UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Métodos multivariados aplicados al análisis de datos para la identificación de factores relacionados a las brechas de género en el contexto de I+D+i+e al interior de la Universidad de Concepción

Por: Javiera Arias Gutiérrez

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción para optar al título de Ingeniero Civil Matemático

Marzo 2023

Concepción, Chile

Profesores Guía: Dra. María Paz Casanova Laudien
y Mag. Jean Paul Navarrete Campos

A Emilia, Mateo y Matilda.

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a todas las personas que han sido parte de mi camino universitario. En primer lugar, a mis profesores guía, María Paz Casanova Laudien y Jean Paul Navarrete Campos, quienes brindaron un valioso apoyo en la estructuración y desarrollo de mi trabajo. También quiero agradecer a mi comisión evaluadora, Katia Sáez Carrillo, Luisa Rivas Calabrán y Daniela De Quevedo Rodríguez, por su tiempo y dedicación en la evaluación de mi memoria de título.

A mi padre, Juan Carlos Arias Sobarzo, quiero agradecerle por su cariño y apoyo en los momentos más difíciles de mi carrera. Sus sabios consejos y enseñanzas han sido esenciales en mi camino hacia el éxito académico.

Agradezco a mi madre, Gladys Gutiérrez Lillo, por su constante preocupación, entrega y cuidado, por haberme educado en mis primeros años de vida y guiarme para poder obtener este importante logro.

A mi amigo, roommate, compañero, y ahora colega, Vicente Marchant Contreras, le agradezco su compañía y fraternidad desde el primer día de clases, cuando aún no sabíamos a lo que nos estábamos enfrentando. A Adonai Angulo Rodríguez, por su amistad y las experiencias vividas juntos. A Claudio Correa Barría, por su confianza y camaradería en episodios cruciales de mi vida univesitaria.

Quiero agradecer a mi novio, Claudio Mansilla Brito, por su apoyo incondicional, afecto, amor y contención durante todo el proceso de mi memoria de título.

A quien fue mi jefa de carrera durante casi toda mi educación universitaria, Mónica Selva Soto, por su compromiso, gestión y apoyo en momentos determinantes de mi formación académica.

No habría sido posible llegar hasta aquí sin ellos.

Resumen

La diferencia que existe entre hombres y mujeres dentro del mundo de la investigación es algo altamente relevante. Esto aborda desde inclinaciones en la asignación de cargos académicos, hasta diferencias desproporcionadas en montos destinados a investigaciones. En este trabajo se busca identificar factores relacionados a las brechas de género dentro de la Universidad de Concepción, específicamente en el contexto de I+D+i+e. Este estudio se hizo a través de métodos multivariados aplicados al análisis de información, sobre una base de datos que contiene proyectos de investigación entre los años 2013 y 2022. Para poder efectuar correctamente estos análisis, es necesario realizar previamente un análisis exploratorio de la información, la cual incluye datos atípicos y datos perdidos, por lo que se aplica un método de imputación. Además, la base de datos contiene en su mayoría valores cualitativos, por lo que se debe formar una integración con variables ficticias. Se realiza un Análisis Descriptivo de los datos, donde se aprecian las variables que tienen una brecha de género más importante, como el número total de proyectos, el financiamiento, el área de investigación, y la edad del participante. Se realiza un Análisis de Conglomerados, separando la base de datos de proyectos liderados por hombres de los liderados por mujeres, donde en ambos casos las componentes principales están relacionadas con los proyectos externos a la universidad. Además, se genera un modelo de Regresión Logística Múltiple para identificar la importancia de cada variable para esta clasificación, donde las más relevantes tienen relación con departamentos, facultades, y cargos académicos, utilizando previamente un método de selección de variables.

Palabras Clave – I+D+i+e (Investigación, Desarrollo, Innovación y Emprendimiento)

Abstract

The difference between men and women in the world of research is highly relevant, ranging from differences in the allocation of academic positions to disproportionate disparities in research funding. This study aims to identify factors related to gender gaps within the Universidad de Concepción, specifically in the context of R+D+i+e. The study utilizes multivariate methods to analyze data from a database containing projects between 2013 and 2022. Prior to conducting these analyses, an exploratory analysis of the information is necessary, which includes addressing outliers and missing data. Furthermore, since the database consists mainly of qualitative values, integration with dummy variables is required. Descriptive analysis is performed on the data, revealing the variables with the most significant gender gap, such as age, area, and funding for projects. Cluster analysis is also conducted, separating the database of projects led by men from those led by women, although both groups exhibit similar principal components, including external projects. Additionally, a multiple logistic regression model is generated to identify the importance of each variable for this classification, with variables related to area and subareas of investigation being the most significant, following the use of a variable selection method.

Keywords – R+D+i+e (Research, Development, Innovation and Entrepreneurship)

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Brechas de género

Las brechas de género son una forma de representar la disparidad entre hombres y mujeres en cuanto a derechos, recursos u oportunidades. Este concepto se aplica a múltiples ámbitos, como el académico, el político o el empresarial. Disminuir las brechas de género incluye justicia y equidad, y aborda todos los campos mencionados, considerando la dimensión cultural.

En las últimas décadas, muchos países han logrado avances significativos hacia la igualdad de género en educación. Aún así, las mujeres continúan obteniendo menores sueldos que los hombres, tienen menos probabilidades de tener cargos de alta jerarquía y, en general, sufren mayores discriminaciones.

En la educación superior, las niñas y mujeres son menos propensas a escoger carreras científicas y tecnológicas. Los campos de estudio que elijen hombres y mujeres jóvenes perpetúan la segregación por género en los mercados laborales. Uno de los motivos por los que existen estas brechas de género, es porque las mujeres actualmente continúan con la carga doméstica diaria, incluyendo el cuidado de los hijos [OCDE, 2014].

1.1.1. Brechas de género dentro de la Universidad de Concepción

La Universidad de Concepción (UdeC) tiene una trayectoria reconocida en I+D+i+e (Investigación, Desarrollo, Innovación y Emprendimiento), donde el plan estratégico institucional 2021-2030 establece, dentro de los valores, la equidad, inclusión y responsabilidad social, destacando en particular, dentro de los principios fundamentales, la equidad de género.

Dentro de la Universidad, existe una brecha de género en el número de profesionales que forman la planta académica, en las subcategorías de investigadoras e investigadores, y en los liderazgos femeninos en proyectos de investigación. Con el fin de precisar, se considerará que las personas investigadoras del cuerpo académico regular de la Universidad son aquellas que tienen al menos una publicación y/o participación en proyectos durante el año 2020.

Un ejemplo de esto, es que en el año 2022 existían 1288 personas en categoría académica regular con jornada completa, pero solamente el 38 % eran mujeres. Además, desde 1986, el premio de "Profesor Emérito" ha sido otorgado a 37 académicos y 5 académicas. Las mujeres que realizan investigación y que ocupan cargos de liderazgo en gestión académica (Decanas, Vicedecanas, Directoras de Departamento y de Programas de Postgrados) son apenas el 33 %.

El número de jornadas completas equivalentes de la planta académica que realizan investigación es 725, de los cuales 275 (38 %) son mujeres y 450 (62 %) son hombres, lo que se condice con la proporción de mujeres y hombres de la institución.

Existe también una distribución dispar por jerarquía académica y género, representada en la Figura 1.1.1:

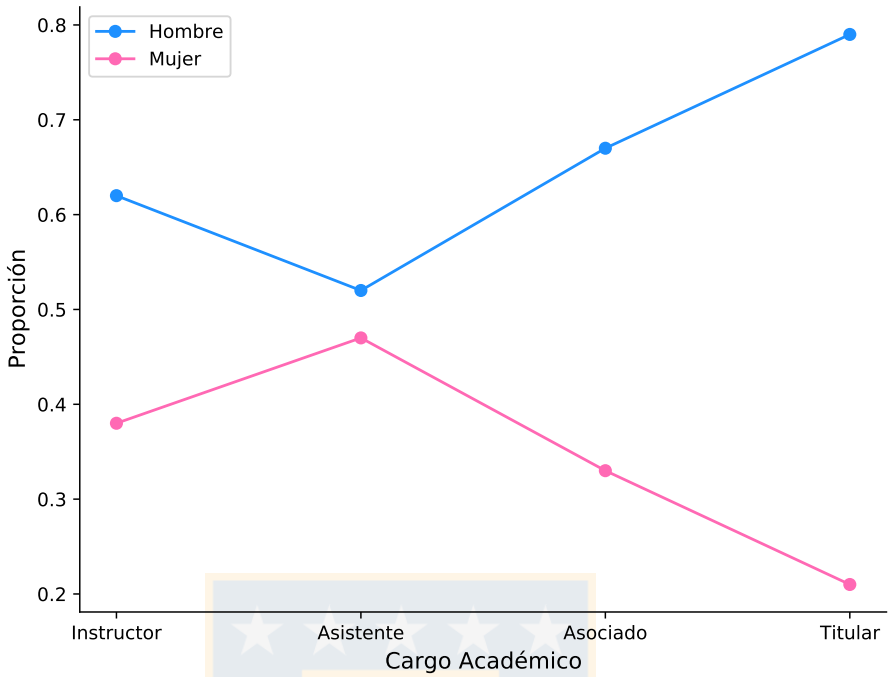


Figura 1.1.1: Distribución jerárquica UdeC
(Fuente: Elaboración propia)

Se encuentra que, el porcentaje de mujeres en las distintas jerarquías académicas es de: 38 % instructores, 47 % asistentes, 33 % asociados y 21 % titulares. Esto marca una desviación de la proporción de mujeres y hombres que forman la planta académica, con un incremento en asistentes pero una disminución a medida que el nivel de la jerarquía es más alto, destacando así la gran brecha para la jerarquía de Titular.

1.1.2. Proyecto InEs de Género

El Proyecto de Innovación en Educación Superior, InES, que se ha adjudicado la Universidad de Concepción, elaborado por la Vicerrectoría de Investigación y Desarrollo (VRID) junto con la Dirección de Equidad de Género y Diversidad, DEGyD, tiene como objetivo general generar oportunidades y capacidades de desarrollo académico, liderazgo y cooperación, que permitan disminuir las brechas de género en los ámbitos de I+D+i+e.

Esta Memoria de Título busca ser un aporte para las investigaciones sobre brechas de género y, en particular, proporcionar un análisis estadístico al Proyecto InES de Género. Dicho análisis se hará sobre una base de datos proporcionada por la

VRID, que contiene los proyectos de investigación de la Universidad entre los años 2013 y 2022.

Para más información sobre el Proyecto InES de Género, revisar *Universidad de Concepción (2021), Formulación Proyecto InES de Género. Vicerrectoría de Investigación y Desarrollo. Documento Interno.*

Un detalle a considerar, es que la base de datos otorgada contiene la categoría `Sexo` como una variable binaria donde los valores posibles son `'Masculino'` y `'Femenino'`. Para realizar un análisis de género, se tuvo que suponer una relación entre las variables `Sexo` y 'Género', por lo que la categoría 'Género' en este análisis también es de forma dicotómica, donde se infiere que `Sexo 'Masculino'` corresponde a 'Género Hombre', y `Sexo 'Femenino'` a Género 'Mujer'.

1.2. Objetivo general

En este trabajo se muestra que existen brechas de género dentro de la Universidad de Concepción, identificando las variables que son más influyentes sobre este fenómeno. Para ello, se aplican técnicas descriptivas para representar los datos y buscar relaciones entre variables: se realiza un Análisis Descriptivo Bivariado para entender la distribución de los datos, se realiza un Análisis de Conglomerados con los proyectos de hombres separados de los proyectos de mujeres, para identificar si se agrupan de manera diferente y qué características influyen, y se construye un modelo de Regresión Logística Múltiple para clasificar las investigaciones según el género de quién las lideró, estimando la probabilidad de que cierto investigador sea hombre o mujer, en base a la selección de las variables más representativas.

1.2.1. Objetivos específicos del Procesamiento de Datos

La información recibida debe ser analizada y estudiada previamente. Para esto, se va a realizar un análisis exploratorio con el fin de entender cómo están distribuidos

los datos, si son coherentes, y si hay datos perdidos. Este proceso se realiza con el propósito de tener la base de datos completa y las variables y las observaciones sean apropiadas.

1.2.2. Objetivos específicos del Análisis Descriptivo

En este análisis, se busca confeccionar un resumen de la información que otorgan los datos, donde lo principal es:

- Representar la información a través de gráficos y medios visuales.
- Extraer las características más representativas.
- Describir tendencias.
- Hallar relaciones simples entre variables.

1.2.3. Objetivos específicos del Análisis de Conglomerados

Dentro del Aprendizaje Automático (*Machine Learning*) no supervisado, se encuentra el Análisis de Conglomerados, el cual etiqueta las observaciones formando distintos grupos (o *clusters*). La idea principal de este análisis, es trabajar los proyectos liderados por hombres como una base de datos diferente a los proyectos liderados por mujeres, para así reconocer si se agrupan de forma similar o no, e identificar qué características influyen en la formación de *clusters* en cada caso.

1.2.4. Objetivos específicos del Análisis de Regresión Logística Múltiple

La construcción del modelo de Regresión Logística Múltiple es una técnica de clasificación binaria que busca estimar la probabilidad de que cierto investigador sea hombre o mujer, en base a sus variables vinculadas a jerarquía e investigación. Además, se aplica un método de selección de variables *boruta*, que mejora el modelo e identifica las variables que tienen mayor influencia sobre esta clasificación de proyectos por género.

Capítulo 2

Marco Teórico

En este capítulo se presentan las principales definiciones y conceptos a utilizar, teoría para estructurar adecuadamente los datos, funcionamiento de modelos, y técnicas estadísticas utilizadas.

2.1. Procesamiento de Datos

Previo a realizar un análisis de los datos, es necesario un proceso de preparación. Esto consiste en utilizar estrategias de limpieza, llenado y transformación (*ETL* por sus siglas en inglés *Extract, Transform and Load*), para posteriormente trabajar con los datos bien estructurados en el análisis.

2.1.1. Análisis Exploratorio de Datos

El Análisis Exploratorio de Datos es un proceso de investigación, cuya finalidad es conocer los datos. Dentro de los objetivos típicos está comprender la distribución de las variables representadas en el conjunto de datos. Además, se pueden encontrar anomalías como valores atípicos u observaciones inusuales, revelar patrones, y comprender posibles relaciones entre variables.

En caso de existir datos atípicos, hay que tomar decisiones sobre qué hacer con ellos, donde las principales opciones son sustituir o descartar. Al sustituir los valores, estos se reemplazan (según cierto criterio) a algún valor cercano, en cambio al descartar, se eliminan estos valores como si no existieran.

2.1.1.1. *Outliers*: Casos atípicos

Los *outliers* o casos atípicos son observaciones que toman valores que, para una o más variables, son muy diferentes al resto de los elementos de la muestra. Estos datos pueden causar problemas en la interpretación de lo que ocurre en una población.

Si bien los datos atípicos distorsionan los resultados del análisis, eliminarlos no es la mejor solución posible, por lo que es necesario identificarlos y tratarlos de forma adecuada. Pueden excluirse si se trata de un error en la construcción de la base de datos o en la medición de la variable, pero de no ser así, son datos potencialmente interesantes en la detección de anomalías. Por tanto, al eliminarlos se podrían modificar los resultados y afectar gravemente a la normalidad, que es una de las condiciones más habituales requeridas para el uso de las técnicas multivariantes.

Una de las formas de determinar si los valores son o no atípicos, es mediante el método de las bisagras de Tukey. Esta técnica estadística, desarrollada por John Tukey, [Tukey, 1977], detecta los valores atípicos en una base de datos. Para aplicar el método, se definen las bisagras de Tukey:

$$\begin{aligned} \text{Bisagra superior} &:= Q_3 + 1{,}5 \times IQR \\ \text{Bisagra inferior} &:= Q_1 - 1{,}5 \times IQR, \end{aligned} \tag{2.1.1}$$

donde IQR es el rango intercuartílico:

$$IQR = Q_3 - Q_1 \tag{2.1.2}$$

Luego, el método de las bisagras de Tukey está descrito siguiendo el Algoritmo 1:

Algorithm 1 Algoritmo del método de las bisagras de Tukey

Require: `database`

1: Ordenar los datos de la `database` de menor a mayor
2: Calcular los cuartiles Q_1, Q_2 y Q_3 de los datos
3: Calcular el rango intercuartículo (IQR)
4: Calcular las bisagras superior e inferior de Tukey
5: Cualquier valor que esté fuera de las bisagras se considera *outlier*

Para visualizar gráficamente los *outliers* utilizando el método de las bisagras de Tukey, se puede utilizar un diagrama de caja (o *boxplot*):

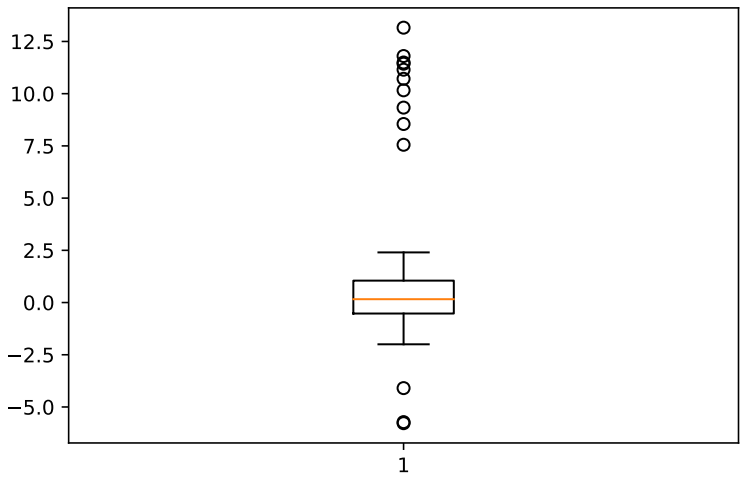


Figura 2.1.1: Ejemplo de Diagrama de caja
(Fuente: Elaboración propia)

En la Figura 2.1.1, se puede observar un *boxplot*, donde el rectángulo central representa los datos entre Q_1 y Q_3, y la línea de color naranjo dentro del rectángulo representa la mediana. Los bigotes (o *whiskers*) representan los valores extremos que están dentro del rango de las bisagras de Tukey. Los valores que se muestran como puntos fuera de las bisagras, son los datos atípicos.

La idea detrás del método de las bisagras de Tukey es que los valores atípicos se alejan significativamente de la mayoría de los datos, lo que resulta en una mayor variabilidad y una dispersión más amplia de los datos. La bisagra superior e inferior de Tukey se definen en función del rango intercuartílico, lo que significa que se basan en la distribución de los datos y no en un umbral arbitrario.

Los valores atípicos se pueden clasificar en dos grandes grupos [Olmo and Mateu, 2003]:

- Verdaderos *outliers*: Forman parte del conjunto de datos, aunque son observaciones que difieren con la tendencia general, tienen una explicación. Se puede intentar suavizar su influencia transformando las variables mediante raíces cuadradas o logaritmos, o también se puede intentar aplicar técnicas de análisis estadístico que sean lo más 'robustas' posibles frente a valores atípicos.
- Falsos *outliers*: No son datos reproducibles y suelen deberse a errores computacionales, de medición, o de transcripción. Para corregir esto, lo

más recomendable es reemplazar el valor o eliminarlo.

Una vez identificados los *outliers*, se pueden tomar decisiones sobre cómo tratarlos en el análisis estadístico. Es importante tener en cuenta que los datos atípicos pueden ser legítimos y representar verdaderas anomalías en los datos, por lo que su eliminación debe ser cuidadosamente considerada y justificada.

2.1.1.2. *Missing Data*: Datos perdidos

En las investigaciones para aplicar técnicas multivariadas, es muy frecuente encontrar *missing Data*, es decir, matrices con valores perdidos. Los motivos de que existan datos faltantes pueden deberse a distintas razones: causas humanas (sobreescritura, eliminación intencional o accidental), errores del sistema, virus, actualizaciones fallidas, cualquier tipo de daño físico del medio de almacenamiento, o simplemente datos con campos desconocidos que no se pudieron completar.

Para solucionar este problema, se aplica un mecanismo de imputación de datos (incluyendo los que antes fueron *outliers*).

2.1.2. Imputación de datos

Dentro de las estrategias más comunes para solucionar el problema de datos faltantes, está la imputación de datos [Little and Rubin, 2014].

Debido a que las mediciones están generalmente altamente correlacionadas, es común que los procedimiento de imputación utilicen información de períodos anteriores o del conjunto de variables para sustituir los datos perdidos. Existen diferentes formas de manejar los datos que faltan. La sustitución de datos faltantes utilizando valores de media, mediana o moda, es una práctica bastante común entre investigadores, pero no se considera un procedimiento apropiado, ya que pueden desperdiciar datos valiosos o reducir la variabilidad del conjunto de datos. Por el contrario, el modelo del vecino más cercano, o *K-Nearest Neighbors (KNN)*, ha demostrado ser efectivo en los experimentos [Kuhn and Johnson, 2013], el cual para procesos de imputación se conoce como Imputación del vecino más cercano o Imputación *KNN*.

El algoritmo *KNN* reemplaza los valores faltantes por valores estimados de los vecinos más cercanos, manteniendo el valor y la variabilidad de sus conjuntos de datos, lo que es más preciso y eficiente que usar los valores promedio. Para

un elemento nuevo en un conjunto de datos, el algoritmo calcula la distancia entre éste y cada uno de los datos existentes, y ordena las distancias de forma ascendente para poder determinar a qué grupo pertenece. La medida de distancia más utilizada es la distancia euclidiana, descrita a continuación para $\mathbb{R}^n$ (2.1.3), que mide una línea recta entre ambos puntos:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2.1.3}$$

El valor K en el algoritmo KNN define cuántos vecinos se verificarán para determinar la clasificación de un punto de consulta específico. Los valores muy bajos de K pueden tener una varianza alta y un sesgo bajo, y los valores muy altos pueden generar un sesgo alto y una varianza más baja. Por lo tanto, la elección óptima de K depende de varios factores, como el tamaño del conjunto de datos, la complejidad del modelo y la distribución de los datos. Una regla empírica que se ha utilizado en la práctica, es considerar la raíz cuadrada del número total de observaciones como un buen número de vecinos para el método KNN [Alomari and Diabat, 2012], [Sharma and Singh, 2015], es decir:

$$K = \sqrt{N}, \tag{2.1.4}$$

donde N es el número total de observaciones.

Si bien esta regla empírica es una buena opción, para una mayor rigurosidad se puede complementar encontrar el valor óptimo de K con una técnica de validación cruzada. La validación cruzada permite evaluar el rendimiento del modelo KNN con diferentes valores de K y seleccionar el valor que minimice el error de predicción en los datos de prueba. Una métrica común para evaluar el rendimiento del modelo KNN es el Error Cuadrático Medio (*RMSE* por sus siglas en inglés *Root Mean Square Error*), que mide la diferencia cuadrática entre los valores imputados por el modelo y los valores reales en los datos de prueba, calculado de la siguiente manera:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{2.1.5}$$

donde, en la ecuación (2.1.5), $\hat{x}_i$ es el valor estimado de x_i, y N es el número total de observaciones.

Luego, se puede encontrar un valor óptimo para K siguiendo el Algoritmo 2:

Algorithm 2 Algoritmo para encontrar un valor óptimo de K

Require: `data`, $a, b \in \mathbb{N}, a < b$

1: Eliminar temporalmente los datos nulos de `data`
2: Separar los datos en `data_imput(size=0.8)` y `data_miss(size=0.2)`
3: **for** K en $[a, b]$ **do**
4: Imputar datos de `data_imput` con K vecinos
5: Calcular el *RMSE* entre los nuevos valores de `data_imput` y `data_miss`
6: **end for**
7: Seleccionar el valor de K correspondiente al $\text{mín}_{\forall K \in (a,b)}\{RMSE\}$

Donde en Algoritmo 2, `data` es la base de datos utilizada, y $[a, b]$ es el intervalo al cual pertenece el número K de vecinos, siendo éste un intervalo cercano a la raíz cuadrada del número total de observaciones.

Así, se puede obtener un valor para K, considerando el que tenga un menor *RMSE*.

2.2. Análisis Descriptivo

Este análisis puede ayudar a comprender algún problema o fenómeno, averiguar rasgos, y recolectar y ordenar información para posteriormente poder describir las relaciones que se dan entre las variables.

En este caso, como la variable 'Género' es la relevante en este estudio, lo más conveniente es realizar un análisis descriptivo bivariado entre 'Género' y todas las demás variables, con la finalidad de determinar si existen diferencias por género en las categorías.

Antes de poder realizar estos análisis, se debe realizar una exploración de los datos [Aldas Manzano and Uriel Jimenez, 2017]. Esto permite visualizar si los datos poseen algún comportamiento inusual, identificar áreas para profundizar, y determinar si satisfacen los supuestos necesarios para realizar el análisis estadístico.

2.2.1. Análisis Bivariado

El Análisis Descriptivo Bivariado estudia la relación entre pares de atributos medidos simultáneamente, que incluye un conjunto de herramientas enfocado en el análisis de dos variables, con el objetivo de determinar las relaciones empíricas entre ellas. Este análisis puede determinar en qué medida es posible predecir el valor de una variable en caso que conozcamos el valor de la otra, es decir, el estudio de la correlación entre dos variables se refiere a un conjunto de relaciones estadísticas que involucran una dependencia entre ellas. Esta relación puede ser representada en manera visual o a través de un conjunto de medidas. En particular, el objetivo de este proyecto es estudiar las relaciones y diferencias entre 2 géneros, por lo que la finalidad es comprender la distribución de valores para cada variable, pero considerando una distinción de los datos según 'Género'. Dentro de las maneras más conocidas para realizar análisis y visualizaciones se encuentran:

- Estadísticas de resumen: Se pueden calcular medidas de tendencia central (media, mediana y moda), medidas de dispersión (rango, rango intercuartílico, desviación estándar), y medidas de localización (cuartiles, deciles), lo que nos indica dónde se encuentra el valor central y ciertos umbrales de interés, y qué tan dispersos están los valores para esa variable.
- Gráficos de distribución de valores: Para variables cuantitativas se pueden generar diagramas de violín, histogramas, gráficos de líneas, entre otros. Por otro lado, para las variables categóricas se crean gráficos de barras (apiladas) y gráficos circulares. En ambos casos considerando el atributo 'Género'.
- Matriz de correlación: Para analizar simultáneamente todos los posibles pares de variables cuantitativos, se calcula el coeficiente de correlación lineal de Pearson entre todos ellos, generando así una matriz de correlación [Hair et al., 2009]. La correlación de Pearson se calcula de la siguiente manera:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \text{ , con } -1 \leq \rho(X,Y) \leq 1 \tag{2.2.1}$$

  Donde si $\rho(X,Y)$ tiene valor -1, entonces las variables X e Y tienen una correlación perfecta negativa, por lo que si una variable aumenta la otra disminuye, en cambio, en cambio, si el valor de $\rho(X,Y)$ es 1, tienen correlación perfecta positiva, es decir, si una aumenta la otra también lo

hace, y si la correlación es 0, entonces no existe relación lineal entre ambas variables.

2.3. Análisis de Conglomerados

El Aprendizaje Automático o *Machine Learning* es una Inteligencia Artificial que implica la construcción de modelos matemáticos para ayudar a comprender la información, permitiendo al sistema aprender directamente de los datos.

Los métodos de *Machine Learning* se clasifican en aprendizaje supervisado y en aprendizaje no-supervisado. El aprendizaje supervisado modela la relación entre las características medidas de los datos y una etiqueta asociada, en cambio, el aprendizaje no-supervisado es un aprendizaje sin etiquetas que busca descubrir patrones en los datos.

El Análisis de Conglomerados o *Cluster Analysis*, es una técnica basada en aprendizaje no-supervisado que tiene por objetivo agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. En este trabajo, se estudiará el Método clásico de partición [Peña, 2002], el cual divide los datos en un número de grupos (*clusters*) ya predeterminados, de manera que:

- Cada elemento pertenezca a uno y a sólo uno de los grupos;
- Cada grupo sea internamente homogéneo.

Las muestras se pueden agrupar utilizando K-Medias o *K-Means*, donde K es el número de grupos prefijado, que a partir de este momento será llamado G para evitar confusiones con el número de vecinos de la sección anterior.

2.3.1. Algoritmo *K-Means*

El algoritmo *K-Means* es un método que permite asignar a cada observación el *cluster* al que se encuentra más cercano. La métrica utilizada es la distancia euclidiana (2.1.3), la cual calcula la distancia entre la observación y los centros de los G grupos, denominados centroides. Los centroides se calculan utilizando también la distancia euclidiana (2.1.3), y es un punto equidistante de los objetos pertenecientes a él.

El algoritmo *K-Means* requiere de las etapas mostradas en el Algoritmo 3:

Algorithm 3 Algoritmo *K-Means*

Require: $x_1, \ldots, x_n$ observaciones

```
Seleccionar m puntos arbitrarios como centroides iniciales.
Sean G'_1, ..., G'_m grupos vacíos.
flag ← 1
while flag == 1 do
    for i en [1, n] do
        Seleccionar ω tal que d(x_i, G_ω) = mín_{j∈[1,m]} d(x_i, G_j).
        Asignar x_i al grupo G_ω.
    end for
    if G_j == G'_j para todo j ∈ [1, m] then
        flag ← 0
    else
        G'_j ← G_j para todo j ∈ [1, m]
        Recalcular los m centroides.
```

Donde ω es el centroide del *cluster* G_ω. Notar que, el algoritmo se repite hasta que los conglomerados sean iguales a los de su iteración anterior (es decir, hasta que los *clusters* se estabilicen).

Así, cada observación pertenece a un grupo G, dado un número prefijado de grupos.

2.3.2. Número de grupos

Uno de los aspectos más importantes para utilizar el algoritmo *K-Means*, es encontrar un valor óptimo para el hiperparámetro G. Para escoger un número adecuado de *clusters*, se puede utilizar el conocido Método del codo (o *Elbow*).

Sea I la inercia de un *cluster*, definido de la siguiente manera:

$$I = \sum_{i=1}^{n} ||x_i - \omega||^2 \tag{2.3.1}$$

La cual calcula la suma de errores cuadráticos dentro del *cluster* (o *intra-cluster*), donde x_i son las observaciones, y ω el centroide del *cluster* al cual pertenece [Pertuz, 2022].

El Método del codo funciona de la siguiente manera: se calcula la inercia para diferentes valores de G, se representan gráficamente los resultados obtenidos, y se

identifica el punto de la curva donde a partir de él exista un cambio de pendiente (lo cual es visualmente como un codo), para determinar un número óptimo de *clusters*, ya que a partir de ese valor agregar más *clusters* no tiene una mejora significativa, es decir, la disminución de la inercia no es idónea.

2.3.3. Análisis de Componentes Principales

En algunos casos en el análisis de datos, cuando existe un número grande de variables, es posible describir con precisión las p variables utilizando un subconjunto de las mismas, a costa de una pequeña pérdida de información. Este proceso se denomina reducción de dimensionalidad [Peña, 2002].

El Análisis de Componentes Principales (PCA), es una técnica estadística de reducción de dimensionalidad, que convierte los datos de alta dimensión en datos de baja dimensión seleccionando las características más importantes. El concepto de mayor información está relacionado con el de mayor variabilidad o varianza, es decir, cuanto mayor sea la variabilidad de los datos se considera que existe mayor información. El PCA utiliza una transformación lineal ortogonal para convertir un conjunto de variables en un conjunto de variables linealmente no correlacionadas, llamadas componentes principales.

Sean $y_1, \ldots, y_p$ combinaciones lineales de las $x_1, \ldots, x_p$ variables originales, es decir,

$$y_j = \sum_{i=1}^{p} a_{ij} x_i = a_j^T x, \forall j \in [1, p], \tag{2.3.2}$$

donde $a_j = (a_{1j}, \ldots, a_{pj})^T$ es un vector de constantes, y $x = (x_1, \cdots, x_p)$.

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima. Los valores de este primer componente se representarán por el vector $y_1 = a_1 x$. Notar que, si las variables originales tienen media cero, entonces y_1 también tendrá media cero. Por lo tanto, su varianza es:

$$Var(y_1) = \frac{1}{n} y_1^T y_1 = \frac{1}{n} a_1^T x^T x a_1 = a_1^T S a_1, \tag{2.3.3}$$

donde n son las observaciones totales, $S = \dfrac{x^T x}{n}$ la matriz de varianzas y covarianzas de las observaciones. Para mantener la ortogonalidad de la transformación, se

impone que:

$$a_j^T \cdot a_j = 1 \tag{2.3.4}$$

Así, para maximizar (2.3.3), se aumenta el módulo del vector a_1 con la condición (2.3.4). Para maximizar una función de varias variables sujeta a una restricción se puede utilizar el método de multiplicadores de Lagrange. Sea f_1 la función y g_1 la restricción, definidas como sigue:

$$f_1(a_1) = a_1^T S a_1 \tag{2.3.5}$$

$$g_1(a_1) = a_1^T a_1 - 1 \tag{2.3.6}$$

Se debe cumplir que:

$$\begin{aligned} \frac{\partial f_1}{\partial a_1} &= \lambda \frac{\partial g_1}{\partial a_1} \\ \implies 2Sa_1 &= \lambda\,(2a_1) \\ \implies Sa_1 &= \lambda a_1 \end{aligned} \tag{2.3.7}$$

Así, a_1 es un vector propio de la matriz S, con valor propio λ. Notar que, como $a_1^T a_1 = 1$, al multiplicar a_1^T por la izquierda se tiene:

$$a_1^T S a_1 = \lambda a_1^T a_1 = \lambda = Var(y_1) \tag{2.3.8}$$

Como (2.3.8) se quiere maximizar, λ será el mayor valor propio de la matriz S, es decir, se busca el atributo que tiene mayor covarianza con el primer componente principal, y el vector propio a_1 define los coeficientes de cada variable en este componente [Peña, 2002].

En el segundo componente principal, se considera como función objetivo la suma de las varianzas y_1 e y_2, donde a_1 y a_2 son los vectores que definen el plano. Es decir, la función objetivo es:

$$f_2(a_1, a_2) = a_1^T S a_1 + a_2^T S a_2 \tag{2.3.9}$$

Y sus restricciones serían $a_1^T a_1 = 1$ y $a_2^T a_2 = 1$. Al derivar f_2 por separado con respecto a a_1 y a a_2, igualando a cero y luego reemplazando, se tiene que λ_1 y λ_2 son los valores propios mayores de la matriz S, y a_1 y a_2 sus vectores propios. Así sucesivamente, se calculan los $y_1, \ldots, y_p$, de manera que cada variable obtenida tendrá menos varianza que la anterior.

2.4. Análisis de Regresión Logística Múltiple

Dentro del aprendizaje supervisado (que trabaja con datos ya etiquetados) se encuentran algoritmos de regresión y de clasificación, los cuales permiten predecir y clasificar variables.

Un modelo de predicción y clasificación es la Regresión Logística. La Regresión Logística es una técnica estadística de clasificación binaria, que mide la relación entre la variable dependiente (que se busca predecir), y las variables independientes. El objetivo es predecir la probabilidad de ocurrencia de la variable dependiente binaria y en función de las otras variables, definida de la siguiente manera:

$$y = \begin{cases} 1, \text{ si el evento ocurre,} \\ 0, \text{ si no.} \end{cases} \tag{2.4.1}$$

Se denomina Regresión Logística Múltiple cuando existe más de una variable independiente.

Los parámetros estimados mediante la Regresión Logística se conocen como *log of Odds*, que representan el logaritmo de la oportunidad de ocurrencia de un evento.

La oportunidad de ocurrencia de un suceso está definido como sigue en la ecuación (2.4.2):

$$Odds(y) = \frac{p(y)}{1 - p(y)} \tag{2.4.2}$$

Donde $p(y)$ corresponde a la probabilidad de que y tome el valor 1 (éxito). Es importante notar que, la oportunidad de ocurrencia *Odds* no es sinónimo de probabilidad. El cálculo de *Odds* compara el número de resultados deseados con

el número de posibles resultados no deseados, el cual puede ser un número en el intervalo $[0, +\infty]$, mientras que el cálculo de probabilidad considera todos los resultados potenciales de un evento, estando sus valores posibles entre $[0, 1]$.

La función que relaciona la variable dependiente con las independientes se conoce como función *sigmoide*, descrita en (2.4.3) y representada en la Figura 2.4.1, la cual puede tomar valores exclusivamente entre 0 y 1.

$$\phi(z) = \frac{1}{1 + e^{-z}} \tag{2.4.3}$$

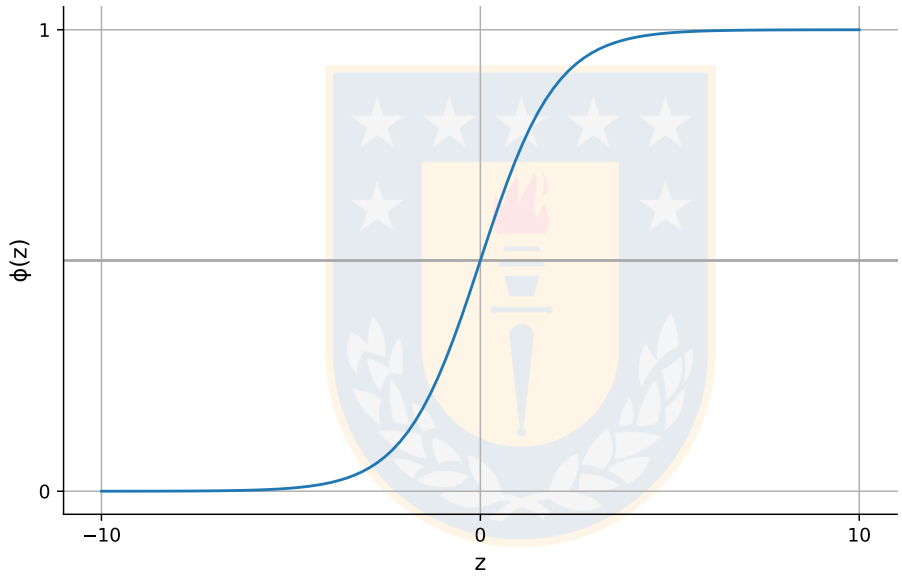


Figura 2.4.1: Función Sigmoide
(Fuente: Elaboración propia)

En la función (2.4.3) es posible notar que z es un número real, el cual cumple que

$$\lim_{z \to -\infty} \phi(z) = 0 \ \wedge \ \lim_{z \to +\infty} \phi(z) = 1 \tag{2.4.4}$$

Una variable cualitativa binaria se puede ajustar a un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, donde p es el número de variables independientes predictoras o *features* [Faraway, 2016]. La regresión logística transforma el valor devuelto por la regresión lineal empleando una función donde su resultado esté entre 0 y 1, en este caso, la función *sigmoide* (2.4.3).

Al reemplazar la función lineal $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ en la variable y de la función *sigmoide*, considerando X un vector de datos con valores $x_1, \ldots, x_p$, se obtiene:

$$\begin{aligned} P(y = 1|X = x) &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \\ \iff P(y = 1|X = x) &= \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \end{aligned} \tag{2.4.5}$$

donde (2.4.5) es la probabilidad de que la variable cualitativa y tome el valor 1 de una variable binaria, dado el predictor X.

Para que sea posible aplicar el modelo, se tienen que cumplir los siguientes supuestos [Bewick et al., 2005]:

1. La variable de respuesta debe ser binaria.
2. Las observaciones deben ser independientes entre sí.
3. El modelo no debe presentar multicolinealidad.
4. No deben existir valores atípicos extremos.
5. La relación entre las variables explicativas y el logaritmo de probabilidades (*logit*) debe ser lineal.
6. El tamaño de la muestra debe ser suficientemente grande.

La relación (2.4.5) no es lineal, siendo ésta una de las condiciones para poder aplicar el modelo. Para solucionar esto, es importante notar que se está trabajando con una clasificación binaria, por lo que la probabilidad de $y = 0$ dado $X = x$ es la probabilidad complementaria de $y = 1$ dado $X = x$:

$$\begin{aligned} P(y = 0|X = x) &= 1 - P(y = 1|X = x) \\ \iff P(y = 0|X = x) &= 1 - \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \\ \iff P(y = 0|X = x) &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \end{aligned} \tag{2.4.6}$$

Así, al dividir la probabilidad (2.4.5) entre (2.4.6) se obtiene:

$$\begin{aligned} \frac{P(y=1|X=x)}{P(y=0|X=x)} &= \frac{\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}} \\ \iff \frac{P(y=1|X=x)}{P(y=0|X=x)} &= e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} \end{aligned} \tag{2.4.7}$$

La expresión (2.4.7), al ser una función exponencial es siempre positiva, por lo que es posible aplicar las propiedades de logaritmo de la siguiente forma:

$$\ln\left[\frac{P\left(y=1|X=x\right)}{P\left(y=0|X=x\right)}\right] = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p \tag{2.4.8}$$

En la ecuación (2.4.8), $\ln\left[\frac{P(y=1|X=x)}{P(y=0|X=x)}\right]$ es la razón de probabilidad *log of Odds*, también conocida como el *logit* del modelo de Regresión Logística. La función *logit* curva la línea de mejor ajuste para transformar el problema de clasificación en un problema de regresión. Debido a la función *logit*, los coeficientes β de la Regresión Logística representan las probabilidades logarítmicas de que una observación esté en la clase objetivo, dados los valores de sus variables x_i.

Así, para obtener la probabilidad de que ocurra un evento se tiene que ajustar el modelo de Regresión Logística Múltiple (2.4.5) al conjunto de datos, estimando los valores $\vec{\beta} = [\beta_0, \ldots, \beta_p]^T$.

2.4.1. Selección de variables

Al momento de ajustar un modelo estadístico, el número de variables predictoras podría ser demasiado grande y no todas las variables necesariamente contribuyen al modelo. El problema de esto, es que además de no ser un número óptimo computacionalmente, las variables que no aportan al modelo pueden ser perjudiciales en la precisión e interpretabilidad.

Es por esto que, es importante identificar qué atributos realmente contribuyen al modelo y así también cuáles son irrelevantes o redundantes.

El algoritmo secuencial *boruta*, es un método de selección de variables que extiende la idea introducida por Stoppiglia H, Dreyfus G, Dubois R, Oussar Y [Stoppiglia et al., 2003], la cual busca determinar la contribución de cada variable comparando

la importancia de las características reales con mediciones aleatorias. En *boruta*, las variables x_i compiten con una versión aleatoria de ellas mismas, llamadas variables sombra o *shadow variables* y denotadas por $\tilde{x_i}$. La importancia que entrega *boruta* es un puntaje numérico calculado de la siguiente manera:

$$z_i = \frac{x_i - \breve{\mu}}{\breve{\sigma}}, \tilde{z}_i = \frac{\tilde{x_i} - \breve{\mu}}{\breve{\sigma}}, \tag{2.4.9}$$

donde $\breve{\mu}$ y $\breve{\sigma}$ corresponden a la media y a la desviación estándar de las importancias de las variables respectivamente. La estimación 2.4.9 es proporcionada utilizando un algoritmo de Bosque Aleatorio o *Random Forest*.

Se define

$$\gamma := \text{máx}\, \tilde{z}_i, \tag{2.4.10}$$

como la mayor importancia $\tilde{z}_i$ de las variables $\tilde{x_i}$.

En este método, se supone la variable x_i como valiosa si y sólo si tiene mayor z_i que γ.

El Algoritmo 4 *boruta* está descrito a continuación:

Algorithm 4 Algoritmo *boruta*

Require: x_i en `data`
flag $\leftarrow 1$
while flag $== 1$ **do**
    flag $\leftarrow 0$
    Para todo i, crear $\tilde{x_i}$
    Aleatorizar $\tilde{x_i}$ por columnas
    Ejecutar un modelo de *Random Forest* en `data` obteniendo z_i y $\tilde{z}_i$ para todo i
    Encontrar $\gamma = \text{máx}\, \tilde{z}_i$
    for i **do**
        if $z_i < \gamma$ **then**
            flag $\leftarrow 1$
            Eliminar x_i de `data`
        end if
    end for
    Eliminar $\tilde{x_i}$ de `data`
end while
return `data` $=0$

Notar que, el Algoritmo 4 tiene iteraciones hasta que ya no exista ninguna variable con importancia menor a γ, es decir, el algoritmo termina cuando $\forall i, z_i \geq \gamma$.

Por lo tanto, el número de variables al que se aplica el modelo de Regresión Logística es óptimo.

2.4.2. Estimación de parámetros

Sea la variable dependiente $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \hat{\varepsilon}$, donde $\beta_0, \ldots, \beta_p$ son los parámetros, $x_1, \ldots, x_p$ las variables independientes, $\hat{\varepsilon}$ el error residual tal que $\hat{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$, y además:

$$y \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2\right) \tag{2.4.11}$$

Así, y tiene distribución normal, con media $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ y varianza σ^2.

La combinación óptima de valores para los parámetros se puede calcular maximizando la probabilidad de que los puntos del conjunto de datos se clasifiquen correctamente, lo que se conoce como estimación de máxima verosimilitud [Johnson, 2000]. Antes de definir la función de máxima verosimilitud, es necesario definir algunos conceptos:

Sea $\theta = [\theta_1, \ldots, \theta_{n-1}]^T$ el vector de parámetros que gobiernan la distribución, tal que $\{f(y;\theta)|\theta \in \Theta\}$, donde Θ es el espacio de parámetros o *parameter space*, el cual es un subconjunto de dimensión finita del espacio Euclidiano.

Sea la función densidad $f : \mathbb{R}^n \to]0, +\infty[$ definida como:

$$f(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\dfrac{(y-\mu)^2}{2\sigma^2}}, \tag{2.4.12}$$

donde $y, \mu, \sigma \in \mathbb{R}$, $\sigma > 0$, y f depende de los parámetros μ y σ^2, correspondiente a la media y a la varianza respectivamente.

Se define la función de máxima verosimilitud de la siguiente manera:

$$L(\theta; y_i) = \prod_{i=1}^{n} f(y_i;\theta), \tag{2.4.13}$$

donde $L(\theta; y_i)$ es la multiplicación de todas las funciones de densidad (2.4.12) que dependen de las observaciones y_i y de los parámetros θ.

El objetivo de la estimación de máxima verosimilitud es encontrar los valores de los parámetros del modelo que maximizan la función de verosimilitud sobre el espacio de parámetros [Myung, 2003].

Para maximizar la función (2.4.13), se pueden aplicar propiedades de logaritmo:

$$\begin{aligned} \ell(\theta; y_i) &= \ln\left[L(\theta|y_i)\right] = \sum_{i=1}^{n} \ln\left[f(y_i;\theta)\right] \\ \implies \ell(\theta; y_i) &= \sum_{i=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2}{2\sigma^2} \right] \end{aligned} \tag{2.4.14}$$

Dado que el logaritmo es una función monótona creciente, el máximo de (2.4.14) tiene el mismo valor de θ que el máximo de (2.4.13).

La función $\ell(\theta; y_i)$ es diferenciable en Θ, por lo que para encontrar un máximo se debe derivar esta función respecto a los parámetros θ [Kane, 1968], obteniendo el siguiente sistema de ecuaciones:

$$\left.\begin{aligned} \frac{\partial}{\partial\beta_0}\left(\ell(\theta)\right) &= 0 \\ \frac{\partial}{\partial\beta_1}\left(\ell(\theta)\right) &= 0 \\ \vdots \quad \vdots \quad &= \vdots \\ \frac{\partial}{\partial\beta_p}\left(\ell(\theta)\right) &= 0 \\ \frac{\partial}{\partial\sigma^2}\left(\ell(\theta)\right) &= 0 \end{aligned}\right\} \tag{2.4.15}$$

Finalmente, resolviendo (2.4.15) se obtiene la estimación de parámetros.

2.4.3. Métricas de desempeño

Para evaluar el desempeño del modelo, se pueden comparar los datos predichos con los datos verdaderos de la muestra de validación (*test*), y una de las maneras más utilizadas es creando una matriz de confusión, para examinar la cantidad de observaciones predichas de forma correcta.

Verdadero Positivo (VP)	Falso Negativo (FN)
Falso Positivo (FP)	Verdadero Negativo (VN)

Cuadro 2.4.1: Matriz de confusión
(Fuente: Elaboración propia)

La diagonal principal del Cuadro 2.4.1 reporta los casos que fueron predichos de forma exitosa.

Se utilizan las siguientes métricas para interpretar de mejor forma los resultados [Kuhn and Johnson, 2013]:

- *Accuracy* (o exactitud): Mide el porcentaje de casos predichos correctamente por sobre el total de casos (2.4.16):

$$\frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{2.4.16}$$

- *Precision* (o precisión): Mide la fracción de predicciones correctas entre las etiquetas positivas. Valores altos significan que el algoritmo predice más resultados relevantes que irrelevantes (2.4.17):

$$\frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.4.17}$$

- *Recall* (o sensibilidad): Mide la fracción de verdaderos positivos predichos por el modelo. Valores altos significan que el algoritmo logra predecir la mayoría de los resultados relevantes (2.4.18):

$$\frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.4.18}$$

- *F1-score* (o valor F1): Representa la media armónica entre *Precisión* y *Recall* (2.4.19):

$$\frac{2 \times \textit{Precisión} \times \textit{Recall}}{\textit{Precisión} + \textit{Recall}} \tag{2.4.19}$$

Capítulo 3

Procesamiento de Datos

En este capítulo, se explica la metodología de trabajo. La base de datos fue trabajada previamente y entregada en formato .xlsx. Para estudiar los datos, graficar, implementar modelos, y todos los análisis necesarios, se utilizará el software Python. Además, como apoyo externo para gráficos interactivos, se utilizará la herramienta Power BI.

3.1. Base de datos

La base de datos utilizada está compuesta por variables numéricas (N) y variables categóricas (C), definidas a continuación:

- (N) `Año Presentación`: Indica el año en el que proyecto fue presentado al respectivo concurso.
- (N) `Año Aprobación`: Indica el año en que el proyecto fue aprobado.
- (N) `Año Inicio`: Indica el año en que comienza la realización del proyecto.
- (N) `Año Término`: Indica el año en que se estima el término del proyecto.
- (N) `Año Término Real`: Indica el año en que finalizó el proyecto.
- (N) `Valor Pesos`: Indica el monto adjudicado para la realización del proyecto.
- (C) `Liderado por UdeC`: Indica si el proyecto es o no liderado por la Universidad de Concepción.

- (C) `Estado:` Indica la etapa en que se encuentra actualmente el proyecto.
- (C) `Clasificación:` Indica si el proyecto es UdeC, nacional o internacional.
- (C) `Subclasificación:` Indica el nombre del organismo que financió el proyecto.
- (C) `Interdisciplinario:` Indica si el proyecto es o no interdisciplinario.
- (C) `Sexo:` Indica el sexo del/la líder del proyecto.
- (N) `Edad Participante:` Indica la edad del/la líder del proyecto.
- (C) `Vigencia UdeC Participante:` Indica la vigencia.
- (C) `Cargo Funcionario:` Indica el cargo que tiene el/la líder del proyecto dentro de la Universidad de Concepción.
- (C) `Es Académico Funcionario:` Indica si el/la líder del proyecto es académico/a dentro de la Universidad de Concepción.
- (C) `Tipo Alumno:` Indica, en caso de que el/la líder del proyecto esté en calidad de alumno/a, si es de pregrado o de postgrado.
- (C) `Carrera/Programa/Repartición:` Indica el área específica a la cual pertenece el proyecto.
- (C) `Facultad/Organismo:` Indica el organismo al que pertenece el proyecto.
- (C) `Campus:` Indica a cuál campus de la Universidad de Concepción pertenece el proyecto.

Se han agregado además las siguientes variables:

- (N) `Duración:` Indica la duración esperada (en años) de cada proyecto, calculado como la diferencia entre `Año Inicio` y `Año Término`.
- (N) `Atraso:` Indica la diferencia (en años) entre el `Año Término Real` y el `Año Término`.

Finalmente, la información personal que no es relevante para el estudio, e incluso podría perjudicar los análisis, es descartada del DataFrame a través de la aplicación de la función de Python `drop`. Las columnas eliminadas corresponden a: `'Código Interno'`, `'Código VRID'`, `'Rut'`, `'Dv Rut'`, `'Nombre Completo'`, `'Nombres'`, `'Paterno'`, `'Materno'`, e `'Investigador Responsable'`.

3.2. Datos faltantes

Se examinarán, primero, todas las variables de la base de datos por separado (excluyendo las variables `Duración` y `Atraso`, ya que son calculadas a partir de otras variables), para estudiar la existencia de datos atípicos y manipularlos adecuadamente en cada ocasión.
En caso de ser errores evidentes de mecanografía, serán reemplazados. Por ejemplo, en la variable `Edad Participante`, existe la observación `'-58'`, y siendo la media de esta variable cercana a 50, el valor es reemplazado por `'58'`.
Otro ejemplo es en la variable `Año Presentación`, que contiene el valor `'7019'`. Éste será reemplazado por `'2019'`. Este reemplazo de datos se hará efectivo aplicando la función `replace` de `Pandas Dataframe` de Python.

Por otro lado, en esta base de datos también existe *Missing Data*, es decir, en la práctica, hay datos donde su valor es `NaN`.

Para estudiar esta situación, se utiliza la librería `missingno` de Python, la que proporciona herramientas para obtener un resumen visual de la integridad del conjunto de datos.

La matriz de nulidad muestra la densidad de datos presentes en la base, como se puede ver a continuación:

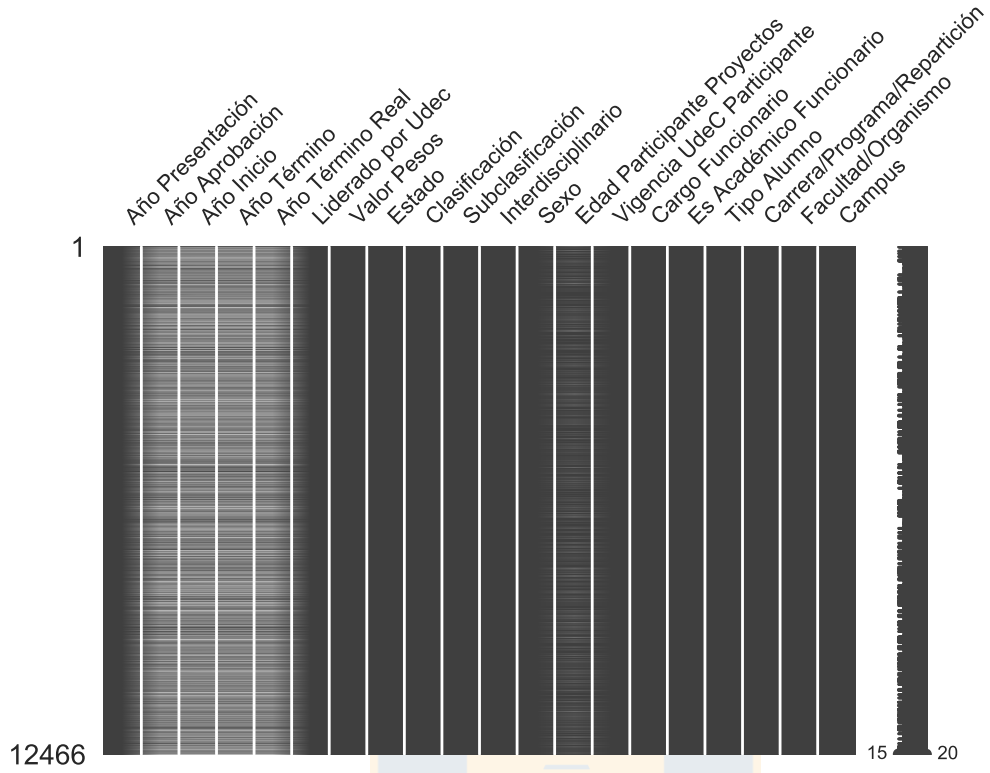


Figura 3.2.1: Matriz de nulidad

De la Figura 3.2.1, es posible notar que la información que incluye `Años` está en su mayoría incompleta, así como `Duración` y `Atraso` que dependen de ellos. Al lado derecho del gráfico se encuentra un resumen de la integridad de los datos, señalando el número de columnas que no presentan datos nulos (15) de las columnas totales (20). Otra manera de visualizar la existencia de datos perdidos es a través de un gráfico de barras:

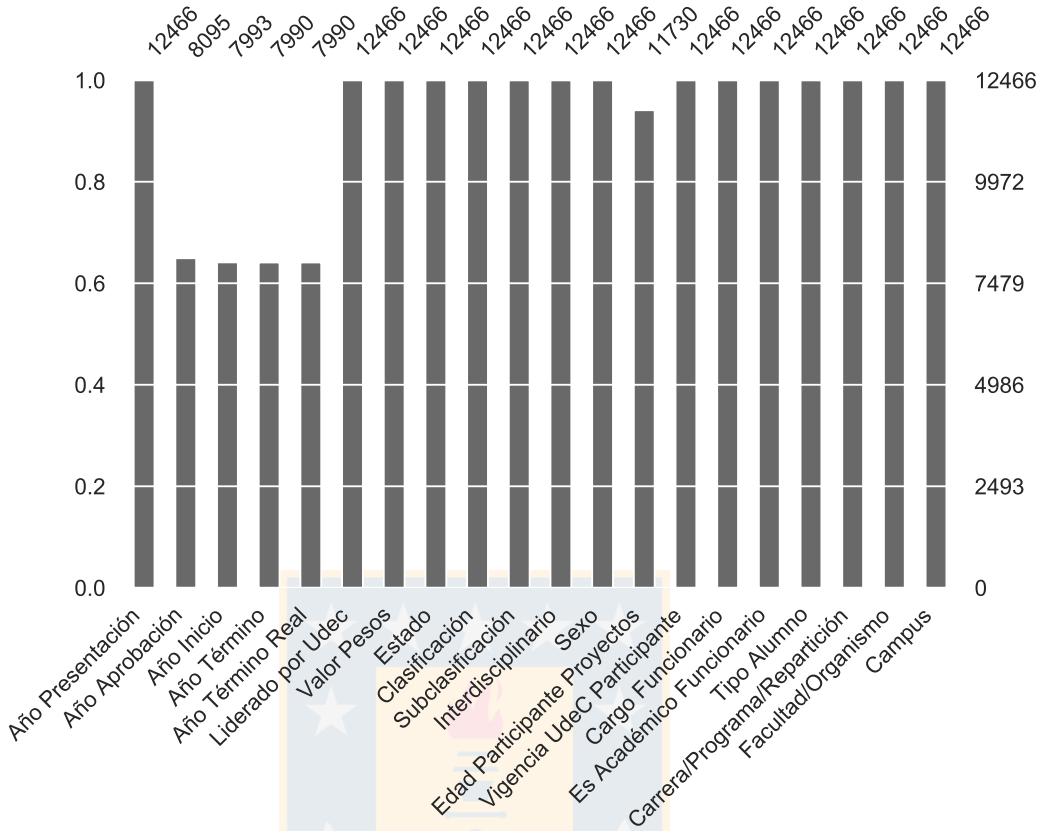


Figura 3.2.2: Gráfico de barras de nulidad

De la Figura 3.2.2 se puede ver con mayor detalle cuáles son las variables que tienen datos faltantes, descritos en la siguiente tabla:

Variable	N° de datos faltantes	% de datos faltantes
`Año Aprobación`	4371	35,06
`Año Inicio`	4473	35,88
`Año Término`	4476	35,91
`Año Término Real`	4476	35,91
`Edad Participante Proyectos`	736	5,90
`Duración`	4476	35,91
`Atraso`	4476	35,91

Cuadro 3.2.1: Cantidad y porcentaje de datos faltantes

Observando el Cuadro 3.2.1 se puede corroborar que las variables cuantitativas compuestas por años tienen un gran porcentaje de valores perdidos, disminuyendo la data verdadera hasta casi un 64 %.

Es importante verificar si la presencia o ausencia de una variable afecta la presencia de otra. Para esto, se genera un mapa de correlación de nulidad entre las variables:

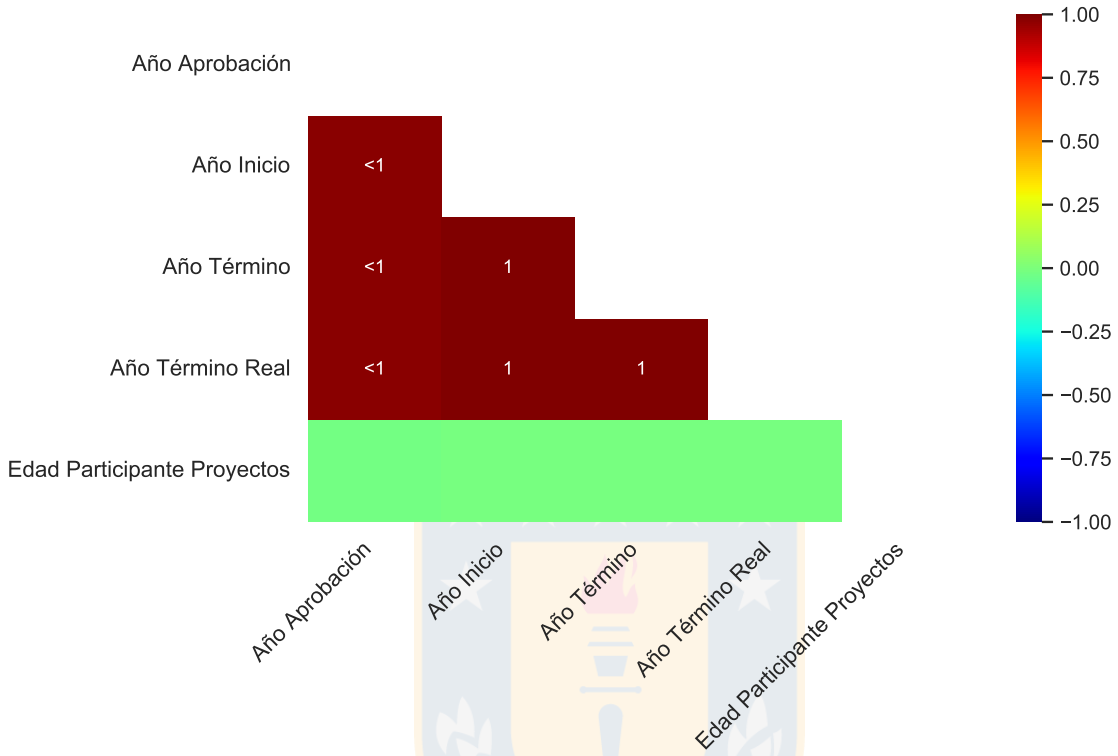


Figura 3.2.3: Matriz de correlación de nulidad

En la matriz de correlación de la Figura 3.2.3, la correlación de nulidad varía de `0` a `1`. Es posible notar que al correlación es cercana a `0` en la variable `Edad Participante Proyectos`, por lo que si esta aparece o no aparece no tiene efecto en las otras variables, en cambio, es cercana a `1` en las variables que incluyen el año, lo que significa que cuando una de estas variables `Año` aparece en la base de datos, las otras también lo hacen.

3.3. Imputación de datos

Antes de implementar este método, hay que ver las condiciones que se deben cumplir. Primero, `KNNImputer` no reconoce valores de datos de texto. Esto es relevante en la base de datos ya que existen muchas variables categóricas que contienen texto.

3.3.1. *Dummy variables:* Variables ficticias

En la base de datos entregada, existen datos categóricos, los cuales como su nombre indica, categorizan información. Una característica de estos datos, es que no tienen un orden. Para la mayoría de los métodos estadísticos, es necesario que los datos de entrada sean matrices numéricas, por lo que se deben transformar las variables categóricas a variables numéricas.

Por ejemplo, en la base de datos existe la variable categórica `Campus`, la cual indica la sede (o campus) UdeC al cual pertenece el proyecto. En esta categoría existen 4 posibles casos: `'Chillán'`, `'Concepción'`, `'Los Ángeles'` y `'No Aplica'`. Estos datos no tienen un orden jerárquico, es decir, es incorrecto asignarles valores numéricos para representarlos, ya que carece de sentido conceptual que `'Chillán'` se encuentre antes que `'Concepción'`, o que , `'No Aplica'` se encuentre después que `'Los Ángeles'`.

Para solucionar esto, se modifican los datos de texto realizando una codificación y creando variables ficticias [Anderson and Semmelroth, 2015], las cuales son variables numéricas que indican si una determinada condición es verdadera o falsa, donde por lo general se asigna un `0` si la condición es falsa, y un `1` si la condición es verdadera. Así, cada categoría se convierte en una nueva columna de datos binarios. Esto se lleva a cabo utilizando las funciones `get_dummies`, `drop` y `concat` de `Pandas`, (proceso también conocido como *one-hot encoding*).

Campus	⟶	Campus_CHILLÁN
		Campus_CONCEPCIÓN
		Campus_LOS ÁNGELES
		Campus_NO APLICA

Cuadro 3.3.1: Ejemplo de dataframe `Campus` con la transformación variables *dummies*

'Concepción'	⟶	0
		1
		0
		0

Cuadro 3.3.2: Ejemplo de `'Concepción'` como variable *dummy*

Al realizar este proceso con todas las variables cualitativas, la base de datos pasa

de tener 21 columnas, a tener 585 columnas, esto debido a la amplia cantidad de categorías.

3.3.2. Normalización

Por otro lado, `KNNImptuer` es un método de imputación basado en la distancia, y requiere que los datos estén normalizados. Para esto, se utiliza el método `MinMaxScaler` de `Scikit-Learn`, el cual escala las variables para que tengan valores entre `0` y `1`. El método de normalización utilizado es el denominado escalado de características, que viene dado por la ecuación (3.3.1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.3.1}$$

donde X' es el valor transformado, X el valor original, X_{min} el valor mínimo, y X_{max} el valor máximo.

3.3.3. Valor de K

Cuando el conjunto de datos tiene sólo variables numéricas y está normalizado (utilizando la ecuación (3.3.1)), se puede hacer la imputación por *KNN*. Se aplica el paquete de imputación `Scikit-Learn` a los datos. Debido a su simplicidad y a su facilidad de implementación, el parámetro K es calculado utilizando el Algoritmo 2 del capítulo anterior, donde el rango posible $[a, b]$ para el número de vecinos será considerado 50 unidades de distancia a la raíz cuadrada de la cantidad observaciones, aproximado al número impar más cercano, es decir:

$$\sqrt{12466} \approx 111, \tag{3.3.2}$$

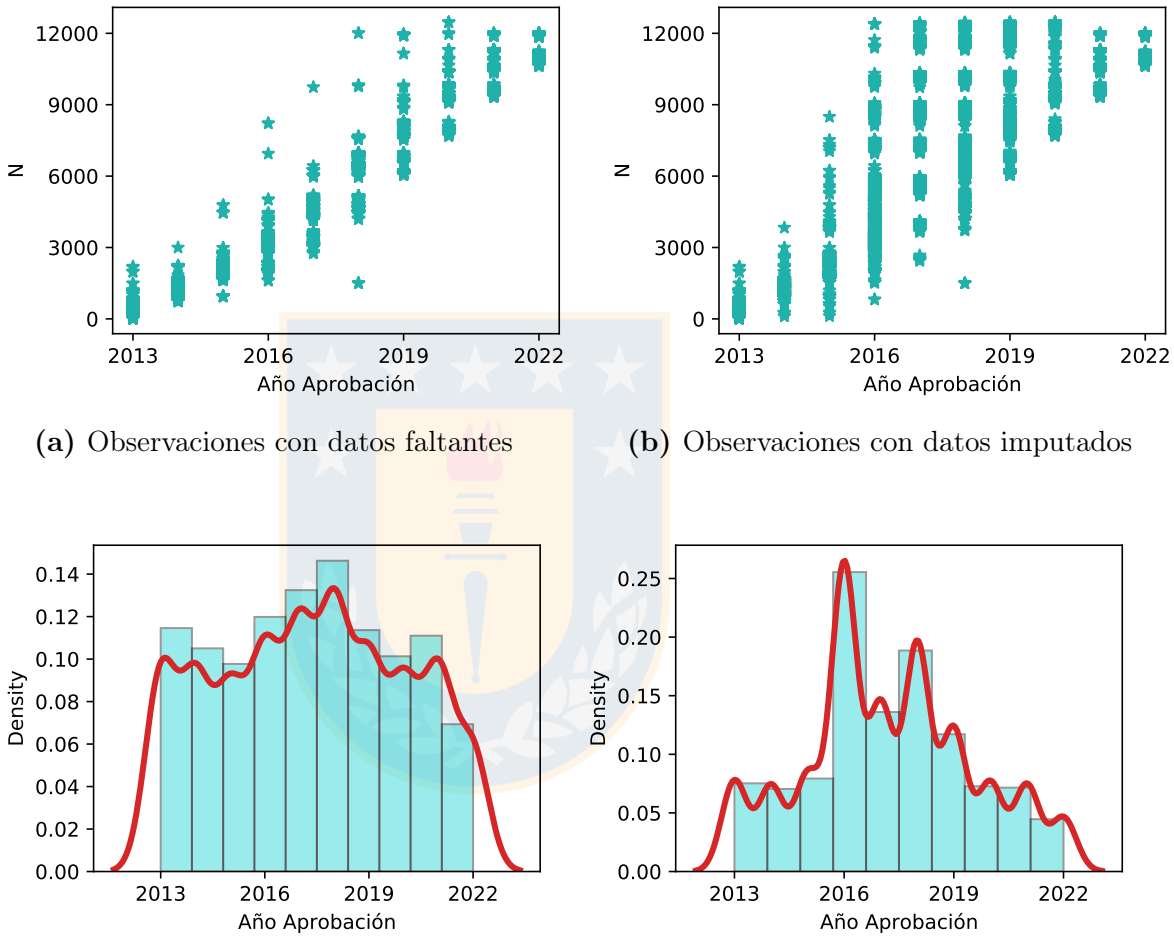
ya que un número muy grande podría reducir la variabilidad de los datos.

Al aplicar el Algoritmo 2 con valores $a = 61$ y $b = 161$, el valor óptimo entregado fue K= 119.

Así, la base de datos ya no tiene valores faltantes, ya que los valores faltantes se han imputado como la media de los valores de los *119-vecinos* más cercanos (medidos por la distancia euclidiana).

3.3.4. Datos imputados

Comparando la distribución de datos de la variables que tenía valores perdidos, con su distribución después de la imputación de datos, se obtienen los siguientes histogramas:



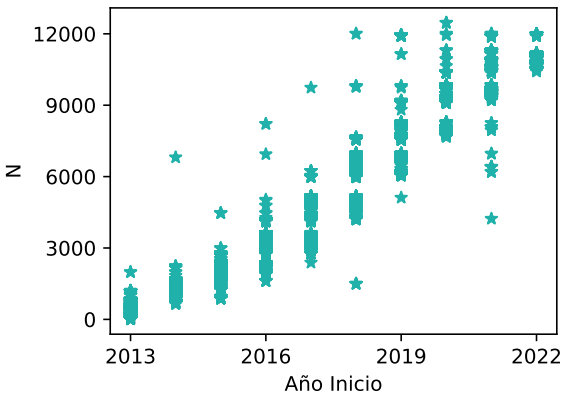
(a) Observaciones con datos faltantes

(b) Observaciones con datos imputados

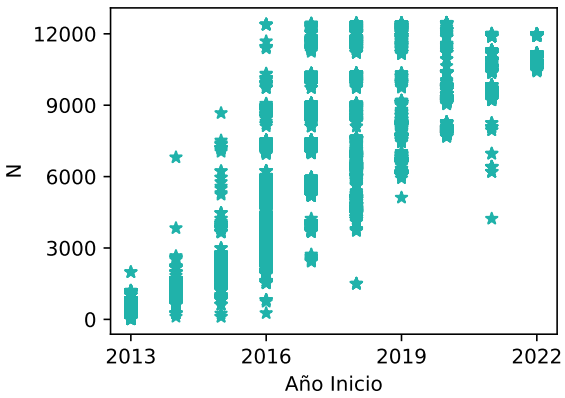
(c) Distribución con datos faltantes

(d) Distribución con datos imputados

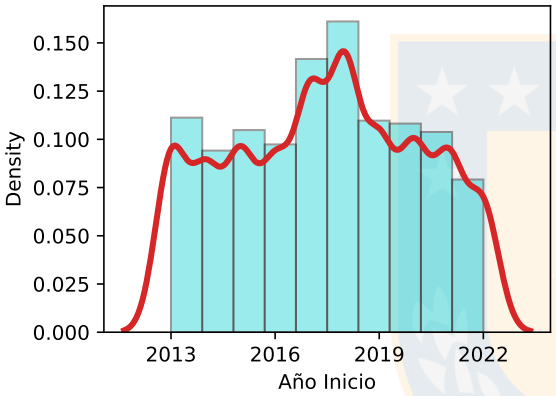
Figura 3.3.1: Proyectos según Año de Aprobación



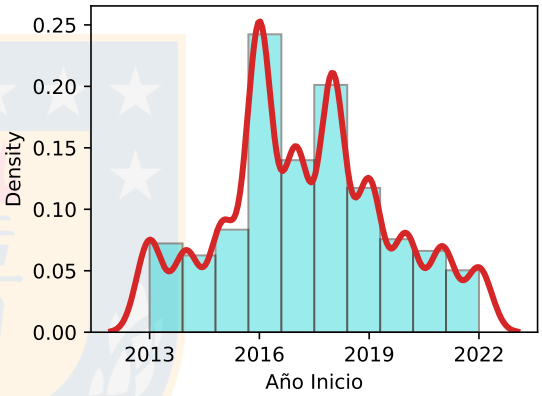
(a) Observaciones con datos faltantes



(b) Observaciones con datos imputados

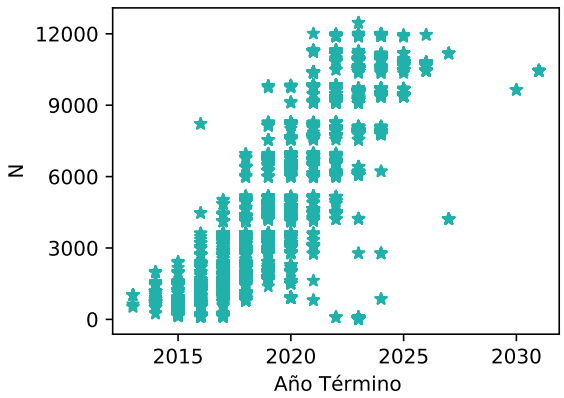


(c) Distribución con datos faltantes

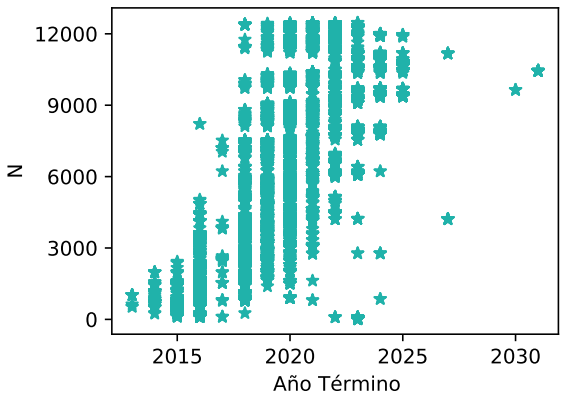


(d) Distribución con datos imputados

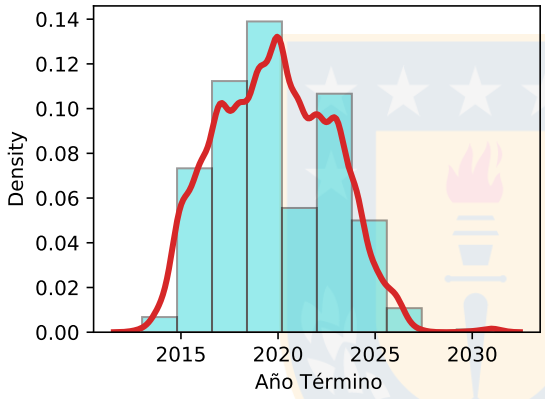
Figura 3.3.2: Proyectos según Año de Inicio



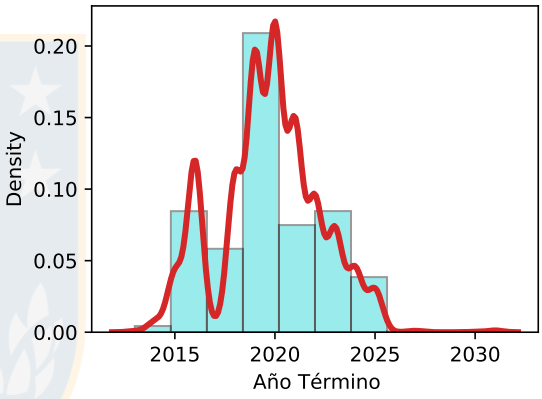
(a) Observaciones con datos faltantes



(b) Observaciones con datos imputados

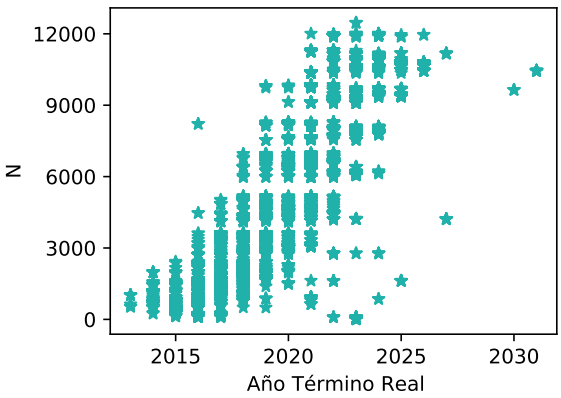


(c) Distribución con datos faltantes

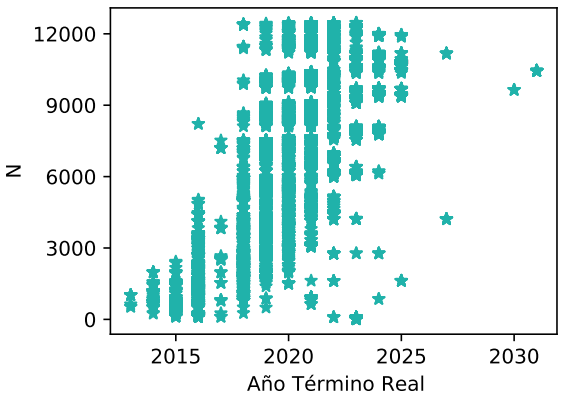


(d) Distribución con datos imputados

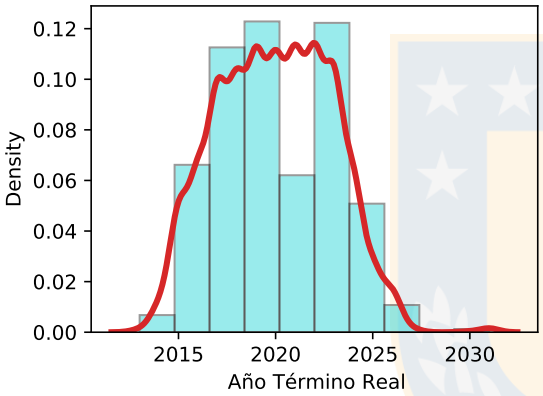
Figura 3.3.3: Proyectos según Año de Término



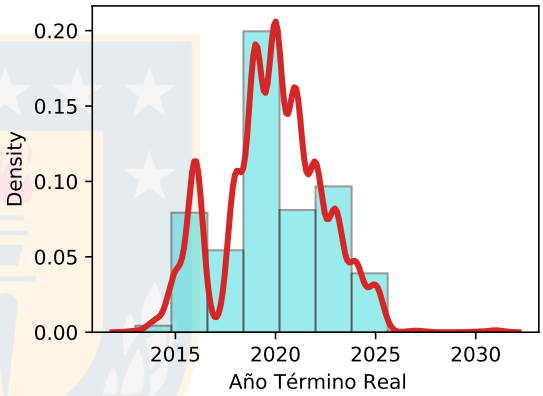
(a) Observaciones con datos faltantes



(b) Observaciones con datos imputados

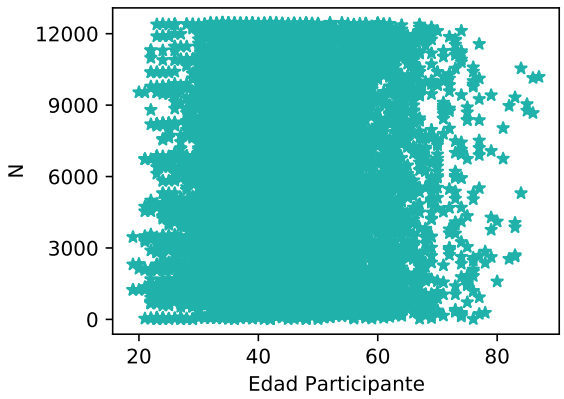


(c) Distribución con datos faltantes

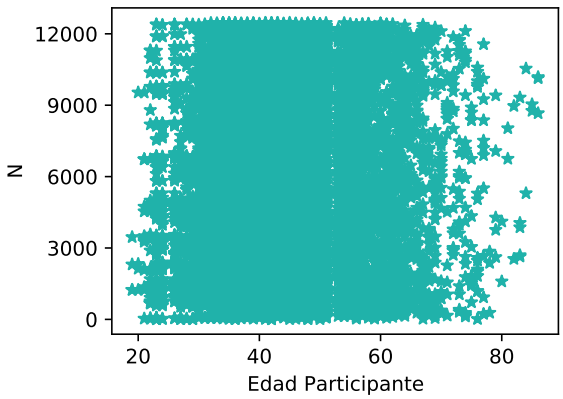


(d) Distribución con datos imputados

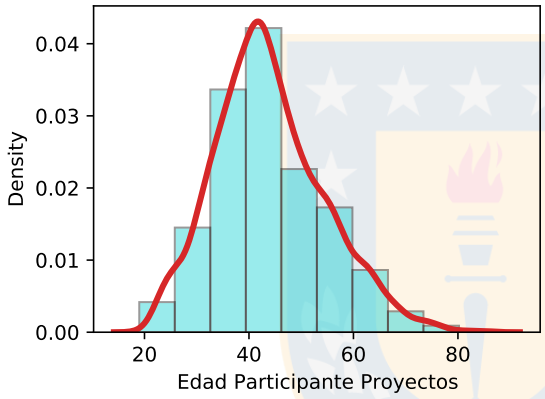
Figura 3.3.4: Proyectos según Año de Término Real



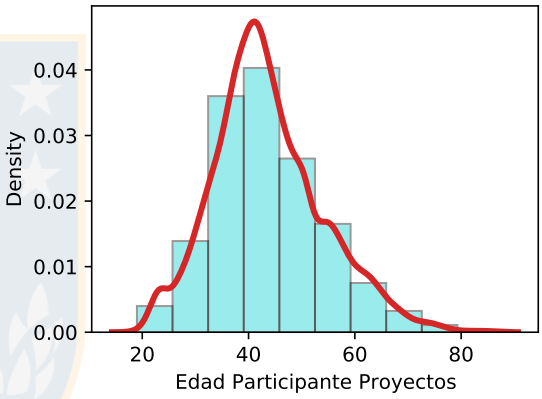
(a) Observaciones con datos faltantes



(b) Observaciones con datos imputados



(c) Distribución con datos faltantes



(d) Distribución con datos imputados

Figura 3.3.5: Proyectos según Edad de Participante

Así, la base de datos está completa, y el análisis de datos posterior se podrá realizar sin necesidad de disminuir muestras. Se puede observar que ya no existen datos vacíos en las Figuras 3.3.6 y 3.3.7:

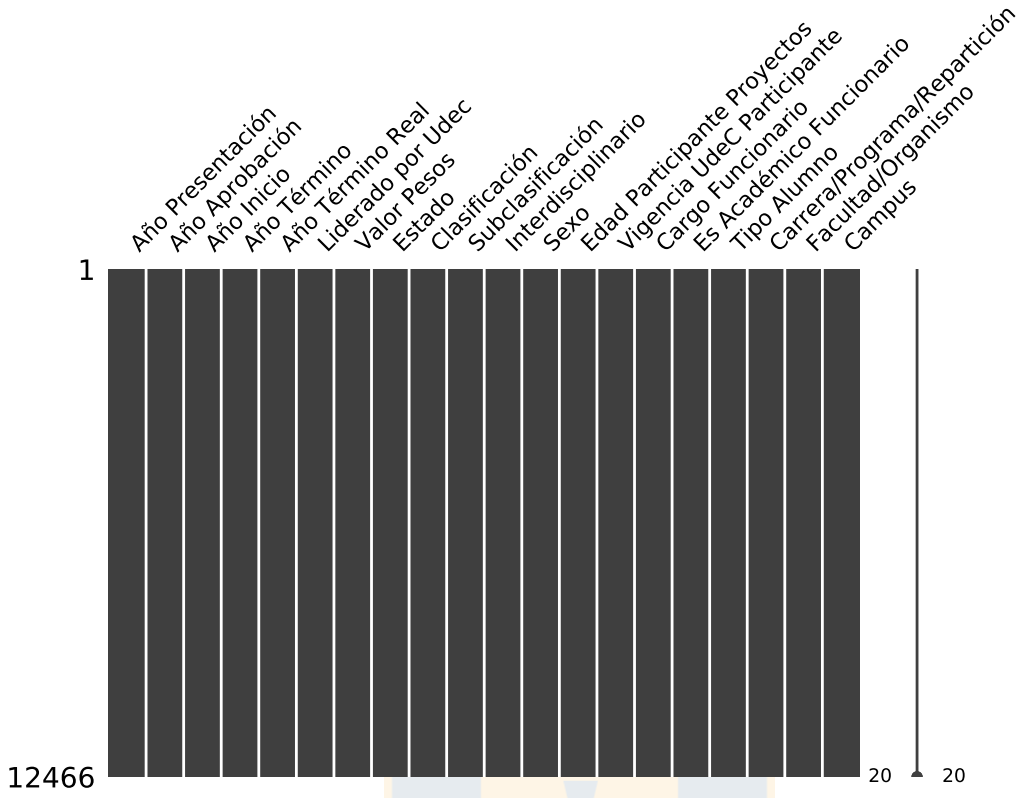


Figura 3.3.6: Matriz de nulidad después de la imputación

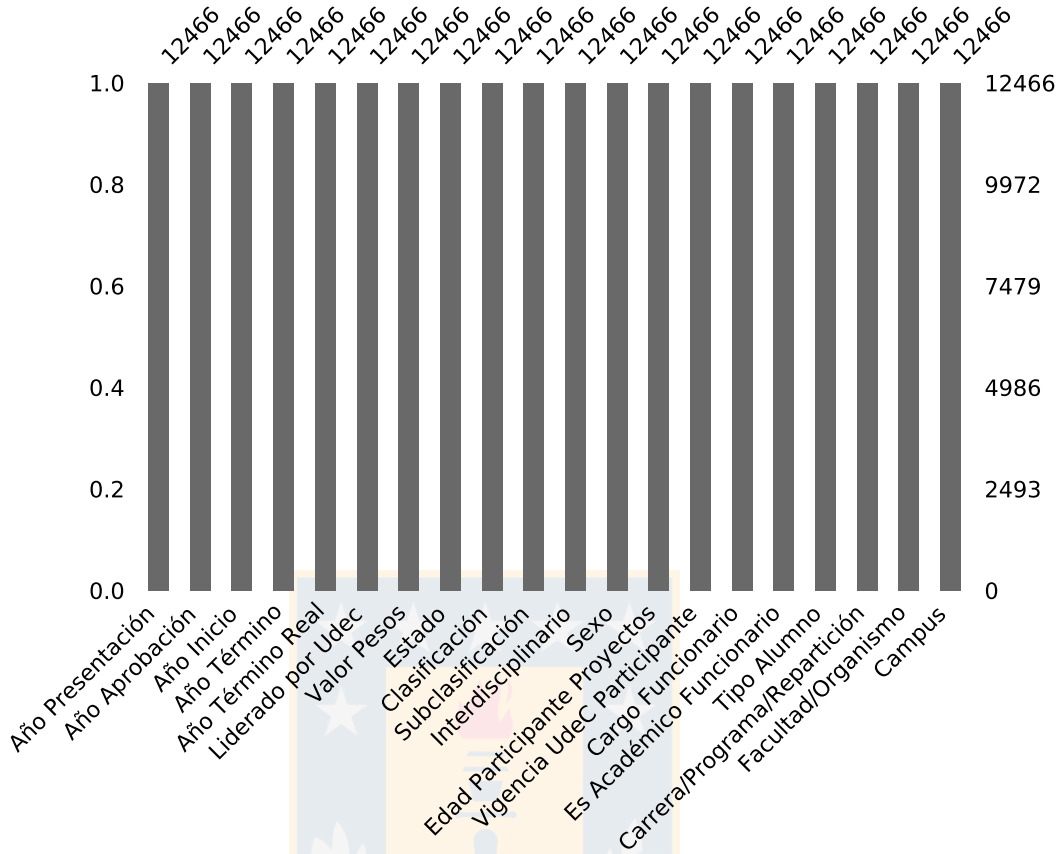


Figura 3.3.7: Gráfico de barras de nulidad después de la imputación

Capítulo 4

Análisis

4.1. Análisis Descriptivo

En esta sección se presenta un resumen de la información que otorgan los datos, se extraen las características más representativas, y se estudian relaciones entre 2 o más variables.

Para visualizar la distribución de la variable categórica 'Género', se genera el siguiente diagrama circular:

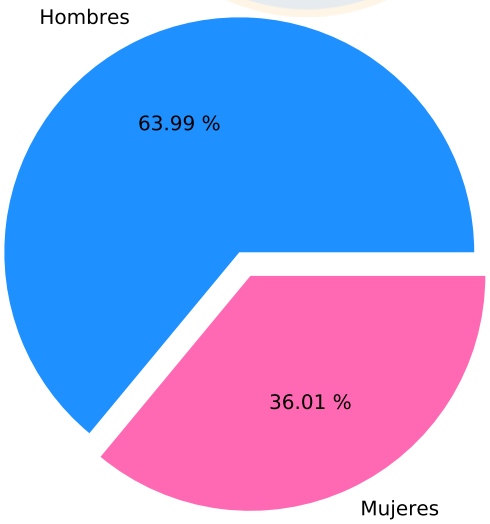


Figura 4.1.1: Distribución de proyectos según el género de su representante

Desde la Figura 4.1.1 es posible ver que para el total de 12466 proyectos, el 64 %

(7977) es liderado por hombres y solo el 36 % (4489) por mujeres.

Dentro de las instituciones que financian los proyectos, a lo largo de los años hay 5 que destacan del resto por su alto aporte monetario (mayor a 200 mil millones de pesos):

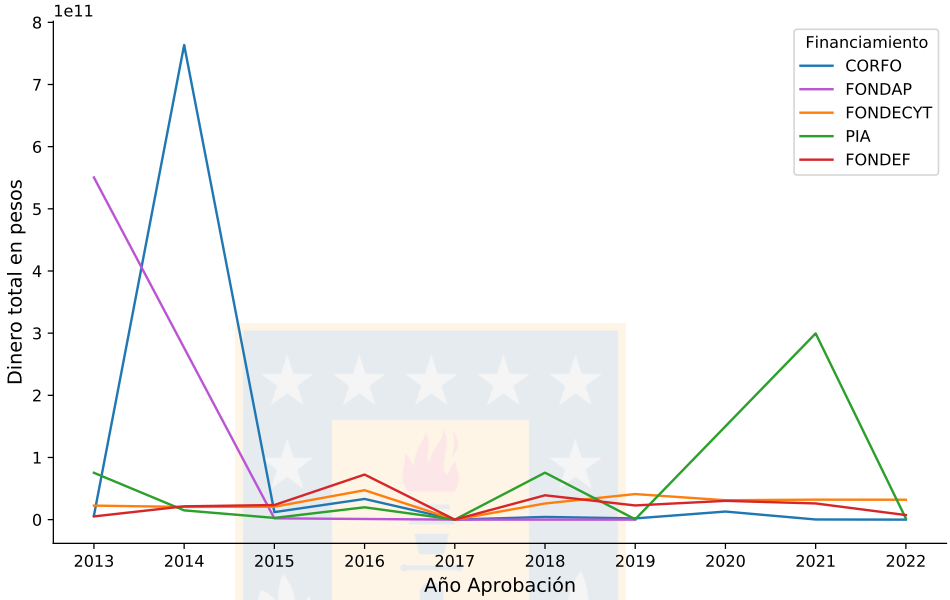


Figura 4.1.2: Distribución de dinero asignado por año de aprobación

En la Figura 4.1.2 se puede ver que en los años 2013 y 2014 hay un gran incremento monetario destinado a los proyectos, principalmente gracias a las instituciones FONDAP (Fondo de Financiamiento de Centros de Investigación en Áreas Prioritarias) y CORFO (Corporación de Fomento de la Producción), otorgando casi 600 mil y 800 mil millones de pesos respectivamente. En al año 2021, el programa PIA (Programa de Investigación Asociativa) destacó por su contribución de aproximadamente 300 mil millones de pesos. Es importante destacar que no todas las instituciones han financiado a proyectos de la Universidad de manera constante. Por ejemplo, como se puede apreciar en la Figura 4.1.2, FONDAP entregó una suma de dinero considerable, pero sólo en el año 2013, y a partir del 2015 no volvió a financiar, a diferencia de otras instituciones como FONDECYT (Fondo Nacional de Desarrollo Científico y Tecnológico) y FONDEF (Fondo de Fomento al Desarrollo Científico y Tecnológico) que han entregado una cantidad menor, generalmente entre 20 mil y 50 mil millones de pesos, pero constante a través de los años.

4.1.1. Análisis Bivariado

El Análisis Descriptivo Bivariado estudia la relación entre pares de atributos medidos simultáneamente, que incluye un conjunto de herramientas enfocado en el análisis de dos variables, con el objetivo de determinar las relaciones empíricas entre ellas.

En esta situación, el género del o la participante del proyecto es sumamente relevante, por lo que primero se estudiarán relaciones entre la variable binaria 'Género' y las demás categorías.

Para analizar la distribución de proyectos según los años de los eventos existentes (es decir, Presentación, Aprobación, Inicio, Término, y Término Real para cada proyecto), se generan los siguientes diagramas comparativos de violín:

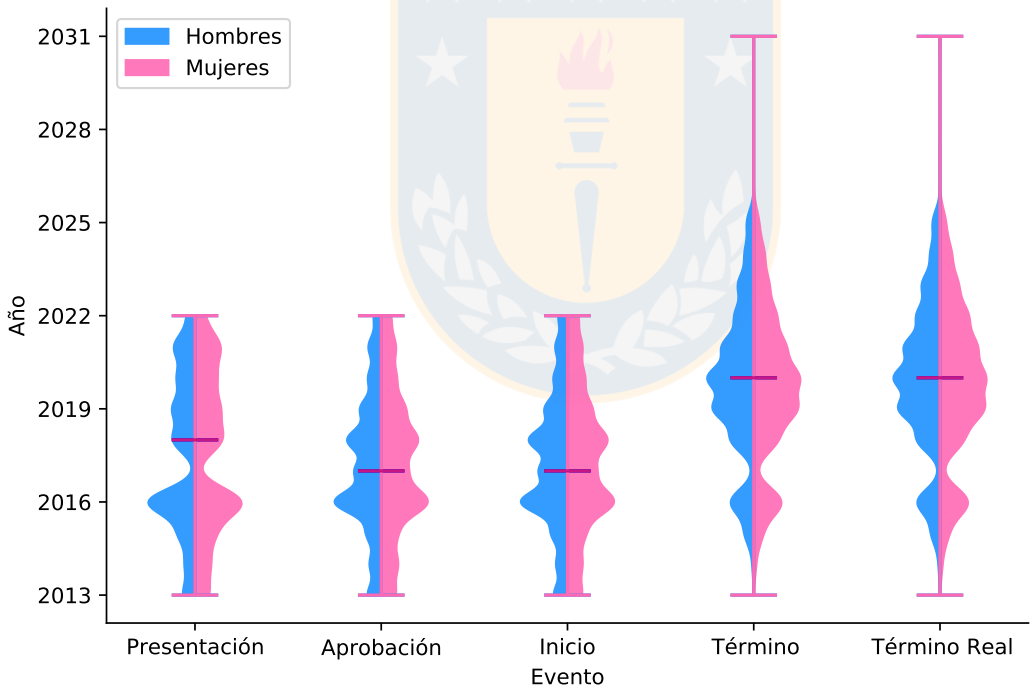


Figura 4.1.3: Distribución de proyectos según fecha de evento

En la anatomía de la Figura 4.1.3, el ancho de densidad de cada gráfico indica la frecuencia del dato. En este caso, del lado derecho de cada diagrama de violín se muestra la distribución de proyectos liderados por mujeres, y del lado izquierdo la distribución de proyectos liderados por hombres. Se observa que no hay una diferencia notoria respecto al género del o la participante del proyecto. La línea

horizontal central refleja la mediana de los datos, la línea horizontal superior el valor máximo y la línea inferior el valor mínimo.

Además, es posible observar que en los datos no hay casos extremos fuera del rango de años posible.

Esto se detalla en la siguiente tabla:

Evento	Media	Moda	Mediana	Mínimo	Máximo
`Año Presentación`	2018	2017	2018	2013	2022
`Año Aprobación`	2017	2016	2017	2013	2022
`Año Inicio`	2017	2016	2017	2013	2022
`Año Término`	2020	2020	2020	2013	2031
`Año Término Real`	2020	2020	2020	2013	2031

Cuadro 4.1.1: Medidas de tendencia central de fechas de eventos

Además, otra variable relevante a considerar es la duración de los proyectos. Del Cuadro 4.1.1 se puede inferir una duración promedio entre 2 y 3 años (considerando las variables `Año Inicio` y `Año Término`), y se puede observar con mayor detalle en el siguiente histograma:

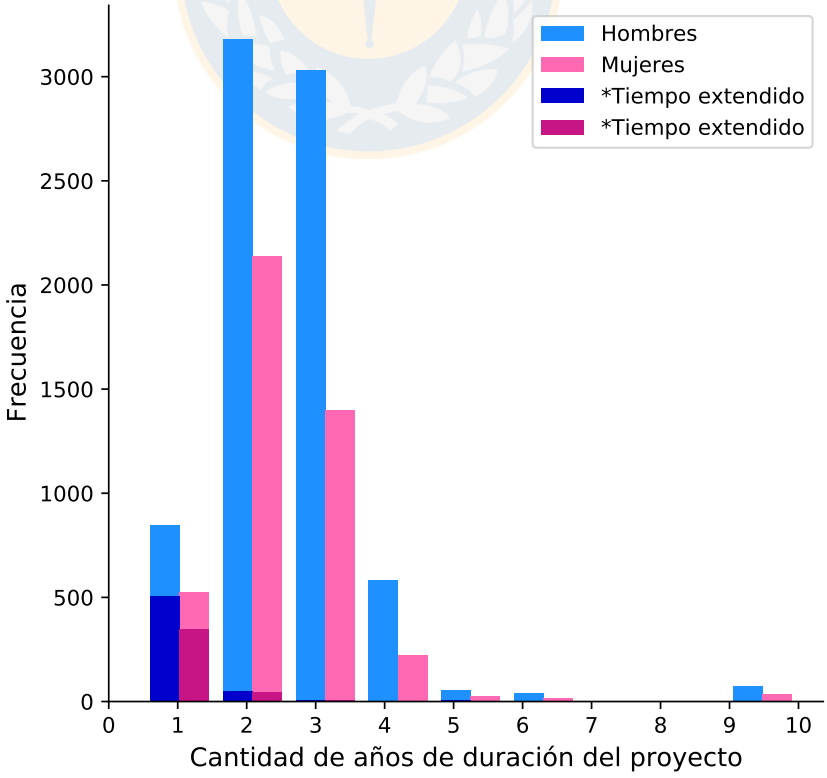


Figura 4.1.4: Distribución de proyectos según su duración

Para la realización de la Figura 4.1.4 también se considera la variable calculada `Atraso`, con la finalidad de representar el tiempo total de la duración de los proyectos y distinguir los casos que tuvieron plazo extendido de los que no lo tuvieron.

Cabe destacar que, si bien para ambos géneros la duración más frecuente de los proyectos es de 2 años, en el caso de los hombres la cantidad de proyectos no presenta diferencias importantes si son 2 ó 3 años, mientras que en el caso de las mujeres la diferencia es muy importante, mostrando una tendencia a una menor duración de los proyectos desarrollados por mujeres.

Para analizar si existe una diferencia de financiamiento según el género, se representa el total de dinero invertido en el siguiente gráfico de barras:

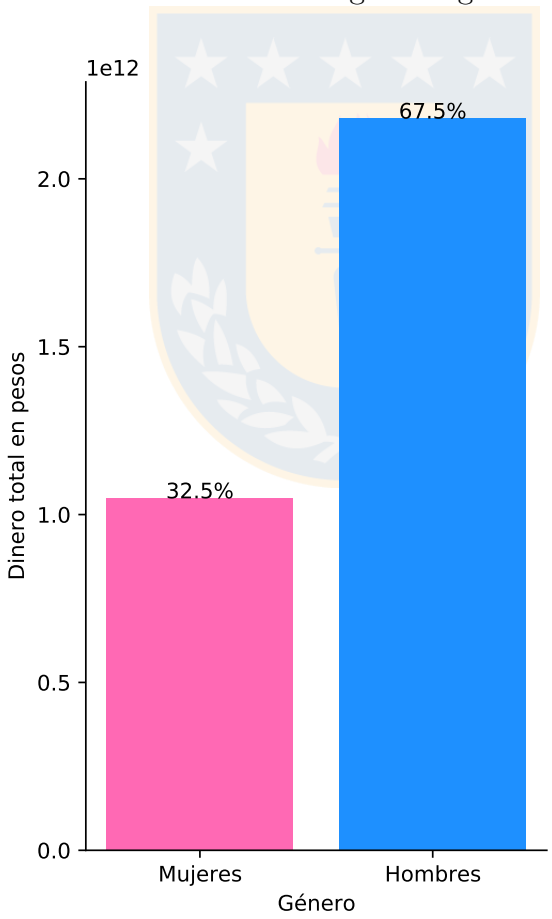


Figura 4.1.5: Distribución de proyectos según el dinero total invertido

A pesar de que la Figura 4.1.5 es clara en términos de la diferencia en los montos asignados a proyectos que lideran hombres y mujeres, es importante notar que

existe también una diferencia en el número de proyectos totales (ver Figura 4.1.1).

De la información analizada surgen algunas preguntas. ¿El dinero es entregado de forma continua?, ¿hay diferencias significativas entre un año y otro? Para responder a esto, se genera una suma de todos los dineros adjudicados por cada año de aprobación. Esta información se puede ver en el siguiente gráfico:

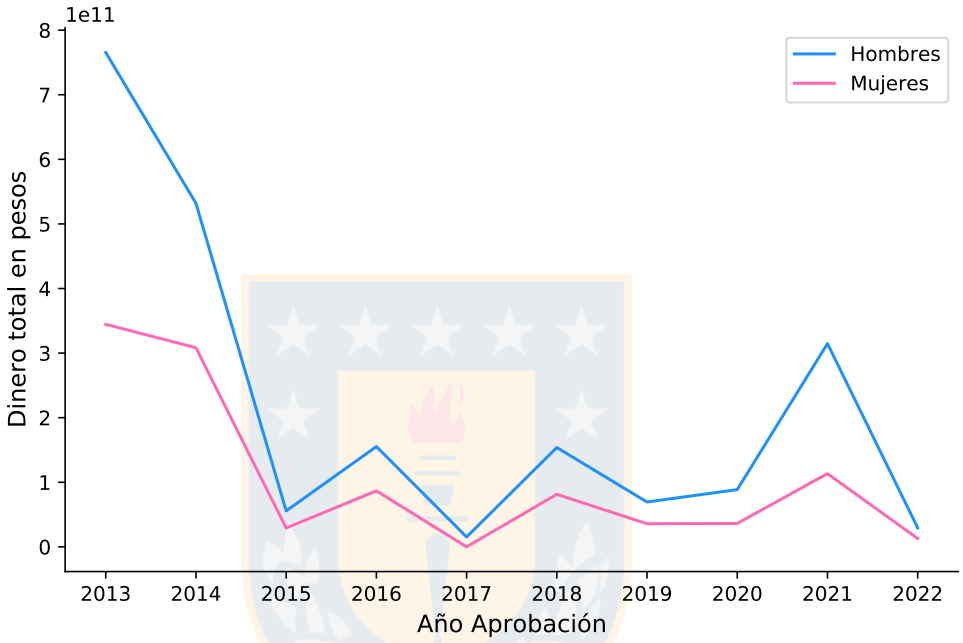


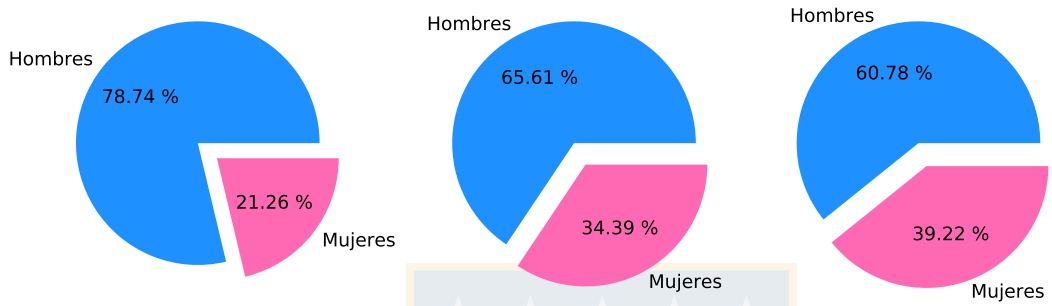
Figura 4.1.6: Distribución de dinero asignado por año de aprobación

Desde la Figura 4.1.6 se aprecia claramente que los montos asignados a proyectos de investigación es bastante variable en el tiempo, presentando máximos en el período del año 2013, con una consiguiente baja hasta 2015 e incrementándose un tanto en 2021, lo que tiene concordancia con el gráfico de la Figura 4.1.2. Destaca en este caso que, si bien durante el período estudiado los montos asignados a proyectos liderados por mujeres es inferior a los montos asignados a proyectos liderados por hombres, las diferencias en este sentido son mayores en el período previo a 2015 y para el año 2021, específicamente.

Dentro de la información entregada, se encuentran las categorías `Cargo Funcionario` y `Tipo Alumno`, que indican el cargo o el grado que tiene la persona responsable del proyecto respectivamente. Existen 70 cargos en la base de datos, pero son 3 los que más destacan: `'Profesor Titular'`, `'Profesor Asociado'` y `'Profesor Asistente'`. Dentro de la jerarquía de la Universidad de Concepción,

el rango académico más alto entre estos es Profesor Titular, seguido de Profesor Asociado, y luego Profesor Asistente.

Se pueden ver los porcentajes de la distribución entre proyectos de hombres y proyectos de mujeres a través de diagramas circulares:

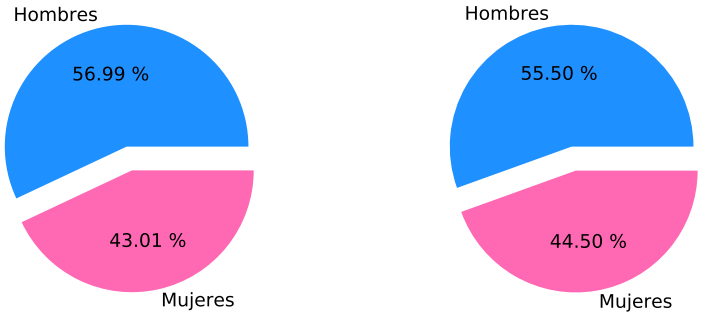


(a) Profesor Titular **(b)** Profesor Asociado **(c)** Profesor Asistente

Figura 4.1.7: Distribución de proyectos según el cargo del participante

De la Figura 4.1.7 es posible notar que hay una diferencia en la distribución de proyectos según del nivel académico. En la Figura 4.1.7a se muestra que solo un 21,26 % de los proyectos de Profesores Titulares son de mujeres, siendo éste el rango más alto en cuanto a cargos académicos. En la Figura 4.1.7c, de Profesores Asistentes, los proyectos de mujeres alcanzan el 39,22 %.

Por otro lado, están las personas responsables que no tienen un cargo académico, ya que son estudiantes de la Universidad. En los siguientes gráficos circulares se muestra el porcentaje en caso de ser alumnos o alumnas de Pregrado o de Postgrado:



(a) Estudiantes de Pregrado **(b)** Estudiantes de Postgrado

Figura 4.1.8: Distribución de proyectos según el tipo de estudiante

De la Figura 4.1.8 se observa que no existe mayor diferencia porcentual en los tipos de estudiantes respecto de su género. Aun así, en ambos casos el porcentaje de proyectos liderados por estudiantes hombres es mayor que el de proyectos liderados por estudiantes mujeres.

Otra característica importante a considerar, es la edad de las personas que participan en los proyectos. En el siguiente gráfico se puede observar la distribución:

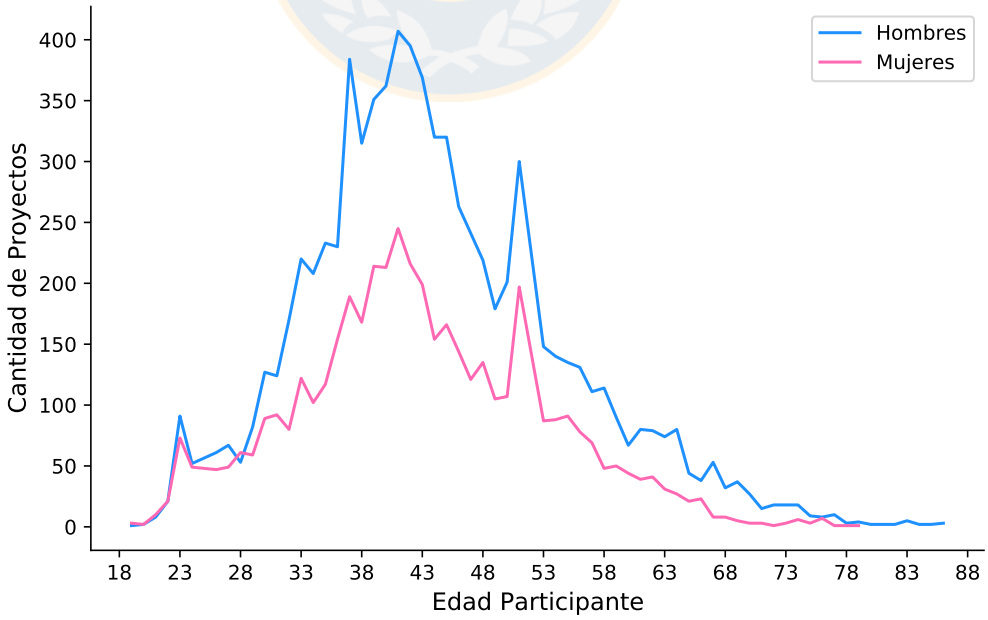


Figura 4.1.9: Distribución de proyectos según la edad del/la participante

De la Figura 4.1.9 se puede observar que la cantidad de proyectos liderados por

personas de hasta 30 años de edad no marcan una diferencia según el género. Sin embargo, a medida que la edad es más cercana a la media, hay un aumento significativo de proyectos liderados por hombres, y la brecha se mantiene hasta aproximadamente los 70 años.

Por otro lado, existen dos categorías importantes que indican la orientación del proyecto. Ellas son `Facultad/Organismo` y `Carrera/Programa/Repartición`. Estas variables contienen una cantidad grande de categorías. En el primer caso, existen 46 Facultades u Organizaciones distintas. Seleccionando las 5 con mayor número de proyectos, se obtiene el siguiente gráfico de barras:

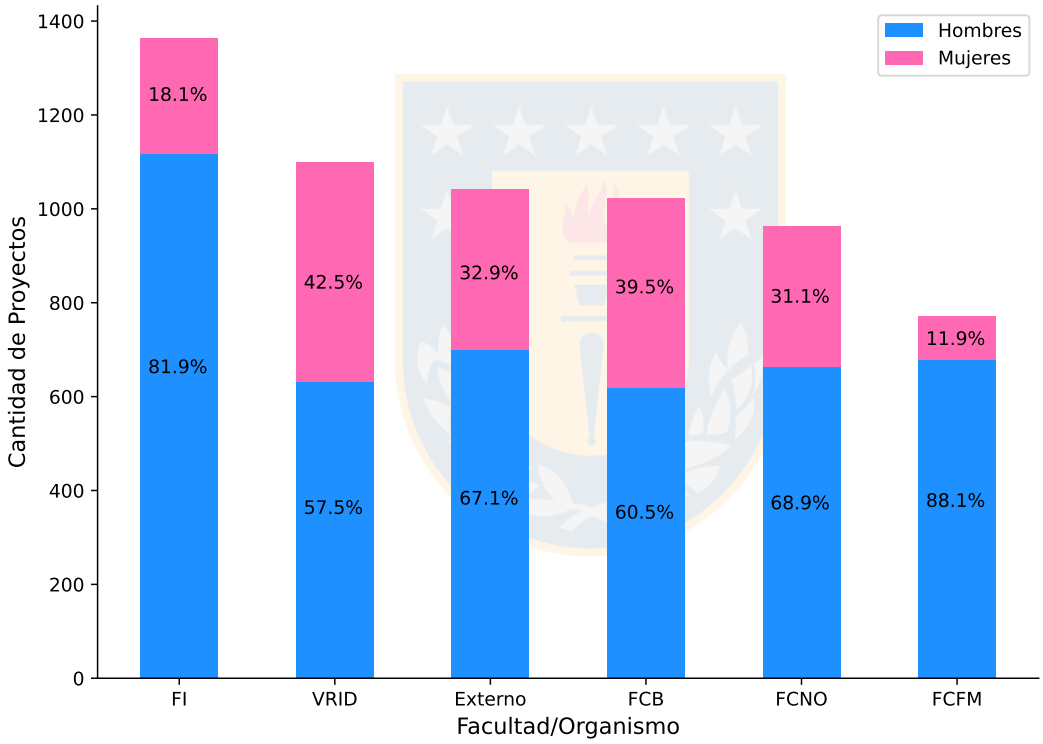


Figura 4.1.10: Distribución de proyectos según Facultad/Organismo

En la Figura 4.1.10, se puede observar que en las categorías `'Facultad de Ingeniería'` (FI) y `'Facultad de Ciencias Físicas y Matemáticas'` (FCFM) existe una proporción muy desequilibrada de proyectos respecto al género del participante, sobretodo en esta última. En cambio, en `'Vicerrectoría de Investigación y Desarrollo'` (VRID), `'Externo'` (proyectos fuera de la universidad), `'Facultad de Ciencias Biológicas'` (FCB), y `'Facultad de Ciencias Naturales y Oceanográficas'` (FCNO), la diferencia es menor,

aunque los proyectos de mujeres siguen sin superar el 43 %.

Por otro lado, la variable `Carrera/Programa/Repartición`, al ser más específica, tiene 351 categorías diferentes. Dentro de las que sobresalen 4:

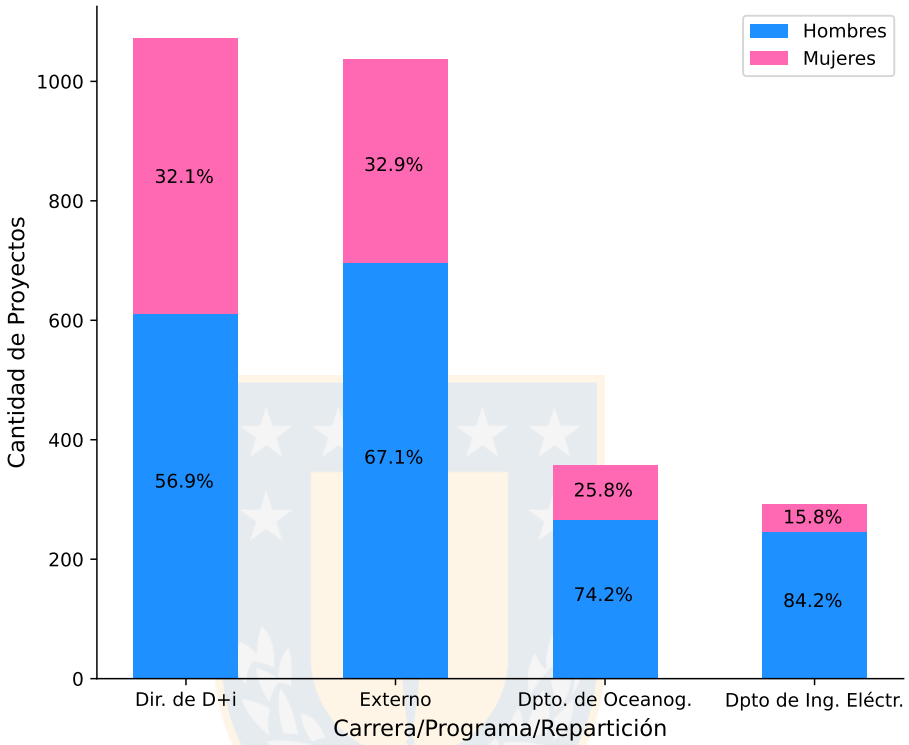


Figura 4.1.11: Distribución de proyectos según Carrera/Programa/Repartición

De la Figura 4.1.11 es posible notar que en las categorías `'Dirección de Desarrollo e Innovación'` (Dir. de D+i) y `'Externo'` (proyectos externos a la universidad), no existe una diferencia mayor según el género del o de la participante del proyecto, pero sí es notable en `'Departamento de Oceanografía'` (Dpto. de Oceanog.) y en `'Departamento de Ingeniería Eléctrica'` (Dpto de Ing. Eléctr.), alcanzando los proyectos de mujeres menos de un 16 % en este último.

Ahora, para todas las variables cuantitativas, se analiza la correlación entre ellas. Este análisis puede determinar en qué medida es posible predecir el valor de una variable en caso que conozcamos el valor de la otra, es decir, el estudio de la correlación entre dos variables se refiere a un conjunto de relaciones estadísticas que involucran una dependencia entre ellas. Esta relación puede ser representada de manera visual generando una matriz de correlación entre todos los pares de variables:

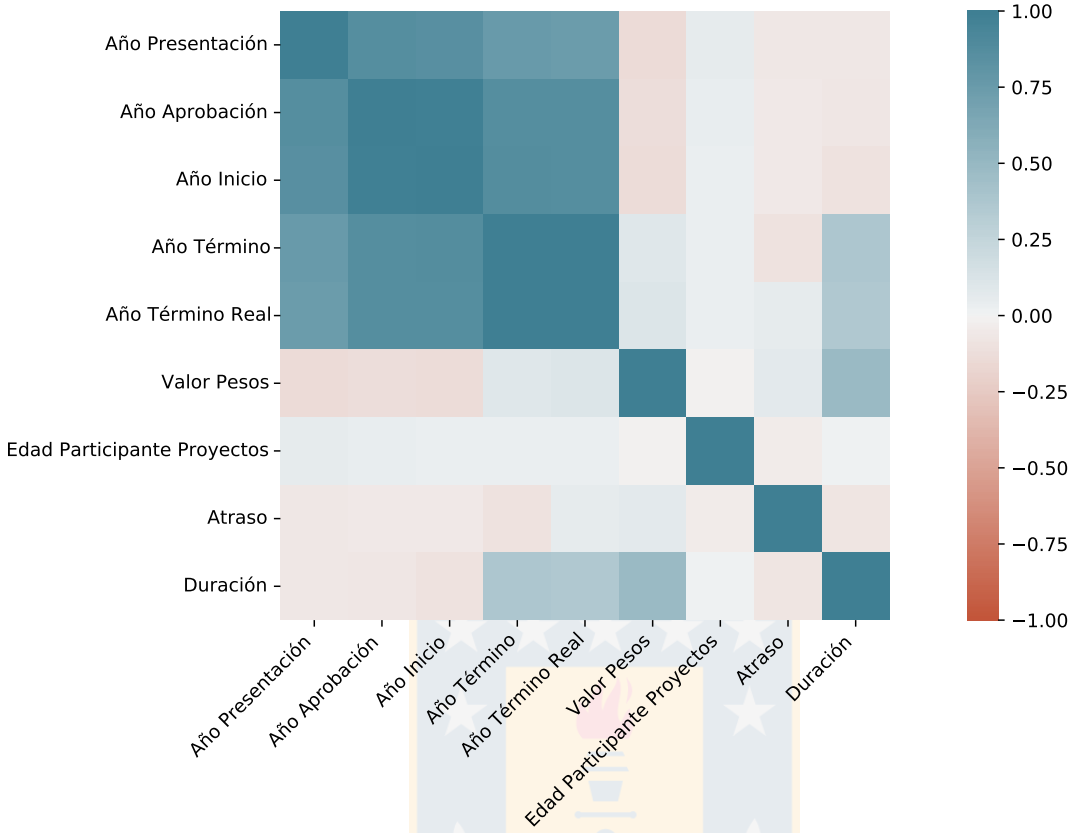


Figura 4.1.12: Matriz de correlación del total de proyectos

De la matriz en la Figura 4.1.12, se observa que existe una correlación muy alta entre las variables de años, alcanzando casi el valor 1 en el caso de `Año Término` con `Año Término Real`. Además, es posible notar que no existen correlaciones negativas significativas entre las variables, y que las otras variables numéricas presentan una correlación cercana a 0 entre sí.

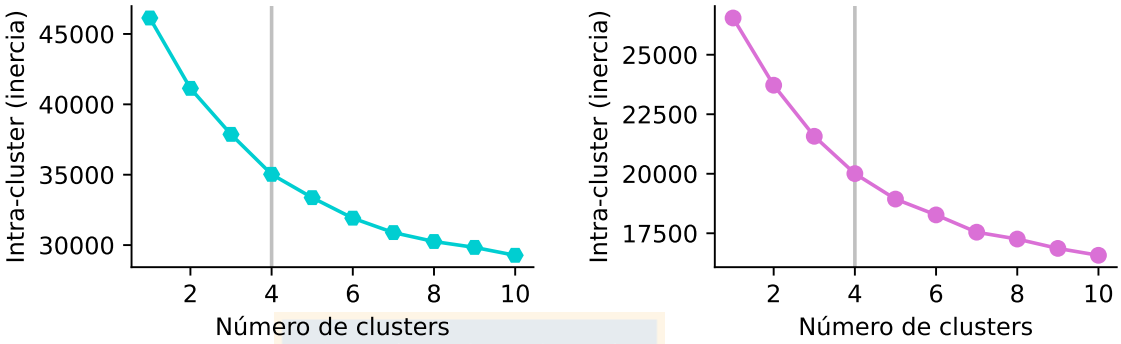
4.2. Análisis de Conglomerados

En este capítulo, la base de datos será particionada de manera que los proyectos liderados por mujeres estarán separados de los proyectos liderados por hombres, para poder comparar ambas bases de datos. Las dimensiones de `data_hombre` y de `data_mujer` son $[7977 \times 498]$ y $[4489 \times 454]$ respectivamente.

Para realizar este análisis en Python, se utiliza principalmente la librería `Scikit-Learn`.

4.2.1. Número de grupos

Utilizando en ambas bases de datos la función `inertia_`, se puede calcular la inercia de los *clusters* entre un rango de números de posibles grupos, en este caso, entre 1 y 10:



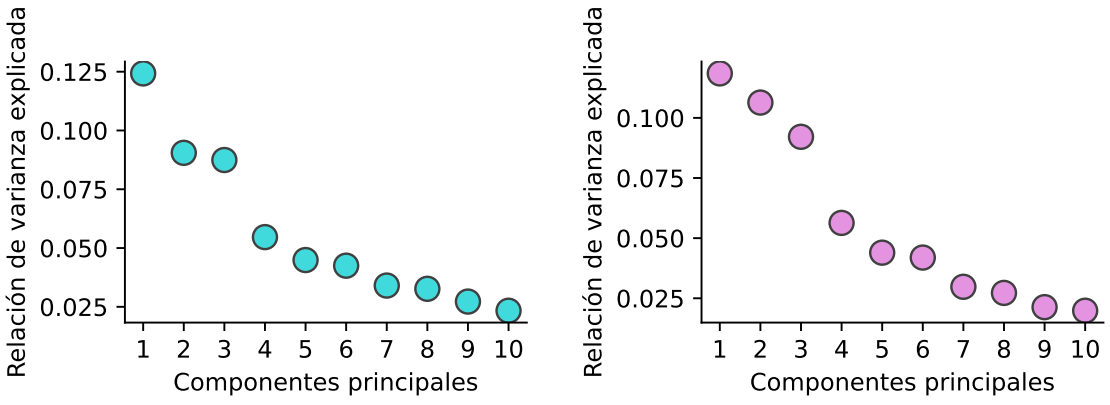
(a) `data_hombre` **(b)** `data_mujer`

Figura 4.2.1: Evolución de la varianza intra-cluster total

En las Figuras (4.2.1a) y (4.2.1b), es posible notar que en el número de *clusters* 4 se forma un pequeño codo, por lo que se va a considerar 4 como un número óptimo de grupos tanto para hombres como para mujeres.

4.2.2. Análisis de Componentes Principales

Desde `Scikit-Learn` se importa `PCA`, para calcular y graficar las varianzas de las 10 primeras componentes principales:



(a) `data_hombre` **(b)** `data_mujer`

Figura 4.2.2: Relación de varianza del vector de componentes principales

En las Figuras (4.2.2a) y (4.2.2b), es posible ver que las primeras 3 componentes principales tienen una varianza notoriamente superior a las demás siguientes en ambos casos. Es por esto que, para graficar los *clusters*, se considerarán las primeras 3 dimensiones, las cuales es posible visualizar en un sólo gráfico.

Es importante recordar que los componentes principales son combinaciones lineales de las variables originales, por lo que para poder interpretar los grupos formados en este nuevo sistema de coordenadas es necesario identificar las variables que más influyen en cada componente principal. A continuación se muestran gráficamente los coeficientes de las variables más representativas en los 3 primeros componentes para el grupo de proyectos de hombres:

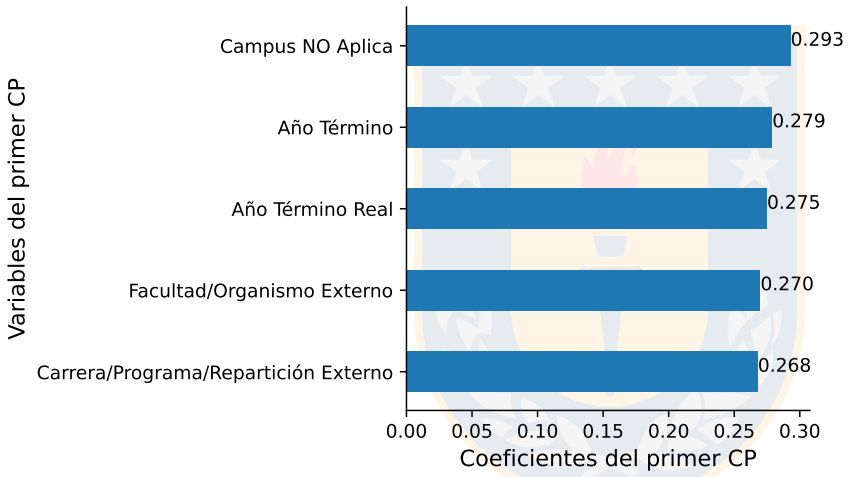


Figura 4.2.3: Coeficientes de las variables más influyentes del primer componente principal de `data_hombre`: 'Año Término de Externos'

Para un mejor entendimiento, las componentes principales tendrán un nombre según sus variables, por ejemplo, en la Figura 4.2.3, `Campus NO Aplica` significa que el proyecto no pertenece a ninguna sede de la Universidad de Concepción, y además, `Facultad/Organismo Externo` y `Carrera/Programa/Repartición Externo` también hacen alusión a que son proyectos externos de la universidad. Por otra parte, las variables `Año Término` y `Año Término Real` especifican que la fecha en que termina un proyecto es relevante. Por ende, considerando lo anterior, se le denominará a esta primera componente principal 'Año Término de Externos'.

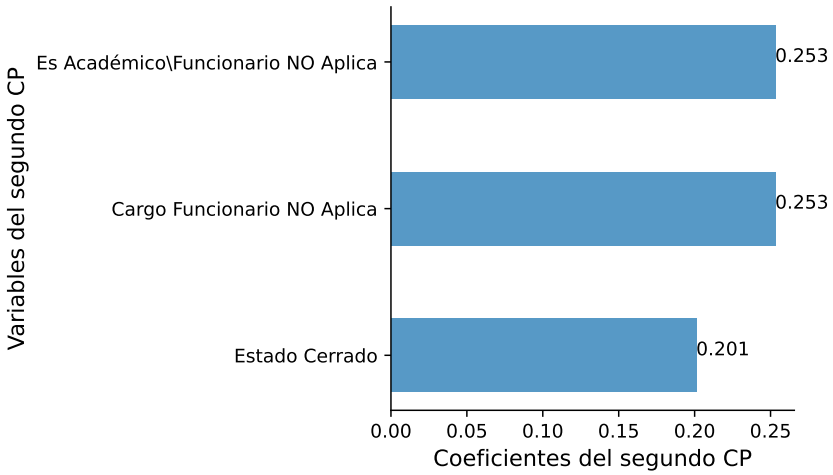


Figura 4.2.4: Coeficientes de las variables más influyentes del segundo componente principal de `data_hombre`: 'Externos Cerrados'

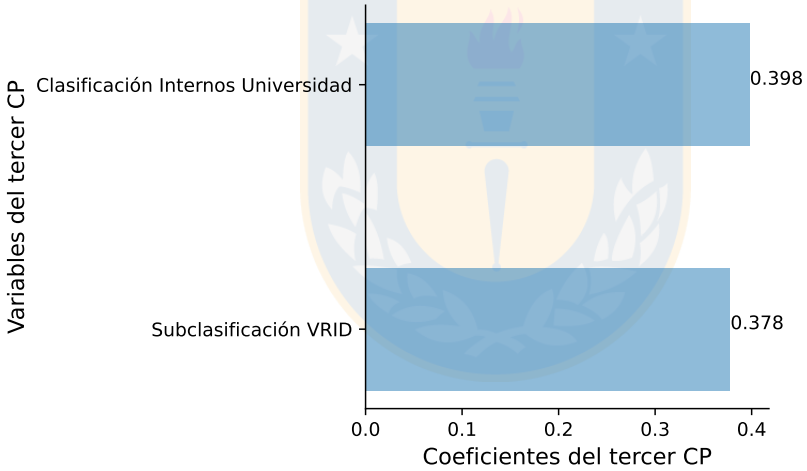


Figura 4.2.5: Coeficientes de las variables más influyentes del tercer componente principal de `data_hombre`: 'VRID'

Análogamente, se obtienen los coeficientes de las componentes principales más significativas para `data_mujer`:

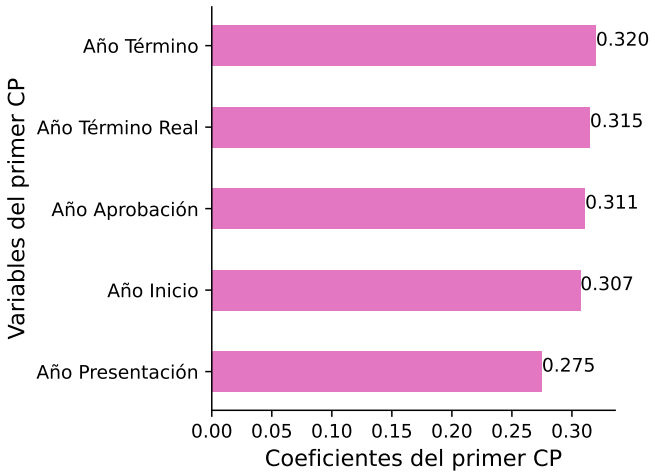


Figura 4.2.6: Coeficientes de las variables más influyentes del primer componente principal de `data_mujer`: 'Años de Eventos'

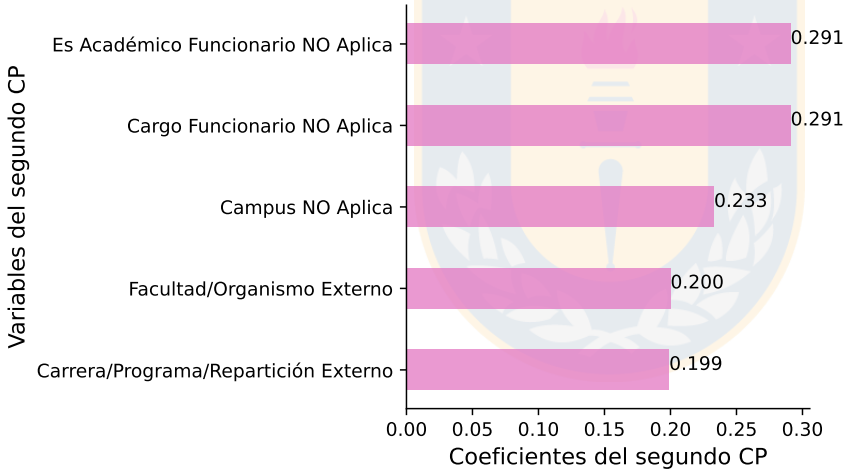


Figura 4.2.7: Coeficientes de las variables más influyentes del segundo componente principal de `data_mujer`: 'Externos'

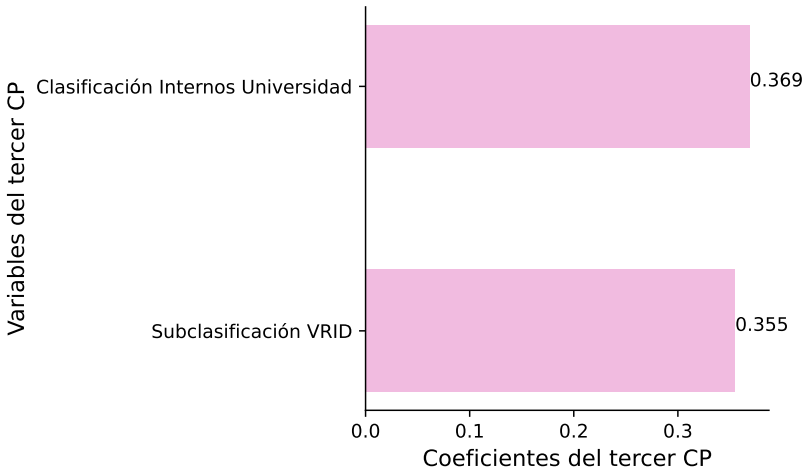


Figura 4.2.8: Coeficientes de las variables más influyentes del tercer componente principal de `data_mujer`: 'VRID'

Así, las dimensiones (o ejes de coordenadas) tienen un nombre más intuitivo para facilitar su interpretación, resumido en el Cuadro 4.2.1:

	`data_hombre`	`data_mujer`
Primer componente principal	Año Término de Externos	Años de Eventos
Segundo componente principal	Externos Cerrados	Externos
Tercer componente principal	VRID	VRID

Cuadro 4.2.1: Nombres de las componentes principales

Se puede observar que la primera componente principal (en ambos casos) tiene relación con los años. Pero examinando las Figuras 4.2.3 y 4.2.6, los proyectos liderados por hombres sólo consideran los años de término, mientras que los proyectos liderados por mujeres consideran los años de los 5 eventos presentes en los datos. Además, en `data_hombre`, los proyectos externos (de 3 variables) son relevantes, mientras que en `data_mujer` es irrelevante.

Por otro lado, en el segundo componente principal sobresalen las variables relacionadas con la procedencia de los proyectos, es decir, si son externos (o internos). Pero, en `data_hombre`, hay una inclinación también a los proyectos que están cerrados. Notar que, el hecho de que un proyecto esté cerrado, significa que su año de término ya concluyó, y los años de término también son variables relevantes en el primer componente principal, que es lo que ha distinguido a los dos primeros componentes principales de `data_hombre` de los dos primeros

componentes principales de `data_mujer`.

Por último, en el tercer componente principal predominan los proyectos relacionados a la Vicerrectoría de Investigación y Desarrollo (VRID), con la variable clasificación, que indica el origen del proyecto, y la variable subclasificación, que muestra el organismo que financió el proyecto, en este caso, la VRID, la cual es la principal fuente de financiamiento interno de la UdeC.

4.2.3. Algoritmo *K-Means*

Importando `KMeans` desde `Scikit-Learn`, se crea un modelo de *K-Means* especificando la cantidad de grupos que se desea encontrar, en este caso, 4 *clusters* (4.2.1). Luego, se llama al método de ajuste del modelo `fit`, el cual ubica y recuerda las regiones donde ocurren los diferentes *clusters*.

Una vez ajustado el modelo, se utiliza el método de predicción `predict`, el cual entrega la etiqueta del *cluster* para cada observación, indicando así a qué conglomerado pertenece cada una.

Finalmente, utilizando `Pyplot`, se puede visualizar el agrupamiento de las muestras, los centroides de los *clusters*, y los puntos donde cada uno representa una muestra y está coloreado según su *cluster* correspondiente:

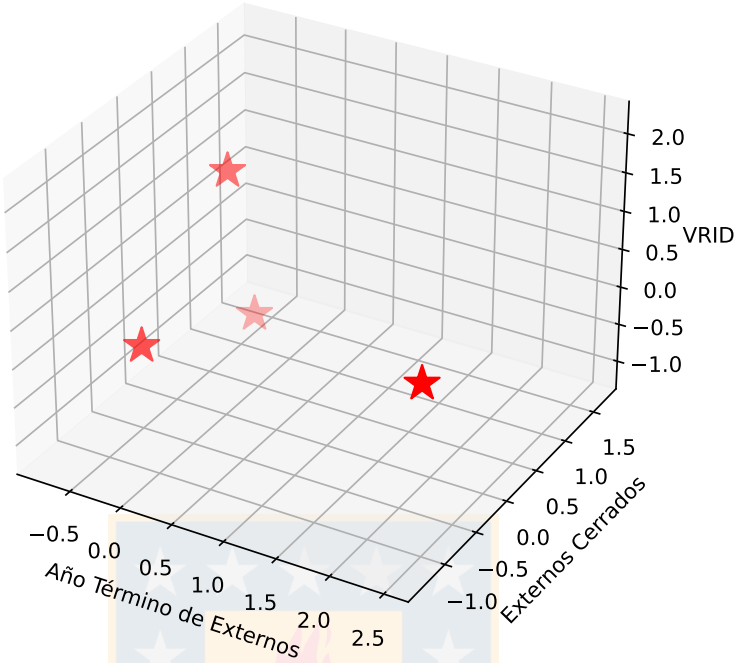


Figura 4.2.9: Centroides de los *clusters* de proyectos liderados por hombres

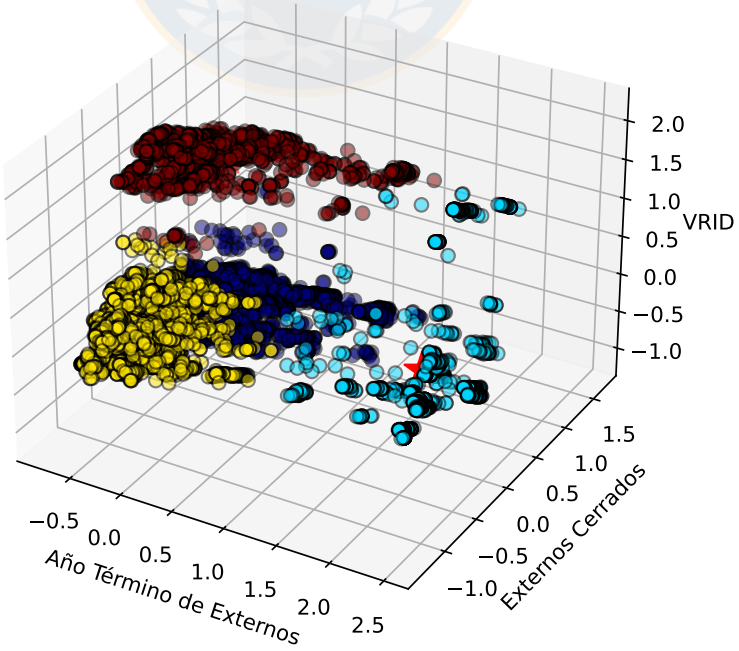


Figura 4.2.10: *Clusters* de los proyectos liderados por hombres

En la Figura 4.2.10, es posible notar que el *cluster* de color rojo (●) tiene principalmente proyectos financiados por la VRID. El *cluster* de color celeste (●) contiene proyectos externos a la universidad que tienen un año de término más alto que el promedio. Los *clusters* de color amarillo (●) y azul (●) abarcan proyectos que no fueron financiados por la VRID, y se diferencian en que los de color amarillo (●) son proyectos que actualmente no están cerrados, mientras que los de color azul (●) sí lo están.

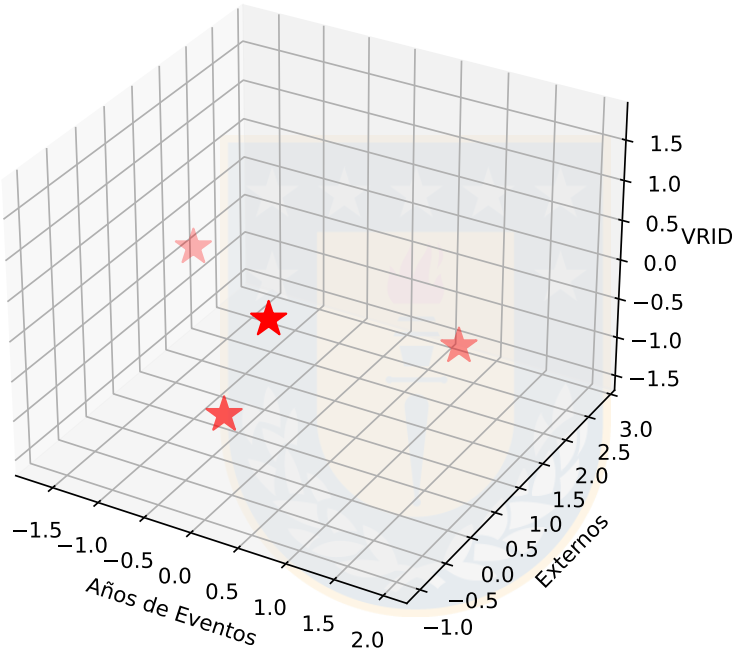


Figura 4.2.11: Centroides de los *clusters* de proyectos liderados por mujeres

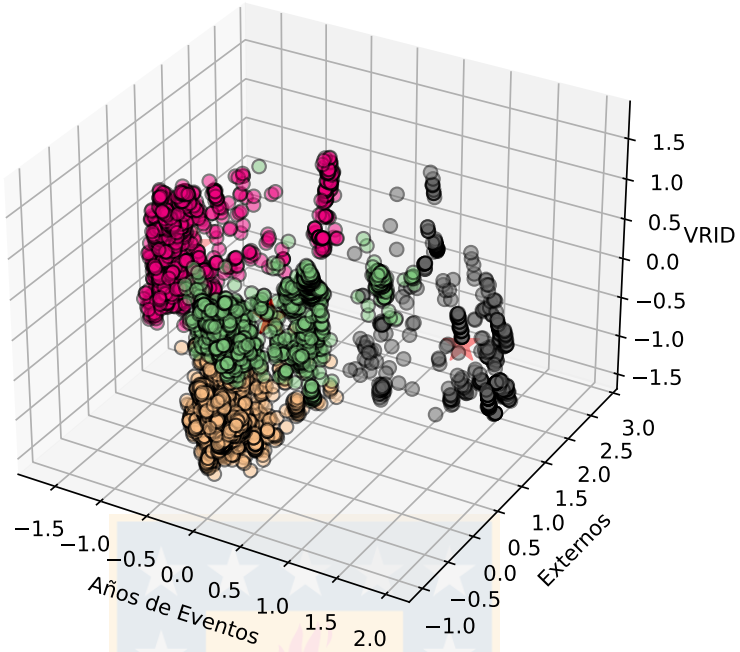


Figura 4.2.12: *Clusters* de los proyectos liderados por mujeres

De la Figura 4.2.12 se puede ver que existe un *cluster* de color damasco (●) cuyos proyectos que pertenecen a él no son externos a la universidad. Los *clusters* de color rosado (●) y gris (●) están a una distancia considerable en el componente principal 'Años de Eventos', siendo el rosado (●) el que contiene proyectos realizados en un período más antiguo, mientras que el gris (●) contiene los proyectos más recientes. Finalmente, es posible notar que el *cluster* de color verde (●) se encuentra centrado con respecto a los 3 primeros componentes principales.

4.3. Análisis de Regresión Logística

La variable dependiente 'Género' se define de la siguiente manera:

$$y_i = \begin{cases} 1, \text{ si el proyecto es liderado por una mujer,} \\ 0, \text{ si el proyecto es liderado por un hombre.} \end{cases} \tag{4.3.1}$$

Antes de poder aplicar el modelo de Regresión Logística a la base de datos, considerando la variable predictora y_i definida anteriormente (4.3.1), es necesario que las variables independientes sean valores numéricos. Para ello, se efectúa

nuevamente una transformación con variables *dummy*, utilizando `get_dummies`, y se normalizan los datos con `scaler` de `Pandas`.

Para entrenar el modelo, se debe separar el conjunto de datos inicial en *train* y *test*, para entrenamiento y prueba respectivamente. La proporción a utilizar es de 0.8 para entrenamiento y 0.2 para prueba [Pedregosa et al., 2011].

Finalmente, se puede realizar la imputación con `LogisticRegression` de `Scikit-Learn`.

4.3.1. Selección de variables

El método selección de variables *boruta* identificó las variables que tenían mayor influencia sobre la predicción de una regresión, considerando como objetivo la variable binaria género. Esto se llevó a cabo con las librerías `Scikit-Learn` y `BorutaPy`, donde se utilizó `RandomForestClassifier` y `BorutaPy` respectivamente, obteniendo así el Cuadro 4.3.1 con 23 variables seleccionadas:

	Variable
	Valor Pesos
	Edad Participante
	Atraso
	Duración
Clasificación	Internos Universidad
	Nacionales
Subclasificación	FONDECYT
	VRID
Vigencia UdeC Participante	Sí
Cargo Funcionario	Profesor Titular
Carrera/Programa/Repartición	Dpto. de Astronomía
	Dpto. de Currículum e Instrucción
	Dpto. de Física
	Dpto. de Ingeniería Ambiental
	Dpto. de Ingeniería de Materiales
	Dpto. de Ingeniería Eléctrica
	Dpto. de Ingeniería Matemática
	Dpto. de Química Analítica e Inorgánica
Facultad/Organismo	Fac. de Ciencias Físicas y Matemáticas
	Fac. de Ciencias Sociales
	Fac. de Enfermería
	Fac. de Farmacia
	Fac. de Ingeniería

Cuadro 4.3.1: Variables seleccionadas

Notar que, en el Cuadro 4.3.1 se encuentran algunas variables que ya habían sido estudiadas en la sección Análisis Descriptivo, ya que mostraban evidencias de que sí existían diferencias entre proyectos de hombres y proyectos de mujeres, y finalmente resultaron ser considerados como buenos predictores para el modelo, como por ejemplo `'Profesor Titular'` en la Figura 4.1.7a, `'Edad Participante'` en la Figura 4.1.9, `'Facultad de Ingeniería'` y `'Facultad de Ciencias Físicas y Matemáticas'` en la Figura 4.1.10, y `'Departamento de Ingeniería Eléctrica'` en la Figura 4.1.11. Además, destacar que el método también seleccionó `'Atraso'` y `'Duración'`, siendo éstas variables calculadas que no estaban presentes originalmente en la base de datos, mostrando así la utilidad de la información que no se encontraba de forma explícita.

4.3.2. Estimación de parámetros

En la Regresión Logística, el efecto que tiene cada variable independiente sobre la variable dependiente se expresa en el coeficiente β correspondiente.

Este valor se puede mostrar utilizando el atributo `coef` de `Scikit-Learn`, representado en la siguiente Figura:

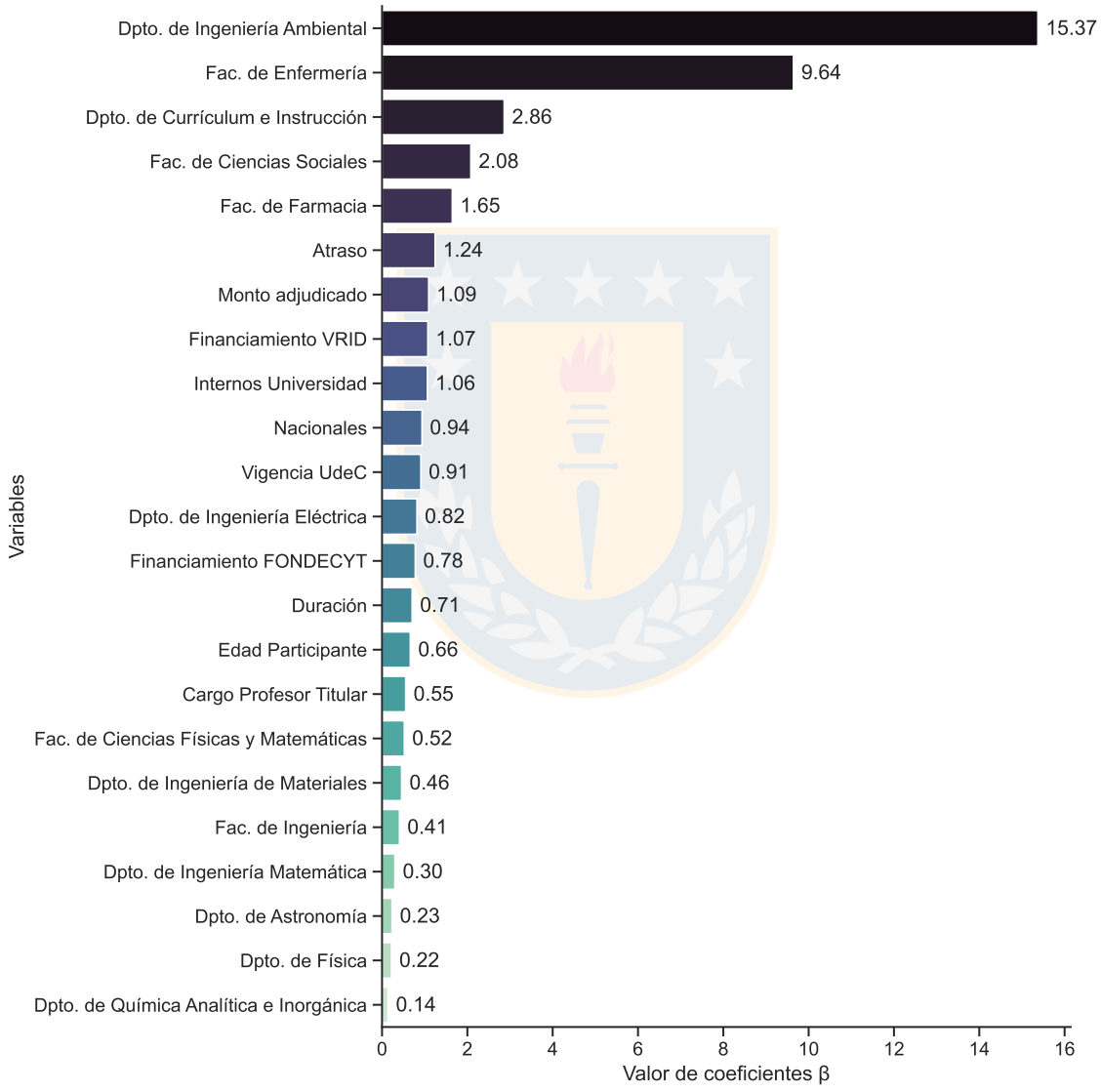


Figura 4.3.1: Gráfico de barras de parámetros β

La Figura 4.3.1 indica que, por un aumento de cada unidad de una variable independiente p, la oportunidad de ocurrencia *odds* de que y sea 1 (ecuación 4.3.1), es β_p veces más a que sea 0, considerando que las demás variables permanecen

constantes. Esta interpretación es diferente para las variables que originalmente eran categóricas y para las variables originalmente numéricas. Esto es porque a las variables cualitativas se les aplicó una transformación utilizando variables *dummy*, y entre 0 y 1 hay siempre una unidad de diferencia.

Por ejemplo, si se agrega un proyecto correspondiente al `Depto. de Ingeniería Ambiental`, la oportunidad de ocurrencia de que sea liderado por una mujer es 15.37 veces mayor a la oportunidad de ocurrencia de que sea liderado por un hombre. Pero para una variable numérica, como `Atraso`, la interpretación sería que si proyecto tiene un año de atraso más que el dato anterior, entonces la oportunidad de ocurrencia de que sea mujer respecto de que sea hombre es 1.24 veces superior. Por otro lado, si se agrega un proyecto correspondiente al `Dpto. de Química Analítica e Inorgánica`, la oportunidad de ocurrencia de que el proyecto sea liderado por una mujer es 0.14 veces la oportunidad de que sea liderado por un hombre, es decir, mucho menor.

Las variables con valores β más grandes indican una tendencia a identificar proyectos liderados por mujeres, y aquellos con valores β más cercanos a cero, una tendencia a identificar proyectos liderados por hombres; mientras que las variables con valores β cercanos a 1, no muestran una tendencia tan fuerte, es decir, eventualmente podrían ser liderados por hombres o por mujeres, aunque siendo las variables significativas, la tendencia existe.

4.3.3. Métricas de desempeño

Para obtener la Matriz de Confusión, se puede utilizar `Scikit-Learn`, que tiene el método `confusion_matrix` en su módulo `metrics`, y permite desarrollar este cruce de información.

1389	214
551	340

Cuadro 4.3.2: Matriz de confusión de la Regresión Logística

Donde en el Cuadro 4.3.2, 1389 (55.7 %) corresponde a los proyectos de hombres

que fueron clasificados correctamente como proyectos de hombres, 214 (8.6 %) son los proyectos de mujeres que fueron clasificados erróneamente como proyectos de hombres, 551 (22.1 %) los proyectos de hombres que fueron clasificados erróneamente como proyectos de mujeres, y 340 (13.6 %) los proyectos de mujeres que fueron clasificados correctamente como proyectos de mujeres.

Además, se calculan las métricas error con `classification_report`:

	precision	recall	f1-score	support
0 (proyectos de hombres)	0.72	0.87	0.78	1603
1 (proyectos de mujeres)	0.61	0.38	0.47	891
accuracy			0.69	2494
macro avg	0.66	0.62	0.63	2494
weighted avg	0.68	0.69	0.67	2494

Cuadro 4.3.3: Métricas de la Regresión Logística

La métrica `accuracy` del Cuadro 4.3.3 indica que la exactitud del modelo es de 0.69. Ya que la variable a predecir es binaria, la medida `accuracy` es 19 % superior a una predicción por azar. Los puntajes de `precision` sugieren que el modelo tiene un desempeño aceptable, y los puntajes de `recall` muestran una inclinación por acertar gran parte de los proyectos de hombres, pero indica que el modelo tiene mayor problema para clasificar los proyectos de mujeres.

Capítulo 5

Conclusión

5.1. Conclusiones

1. Este estudio aporta evidencia sólida sobre la efectividad de las técnicas de *Machine Learning* para abordar problemas sociales, especialmente en el contexto de las brechas de género.

2. A través del Análisis Descriptivo, se evidenciaron diferencias significativas entre hombres y mujeres en ciertas variables de la base de datos: los hombres presentan una cantidad mayor de proyectos con mayor financiamiento total, una tendencia a una mayor duración de proyectos y ocupan la mayoría de los cargos de mayor jerarquía. Además, los hombres tienen en promedio más proyectos y existen áreas de investigación donde la proporción de hombres es mucho mayor, como en la Facultad de Ciencias Físicas y Matemáticas, Facultad de Ingeniería, Departamento de Oceanografía y Departamento de Ingeniería Eléctrica.

3. La decisión de calcular variables resultó ser una estrategia efectiva para obtener información valiosa que no estaba explícitamente disponible en la base de datos, ya que tuvo un impacto significativo en el Análisis Descriptivo y en la estimación de parámetros de la Regresión Logística.

4. El Análisis de Conglomerados mostró leves diferencias de agrupamiento entre proyectos de hombres y mujeres, con componentes principales relacionadas a las variables que discriminan entre proyectos internos y externos, el año

de término, y la VRID.

5. Utilizar el método de selección de variables *boruta* sirvió para identificar los atributos que tenían mayor influencia sobre la base de datos con respecto al género, pero fueron los parámetros β del modelo de Regresión Logística los que indicaron la importancia de cada una de las variables seleccionadas.

6. La oportunidad de ocurrencia de que un proyecto nuevo sea liderado por una mujer es significativamente mayor si éste pertenece al Departamento de Ingeniería Ambiental o a la Facultad de Enfermería, mientras que es significativamente menor si el proyecto pertenece al Departamento de Ingeniería Matemática, Astronomía, Física, o al de Química Analítica e Inorgánica.

7. En los resultados obtenidos se pudo identificar cuáles variables están más presentes en las brechas de género, por lo que el objetivo principal de los análisis fue satisfactorio.

5.2. Trabajos futuros

1. Se puede replicar este mismo análisis cada cierto período para poder estudiar con mayor detalle los avances en la disminución de las brechas de género.

2. La base de datos trabajada se puede complementar con bases de datos de matrículas de estudiantes, docentes por facultad, entre otros, para poder tener una mayor perspectiva.

3. Para realizar un estudio con menor margen de error, sería apropiado incluir la variable género como categoría a las nuevas bases de datos.

4. Una buena práctica sería automatizar este proceso de análisis, para que se generen reportes estadísticos respecto a las brechas de género, con las bases de datos actualizadas.

Bibliografía

Aldas Manzano, J. and Uriel Jimenez, E. (2017). *Análisis multivariante aplicado con R. 2ª ed.* Ediciones Paraninfo, S.A.

Alomari, M. and Diabat, A. (2012). Prediction of oil price by k-nearest neighbor (knn) method. *Expert Systems with Applications*, 39(18):13407–13411.

Anderson, A. and Semmelroth, D. (2015). *Statistics for Big Data For Dummies*. –For dummies. Wiley.

Bewick, V., Cheek, L., and Ball, J. (2005). Statistics review 14: Logistic regression. *Critical care (London, England)*, 9:112–8.

Faraway, J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Hair, J., Anderson, R., and Babin, B. (2009). *Multivariate Data Analysis*. Prentice Hall, 7 edition.

Johnson, D. (2000). *Métodos multivariados aplicados al análisis de datos*. Soluciones empresariales. Thomson.

Kane, E. (1968). *Economic Statistics and Econometrics: An Introduction to Quantitative Economics*. Harper international edition. Harper & Row.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Little, R. and Rubin, D. (2014). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100.

OCDE (2014). *Cerrando las brechas de género Es hora de actuar: Es hora de actuar*. OCDE iLibrary. Corporación de Investigación, Estudio y Desarrollo de la Seguridad Social.

Olmo, M. and Mateu, J. (2003). *Geoestadística y modelos matemáticos en hidrogeología*. Colälecció Medi Ambient Series. Universitat Jaume I.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Pertuz, C. (2022). *Aprendizaje automático y profundo en python: Una mirada hacia la inteligencia artificial*. Ediciones de la U.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill Interamericana de España S.L.

Sharma, A. and Singh, S. P. (2015). A feature selection and classification model for intrusion detection system using knn and svm. *International Journal of Computer Science and Network Security*, 15(8):62–69.

Stoppiglia, H., Dreyfus, G., Dubois, R., and Oussar, Y. (2003). Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.