Proyecto de Riesgo ANID

Diplomado en Data Science

Andrea Araya (14.526.353-0) Pablo Bellei (15.088.206-0)
Luis Cuello (18.002.988-5) Gino Benedetti (17.349.122-0)
Gabriela Ossa (19.667.946-4)

2021-07-12

Contents

PRIMER REPORTE

1. Introducción

La Subdirección de Proyectos de Investigación como parte de la Agencia Nacional de Investigación y Desarrollo (ex CONICYT), es la encargada de la administración de los tres principales fondos de investigación individual en Chile (Regular, Posdoctorado e Iniciación en investigación) y también de realizar el seguimiento de los proyectos adjudicados, para asegurar la correcta ejecución de los mismos, específicamente en lo relacionado a las exigencias académicas.

1.1. Descripción de la problemática

En el proceso de evaluación de los proyectos se identifican falencias en la ejecución, es decir, hay proyectos que presentan dificultades para cumplir con las exigencias académicas mínimas para su adecuado cierre, tales como: la presentación de un *paper* o manuscrito académico aceptado, publicado o en prensa, el cumplimiento de las exigencias éticas a través de un informe de seguimiento ético y/o bioético, la realización de alguna actividad de divulgación científica, entre otras.

Por lo tanto, se vuelve necesario para la Subdirección diseñar un método que identifique patrones (variables) de no cumplimiento académico de los proyectos, y así poder crear un clasificador de riesgo, con la finalidad de alertar de forma temprana a los evaluadores, e implementar medidas de acompañamiento a los y las investigadoras para lograr el cumplimiento de las exigencias académicas y así, puedan finalizar exitosamente su investigación.

1.2. Preguntas a resolver

Las principales preguntas a responder serán: ¿Cuáles son los proyectos que están más propensos a presentar dificultades para culminar con éxito su investigación?, ¿Es posible asignar una clasificación de riesgo a los proyectos de acuerdo a ciertos patrones?, ¿Existen variables que ayuden a predecir el riesgo de no cumplimiento de un proyecto?

1.3. Hipótesis

El modelo analítico construido a partir de la información disponible, podrá predecir el incumplimiento académico de los proyectos a partir de la clasificación de riesgo asociada a ciertas variables.

2. Alcance del proyecto

La información se encuentra almacenada en una base de datos administrada por la Subdirección de Proyectos de Investigación. Se cuenta con datos de los proyectos aprobados desde el año 1991 hasta la fecha, lo que se encuentran desagregados por proyecto de investigación, en los tres principales fondos de investigación individual: Regular, Postdoctorado e Iniciación en investigación.

2.1. Variables

Item	Variables	Descripción de la variable
Identificación del proyecto	Folio	
	Año	
	Duración	
	Instrumento	Fondo de investigación (regular, post-doctorado o iniciación)
Identificación Institución	Nombre	
	Tipo	
	Principal	Institución principal asociada
	Secundaria	Institución secundaria asociada
	Región	
	Comuna	
Identificación del investigador principal y asociados	RUT	
	Calidad del Investigador	Relación entre el investigador y el proyecto
	Sexo	
	Nacionalidad	
	País	
	Comuna	
	Institución	
Carta de Adjudicación	Puntaje de proyecto	
	Puntaje de corte	
Estado de proyecto	Fecha de inicio	
	Fecha de término	
	Estado proyecto	Resultado de la postulación
Presupuesto	Presupuesto solicitado	
	Presupuesto asignado	
Bioética	Estado bioética	Estado de cumplimiento bioético
	Fecha revisado	
	Fecha actualización	
Proyecto	Disciplina	Tipo de disciplina según OCDE
	Área	Área del conocimiento
	Cambio de disciplina	
Grupo de evaluación	Código de grupo	Grupo de evaluadores del informe académico
	Código de área del grupo	
Situaciones especiales	Cambio institución	
	Modificación académica	
Informe académico	Estado etapa	Resultado de la evaluación al informe

Las variables se encuentran en una base de datos relacional y se necesita una revisión de esta para poder extraer muestras, transformarlas y poder analizarlas.

A continuación se presenta un modelo relacional de las variables.


hst_dis_asoc
folioproy int
dis_asoc int
f_act_disc date
c_disciplina int
dis_asoc
c_disciplina int
folioproy int
gr_rel_disc int
disciplina_ocde
c_disciplina int
gl_disciplina string
disc_asoc_ocde
folioproy int
c_disciplina int
est_etapa
c_est_etapa int
eval_concurso
folioproy int
rut_persona int
puntaje decimal
disciplina
c_disciplina int
gl_disciplina string
c_grup_est int
cartalinea
folioproy int
nolugar int
n_espera int
puntaje_proy decimal
puntaje_corte decimal
patrocinante
c_instit int
c_t_patrocin int
gr_imp_patroc int
hst_est_proy
folioproy int
c_est_proyecto int
est_proy
c_est_proy int
gl_est_proy string
institucion
c_instit int
c_t_instit int
nom_instit string
nom_completo string
c_region int
c_comuna int
c_ciudad int
c_pais int
rut_instit int
dv_rut int
hst_etapanual
folioproy int
agno_etapa int
c_est_etapa int
f_est_etapa date
calidad_rel
folioproy int
rut_persona int
c_t_relac int
c_instit int
proyecto
folioproy int
c_t_concurso int
agno_concurso int
c_consejo int
c_area int
c_t_proyecto int
duracion int
c_region int
c_est_proyecto int
f_ini_proy date
f_fin_proy date
situacion_especial
c_situacion int
folioproy int
agno_etapa int
c_t_situacion int
c_est_situacion int
instemp
c_instit int
nivel1 int
nivel2 int
nivel3 int
nivel4 int
hst_situacion_especial
c_situacion int
folioproy int
agno_etapa int
c_t_situacion int
c_est_situacion int
f_act_situac date
proy_bioetica
folioproy int
c_est_bioetica int
f_revisado date
f_act_bio date
item_presup
c_it_presup int[pk]
folioproy int
agno_etapa int
m_sol_ipresup decimal
m_asi_ipresup decimal
persona
rut_persona int
f_nacimiento date
sexo char
c_nacion int
c_idioma int
c_comuna int
c_ciudad int
c_region int
c_pais int
c_est_civil int
dedicacion
rut_persona int
folioproy int
agno_etapa int
ct_horas int
c_t_financ int
etapanual
folioproy int
agno_etapa int
c_est_etapa int
f_est_etapa date
f_ini_est date
f_fin_est date
miembro_g_e
c_grupo_est int
grupo_est
c_grupo_est int
c_area_grupo int
dbdiagram.io

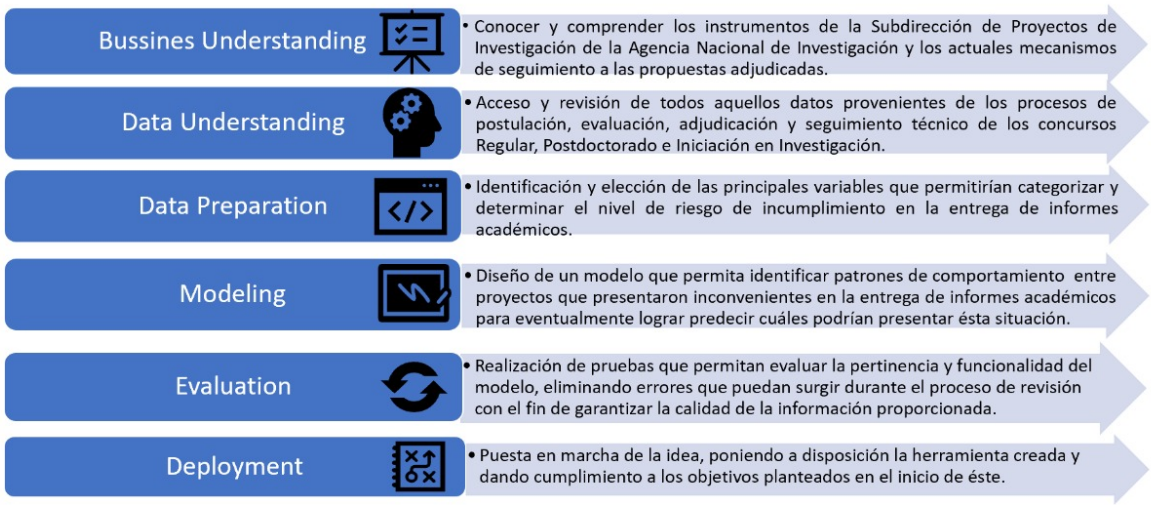
3. Descripción de metodología

En la Subdirección de Proyectos de Investigación existe información respecto de los proyectos adjudicados, la cual se recolecta a través de los procesos de postulación, evaluación, adjudicación y seguimiento técnico. A partir de esta información histórica, se buscará identificar las variables que puedan incidir en el incumplimiento académico de un proyecto y de esta manera diseñar un modelo de clasificación de riesgo.

Esta información se tratará de manera anónima para resguardar la confidencialidad de los/as investigadores/as, para estos efectos se considera la transformación de las variables de identificación originales (RUT, folio) a una nueva variable de identificación para efectos de procesamiento de la información.

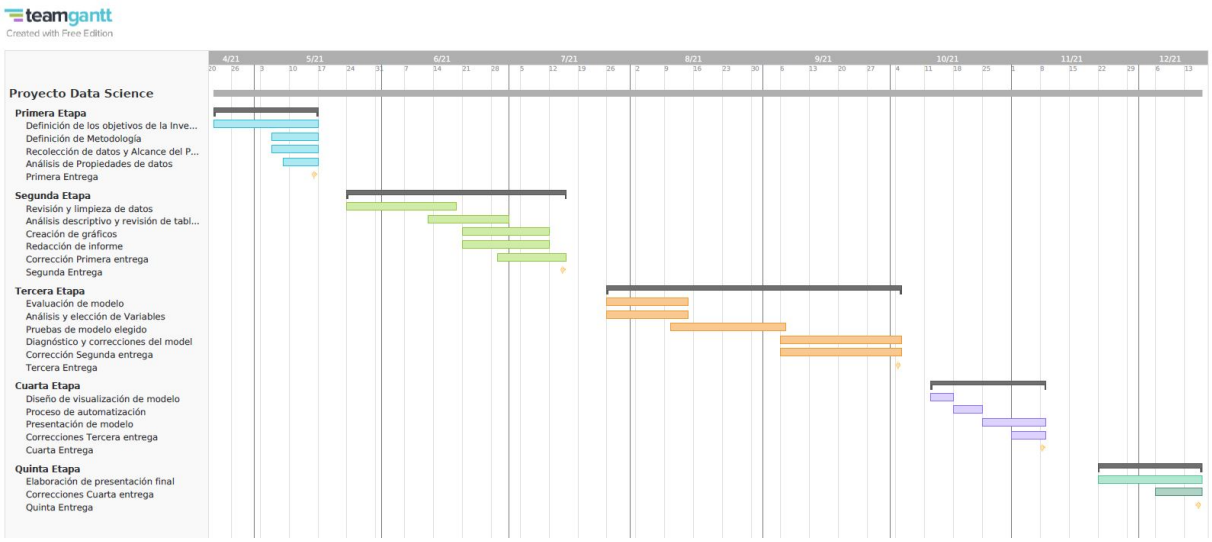
Para abordar este proyecto se utilizará el método CRISP-DM (Cross Industry Standard Process for Data Mining), que desglosa el proceso en seis fases: 1) comprensión del negocio, 2) comprensión de los datos, 3) preparación de los datos, 4) modelado, 5) evaluación de los resultados y 6) puesta en marcha. En las primeras etapas, los esfuerzos estarán puestos en conocer y comprender los mecanismos de control y seguimiento de los proyectos adjudicados, para lograr identificar las variables que pueden ser utilizadas para evaluar el riesgo de incumplimiento. Una vez identificadas las variables, se procederá a la preparación de los datos para su posterior análisis y modelación. Finalmente, se evaluarán los resultados respecto de los objetivos iniciales y la aplicación de la herramienta por parte de los profesionales de la Subdirección.

Los datos necesarios para desarrollar este proyecto serán obtenidos directamente desde la Subdirección. Estos se encuentran almacenados en una base de datos institucional y el periodo disponible es desde el año 1991 a la fecha. En este momento, se cuenta con la autorización para llevar a cabo esta iniciativa y utilizar los datos disponibles.



4. Carta Gantt

A continuación un cronograma con las principales actividades consideradas en el proyecto.



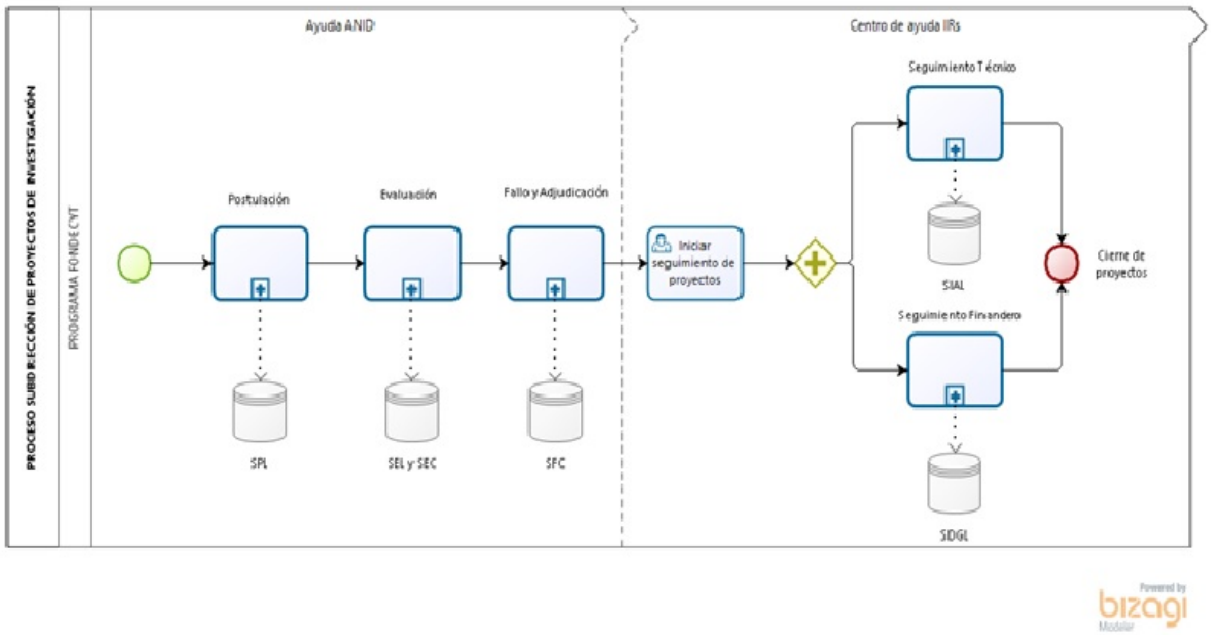
5. Anexo

Se incorporó la carta gantt como se sugirió en la primera entrega.

SEGUNDO REPORTE

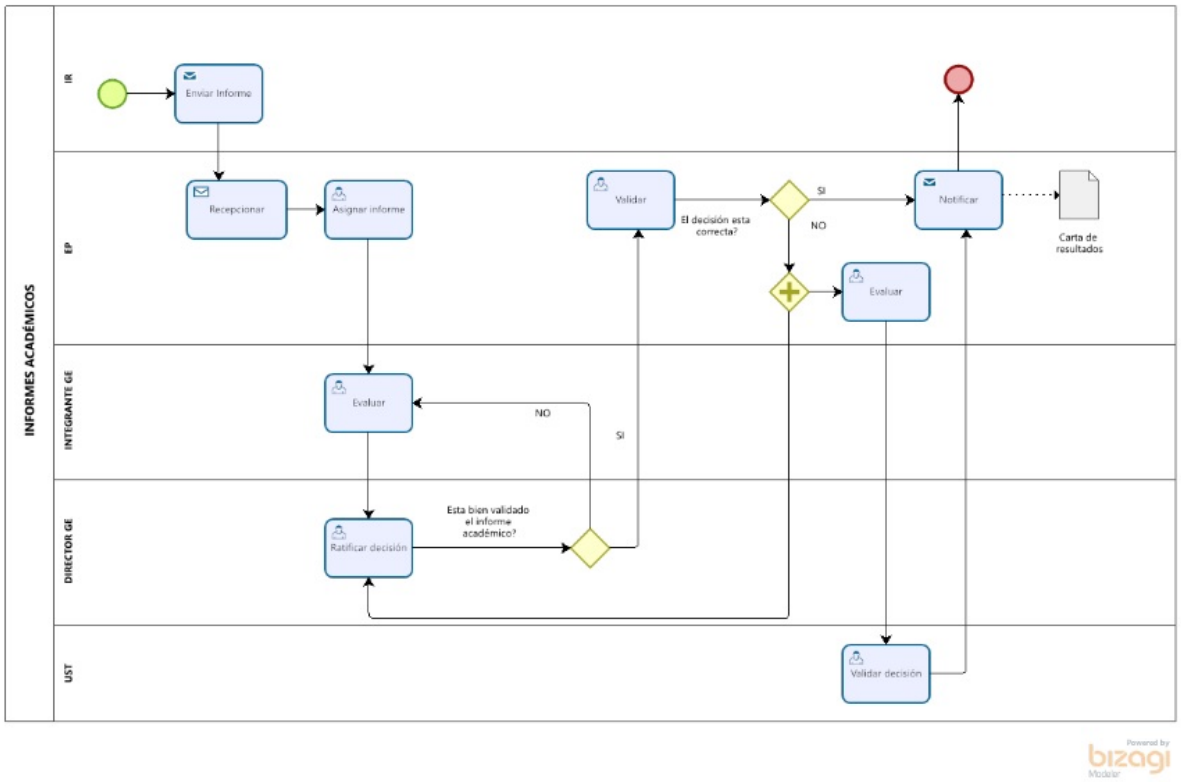
1. Comprensión del negocio

El ciclo de un proyecto que postula a los instrumentos de FONDECYT de la ANID se gráfica en la siguiente imagen. Se inicia con el proceso de postulación, seguido por la evaluación y posterior adjudicación. Si un proyecto resulta adjudicado, entonces entra a un proceso de seguimiento, el cual se estructura por dos aspectos: el aspecto técnico y el aspecto financiero. En este trabajo nos acotaremos al seguimiento técnico.



Parte de la evaluación técnica incluye la elaboración y entrega de un informe académico, el cual es evaluado por la ANID, y a partir de su resultado se puede clasificar el proyecto como 'aprobado', 'rechazado' o 'en

proceso'. A continuación se presenta un flujo con los detalles respecto de la revisión y fallo de los aspectos comprometidos.



2. Comprensión de los datos

Los datos están estructurados en nueve tablas, todas disponibles en formato excel. A continuación, se identifica cada tabla, con la cantidad de variables y registros que contienen. Previamente, cada tabla fue innominada a requerimiento de la Institución. En total existen 55 variables únicas. La *primary key* de todas las tablas corresponde al código que identifica a cada proyecto, denominado `cod_folio`.

Nombre tabla	Número de variables	Número de registros
Data general	25	14.850
Etapa	8	47.253
Bioetica	11	6.010
Disciplinas	13	4.093
Miembros	8	3.198
Monto	5	14.874
Situacion especial	13	4.221
Cambios_a	9	40.741
Cambios_b	8	40.083

En la siguiente figura se muestra un modelo relacional para los datos.


ge_miembros
cod_rut Int
gl_disciplina String
gl_grup_est String
gl_miembro String
f_nacimiento Date
f_ini_vge Date
f_fin_vge Date
sexo Char
carta_notas
cod_folio Int
c_est_proyecto Int
c_t_proyecto Int
gl_grup_est String
fecha_carta Date
puntaje_proy Int
nolugar Int
puntaje_corte Int
situaciones_especiales
cod_folio Int
c_t_proyecto Int
gl_grup_est String
gl_est_situacion String
gl_t_situacion String
f_fin_proy Date
f_ini_situacion Date
agno_etapa Int
agno_concurso Int
nom_institt String
titulo String
sexo Char
data_gral
cod_folio Int
cod_rut Int
c_t_relac Int
c_region_IR Int
c_disciplina Int
c_region_ejecucion Int
c_region_institucion Int
gl_nacion String
gl_est_proy String
gl_tiproy String
gr_rel_disc Int
f_ini_proy Date
f_fin_proy Date
f_nacimiento Date
agno_concurso Int
sexo Char
duracion String
titulo String
grupo_estudio String
disciplina String
institucion String
facultad String
depto_unidad String
sector_aplicacion String
ppto_asig_miles Int
hist_etapa
cod_folio Int
c_t_proyecto Int
gl_est_etapa String
gl_grup_est String
f_est_etapa Date
f_fin_proy Date
agno_etapa Int
duracion String
cambios_disciplinas
cod_folio Int
cod_disc_nueva Int
c_est_proyecto Int
cod_rut Int
c_disciplina Int
c_grup_nuevo Int
c_grup Int
f_act_disc Date
disc_nueva String
ge_nuevo String
agno_concurso Int
disciplina String
ge String
bioetica
cod_folio Int
c_t_proyecto Int
gl_grup_est String
gl_est_bioetica String
agno_concurso Int
nom_institt String
fac_institt String
f_revisado Date
f_act_bio Date
f_ini_proy Date
f_fin_proy Date
monto_solicitado
cod_folio Int
c_t_proyecto Int
c_est_proyecto Int
agno_concurso Int
sum Int

Inicialmente trabajaremos con dos tablas, una que contiene los datos generales del proyecto denominada `data_gral` y otra llamada `etapa`, que incluye datos específicos sobre el avance del proyecto.

- La tabla `data_gral` incluye las siguientes variables: año proyecto, duración, fecha de término, tipo de concurso, disciplina, universidad, entre otras.
- La tabla `etapa` incluye las siguientes variables: fecha actualización del estado de etapa, el estado (aprobado, rechazado o en ejecución), entre otras.

3. Preparación de los datos

3.1 Carga y limpieza de datos

Se utilizaron las siguientes funciones para la carga y limpieza de datos:

- Carga de datos : `readxl::read_excel`
- Visualización formato variables : `dplyr::glimpse`
- Visualización de datos faltantes : `naniar::vis_miss`

Los principales problemas detectados en la limpieza y su solución se presentan a continuación:

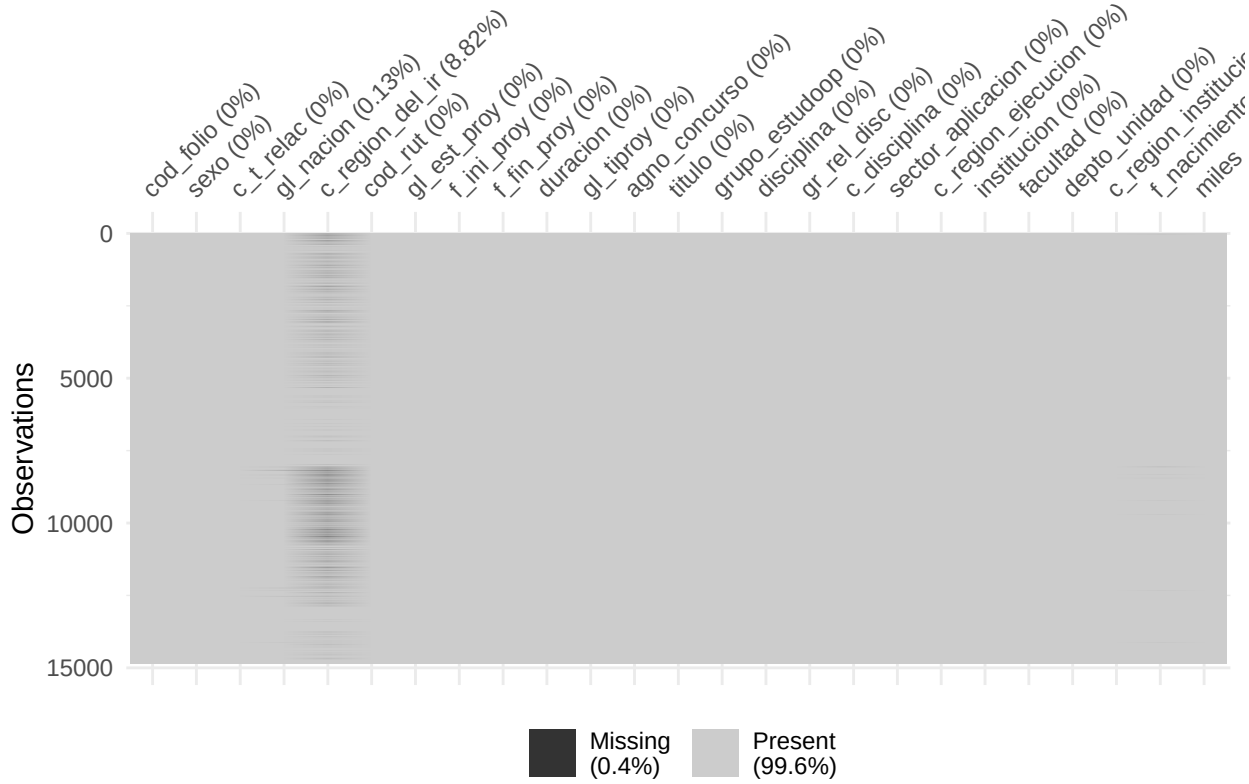
Problema	Solución
Variables carácter cargan como numérica	Se aplica función `as.character`
Nombres campos con espacios	Se aplica función `janitor::clean_names`
Entradas con datos faltantes	Se aplica `na.omit`

3.1.1 Limpieza tabla Data General

A continuación se muestra el retorno de las funciones `dplyr::glimpse` y `naniar::vis_miss` aplicadas a la tabla `data_gral` después de la limpieza.

```
## Rows: 14,850
## Columns: 25
## $ cod_folio            <chr> "1061", "1062", "1063", "1064", "1065", "1066", "~
## $ sexo                 <chr> "M", "M", "F", "F", "M", "M", "M", "M", "M", "M",~
## $ c_t_relac            <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",~
## $ gl_nacion            <chr> "CHILE", "CHILE", "CHILE", "CHILE", "CHILE", "CHI~
## $ c_region_del_ir      <chr> NA, NA, "13", "12", "13", NA, "13", "13", "13", "~
## $ cod_rut              <chr> "1978", "1910", "1256", "3501", "1614", "1576", "~
## $ gl_est_proy          <chr> "APROBADO", "APROBADO", "APROBADO", "APROBADO", "~
## $ f_ini_proy           <dttm> 2006-03-15, 2006-03-15, 2006-03-15, 2006-03-15, ~
## $ f_fin_proy           <dttm> 2009-03-15, 2008-03-15, 2009-03-15, 2010-03-15, ~
## $ duracion             <dbl> 3, 2, 3, 4, 3, 3, 4, 3, 3, 3, 3, 4, 4, 4, 3, 3, 4~
## $ gl_tiproy            <chr> "REGULAR", "REGULAR", "REGULAR", "REGULAR", "REGU~
## $ agno_concurso        <dbl> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2~
## $ titulo               <chr> "ESTUDIO DE LA RESPUESTA A EMBUTICION DE ACEROS D~
## $ grupo_estudoop       <chr> "INGENIERIA 1", "CS. ECONOM/ADMI", "SOCIOLOGIA CS~
## $ disciplina           <chr> "INGENIERIA DE MATERIALES", "FINANZAS", "CAMBIO S~
## $ gr_rel_disc          <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1",~
## $ c_disciplina         <chr> "86", "217", "164", "159", "43", "93", "36", "36"~
## $ sector_aplicacion    <chr> "TECNICAS DE MANUFACTURAS Y DE PROCESOS.", "CONOC~
## $ c_region_ejecucion   <chr> "13", "13", "13", "12", "13", "13", "13", "13", "~
## $ institucion          <chr> "UNIV.DE SANTIAGO DE CHILE", "UNIV.ADOLFO IBANEZ"~
## $ facultad             <chr> "FAC.DE INGENIERIA", "ESCUELA DE NEGOCIOS-SANTIAG~
## $ depto_unidad         <chr> "DEPTO.ING. METALURGICA", "ESCUELA DE NEGOCIOS-SA~
## $ c_region_institucion <chr> "13", "13", "13", "12", "13", "13", "13", "13", "~
```

```
## $ f_nacimiento         <dttm> 1960-02-01, 1972-09-08, 1946-11-26, 1972-09-08, ~
## $ miles_de_ppto_asig   <dbl> 32543, 16802, 28100, 124425, 116028, 45715, 84791~
```



Notar los siguientes aspectos relevantes: a) que en total se identificaron 14.850 proyectos para el periodo 2006-2021, b) que pueden ser clasificados por tipo de proyecto `gl_tiproy` (iniciación, postdoctorado y regular), y c) que la duración de cada uno puede variar entre uno y cuatros años, los cuales denominaremos etapas. [1]

3.1.2 Limpieza tabla Etapa

La tabla `etapa` incluye tantas entradas con el mismo número de folio como años de duración del proyecto. Por ejemplo, un mismo proyecto puede aparecer hasta cuatro veces (duró cuatro años) en dicha tabla, con dos entradas 'en ejecución', una 'rechazado' y otra 'aprobado', pero todas con diferentes fechas de actualización. Es por ello, que primero se tuvo que eliminar todos excepto el último registro por código de folio. Y también eliminar las filas con datos faltantes, porque no se sabe si esos proyectos se repiten dentro de la tabla pero para otro año.

Además, fue necesario re-clasificar las categorías que toma la variable `gl_est_etapa`, dado que tiene 24 categorías y estas no son informativas por si mismas. De este modo, la nueva variable `etapas_agrupadas` solo toma tres valores `APROBADO`, `RECHAZADO` o `EN PROCESO`. Se entenderá por:

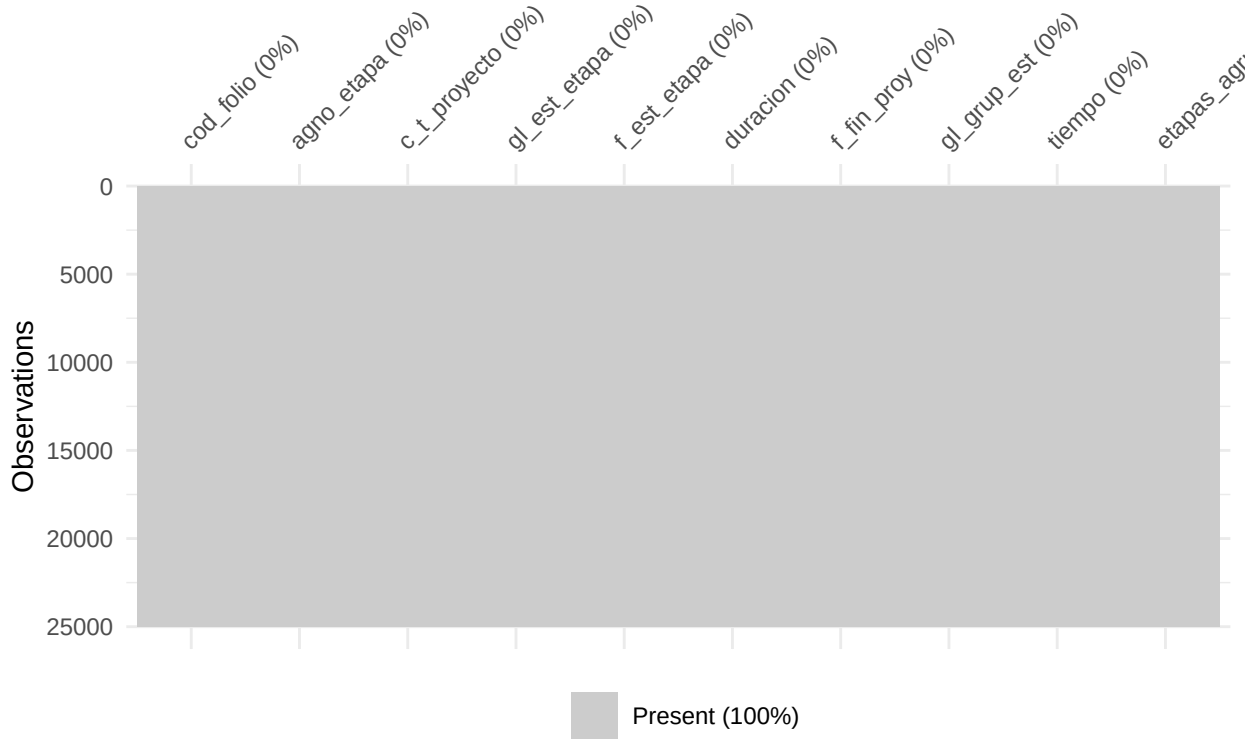
- `APROBADO`: cumple con las exigencias académicas establecidas en conformidad a las bases del concurso.
- `RECHAZADO`: no cumple con las exigencias académicas establecidas en conformidad a las bases del concurso, por lo tanto, se exigen rectificaciones que puedan cambiar esta situación.
- `EN PROCESO`: aquellos proyectos vigentes que se encuentran en ejecución, cuyo informe académico se encuentra recibido o en evaluación.

[1] Recordar que la tabla etapa actualiza el estado del proyecto anualmente

Por último, se añade una columna tiempo que equivale a la diferencia de tiempo límite inicial del proyecto menos el tiempo que tomó el proyecto en terminar.

A continuación se muestra el retorno de las funciones `dplyr::glimpse` y `naniar::vis_miss` aplicadas a la tabla `etapa` después de la limpieza.

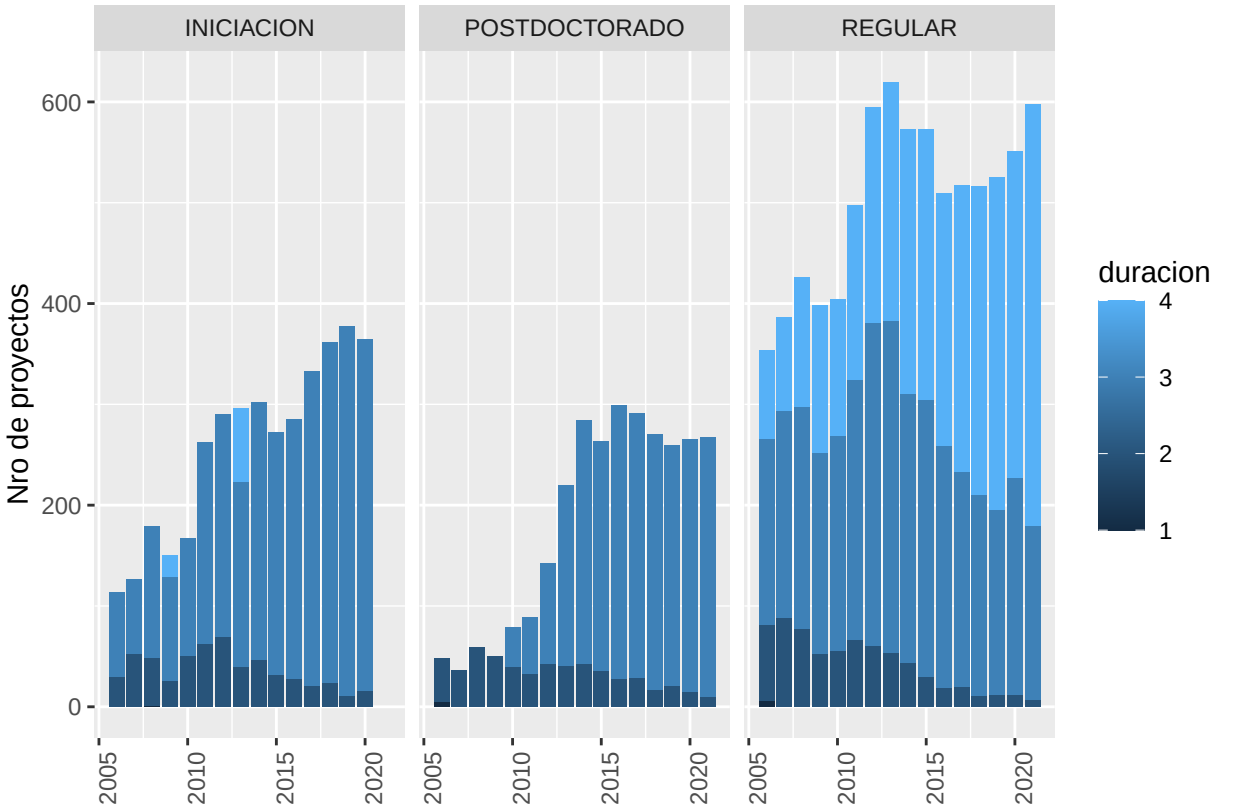
```
## Rows: 25,016
## Columns: 10
## $ cod_folio        <chr> "1061", "1061", "1061", "1062", "1062", "1063", "1063~
## $ agno_etapa       <dbl> 2006, 2007, 2008, 2006, 2007, 2006, 2007, 2008, 2006,~
## $ c_t_proyecto     <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1"~
## $ gl_est_etapa     <chr> "APROBADA", "APROBADA", "APROBADA", "APROBADA", "APRO~
## $ f_est_etapa      <dttm> 2007-05-07, 2008-04-11, 2009-04-14, 2007-05-07, 2012~
## $ duracion         <dbl> 3, 3, 3, 2, 2, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3,~
## $ f_fin_proy       <dttm> 2009-03-15, 2009-03-15, 2009-03-15, 2008-03-15, 2008~
## $ gl_grup_est      <chr> "INGENIERIA 1", "INGENIERIA 1", "INGENIERIA 1", "CS. ~
## $ tiempo           <drtn> 678 days, 338 days, -30 days, 313 days, -1453 days, ~
## $ etapas_agrupadas <chr> "APROBADO", "APROBADO", "APROBADO", "APROBADO", "APRO~
```



Luego de la limpieza, la tabla `etapa` está lista para ser usada más adelante.

4. Análsis exploratorio

Primero, se grafican las variables de la tabla `data_gral` que se creen importantes, con la finalidad de ver cómo se relacionan entre sí. Estas son el número total de proyectos por año, desagregados por tipo de proyecto y duración.



Luego, mirando sólo la tabla `etapa`, se quiere conocer la frecuencia según la clasificación estado de proyecto. Esto es:

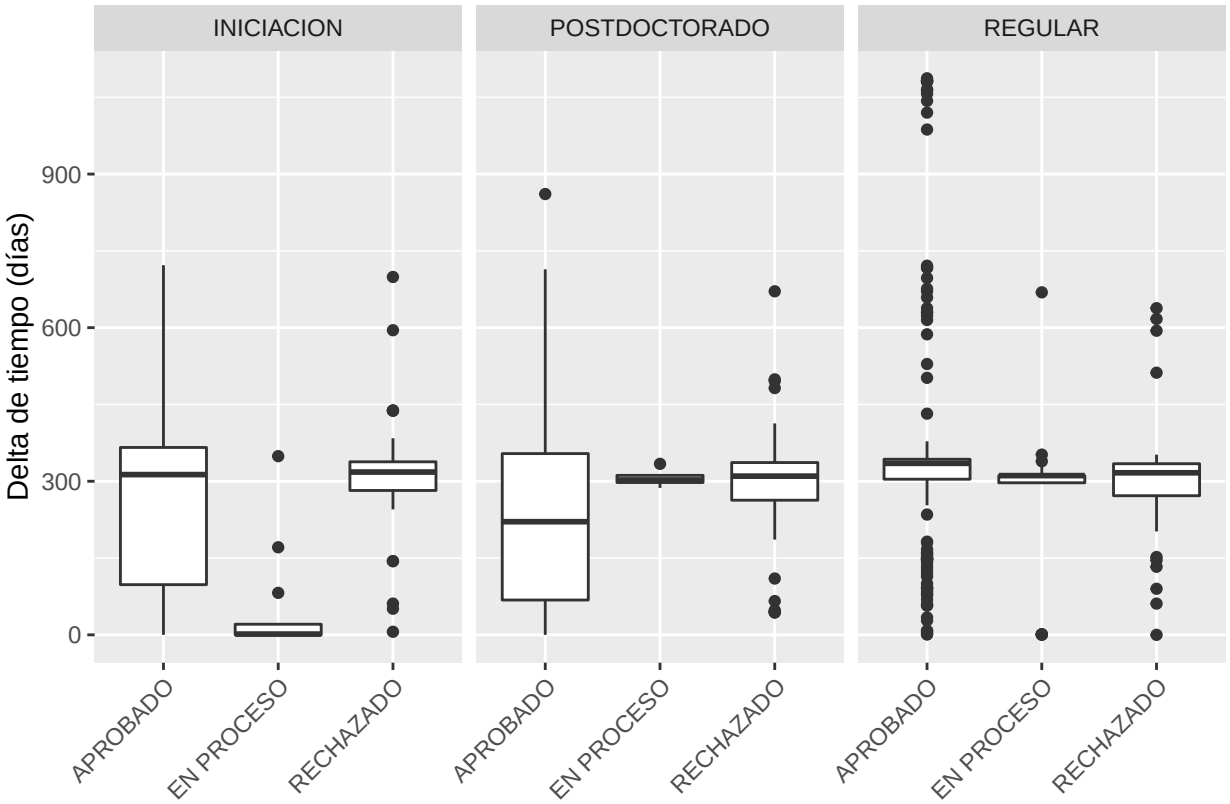
Var1	Freq
APROBADO	10189
EN PROCESO	173
RECHAZADO	642

En este punto, se ha extraído información relevante sobre ambas tablas por separado, pero ha llegado el momento de unirlas y responder preguntas más interesantes. Entonces, se crea una nueva tabla llamada `etapas_gral`, que es el resultado de un *inner_join* entre `data_gral`y `etapa`. La nueva tabla de análisis contiene 10.404 proyectos.

Verificación de supuestos:

1. Los proyectos en la categoría de rechazado se exceden en la duración oficial del proyecto.

Para verificar el supuesto, se realizó una gráfica del delta de tiempo desde el término de la última etapa del proyecto versus la fecha de termino del proyecto. Tiempo estimado en días. Además, se muestran los principales estadísticos de resumen.



gl_tiproy	etapas_agrupadas	mean_tiempo	max_tiempo	std_tiempo
INICIACION	APROBADO	304.67568	722	225.61522
INICIACION	EN PROCESO	46.07143	349	99.53157
INICIACION	RECHAZADO	310.07692	699	151.36893
POSTDOCTORADO	APROBADO	266.60684	861	220.01477
POSTDOCTORADO	EN PROCESO	306.50000	334	19.77372
POSTDOCTORADO	RECHAZADO	294.82857	671	138.42124
REGULAR	APROBADO	341.77620	1087	169.83027
REGULAR	EN PROCESO	286.83784	669	117.39878
REGULAR	RECHAZADO	305.00000	638	125.81345

Este supuesto no muestra una clara relación entre exceso en el tiempo de fin de proyecto, dado que muchos proyectos aprobados exceden por mucho la fecha límite.

2. La macro-zona de Chile influye en la posibilidad de caer en incumplimiento.

Para verificar la hipótesis, se caracterizaron las regiones de ejecución del proyecto en cuatro zonas. La tabla de análisis contiene 10.404 proyectos y se presenta a continuación.

	norte	centro	centro-sur	sur
APROBADO	322	7393	1378	1096
EN PROCESO	6	126	23	18
RECHAZADO	20	453	100	69

De los datos obtenidos no se puede concluir que la pertenencia del proyecto a una macro-zona particular influya en un posible incumplimiento.

3. ¿El incumplimiento Será una función del tiempo?

Para evaluar esta hipótesis, los proyectos fueron desagregados por estado y año. A continuación se presenta una tabla con el porcentaje de proyectos aprobados, rechazados y en proceso, por año.

	APROBADO	EN PROCESO	RECHAZADO
2006	0.3	0.0	0.0
2007	1.4	0.0	0.2
2008	4.3	0.0	1.6
2009	5.3	0.0	2.0
2010	5.5	0.0	2.0
2011	5.5	0.0	3.1
2012	6.7	0.0	2.5
2013	7.9	0.0	2.3
2014	9.2	0.0	3.6
2015	9.8	0.0	8.7
2016	10.7	0.6	10.6
2017	9.6	2.9	15.3
2018	9.5	7.5	17.6
2019	7.9	29.5	16.7
2020	6.5	59.5	13.9

En el periodo de análisis 2006-2020, se observa que el porcentaje de proyectos rechazados se va incrementando, concentrándose la mayor cantidad entre 2005 y 2020. Será relevante considerar este elemento en las siguientes etapas del proyecto, ya que pueden existir cambios en los procesos de evaluación que se deberían tener en cuenta.

4. Siguientes pasos

Avanzar en la construcción del panel de datos con todas las variables que puedan ayudar a explicar y predecir el incumplimiento de los proyectos de investigación.