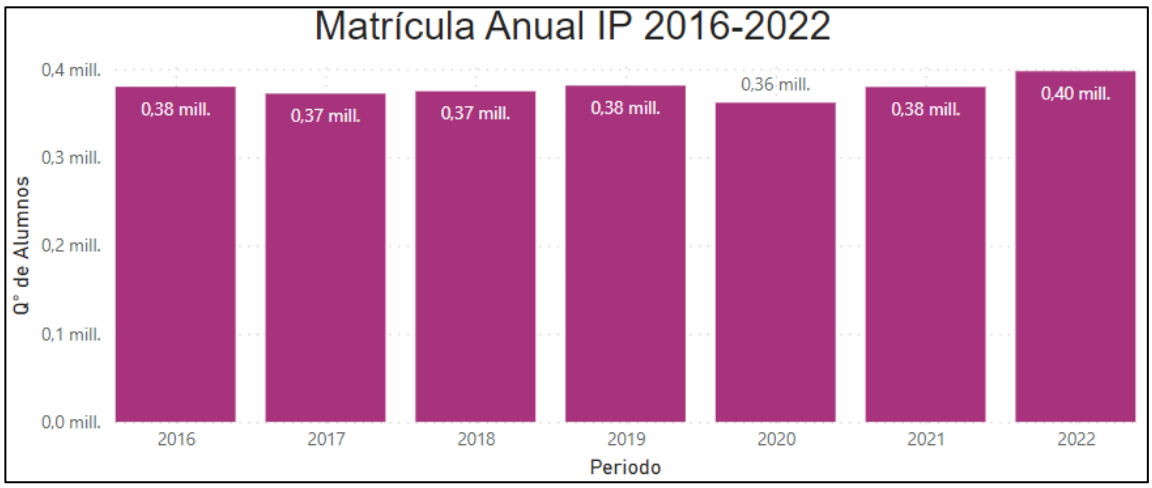
Como se menciona en el capítulo anterior, los IP son Instituciones de Educación Superior (IES) privadas nacidas en el DFL n.º 5 del Ministerio de Educación en 1981 en la reforma de la Educación Superior y están capacitados para entregar títulos profesionales o técnicos de nivel superior.

[1] Diccionario de la Lengua Española, - Edición del Tricentenario. Disponible en: https://dle.rae.es/retenci%C3%B3n?m=form (Accessed: January 20, 2023).

El único requisito académico para ingresar a un Instituto Profesional es que el alumno tenga su licencia de enseñanza media.

Respecto del centro de nuestra investigación, las matrículas en el subsegmento de los Institutos Profesionales en términos anuales en el año 2022 alcanzaron los 397.504 alumnos, lo cual representa un aumento de un 4,7% respecto del periodo anterior (379.674).

A continuación, grafico N°1 con la evolución de la matrícula anual:




Fuente: elaboración propia en base a datos de matrícula de alumnos en chile.

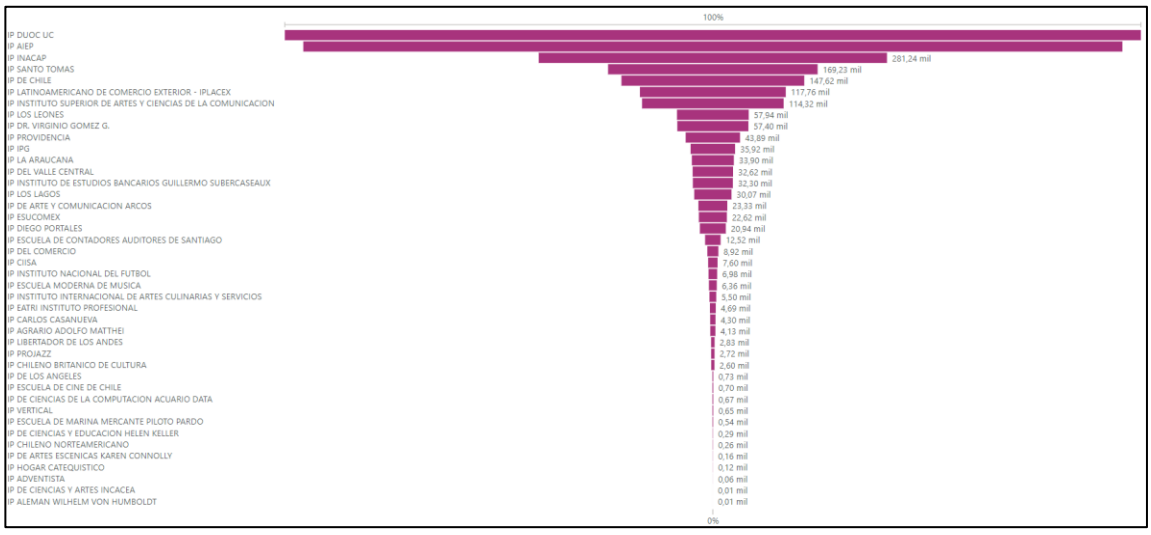
En términos relativos los alumnos pertenecientes a este subsegmento se han mantenido en línea entre los años 2016 y 2022. El periodo con menor cantidad de matrículas fue el periodo 2020 con 361.862, lo que tiene directa relación con el comienzo de la crisis

sanitaria que vivimos, la cual tiene especial énfasis en las carreras técnico profesionales, ya que muchas de sus carreras poseen un componente practico importante al interior de sus mallas académicas. Si comparamos la cantidad de alumnos del año 2016 con los del 2022 podemos apreciar un crecimiento de un 4,7% lo que reafirma lo mencionado anteriormente respecto de la baja variabilidad de los alumnos matriculados.

## 4.2. Instituciones de Educación Superior del subsegmento IP

En el 2022 existen 32 IES que imparten sus carreras en Chile las que agrupan a más de 397 mil alumnos. Estas Instituciones imparten carreras técnicas, pero además pueden entregar formación profesional, aunque no otorgan grados académicos (licenciatura, bachiller o posgrados).

A continuación, grafico N°2 con el número de matriculados históricos por IES:




Fuente: elaboración propia en base a datos de matrícula de alumnos en Chile.

Para el año 2022, IP Duoc UC es el Instituto Profesional con mayor cantidad de alumnos superando los 97 mil alumnos matriculados, seguido de IP AIEP el cual posee más de 94 mil alumnos. En tercer lugar, encontramos a IP Inacap con casi 47 mil alumnos. Las 3 IES mencionadas anteriormente concentran 238.794 alumnos matriculados lo que equivale a un 60,01% del total del subsegmento IP. Dado lo anterior podemos entender que existe un alto grado de concentración en las matrículas en los IP.

Lo anterior en parte se debe que estas 3 IES poseen Sedes en todo Chile y específicamente Duoc UC se concentra mayoritariamente en la Región Metropolitana. Respecto de esta Institución podemos mencionar que posee el nivel de acreditación más alto de todo el subsegmento IP (7 años), inclusive mayor que muchas Universidades y CFT´S.

Respecto del total de alumnos matriculados en el periodo de estudio, Duoc UC es la IES con mayor cantidad de alumnos matriculados con 691.256 alumnos, lo cual representa un 26,11% del total de matriculados. En segundo lugar, se encuentra IP AIEP con 661.412 y finalmente IP Inacap con 281.236. En total estas 3 IES concentran un 61,73% del total de la matricula entre los años 2016-2022, observándose nuevamente un alto grado de concentración de las IES que comparten el subsegmento de los IP.

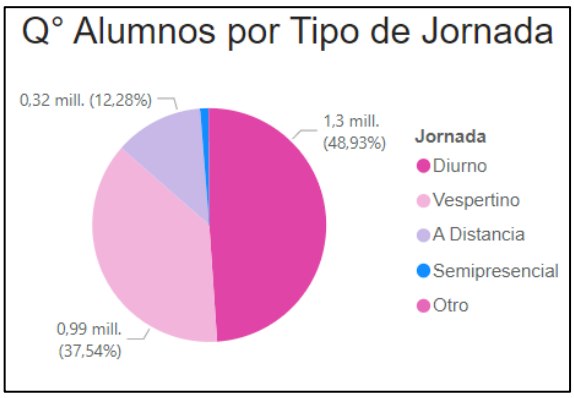
## 4.3. Tipo de Jornada en los Estudios

Si bien las carreras poseen distintos programas de estudio (duración, mallas académicas, objetivos y otras particularidades), también poseen diferentes tipos de jornadas, siendo las más conocida la jornada diurna.

Esta propuesta de tener diferentes jornadas es para hacer más inclusivos los programas de estudio, ya que hay un segmento de estudiantes que son; padres, madres o trabajadores, los que, por diferentes contextos, se encuentran imposibilitados de asistir y/o matricularse en un programa de estudio durante el día (jornada diurna).

No obstante lo anterior, también existen IES que poseen jornadas a distancia y/o semipresenciales. Este tipo de jornada fue la que todas las Instituciones tuvieron que aplicar obligatoriamente durante la pandemia (distancia), ya que por restricciones sanitarias no se podía asistir a clases, por lo que la mayoría de las cátedras (ramos) fueron dictados de manera online, hasta que el Ministerio de Salud (MINSAL) y de Educación (MINEDUC) indicaron que se permitía asistir a clases presenciales.

A continuación, grafico N°3 con el número total de alumnos matriculados por tipo de jornada, para los años 2016-2022:




Fuente: elaboración propia en base a datos de matrícula de alumnos en Chile.

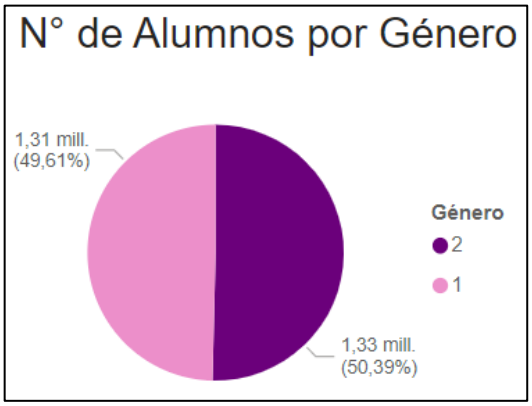
En el año 2022 el 45,6% del total de alumnos de ese periodo se encontraban matriculados en una jornada diurna mientras que el 31,1% lo hacían en una jornada vespertina (123 mil alumnos) y el 23,3% restante en otras modalidades (semipresencial y a distancia).

Respecto de la base total el 48,93% lo hace en una jornada diurna, el 37,54% en vespertino y el 12,28% posee una jornada a distancia. Dado lo anterior podemos inferir que el tipo mayoritario de jornada es la diurna, para todos los periodos entre el 2016 y el 2022.

## 4.4. Tipos de Alumnos por Género

Respecto de los alumnos por género corresponde al sexo del matriculado el cual consiste en 2 variables de un campo numérico. Por lo que se asigna el valor 1 a hombres y 2 a mujeres.

A continuación, grafico N°4 con el número total de alumnos matriculados por tipo de género, para los años 2016-2022:




Fuente: elaboración propia en base a datos de matrícula de alumnos en Chile.

En relación al número de alumnos matriculados por género desde el año 2016 hasta el 2022, podemos visualizar que la estadística es bastante contundente en relación a la paridad de géneros. Donde el 50,39% representan a los matriculados de sexo femenino y el 49,61% restante son del sexo masculino del total de la base.

Respecto del periodo 2022 la estadística no cambia mucho, el 50,54% son alumnos matriculados de sexo femenino y el 49,46% son matriculados de sexo masculino.

## 4.5. Metodología

### Modelos de deserción estudiantil en la educación superior

La regresión logística, es un modelo lineal para clasificación. La regresión logística también se conoce en la literatura como RL o Logit. El resultado puede ser utilizado para predecir como se comportará la variable dependiente en función de las variables independientes. La regresión logística (RL) puede manejar indistintamente variables predictoras tanto categóricas como continuas y en lugar de considerarlo como un caso especial del modelo LGL o Logit, puede considerarse como un híbrido entre el modelo Logit y el MCO (Menard, 2010).

Existen distintos tipos de regresión logística, principalmente podemos encontrar en la literatura 3 enfoques, los que mencionamos a continuación:

### Regresión logística binaria (RL)

Es una técnica estadística la cual tienen como objetivo comprobar las relaciones causales cuando la variable dependiente es nominal. Los coeficientes pueden ser utilizados para estimar las probabilidades de cada variable independiente del modelo. El resultado de la función logística devuelve un rango de valores entre 0 y 1, aquí es donde el modelo binario redondea las estimaciones a los valores más cercanos.

**Regresión logística multinomial (RLM)**

Es otra técnica, la cual es comúnmente utilizada para problemas que tienen varios resultados posibles, es decir para problemas que sean multiclase, siempre que la cantidad de resultados sea finita. El modelo también supone que la variable dependiente no puede ser perfectamente pronosticada a partir de las variables independientes. El análisis divide la variable dependiente en una serie de comparaciones entre dos categorías, esto se realiza con una variable de referencia. La regresión multinomial agrupa también los resultados a los valores más cercanos.

**Regresión logística ordinal (RLO)**

Es otro tipo regresión multinomial la cual busca una función de unión para relacionar de forma lineal a las variables explicativas con la razón de probabilidad entre la probabilidad acumulada hasta la categoría i de la variable ordinal, y la probabilidad que la variable tome un valor mayor que la categoría i (Agresti, 1990; Hosmer & Lemeshow, 2000).

El modelo que utilizaremos es un modelo de regresión logística, específicamente el modelo de la librería *scikit learn* (*sklearn.linear_model.LogisticRegression*).

El modelo se basa en un modelo clásico de regresión logística que matemáticamente en su forma más sencilla, cuando tenemos sólo una variable predictora podemos resumir como:

$$P(Y) = \frac{1}{1 + e^{-(a+bx)}}$$

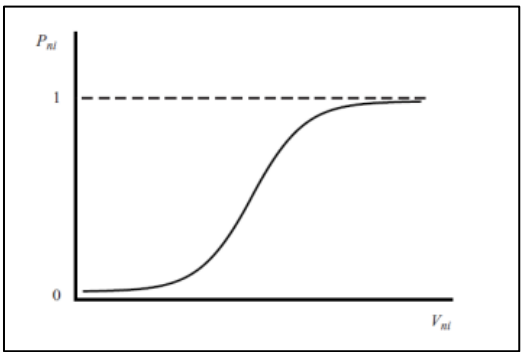
Donde:

P= Probabilidad de ocurrencia de la variable dependiente Y.

e= Es la base de los logaritmos naturales.

-(a+bx)= es finalmente la regresión lineal simple.

Clásicamente podemos graficar la solución con forma sigmoidea o función-S, tal y como se aprecia en la figura N° 1 a continuación[2].




En Python, la regresión logística se implementa por medio de *LogisticRegression*. Esta implementación puede adaptarse a la regresión logística binaria, uno contra el resto o multinomial con opción, o regularización Elastic-Net.

[2]Figura obtenida desde Econometrics Laboratory, UC Berkeley. https://eml.berkeley.edu/books/choice2nd

En relación a los denominados “solucionadores” implementados que podemos encontrar al interior de la función (método) *LogisticRegression* son los siguientes:

Liblinear utiliza un algoritmo de descenso de coordenadas (CD). El problema de optimización se desarrolla "uno contra el resto", por lo que se entrenan clasificadores binarios separados para todas las clases.

"lbfgs", "sag" y "newton-cg" solo admiten regularización o no regularización, y convergen más rápido para datos de alta dimensionalidad.

El solucionador de "sag" utiliza el descenso del gradiente promedio estocástico. Es más rápido que otros solucionadores para grandes conjuntos de datos, cuando tanto la cantidad de muestras como la cantidad de características son grandes[3].

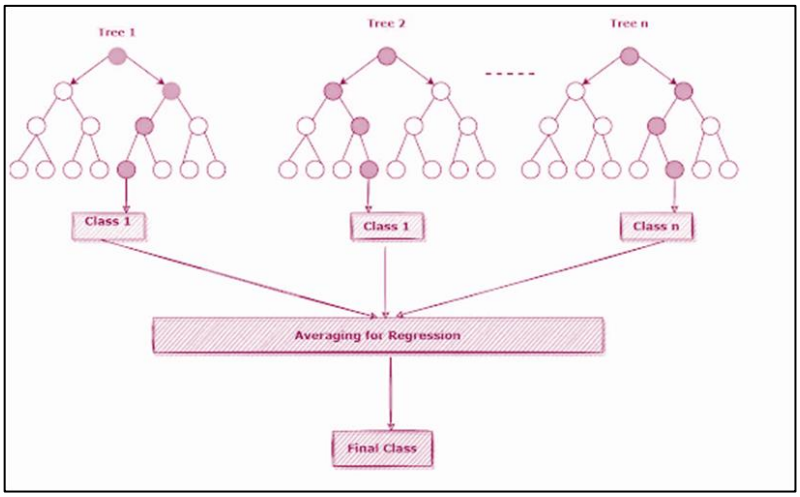
[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

**Modelo Random Forest**

Los bosques aleatorios son un método de aprendizaje para la clasificación y regresión. En materia de funcionamiento el algoritmo se basa en generar una serie de árboles de decisión individuales (ensemble), donde se entrenan individualmente con una muestra aleatoria extraída de los datos de entrenamiento originales, desde este punto nace su nombre de bosques aleatorios o *Random Forest.*

Para las tareas de clasificación, la salida del *Random Forest* es la clase seleccionada por la mayoría de los árboles. Para tareas de regresión, se devuelve la predicción media o promedio de los árboles individuales.

En forma gráfica podemos ver al modelo de *Random Forest* como la figura N° 2 que mostramos a continuación[4].




[4]Figura obtenida de Chaudhary, M. (2022, 28 de junio). Random Forest Algorithm - How It Works and Why It Is So Effective. Hire the World's Most Deeply Vetted Remote Developers | Turing. https://www.turing.com/kb/random-forest-algorithm

El primer algoritmo de *Random Forest* fue creado en 1995 por Tin Kam Ho donde utilizó el método del subespacio aleatorio[5].

En el 2001 Leo Breiman lo propone un nuevo multiclasificador, el que se basaba en una mejora del método Bootstrap Aggregating (Leo Breiman 1994).

Nuestra implementación la realizaremos utilizando la librería de *scikit-learn*. La aplicación del modelo *Random Forest*, cada árbol del conjunto se construye a partir de una muestra aleatoria extraída con reemplazo del conjunto de entrenamiento. Luego se divide cada nodo durante la construcción de un árbol, la mejor división se encuentra entre todas las entradas o un subconjunto aleatorio de tamaño *max_features*[6].

El objetivo de esta aleatoriedad es reducir la varianza del estimador del modelo. A partir de lo anterior es que los modelos *Random Forest* consiguen bajos indicadores de varianza al combinar los árboles.

"La implementación de *scikit-learn* combina clasificadores promediando su predicción probabilística, en lugar de permitir que cada clasificador vote por una sola clase"[7].

[5] Tin Kam Ho. 1995. Random decision forests. In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1 (ICDAR '95). IEEE Computer Society, USA, 278.

[6] Ensemble methods. (s.f.). scikit-learn. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees

[7] Ensemble methods. (s.f.). scikit-learn. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees

# 5. Resultados

## 5.1. Limpieza de datos

Luego de realizar la exploración de datos a nivel macro (para todo el conjunto de datos), para abordar y resolver el problema, pasamos a entender que hay detrás de la información que se encuentra en cada una de las líneas y como estos datos afectan a nuestro modelo, por lo que procederemos a limpiar los datos, para eliminar posibles valores en blanco, *outliers* y otros valores que consideremos que no aportan a nuestro modelo.

Como se mencionó en el capítulo anterior el periodo de investigación abarca desde el año 2016 hasta el 2022, para esto se trabajó la limpieza de datos en Python, específicamente con la librería Pandas.

Pandas es una librería de Python, que se utiliza para el análisis y manejo de datos. “Pandas pretende ser el bloque de construcción fundamental de alto nivel para realizar análisis prácticos de datos del mundo real en Python. Además, tiene como objetivo convertirse en la herramienta de manipulación/análisis de datos de código abierto más poderosa y flexible disponible en cualquier idioma"[8].

[8] Data structures for statistical computing in python, McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010. https://pandas.pydata.org/about/

## 5.2. Revisión de Missings Values

La base consta de 52 columnas las que mostramos en la tabla N°1:

| | | | |
|---|---|---|---|
| cat_periodo | tipo_inst_3 | dur_total_carr | area_conocimiento |
| id | cod_inst | region_sede | cine_f_97_area |
| codigo_unico | nomb_inst | provincia_sede | cine_f_97_subarea |
| mrun | cod_sede | comuna_sede | area_carrera_generica |
| gen_alu | nomb_sede | nivel_global | cine_f_13_area |
| fec_nac_alu | cod_carrera | nivel_carrera_1 | cine_f_13_subarea |
| rango_edad | nomb_carrera | nivel_carrera_2 | acreditada_carr |
| anio_ing_carr_ori | modalidad | requisito_ingreso | acreditada_inst |
| sem_ing_carr_ori | jornada | vigencia_carrera | acre_inst_desde_hasta |
| anio_ing_carr_act | version | formato_valores | acre_inst_anio |
| sem_ing_carr_act | tipo_plan_carr | valor_matricula | costo_proceso_titulacion |
| tipo_inst_1 | dur_estudio_carr | valor_arancel | costo_obtencion_titulo_diploma |
| tipo_inst_2 | dur_proceso_tit | codigo_demre | forma_ingreso |

Fuente: elaboración propia en base a datos de matrícula de alumnos en chile.

Esta información nos ayudará a entender las columnas y revisar los datos al interior de cada una de ellas.

Para revisar la cantidad de la base total de valores missing, nan y null se concatenarán todas las bases (por año) y se sumarán los valores que cumplan con la descripción anterior, siendo el resultado el siguiente:

A continuación, tabla n°2 con el número total de missing, nan y null por periodo, para los años 2016-2022:

| Nombre_Columna | Missings_2016 | Missings_2017 | Missings_2018 | Missings_2019 | Missings_2020 | Missings_2021 | Missings_2022 | Total |
|---|---|---|---|---|---|---|---|---|
| mrun | 1.473 | 1.564 | 1.564 | 1.564 | 1.564 | 1.564 | 1.564 | 10.857 |
| sem_ing_carr_ori | 61.237 | 67.739 | 67.739 | 67.739 | 67.739 | 67.739 | 67.739 | 467.671 |
| anio_ing_carr_act | 1.045.668 | 990.556 | 990.556 | 990.556 | 990.556 | 990.556 | 990.556 | 6.989.004 |
| sem_ing_carr_act | 1.003.675 | 968.855 | 968.855 | 968.855 | 968.855 | 968.855 | 968.855 | 6.816.805 |
| cod_sede | 45 | - | - | - | - | - | - | 45 |
| cod_carrera | 15.040 | 16.442 | 16.442 | 16.442 | 16.442 | 16.442 | 16.442 | 113.692 |
| version | 15.040 | 16.442 | 16.442 | 16.442 | 16.442 | 16.442 | 16.442 | 113.692 |
| requisito_ingreso | 19.782 | 3 | 3 | 3 | 3 | 3 | 3 | 19.800 |
| formato_valores | 1.247.178 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 8.736.936 |
| valor_matricula | 266 | 992 | 992 | 992 | 992 | 992 | 992 | 6.218 |
| valor_arancel | 618 | 970 | 970 | 970 | 970 | 970 | 970 | 6.438 |
| codigo_demre | 12.288 | 7.962 | 7.962 | 7.962 | 7.962 | 7.962 | 7.962 | 60.060 |
| acre_inst_desde_hasta | 108.727 | 528 | 528 | 528 | 528 | 528 | 528 | 111.895 |
| acre_inst_anio | 108.727 | 120.416 | 120.416 | 120.416 | 120.416 | 120.416 | 120.416 | 831.223 |
| costo_proceso_titulacion | 224.764 | 1.758 | 1.758 | 1.758 | 1.758 | 1.758 | 1.758 | 235.312 |
| costo_obtencion_titulo_diploma | 137.650 | 1.376 | 1.376 | 1.376 | 1.376 | 1.376 | 1.376 | 145.906 |
| forma_ingreso | 1.247.178 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 1.248.293 | 8.736.936 |

Fuente: elaboración propia en base a datos de matrícula de alumnos en chile.

A partir de la tabla podemos identificar las columnas que poseen una gran cantidad de información faltante y si esta será útil o es relevante para nuestro estudio. Por ejemplo, mrun nos resultará vital para poder identificar a los alumnos que deserten entre un periodo y otro, lo cual no ocurre con las columnas sem_ing_carr_ori, anio_ing_carr_act, sem_ing_carr_act, cod_sede, cod_carrera, versión, requisito_ingreso, formato_valores por lo que podemos solo podremos descartar los valores que se encuentran en columnas que son necesarias y relevantes para nuestro análisis. Dado lo anterior descartaremos las siguientes columnas que no poseen información mrun, cod_inst, cod_carrera, valor_matricula y valor_arancel, como se mencionó en el ejemplo mrun resulta vital para identificar a los alumnos matriculados y servirá como llave al igual que cod_inst la cual nos indica el código que posee la IES al interior de las bases, lo mismo sucede con

cod_carrera, que nos indica el código de la carrera del alumno. Estas variables resultan clave para generar un identificador o llave al interior de la base.

En relación a valor_matricula y valor_arancel serán necesarias para el final del estudio, para incorporar alguno de los impactos financieros que tienen que incurrir los alumnos para estudiar.

La segunda parte de la limpieza se realiza revisando el documento Esquema de Registro Matrícula de Educación Superior 2007-2022 por estudiante, bases públicas con MRUN donde se indican los valores y las conversiones de cada uno de los datos. Por medio de esto se logra identificar aquellos valores que podemos considerar como *outliers* y proceder a revisar si se elimina o se convierte simplemente en 0.

La tercera parte de la limpieza fue dar un formato correcto a las columnas, como por ejemplo establecer la columna fec_nac_alu como formato *datetime*. Las otras columnas se prepararon con formato str e int para facilitar la lectura de pandas.

Finalmente es relevante mencionar que para nuestro estudio se eliminó una cantidad inmaterial de filas, las que representaban aproximadamente un 0,12% de la data total, estas podrían tener un impacto negativo al ejecutar los modelos.

## 5.3. Preparación de la data

Luego de realizar la limpieza de la data, procedemos a prepararla para su concatenación debido a que son 7 bases las que contienen información anualizada de la matrícula.

Estableciendo alumnos por tipo: Se establecieron los tipos de alumnos, es decir se identificaron los alumnos de inicio y continuidad, para poder discernir quienes continúan estudiando la misma carrera. Es importante mencionar que para este estudio se utilizó el parámetro anio_ing_carr_ori para identificar el tipo de alumno y de esta manera etiquetarlo durante el estudio completo. No se consideraron aquellos alumnos que cambian de carrera como alumnos de continuidad, ya que en estricto rigor estos desertaron de una carrera para comenzar en otra, no siendo este el objetivo del estudio.

Estableciendo alumnos por tipo de comuna_sede: Para facilitar la dimensionalidad del estudio se seleccionó a la región metropolitana como *dummie*, entendiendo que es la región que más aporta al número de alumnos matriculados y es la región que mayor cantidad de habitantes posee al día de hoy, concentrando casi el 40,1% de la población total.

También se estableció como *dummie* la columna acreditada_carr (acreditación de carrera) y acreditada_inst (acreditación Institución) ambos para identificar si la carrera y la

institución se encuentran debidamente acreditadas (esta última se consideró para el modelo).

Para poder generar una llave única que identifique a los alumnos y su respectiva carrera durante el periodo de estudio se desarrolló la llave de alumnos, institución, sede, carrera, modalidad conocida en el sector de las IES como llave Alumnos-SIES con esta información podemos hacer seguimiento de los alumnos que continúan estudiando su misma carrera. Mediante esta metodología se estableció un contador para cada uno de los Alumnos-SIES por año.

## 5.4. Filtrado de Datos

Antes de proceder a la concatenación de las bases, se filtró la información para reducir su dimensionalidad, ya que como mencionamos en el inicio del estudio solo se considerarán aquellas carreras de 8 semestres, por simplicidad y para hacer comparables, en su medida los planes de estudio. Lo anterior comprendiendo la complejidad y las particularidades que posee cada plantel de estudios, región y formación académica, además de una serie de factores que son propios de la idiosincrasia de cada IES.

También se filtran aquellos alumnos matriculados solo con modalidad presencial. Si bien, existen 3 tipos de modalidades (Presencial, Semipresencial y No presencial) el 84,25% estudia en modalidad presencial, así también se descarta posibles *outliers* y ruido en la información y nuevamente para hacer comparables a los alumnos en el estudio en cuestión.

Finalmente, como el objetivo es medir la permanencia (deserción) de los alumnos matriculados desde el año 2016 y para hacer comparable el estudio se definió considerar solo aquellos alumnos que pertenecen al subsector de Institutos Profesionales, no considerando las Universidades y los Centros de Formación Técnica.

## 5.5. Concatenado de bases

Una vez que tenemos la data limpia y filtrada con las columnas que consideramos relevantes en el estudio, concatenamos la información de todos los años para general la base definitiva y proceder a nuevamente limpiar, revisar, analizar y filtrar la información.

## 5.6. Seguimiento del avance de la continuidad

Para hacer seguimiento a la continuidad, el inicio se estableció en el año 2016 y también se definieron aquellos alumnos que efectivamente pertenecen a la continuidad (medido mediante Rut, IES y carrera). De esta forma se realizó un seguimiento y conteo de los semestres (medido en años[9]) que efectivamente los alumnos continúan estudiando. Debido a que no contamos con la información de titulación, así como tampoco con la información de egreso efectivo, por simplicidad del análisis se consideró que aquellos alumnos que

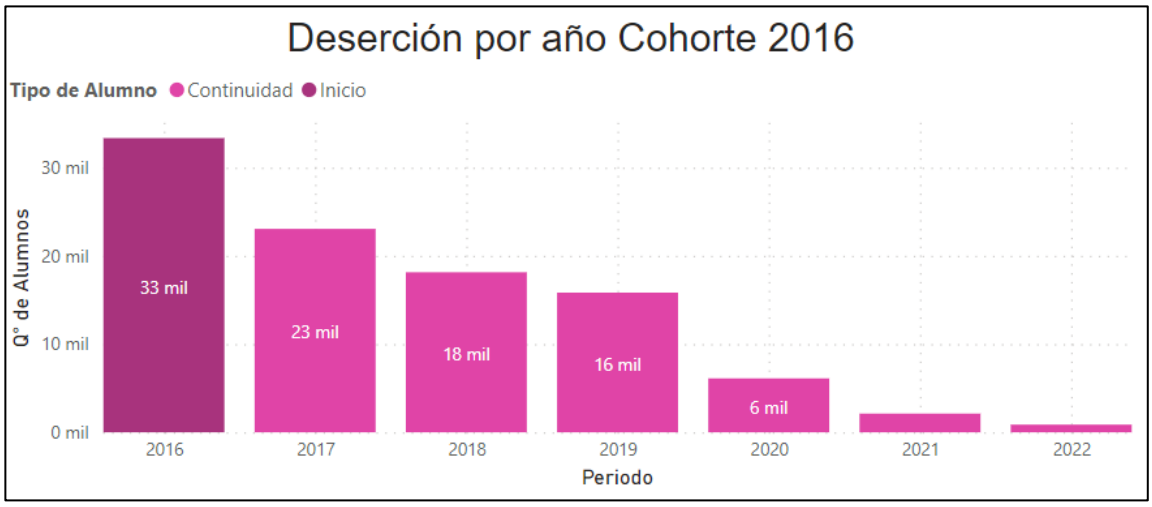
[9] En el presente estudio se optó por considerar el año y no los semestres, ya que no se contaba con la información de matrícula semestral, debido que esta información es de carácter privado de las IES.

llegan hasta el octavo semestre (4to año) permanecieron los 8 semestres y se encuentran ad portas de titularse.

## 5.7. Análisis exploratorio de la data preparada

Respecto del objetivo de nuestra investigación, las matrículas de las carreras de 8 semestres de duración, considerando el año 2016 como inicio (y al interior de los Institutos Profesionales), posee una deserción inter-anual para el primer año de un 30,86%, debido a esto en el segundo año, es decir en el 2017 tenemos 23.058 alumnos matriculados.

A continuación, grafico n°5 con la evolución de la deserción (medido por la permanencia) por año de la cohorte 2016.



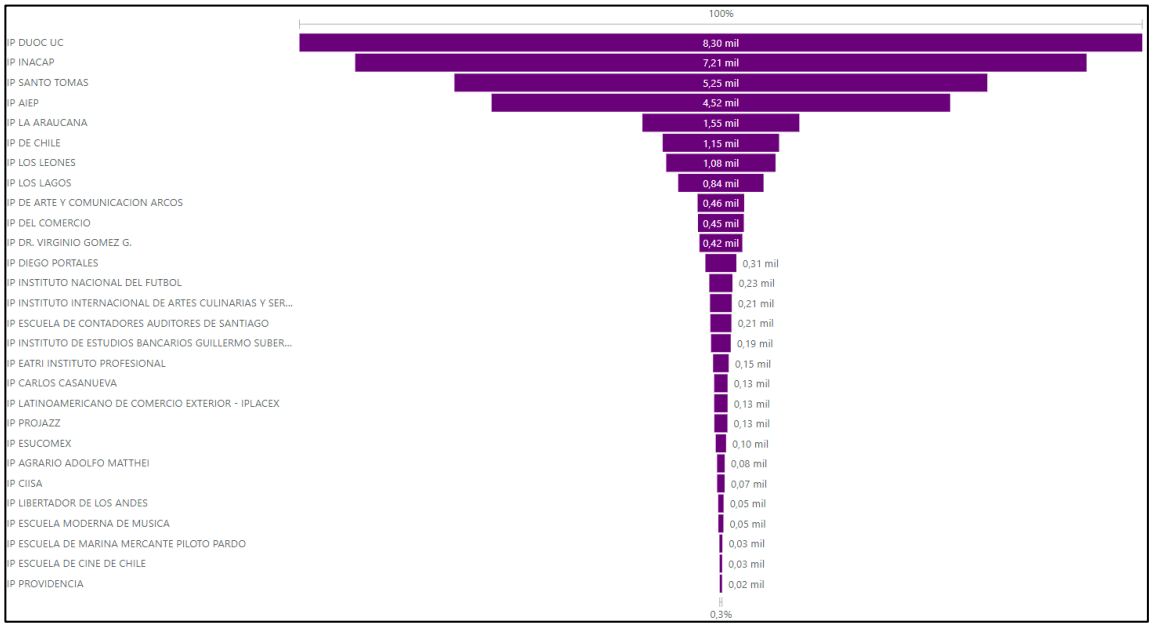

Fuente: elaboración propia en base a datos de matrícula de alumnos en chile subsegmento IP.

El grafico muestra como los alumnos van desertando en el paso de los años, se incluyó solo para fines gráficos los periodos 2021 y 2022, donde aún existe permanencia de algunos alumnos (atrasos académicos).

La tendencia es clara, nos muestra una baja promedio de 23,04% en los primeros 4 años de estudio, es decir durante el primer año se matriculan 33.351 de los cuales solo llegan, en términos porcentuales, al cuarto año (8vo semestre) un 47,47% del total de alumnos de la cohorte 2016.

Respecto a las IES que poseen alumnos matriculados con los parámetros definidos contabilizamos 28 y muestran una concentración similar a la que pudimos ver anteriormente cuando mostramos el total del subsegmento de los IP´S. Respecto del total de alumnos matriculados, nuevamente Duoc UC es la IES con mayor cantidad de alumnos matriculados con 8.304 (periodo 2016), lo cual representa un 24,90% del total de matriculados. En segundo lugar, se encuentra Inacap con 7.206 y finalmente IP Santo Tomás con 5.250. En total estas 3 IES concentran un 62,25% del total de la matricula periodo 2016, observándose nuevamente un alto grado de concentración de las IES que comparten el subsegmento de las IP´S.

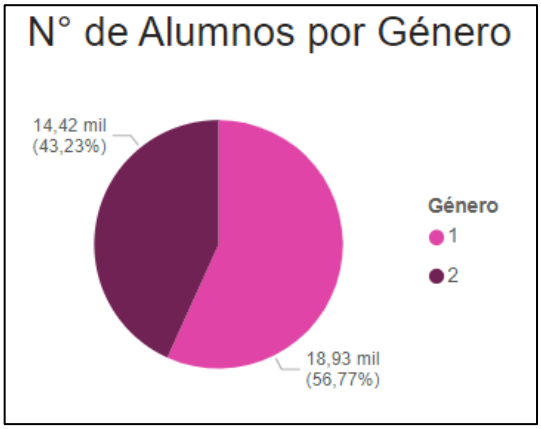
A continuación, grafico N°6 con el número de matriculados históricos por IES:




Fuente: elaboración propia en base a datos de matrícula de alumnos en chile subsegmento IP.

Respecto de los alumnos por género corresponde al sexo del matriculado el cual consiste en 2 variables de un campo numérico. Por lo que se asigna el valor 1 a hombres y 2 a mujeres.

A continuación, grafico N°7 con el número total de alumnos matriculados por tipo de género, para el año 2016:




Fuente: elaboración propia en base a datos de matrícula de alumnos en Chile.

En relación al número de alumnos matriculados por género, a diferencia del análisis exploratorio macro, se visualiza una diferenciación en relación a la paridad de géneros. Donde el 56,77% representan a los matriculados de sexo masculino y el 43,23% restante son del sexo femenino.

Al comparar la permanencia a través de los años podemos mencionar que para el 2017 el porcentaje de matriculados de sexo masculino cae 1,07 pp (es decir 55,70%). Respecto de la hipótesis comprobamos en términos numéricos mediante tabla de frecuencias, que efectivamente desertan más hombres que muejeres al 8vo semestre.

| Género | Total 2016 | Total 2016 (%) | Total 8vo Sem. | Total Desersión | Total Desersión (%) |
|---|---|---|---|---|---|
| Hombres | 18.932 | 56,77% | 8.407 | 10.525 | 57,48% |
| Mujeres | 14.419 | 43,23% | 6.634 | 7.785 | 42,52% |
| **Total** | **33.351** | **100,0%** | **15.041** | **18.310** | **100,0%** |

Fuente: elaboración propia en base a datos de matrícula de alumnos en Chile.

Dado lo anterior, se puede apreciar un aumento de deserción de los alumnos de género masculino de 0,72 puntos porcentuales. Respecto de la cantidad de alumnos que permanecen efectivamente hasta el 8vo semestre de estudios son 15.041 lo que representa el 45,1% mientras los que desertan son 18.310 un 54,9% de la cohorte del año 2016.

## 5.8. Etiquetado de datos

Como se obtuvieron los alumnos que efectivamente continuaron sus estudios se procedió a generar un contador que nos permitiera saber cuántos periodos estudiaron los alumnos, con esto pudimos determinar aquellos alumnos que efectivamente avanzaron académicamente y aquellos que no permanecieron (desertaron) en el tiempo. Esta nueva base adicional que sumó los periodos también se llevó a una *dummie* que permite saber en el año 2019 quien permaneció y quien deserto.

## 5.9. Separación de la data

Mediante la librería *sklearn* se utilizó *model_selection* y *train_test_split* para hacer la separación de la data en: sets de testeo y entrenamiento, donde los porcentajes que se utilizaron fueron en razón de 70% y 30%, utilizando también *random_state*=5 para que los resultados no cambien y se utilice la misma configuración (=5).

## 5.10. Análisis de Frecuencias

Para poder entender de mejor manera los datos, utilizamos la librería *funpymodeling*[10], mediante esta librería podremos realizar el análisis de frecuencia de nuestras variables cualitativas, a continuación, dejamos las tablas de frecuencia para las principales variables:

**Tabla de frecuencia Rango de Edad:**

| Rango_Edad | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| 15 a 19 años | 12.591 | 37,8% | 37,8% |
| 20 a 24 años | 12.098 | 36,3% | 74,0% |
| 25 a 29 años | 4.457 | 13,4% | 87,4% |
| 30 a 34 años | 2.184 | 6,5% | 93,9% |
| 35 a 39 años | 1.082 | 3,2% | 97,2% |
| 40 y más años | 939 | 2,8% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

[10] funpymodeling. (2020, 16 septiembre). PyPI. https://pypi.org/project/funpymodeling/

## Tabla de frecuencia por nombre de IES

| Nombre Institución | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| IP DUOC UC | 8.304 | 24,9% | 24,9% |
| IP INACAP | 7.206 | 21,6% | 46,5% |
| IP SANTO TOMAS | 5.250 | 15,7% | 62,2% |
| IP AIEP | 4.521 | 13,6% | 75,8% |
| IP LA ARAUCANA | 1.546 | 4,6% | 80,4% |
| IP DE CHILE | 1.147 | 3,4% | 83,9% |
| IP LOS LEONES | 1.077 | 3,2% | 87,1% |
| IP LOS LAGOS | 841 | 2,5% | 89,6% |
| IP DE ARTE Y COMUNICACION ARCOS | 459 | 1,4% | 91,0% |
| IP DEL COMERCIO | 452 | 1,4% | 92,4% |
| IP DR. VIRGINIO GOMEZ G. | 422 | 1,3% | 93,6% |
| IP DIEGO PORTALES | 305 | 0,9% | 94,5% |
| IP INSTITUTO NACIONAL DEL FUTBOL | 228 | 0,7% | 95,2% |
| IP INSTITUTO INTERNACIONAL DE ARTES CULINARIAS | 214 | 0,6% | 95,9% |
| IP ESCUELA DE CONTADORES AUDITORES DE SANTIAGO | 207 | 0,6% | 96,5% |
| IP INSTITUTO DE ESTUDIOS BANCARIOS GUILLERMO S. | 194 | 0,6% | 97,1% |
| IP EATRI INSTITUTO PROFESIONAL | 154 | 0,5% | 97,5% |
| IP CARLOS CASANUEVA | 132 | 0,4% | 97,9% |
| IP IPLACEX | 132 | 0,4% | 98,3% |
| IP PROJAZZ | 129 | 0,4% | 98,7% |
| IP ESUCOMEX | 102 | 0,3% | 99,0% |
| IP AGRARIO ADOLFO MATTHEI | 75 | 0,2% | 99,2% |
| IP CIISA | 74 | 0,2% | 99,5% |
| IP LIBERTADOR DE LOS ANDES | 53 | 0,2% | 99,6% |
| IP ESCUELA MODERNA DE MUSICA | 50 | 0,1% | 99,8% |
| IP ESCUELA DE MARINA MERCANTE PILOTO PARDO | 29 | 0,1% | 99,9% |
| IP ESCUELA DE CINE DE CHILE | 25 | 0,1% | 99,9% |
| IP PROVIDENCIA | 23 | 0,1% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

## Tabla de frecuencia por Jornada

| Jornada | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| Diurno | 20.470 | 61,4% | 61,4% |
| Vespertino | 12.857 | 38,6% | 99,9% |
| Otro | 24 | 0,1% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

## Tabla de frecuencia por Región

| Región Sede | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| Metropolitana | 15.225 | 45,7% | 45,7% |
| Valparaíso | 3.041 | 9,1% | 54,8% |
| Biobío | 2.754 | 8,3% | 63,0% |
| Antofagasta | 1.928 | 5,8% | 68,8% |
| Maule | 1.721 | 5,2% | 74,0% |
| Los Lagos | 1.647 | 4,9% | 78,9% |
| Lib. Gral B. O'Higgins | 1.385 | 4,2% | 83,1% |
| La Araucanía | 1.363 | 4,1% | 87,1% |
| Coquimbo | 1.102 | 3,3% | 90,5% |
| Tarapacá | 871 | 2,6% | 93,1% |
| Los Ríos | 753 | 2,3% | 95,3% |
| Ñuble | 595 | 1,8% | 97,1% |
| Atacama | 444 | 1,3% | 98,4% |
| Magallanes | 263 | 0,8% | 99,2% |
| Arica y Parinacota | 238 | 0,7% | 99,9% |
| Aysén | 21 | 0,1% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

## Tabla de frecuencia por área de conocimiento

| Área de Conocimiento | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| Tecnología | 13.728 | 41,2% | 41,2% |
| Administración y Comercio | 8.781 | 26,3% | 67,5% |
| Arte y Arquitectura | 3.860 | 11,6% | 79,1% |
| Ciencias Sociales | 3.228 | 9,7% | 88,7% |
| Educación | 2.559 | 7,7% | 96,4% |
| Agropecuaria | 602 | 1,8% | 98,2% |
| Salud | 247 | 0,7% | 99,0% |
| Ciencias Básicas | 192 | 0,6% | 99,5% |
| Humanidades | 154 | 0,5% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

## Tabla de frecuencia por acreditación de carrera

| Acreditación Carrera | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| 1 | 19.245 | 57,7% | 57,7% |
| 0 | 14.106 | 42,3% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

**Tabla de frecuencia por acreditación de Institución**

| Acreditación Institución | Frecuencia | % del Total | % Acumulado |
|---|---|---|---|
| 1 | 31.016 | 93,0% | 93,0% |
| 0 | 2.335 | 7,0% | 100,0% |
| **Total** | **33.351** | **100,0%** | **-** |

## 5.11. Establecimiento de los modelos

La selección de las variables dependientes e independientes fue realizada acorde a los resultados que buscamos, identificar aquellos alumnos que permanecen y desertan en nuestra base a partir del periodo 2016, como mencionamos en pasos previos, etiquetamos la data, por lo que ya sabemos quiénes son los que desertan y quienes permanecen así que en la selección de variables dependientes dejaremos la data etiquetada. Por otra parte, las variables independientes seleccionadas son las siguientes: gen_alu, jornada_dummie, edad, region_dummie y carrera genérica (como *dummie*[11]), acreditada_inst, valor_arancel, valor_matricula.

## 5.12. Selección de modelos

Los modelos seleccionados para este estudio son el modelo Logit y el *Random Forest*, ambos modelos son utilizados para realizar este tipo de estudio, debido a su simplicidad ya que nos entregan resultados que son binarios y trabajan bien con variables categóricas, como son las que tenemos disponibles y utilizaremos más adelante.

[11] Se utilizo la librería sklearn mediante la función de OneHotEncoder para transformar la variable a una dummie. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

**Resultados de la aplicación del modelo Logit**

Para obtener los resultados más detallados se utilizó el F1 score, que mide la exactitud del modelo combinando la *precision* y *recall* en un resultado. Esto permite hacer comparaciones con otros modelos.

Donde:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

A continuación, mostramos los resultados del modelo el cual obtuvo un 84% de precisión, medido por el F1.

| Logit Model | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0,84 | 1,00 | 0,91 | 19.571 |
| 1 | 0,00 | 0,00 | 0,00 | 3.775 |
| **Accuracy** | | | **0,84** | **23.346** |
| Macro Avg | 0,42 | 0,50 | 0,46 | 23.346 |
| Weighted Avg | 0,70 | 0,84 | 0,76 | 23.346 |

Fuente: elaboración propia en base a resultados del estudio.

Resultados de la aplicación del modelo *Random Forest*

Por otra parte, en el modelo de *Random Forest* los resultados del F1 score fueron de un 81% de precisión.

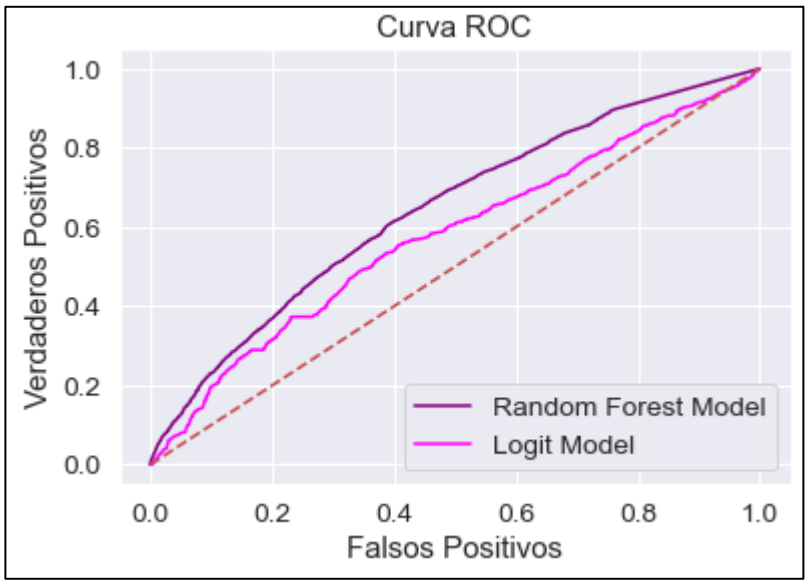
| Random Forest Model | Precision | Recall | F1 Score | Support |
|---|---|---|---|---|
| 0 | 0,85 | 0,93 | 0,89 | 19.571 |
| 1 | 0,32 | 0,18 | 0,23 | 3.775 |
| **Accuracy** | | | **0,81** | **23.346** |
| Macro Avg | 0,42 | 0,50 | 0,56 | 23.346 |
| Weighted Avg | 0,70 | 0,84 | 0,78 | 23.346 |

Fuente: elaboración propia en base a resultados del estudio.

## 5.13. Comparación de Modelos

Para comparar gráficamente utilizamos la curva de ROC (*Receiver Operating Characteristic*), es una representación gráfica, la cual se construye mediante la unión de los puntos, midiendo la sensibilidad del modelo. Lo anterior nos proporcionó una herramienta más para discernir que modelo es mejor (posiblemente óptimo). Para la comparación se utilizó *sklearn* con los modelos previamente utilizados.

A continuación, se presenta el grafico N° 8 con la comparación de ambos modelos:




Elaboración propia.

Se aprecia claramente que el modelo *Random Forest* resulta adecuado para nuestro análisis, ya que se encuentra por sobre el modelo *Logit*. Además, al revisar los resultados del F1 score podemos apreciar que, si bien el modelo *Logit* posee mayor F1, al revisar la composición del resultado notamos que solo funciona para predecir un resultado, mientras que el modelo *Random Forest* posee un F1 más bajo, este si posee predicción para ambos resultados (89% y 23%).

# 6. Conclusiones

Se utilizaron dos modelos de *Machine Learning* para predecir la deserción estudiantil en cuarto año (8vo semestre), al interior de los Institutos Profesionales en Chile. Las variables seleccionadas que fueron incorporadas en el modelo son: género del alumno, edad, jornada, rango de edad, región, tipo de carrera, acreditación de la Institución, los valores de arancel y matrícula.

El resultado reflejó una precisión de las predicciones de un 84% para el modelo Logit y un 81% para el modelo *Random Forest*.

Es relevante destacar la cantidad de alumnos que desertan año a año en los distintos planteles del País (promedio 23,04%[12]). Es el 45,1% de los alumnos que comenzaron a estudiar el año 2016 y llegaron hasta último año (permanencia).

Muchos de los estudios, indican que no existe una receta, tampoco un manual para elegir las variables adecuadas, que se ajusten de mejor manera al modelo y que este tenga mayor capacidad de predicción. En relación a nuestro ejercicio por una parte los resultados de los F1 score muestran buenos resultados, pero al analizar los resultados de la Curva de ROC concluimos que ambos modelos resultan ser "regulares" respecto de la relación sensibilidad versus especificidad.

[12] Subsector IP, para los años 2016-2020.

## 6.1. Limitaciones del estudio

No considerar aspectos socioeconómicos de los alumnos; quintil de ingreso, ingresos monetarios y autónomos del hogar, clasificación de socioeconómica de los padres y del alumno, tipo de vivienda, otros.

No considerar aspectos sociales de los alumnos; edad de los padres, personas en el hogar, nivel de escolaridad de padres, comuna de residencia, colegio de los padres, otros.

No considerar aspectos académicos de los alumnos; avance por créditos, promedio de notas, notas por ramo, asistencia, notas PSU, nem colegio, colegio, programa RAP, PACE, otros.

No considerar aspectos de Financiamiento Estudiantil de los alumnos; financiamiento por tipo, CAE, Becas, Gratuidad, otros beneficios, morosidad, otros.

# Bibliografía

Universidad Privada de San Juan Bautista, Facultad de Ciencias de la Salud, Dr. Willfredo Erwin Gardini Tuesta, Material de la clase 13, Regresión Logística, Análisis de regresión, Primer semestre 2021. https://www.scribd.com/document/526566398/CLASE-13-Regresion-Logistica

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. scikit-learn
https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados, Patricio E. Ramírez y Elizabeth E. Grandón.

Ministerio de Educación – SIES (2014), Deserción en la educación superior en Chile, Ministerio de Educación.

González, L. E., Estudio sobre la repitencia y deserción en la educación superior chilena, Digital Observatory for higher education in Latin America and The Caribbean. IESALC – UNESCO (2005)

¿Dónde estudiar? - Subsecretaría de Educación Superior. (s.f.). Subsecretaría de Educación Superior. https://educacionsuperior.mineduc.cl/donde-estudiar/

Grandón, E. y G. Vargas, Movilidad social intergeneracional: Una mirada de la contribución social de las universidades. El caso de los titulados de la Universidad del Bío-Bío, Chile, Centro Interuniversitario de Desarrollo CINDA, Colección Gestión Universitaria, 275-294 (2012).

Prieto Arciniegas, C. (2015). Uso de regresión logística para predecir deserción estudiantil temprana. Uniandes.

Biblioteca del Congreso Nacional | Ley Chile. (s.f.). www.bcn.cl/leychile. https://www.bcn.cl/leychile/navegar?idNorma=1118991&idParte=9917404&idVersion=2018-05-29

Biblioteca del Congreso Nacional | Ley Chile. (s.f.-b). www.bcn.cl/leychile. https://www.bcn.cl/leychile/navegar?idNorma=30330

Ley Fácil - Biblioteca del Congreso Nacional de Chile - BCN. https://www.bcn.cl/leyfacil/recurso/jornada-parcial-alternativa-para-estudiantes-trabajadores

Historia de la Educación Técnico-Profesional - Educación Técnico Profesional. (s.f.-a). https://www.tecnicoprofesional.mineduc.cl/secretaria-tecnico-profesional/historia/

MI FUTURO | Ministerio de Educación de Chile. (s.f.). https://www.mifuturo.cl/tipos-de-institucion/

Data structures for statistical computing in python, McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010. https://pandas.pydata.org/about/

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Ho, T. K. (1995). Random Decision Forest. Retrieved from http://cm.bell-labs.com/cm/cs/who/tkh/papers/odt.pdf

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. https://doi.org/10.1109/34.709601

Pablo Casas funpymodeling. (2020, 16 septiembre). MIT License (MIT) PyPI. https://pypi.org/project/funpymodeling/