UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Ingeniería en Software

"Desarrollo de un modelo inteligente de análisis de datos que prediga los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi"

Autor:
Ing. Rosero Valdiviezo, José Luis

Director:
Ing. Gallardo Corrales, Diego Eduardo

Latacunga, 2022

Planteamiento del Problema

En la actualidad se ha incrementado la demanda al acceso a la Educación Superior Pública

En el 1P del 2020 las IES ofertaron alrededor de 113,072 cupos siendo un 31% más que el 1P del 2019

Alta demanda de cupos a las IES con una alta tasa de deserción en los primeros niveles que a la larga afecta a la eficiencia terminal

Formulación del Problema

¿ Cómo predecir los posibles casos de deserción de estudiantes en los primeros niveles del Instituto Superior Tecnológico Cotopaxi ?

Objetivos de la Investigación

Objetivo General

Desarrollar un modelo inteligente de análisis de datos por medio de un algoritmo de Machine Learning que permita la identificación de patrones de comportamiento o causalidad en los datos de forma temprana, sobre los posibles casos de deserción estudiantil y aplicarlo en un caso de estudio.

Objetivos Específicos

1. Determinar las variables de causalidad que formarán parte del set de datos
2. Determinar los requerimientos que debe cumplir el set de datos y el modelo inteligente de análisis de datos
3. Validar el modelo inteligente de análisis de datos con set de datos de diferentes periodos académicos

Variables de la Investigación

Variable Independiente

Desarrollar un modelo inteligente de análisis de datos.

Variable Dependiente

Predecir los posibles casos de deserción en etapas tempranas de los estudiantes de los primeros niveles en el Instituto Superior Tecnológico Cotopaxi

- **Deserta**
- **NO Deserta**

Metodología de Investigación

Tipo de Investigación

Aplicada

- **Se centra en solucionar un problema concreto**

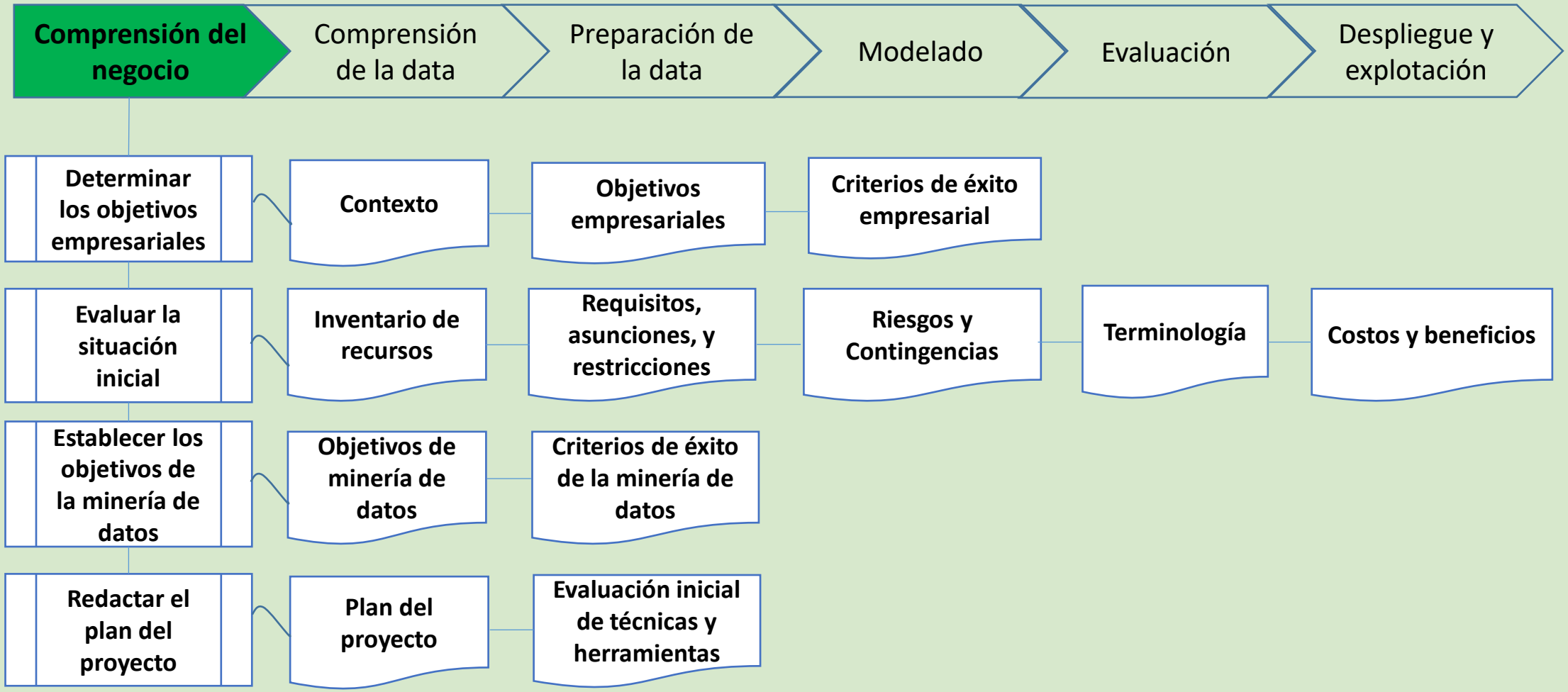
Niveles de Investigación

Investigación Exploratoria

Investigación Descriptiva

Metodología CRISP-DM

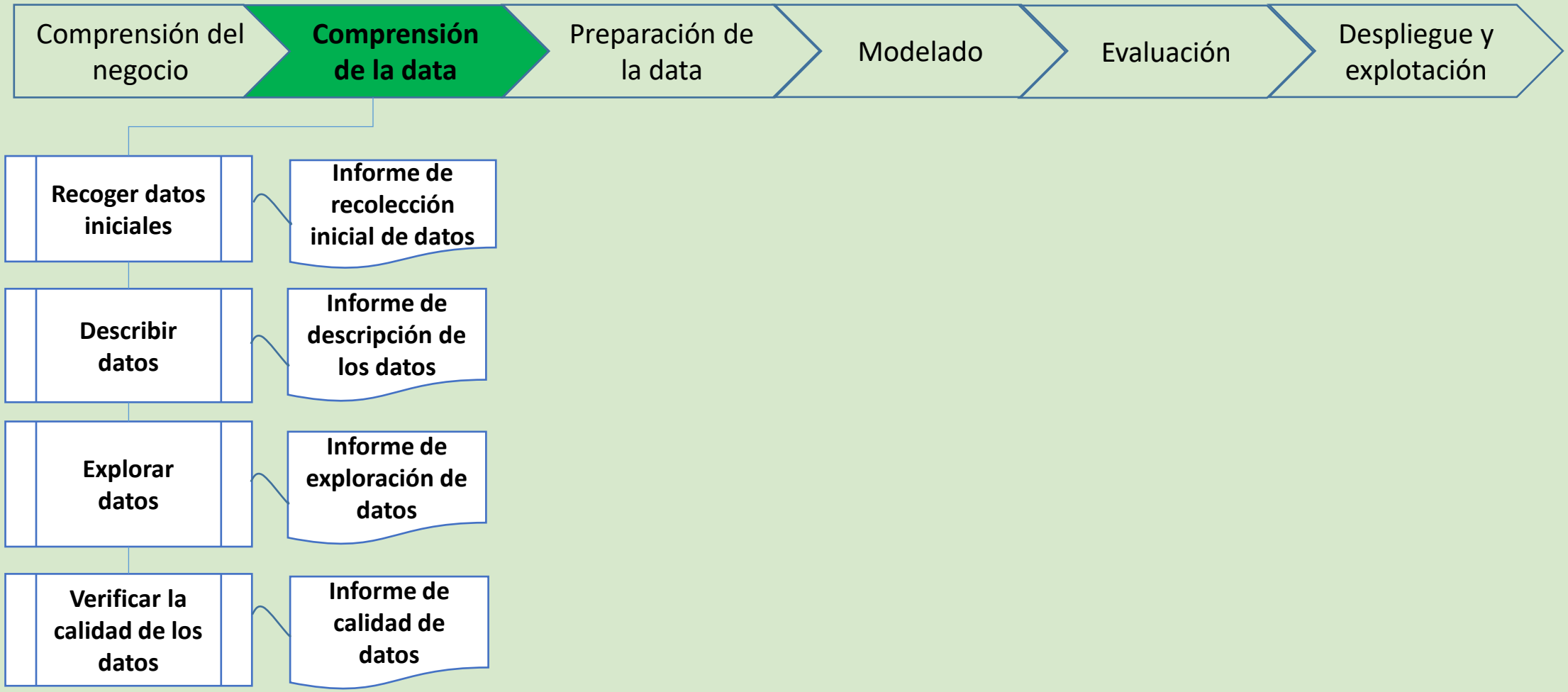
Fase 1. Comprensión del negocio



Planificación y seguimiento de la Metodología

Fase	Semanas	Personal
1. Comprensión del negocio	3	Director Bienestar, Investigador
2. Comprensión de la data	4	Director Tics, Investigador
3. Preparación de datos	4	Investigador
4. Modelado	4	Investigador
5. Evaluación	2	Director Bienestar, Investigador
6. Implementación o Despliegue	1	Departamento de TI, Investigador

Fase 2. Comprensión de los datos

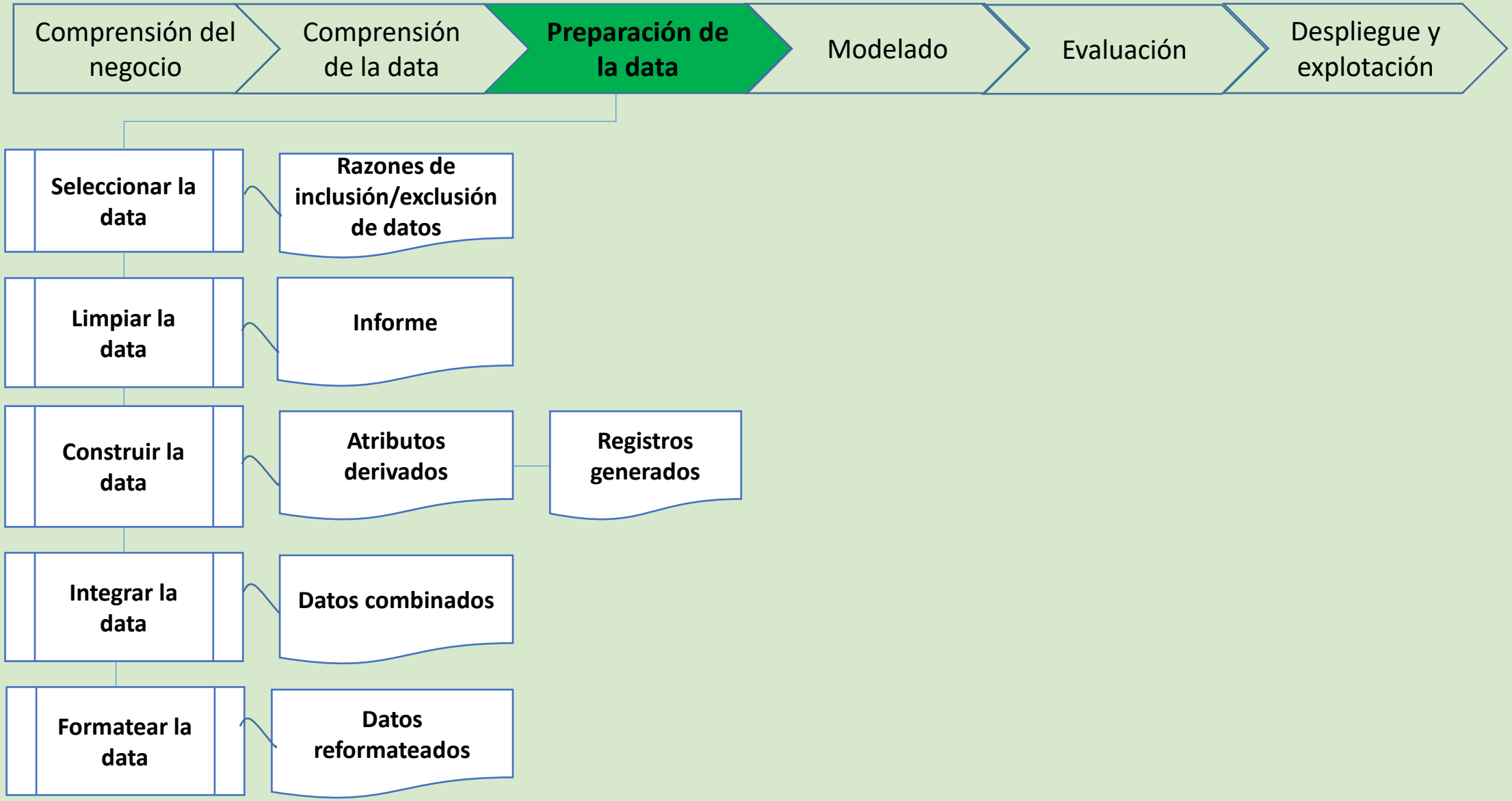


Objetivo Específico 1. Determinar las variables de causalidad que formarán parte del set de datos

Lista de Variables del DataSet

No	Variable
1	Edad
2	Tiene_discapacidad
3	Formación_padre
4	Formación_madre
5	Estado_civil
6	Genero
7	Etnia
8	Region_pais
9	Bono_desarrollo
10	Ocupación
11	Destino_ingresos
12	Ingreso_total_hogar
13	Cantidad_miembros_hogar
14	Desercion

Fase 3. Preparación de los datos

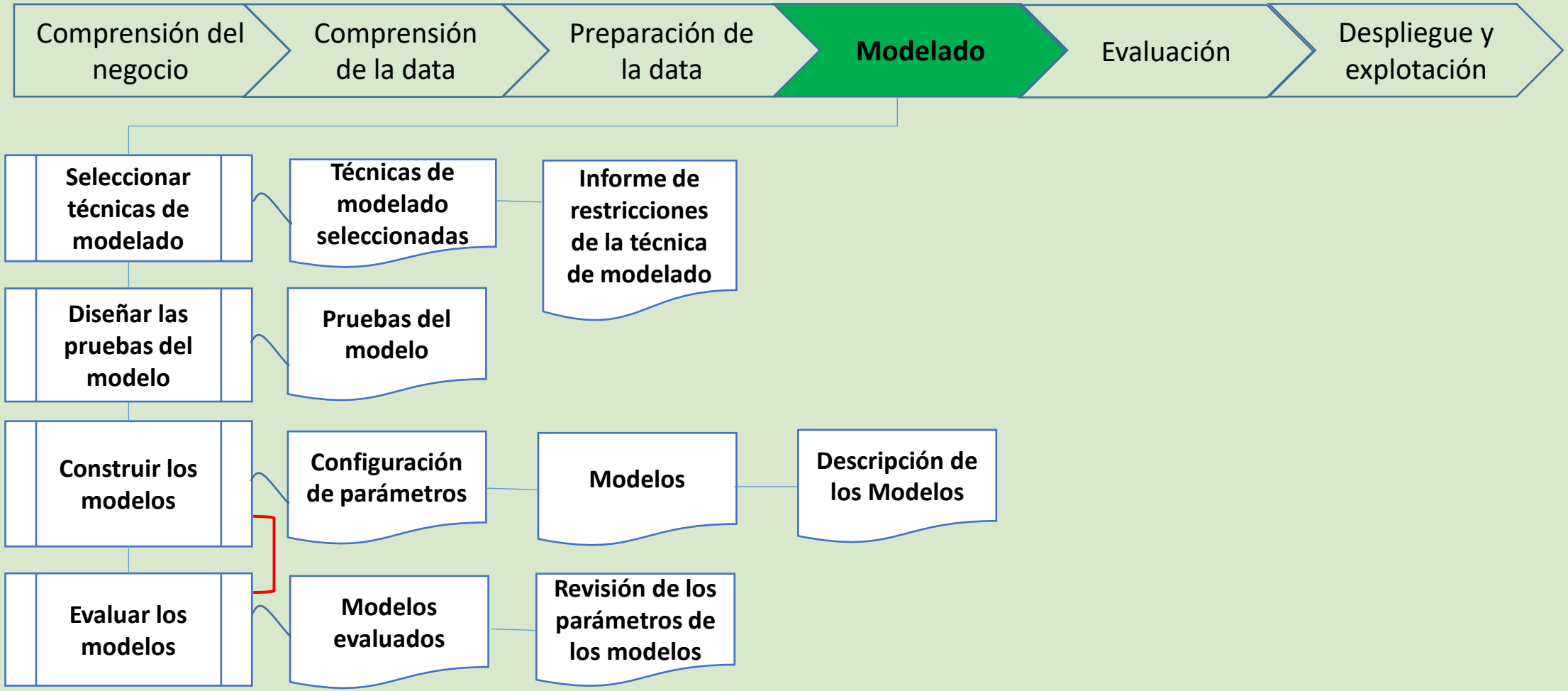


Lista de Variables del DataSet

```
raw_data2.head()
```

	edad	tiene_discapacidad	formacion_padre	formacion_madre	estado_civil	genero	etnia	region	bono_desarrollo	ocupacio
0	24	0	1	1	3	2	2	3	0	
1	44	0	1	1	3	1	2	3	0	
2	22	0	2	2	3	2	2	3	0	
3	21	0	1	1	3	2	2	3	0	
4	21	0	1	2	3	2	2	3	0	

Fase 4. Modelado



Modelos

```
X = raw_data2.drop('desercion', axis=1).values
y = raw_data2['desercion'].values
print('X shape: {}'.format(np.shape(X)))
print('y shape: {}'.format(np.shape(y)))

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
X shape: (2420, 13)
y shape: (2420,)
```

```
dt = tree.DecisionTreeClassifier(criterion='entropy', max_depth=5)
```

```
dt.fit(X_train, y_train)
```

```
##  Realiza la predicción
y_pren = dt.predict(X_test)
```

```
# Graficar la matriz de confusion
cm = confusion_matrix(y_test, y_pren)
cm_norm = cm/cm.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix(cm_norm, classes=dt.classes_, title='Matriz de Confusion')
print('Matriz de Confusión')
print(cm)
X_test.shape
```

```
Matriz de Confusión
[[319   7]
 [126  32]]

(484, 13)
```

```
rf = RandomForestClassifier(n_estimators=100,
                            criterion='entropy',
                            max_depth=5,
                            max_features="sqrt",
                            bootstrap=True,
                            max_samples= 2/3,
                            oob_score=True)
```

```
rf.fit(X_train, y_train)
```

```
##  Prediccion
prediction_test = rf.predict(X=X_test)
```

```
# Matriz de Confusión
cm_rf = confusion_matrix(y_test, prediction_test)
cm_norm_rf = cm_rf/cm_rf.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix(cm_norm_rf, classes=rf.classes_, title='Matriz de Confusión')
print('Matriz de confusion')
print(cm_rf)
X_test.shape
```

```
Matriz de confusion
[[316  10]
 [116  42]]

(484, 13)
```

```
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test  = scaler.transform(X_test)
```

```
mlp=MLPClassifier(hidden_layer_sizes=(10,), max_iter=500, solver='lbfgs')
```

```
mlp.fit(X_train, y_train)
```

```
y_pred_nn = mlp.predict(X_test)
```

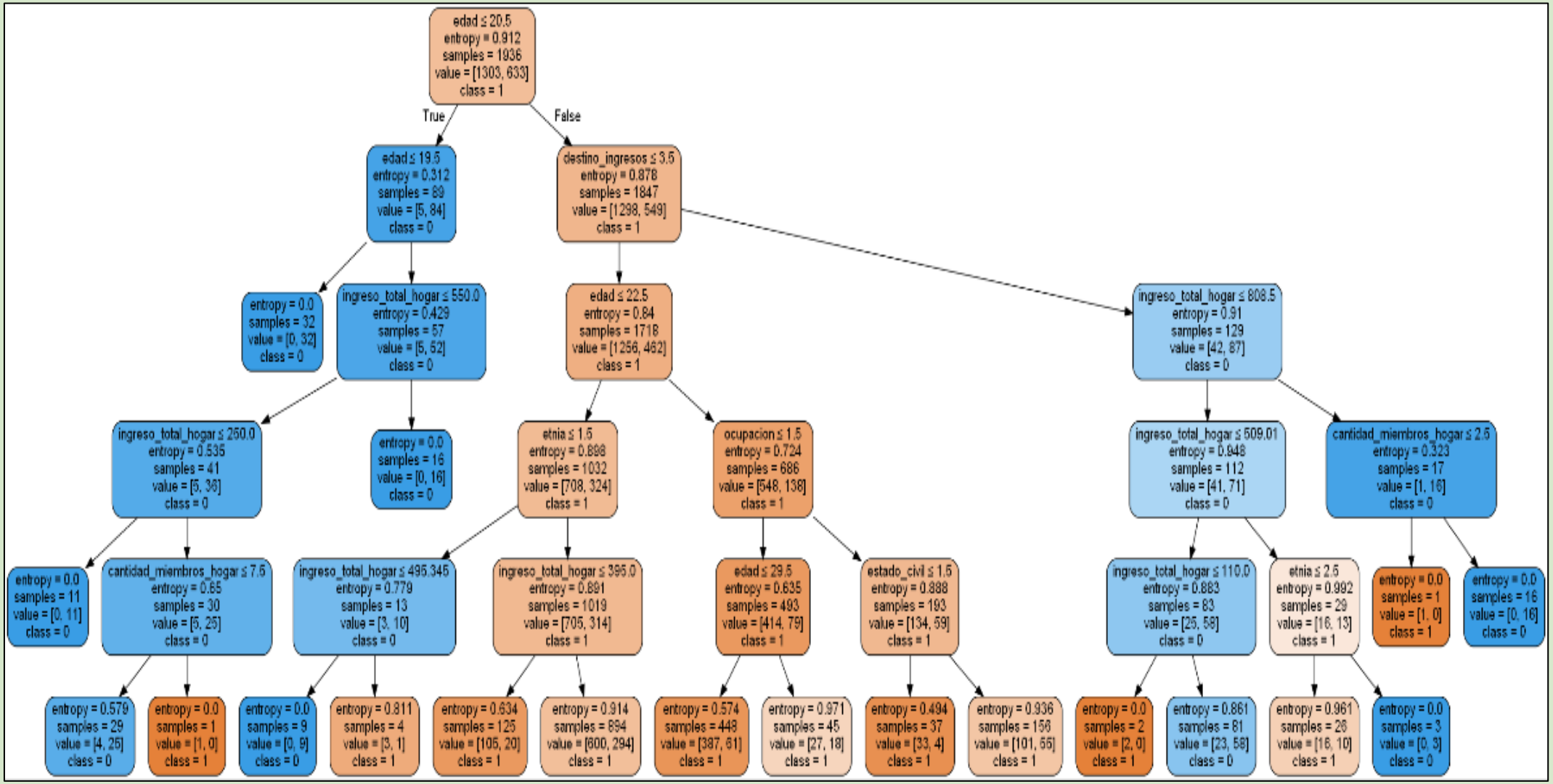
```
#matriz de confusion
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix

cm_nn = confusion_matrix(y_test, y_pred_nn)
cm_norm_nn = cm_nn/cm_nn.sum(axis=1)[:, np.newaxis]
plt.figure()
plot_confusion_matrix1(cm_norm_nn, classes=rf.classes_, title='Matriz de Confusión')
print(cm_nn)
X_test.shape
```

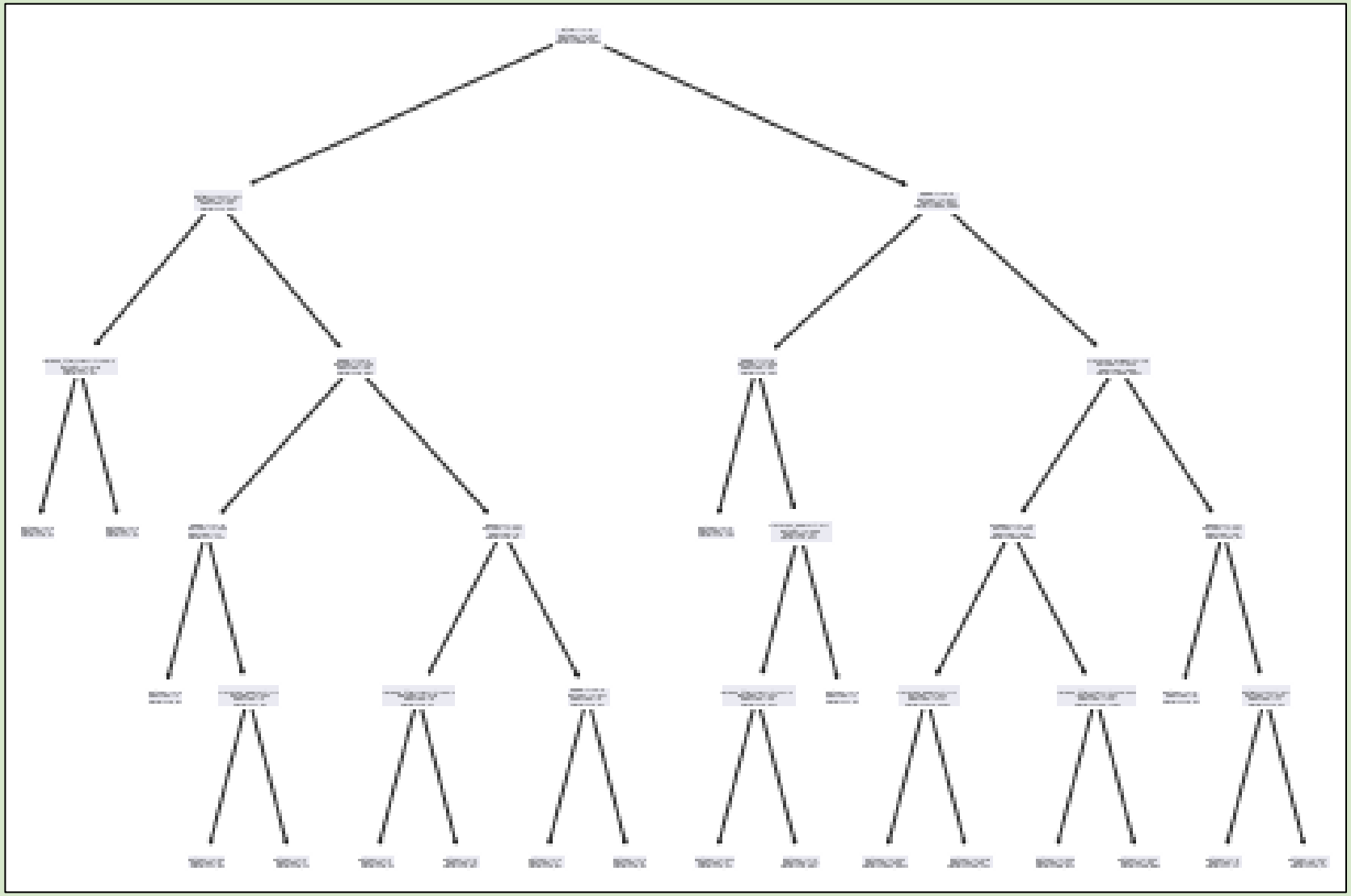
```
[[299  27]
 [110  48]]

(484, 13)
```

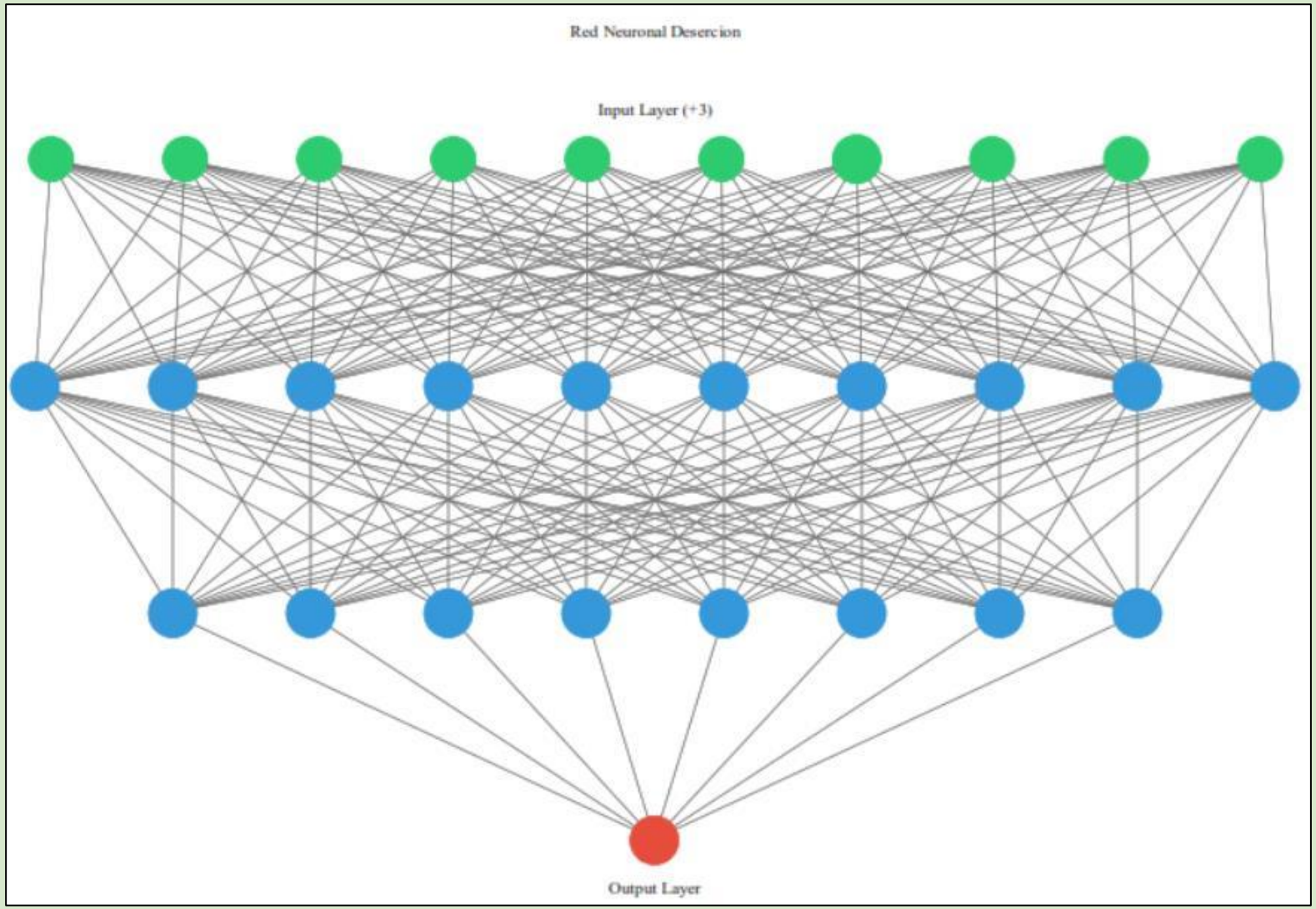
Modelo – Árbol de decisión



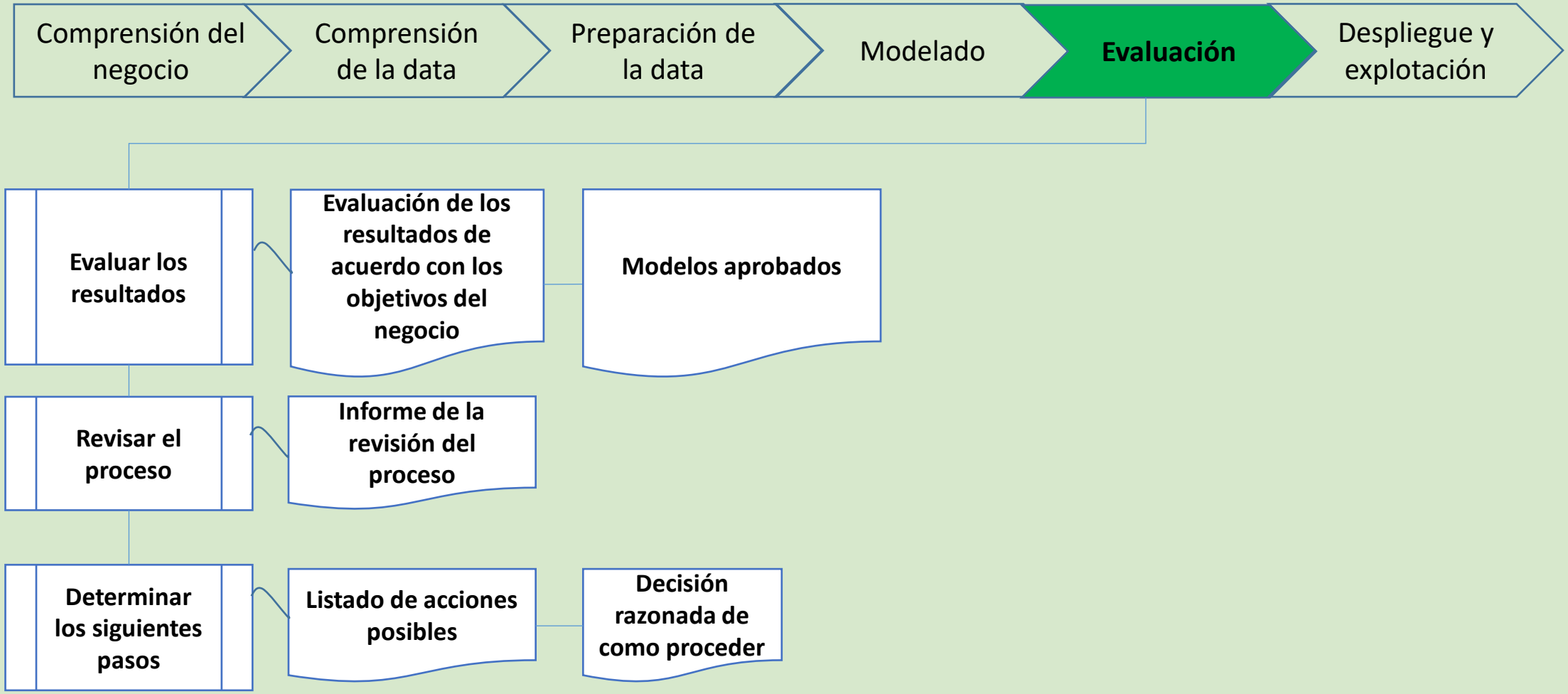
Modelo – Bosques aleatorios



Modelo – Redes Neuronales



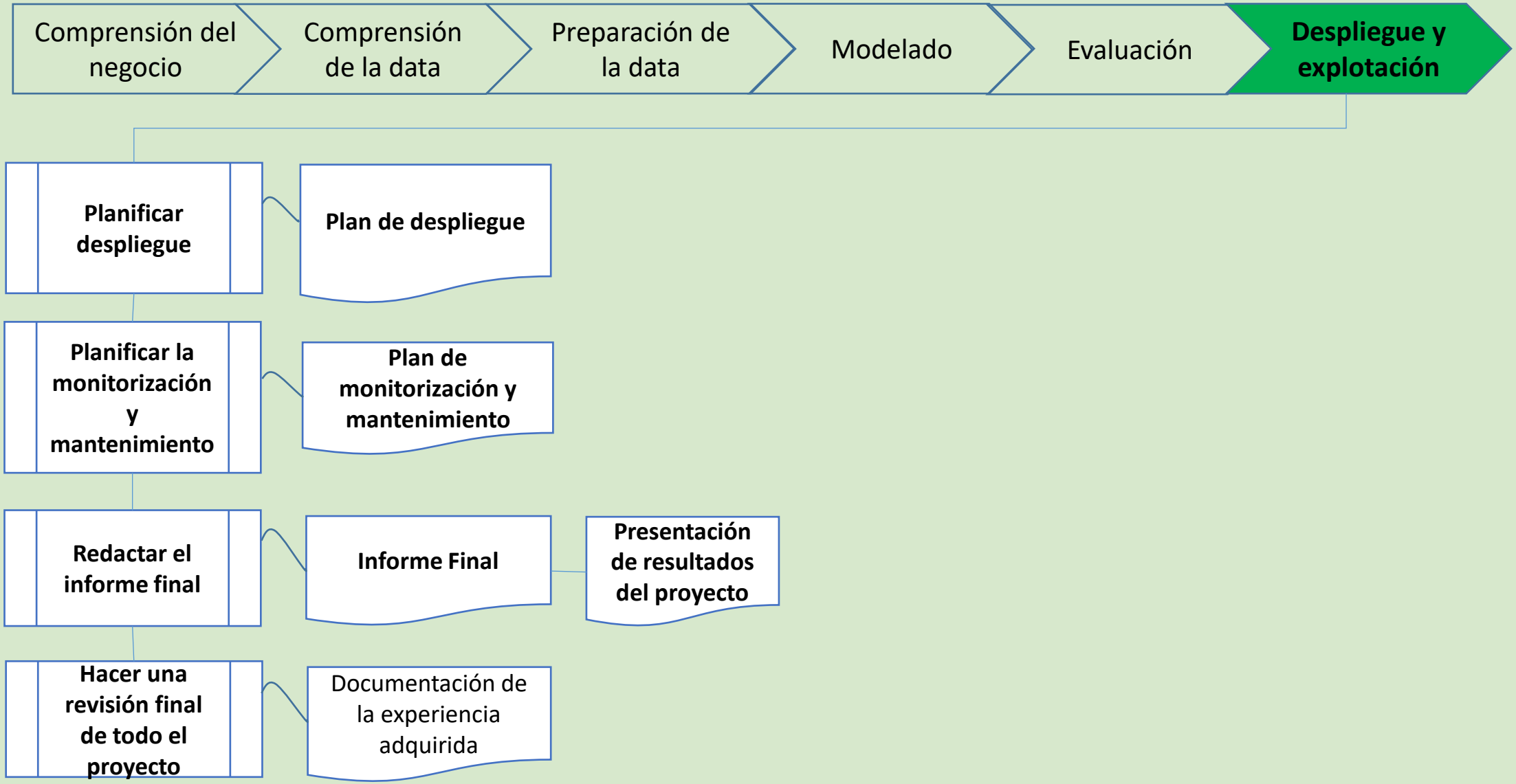
Fase 5. Evaluación



Evaluación

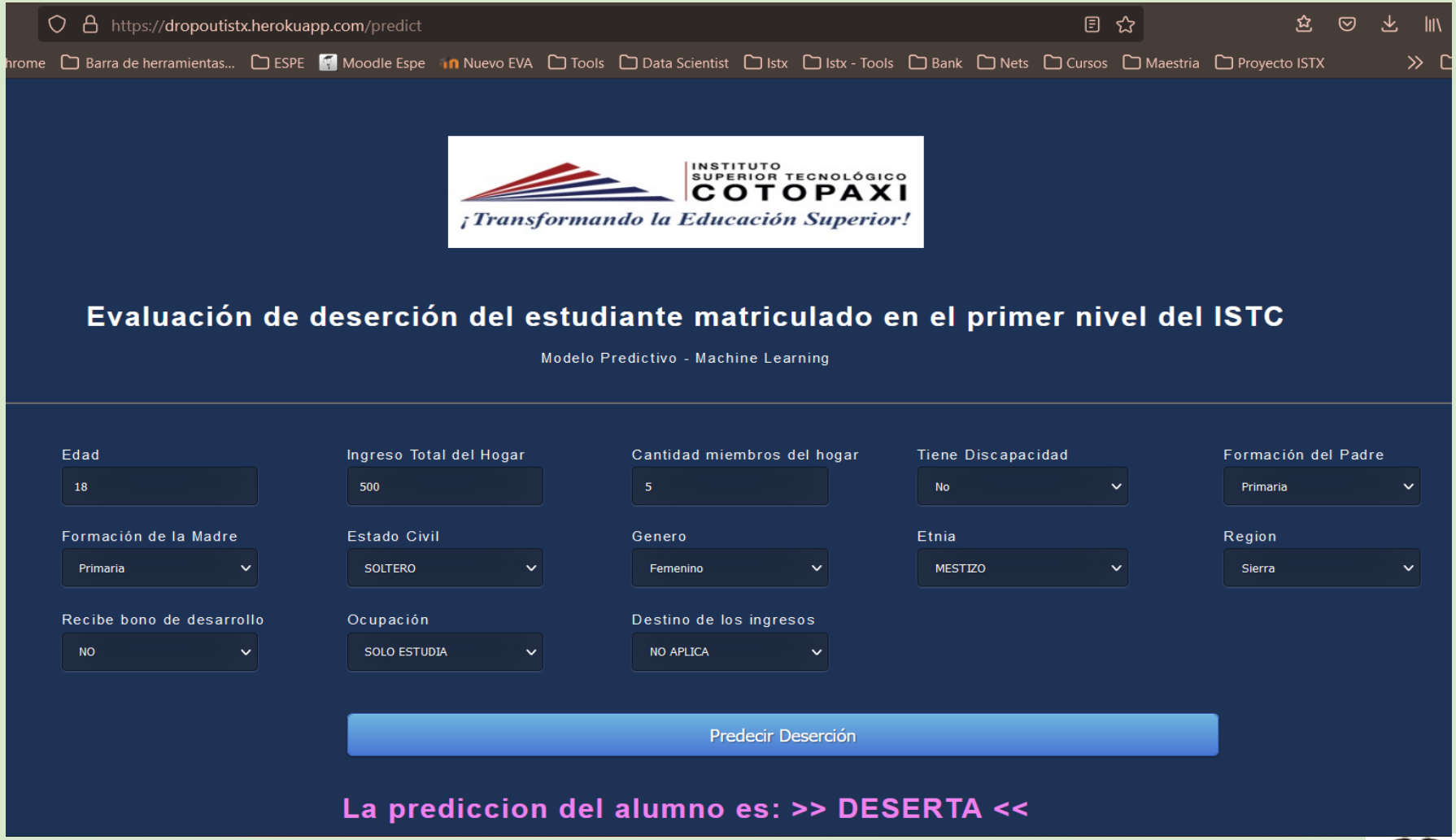
Modelo/Métrica	Exactitud	Precisión	Sensibilidad	Especificidad	Tasa de error	Vp	Vn	Fp	Fn	Roc	F1 Score
ML1. Árboles de decisión	0.7252	0.7169	0.9785	0.2025	0.2748	319	32	126	7	0.6208	0.8274
ML2. Random Forest	0.7397	0.7315	0.9693	0.2658	0.2603	316	42	116	10	0.6176	0.8337
ML3. Redes Neuronales	0.7169	0.7311	0.9172	0.3038	0.2831	291	39	128	27	0.6104	0.8136

Fase 6. Despliegue y Explotación



Objetivo General. Desarrollar un modelo inteligente de análisis de datos por medio de un algoritmo de Machine Learning que permita la identificación de patrones de comportamiento o causalidad en los datos de forma temprana, sobre los posibles casos de deserción estudiantil

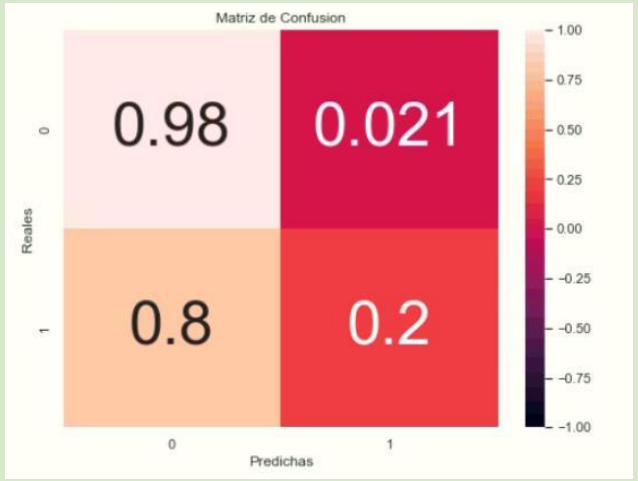
Despliegue



Objetivo específico 3. Validar el modelo inteligente de análisis de datos con set de datos de diferentes periodos académicos

Resultados

Matriz de confusión – Árboles de decisión

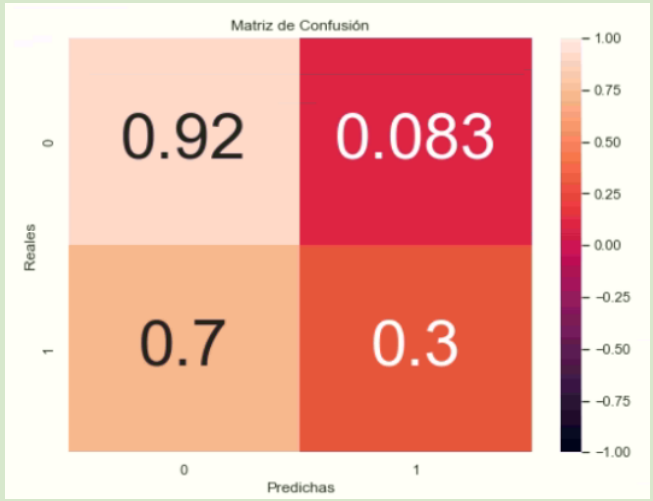


351 registros fueron clasificados correctamente y 133 fueron clasificados erróneamente.

Matriz de confusión – Bosques aleatorios

358 registros fueron clasificados correctamente y 126 fueron clasificados erróneamente.

Matriz de confusión – Redes Neuronales



347 registros fueron clasificados correctamente y 137 fueron clasificados erróneamente.

Conclusiones

- Se acepta la hipótesis planteada Ho al desarrollar un modelo inteligente de análisis de datos, se genera una predicción de los alumnos con riesgo de deserción en etapas tempranas al iniciar sus estudios en los primeros niveles en el Instituto Superior Tecnológico Cotopaxi con lo cual se da cumplimiento al objetivo general.
- Las variables de causalidad de mayor incidencia identificadas para predecir la deserción estudiantil en los primeros niveles son: edad, destino ingresos, y cantidad de miembros en la familia.
- La información proporcionada por parte de Instituto Superior Tecnológico Cotopaxi sirvió de base fundamental para la construcción del DataSet que cumplió con los estándares requeridos para ser utilizado en los modelos seleccionados para realizar la predicción de deserción.

Conclusiones

- Se validaron los tres modelos de clasificación propuestos con sus respectivas métricas, siendo el modelo de bosques aleatorios el que registró un ranking superior en sus métricas respecto a los otros modelos.
- La metodología CRISP-DM por su completitud en todas sus fases apalancó el desarrollo del presente trabajo de investigación, además que facilitó la resolución de los objetivos e hipótesis planteada.
- Las técnicas de clasificación binaria de Machine Learning son ideales para su aplicación en un conjunto de datos y poder predecir posibles eventos futuros.
- La API implementada en el Instituto Superior Tecnológico Cotopaxi para la evaluación de deserción de los estudiantes matriculados en primer nivel contribuye a detectar los posibles casos de deserción mediante la emisión de una alerta, realizar por parte de la Unidad de Bienestar Institucional un seguimiento, monitoreo e intervención en los casos identificados para prevenir la deserción y mejorar los indicadores de retención estudiantil.

Recomendaciones

- Se recomienda la incorporación de más variables numéricas y categóricas al sistema académico del Instituto, dichas variables se encuentran disponibles en el punto 3 de la guía de registro de Institutos públicos (GRIPP) de la SENESCYT y fijarlas en la base de datos de ingreso obligatorio, con ello se podría actualizar el DataSet en el modelo de predicción.
- Se recomienda validar y depurar el DataSet antes de ser utilizado en un algoritmo de clasificación para evitar alteraciones en los scores de sus métricas de validación.
- En el ciclo de CRISP-DM, se recomienda destinar el mayor tiempo posible a las primeras fases de la metodología para garantizar data de calidad en la implementación de los modelos.

Recomendaciones

- Se sugiere realizar validaciones periódicas a la base de datos del Instituto con el objetivo de mantener data de calidad para futuros proyectos de Machine Learning.
- Se sugiere que la Unidad de Tics emita reportes de control de los datos del estudiante y enviarlos a la Unidad de Bienestar Institucional o Coordinación Estratégica para su gestión y actualización de la data del estudiante.
- Se recomienda la incorporación del modelo desarrollado al sistema SIGA, para que las predicciones se ejecuten a todo un conjunto de datos en específico y no de manera individual.
- Como trabajos futuros se recomienda trabajar sobre el modelo desarrollado, unificando la información del estudiante con sus calificaciones y asistencias al terminar el primer parcial de un periodo académico.

Gracias por su atención.

Ing. José Luis Rosero V.