Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Puebla

Analítica de datos y herramientas de inteligencia artificial II (Gpo 101)

Actividad AG_4.2

Estudiantes:

María Matanzo Hermoso | A01737554

Marco Cornejo Cornejo | A01276411

Jorge Alberto Cortes Sánchez | A01736236

Eduardo Torres Naredo | A01734935

Laisha Fernanda Puentes Angulo | A01736397

19/10/2025

Reporte de Hallazgos: Actividad 4.2 - Regresión Logística (Datos Forvia)

Este reporte detalla el proceso de **limpieza de datos**, **conversión de variables** y la aplicación de **cinco modelos de Regresión Logística** utilizando el conjunto de datos de Forvia.

1. Limpieza y Preparación de Datos

El archivo projectos_forvia.csv presentaba valores nulos, los cuales se trataron mediante la eliminación de columnas con una gran cantidad de datos faltantes o mediante la imputación.

1.1. Tratamiento de Valores Faltantes (NaNs)

Se identificaron y eliminaron las siguientes columnas debido a su alta proporción de valores nulos o por no ser adecuados para el análisis de regresión logística.

- Actual end date (246 nulos)
- Closed (245 nulos)
- Project target phase (174 nulos)
- Actual Go Live date (198 nulos)

Para las demás columnas con pocos valores faltantes, se aplicó la imputación utilizando el método de propagación hacia adelante (ffill) y hacia atrás (bfill) o mediante un valor constante:

- Las columnas Number, Active, y Project Name se imputaron usando bfill y ffill.
- Las columnas Project Type, Geographical scope, Project manager, y State se imputaron usando bfill y ffill.
- Percent complete se imputó usando bfill y ffill.
- Project size, Project organization, y Planned Go Live date se imputaron usando bfill y ffill.
- Domain se rellenó con el valor "Global".
- BG se imputó usando bfill y ffill.
- Domain Path se rellenó con el valor "/".
- Project type se rellenó con el valor "REGULAR".
- Recurrent activity se rellenó con el valor "FALSO".
- On-hold se imputó usando bfill y ffill.
- Last WAR, Project Health, y Actual start date se imputaron usando bfill y ffill.

Al finalizar, el dataframe “limpiado” quedó sin valores nulos en las columnas seleccionadas para el análisis.

2. Conversión de Variables Categóricas a Numéricas (Dicotómicas)

Para facilitar la aplicación de la Regresión Logística, las variables categóricas fueron codificadas y luego transformadas a un formato dicotómico (0 o 1).

2.1. Codificación de Frecuencias (Variables Categóricas)

Las variables categóricas como: Project Type, Geographical scope, Project manager, State, Project size, Project organization, BG, Planned start date, Actual start date, Project Health, y On-hold y se convirtieron a valores numéricos enteros basados en su frecuencia o un orden asignado.

2.2. Binarización a Variables Dicotómicas

Las variables numéricas o codificadas se convirtieron a dicotómicas usando el **percentil 50 (mediana)** como umbral para las variables Percent complete, Planned start date, y Actual start date (codificadas)

3. Análisis de Regresión Logística

Se entrenaron cinco modelos de Regresión Logística, aplicando **escalado estándar** (Standard Scaler) a las variables independientes y una división de datos de **70% para entrenamiento y 30% para prueba**.

Caso 1: Predicción de Planned start date

1 si ≥ 28.40; 0 si < 28.40
Precisión del modelo (precision):

```
Precisión del modelo label 1:
0.5277777777777778
```

```
Precisión del modelo label 0:
0.6052631578947368
```

Sensibilidad del modelo (recall):

```
Sensibilidad del modelo label 1:
0.5588235294117647
```

```
Sensibilidad del modelo label 0:
0.575
```

Exactitud (Accuracy):

```
Exactitud del modelo:
0.5675675675675675
```

Puntaje F1 (F1-score):

```
Puntaje F1 del modelo label 1:
0.5428571428571428
```

```
Puntaje F1 del modelo label 0:
0.5897435897435898
```

Matriz de confusión:

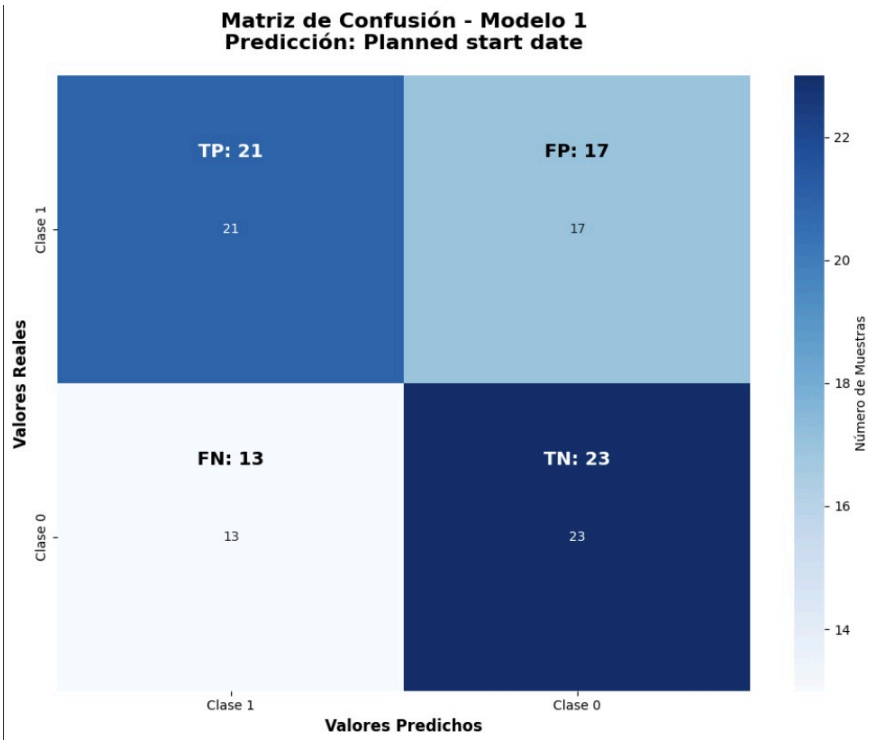
```
Matriz de Confusión:
[[23 17]
 [15 19]]
```

TP (Clase 0): 23

FP (Clase 0): 17

FN (Clase 1): 15

TN (Clase 1): 19



Hallazgos: El modelo presenta un Accuracy moderado del **56.76%**, apenas superior a una conjetura al azar. Las métricas de Sensibilidad y Precisión son similares entre las clases, lo que indica un desempeño pobre pero equilibrado.

Caso 2: Predicción de Actual start date

Variables: X: Geographical scope, Planned start date, Percent complete (originales); Y: Actual start date (dicotómica)

Precisión del modelo (precision):

```
Precisión del modelo label 1:
0.675
```

```
Precisión del modelo label 0:
0.5882352941176471
```

Sensibilidad del modelo (recall):

```
Sensibilidad del modelo label 1:
0.6585365853658537
```

```
Sensibilidad del modelo label 0:
0.6060606060606061
```

Exactitud (Accuracy):

```
Exactitud del modelo:
0.6351351351351351
```

Puntaje F1 (F1-score):

```
Puntaje F1 del modelo label 1:
0.6666666666666666
```

```
Puntaje F1 del  label 0:
0.5970149253731343
```

Matriz de confusión:

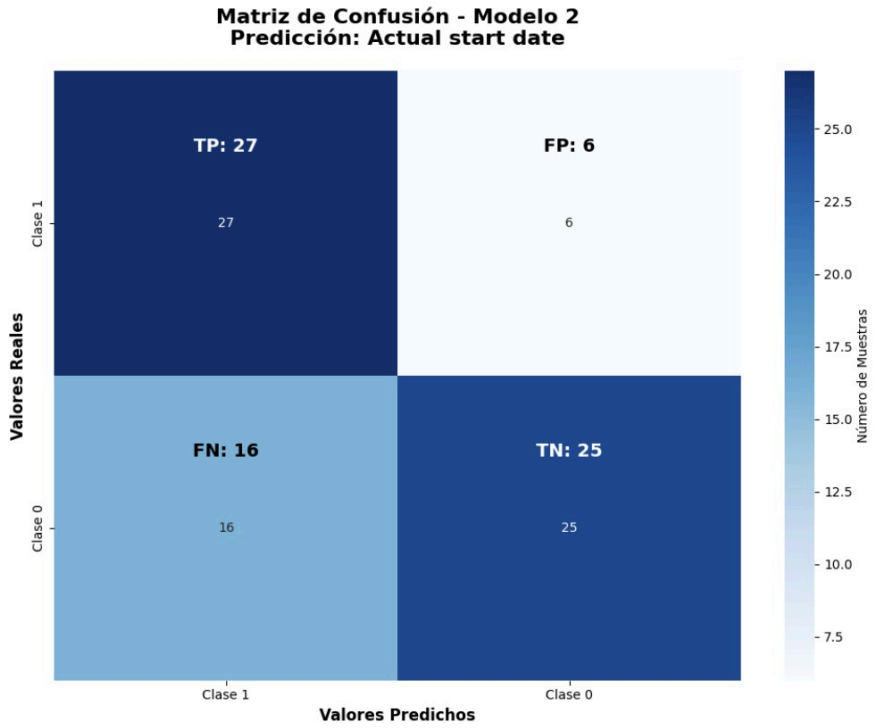
```
Matriz de Confusión:
[[20 13]
 [14 27]]
```

TP (Clase 0): 20

FP (Clase 0): 13

FN (Clase 1): 14

TN (Clase 1): 27



Hallazgos: Este modelo es el **mejor de los cinco** con UN Accuracy **del 63.51%**. Muestra una mejor capacidad de predicción para la Clase 1 67.50% de Precisión y 65.85% de Sensibilidad-Recall), pero un desempeño aceptable en la Clase 0.

Caso 3: Predicción de Percent complete

Variables: x: Geographical scope (original); y: Percent complete (dicotómica)

Precisión del modelo (precision):

```
Precisión del modelo label 1:
0.6666666666666666
```

```
Precisión del modelo label 0:
0.5471698113207547
```

Sensibilidad del modelo (recall):

```
Sensibilidad del modelo label 0:
0.8055555555555556
```

```
Sensibilidad del modelo label 1:
0.3684210526315789
```

Exactitud (Accuracy):

```
Exactitud del modelo:
0.581081081081081
```

Puntaje F1 (F1-score):

```
Puntaje F1 del modelo label 1:
0.4745762711864407
```

```
Puntaje F1 del modelo label 0:
0.651685393258427
```

Matriz de confusión:

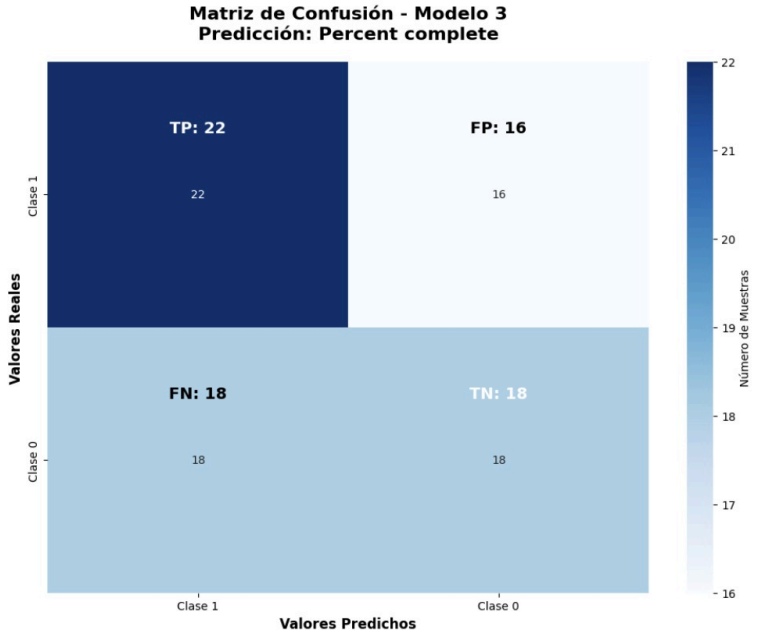
```
Matriz de Confusión:
[[29  7]
 [24 14]]
```

TP (Clase 0): 29

FP (Clase 0): 7

FN (Clase 1): 24

TN (Clase 1): 14



Hallazgos: El Recall **de la Clase 1 es muy baja** 36.84%, lo que implica que el modelo falla en identificar la mayoría de los proyectos con alto porcentaje de completado (Clase 1). El modelo está sesgado a predecir la Clase 0.

Caso 4: Predicción de Geographical scope

Variables: X: Percent complete, Actual start date (originales); y: Geographical scope (dicotómica)

Precisión del modelo (precision):

```
Precisión del modelo label 1:
0.509090909090909
```

```
Precisión del modelo label 0:
0.631578947368421
```

Sensibilidad del modelo (recall):

```
Sensibilidad del modelo label 0:
0.3076923076923077
```

```
Sensibilidad del modelo label 1:
0.8
```

Exactitud (Accuracy):

```
Exactitud del modelo:
0.5405405405405406
```

Puntaje F1 (F1-score):

```
Puntaje F1 del modelo label 1:
0.6222222222222222
```

```
Puntaje F1 del modelo label 0:
0.6222222222222222
```

Matriz de confusión:

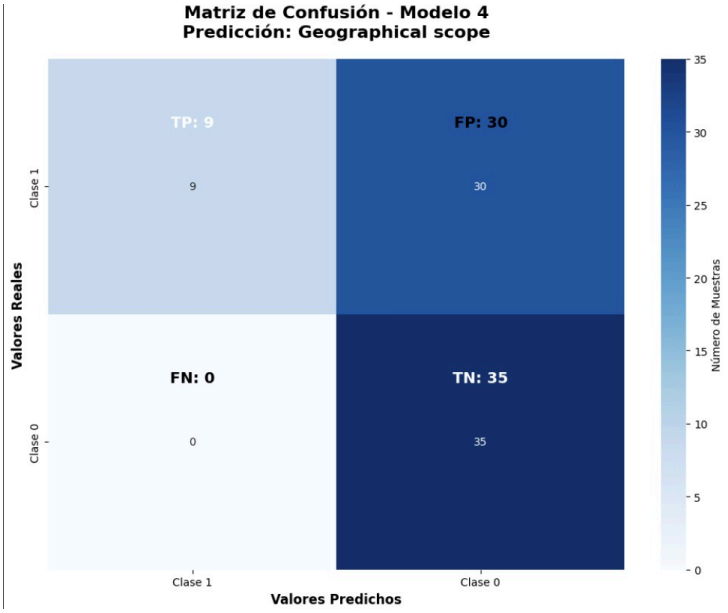
```
Matriz de Confusión:
[[12 27]
 [ 7 28]]
```

TP (Clase 0): 12

FP (Clase 0): 27

FN (Clase 1): 7

TN (Clase 1): 28



Hallazgos: El Recall **de la Clase 1 es alta** 80.00%, pero la **de la Clase 0 es muy baja** 30.77%. Esto indica que el modelo clasifica la mayoría de las muestras como Clase 1, independientemente de la realidad, lo que resulta en una **Exactitud baja** con un 54.05%.

Caso 5: Predicción de Project Manager

Variables: X: Percent complete, Actual start date (originales); Y: Project manager (dicotómica)

Precisión del modelo (precision):

```
Precisión del modelo label 1:
0.6129032258064516
```

```
Precisión del modelo label 0:
0.5581395348837209
```

Sensibilidad del modelo (recall):

```
Sensibilidad del modelo label 0:
0.6666666666666666
```

```
Sensibilidad del modelo label 1:
0.5
```

Exactitud (Accuracy):

```
Exactitud del modelo:
0.581081081081081
```

Puntaje F1 (F1-score):

```
Puntaje F1 del modelo label 1:
0.5507246376811594
```

```
Puntaje F1 del modelo label 0:
0.5507246376811594
```

Matriz de confusión:

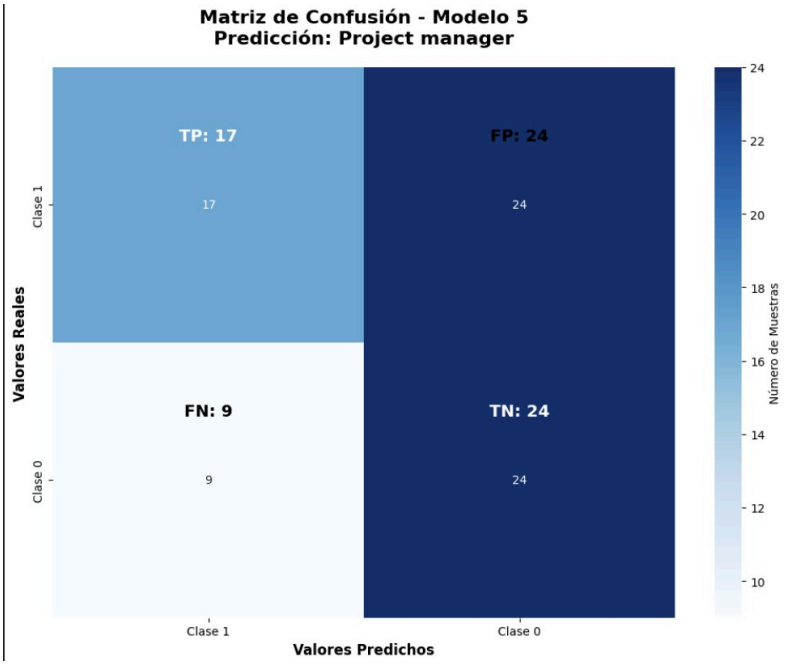
```
Matriz de Confusión:
[[24 12]
 [19 19]]
```

TP (Clase 0): 24

FP (Clase 0): 12

FN (Clase 1): 19

TN (Clase 1): 19



Hallazgos: El modelo presenta un Accuracy moderada con un 58.11%. El Recall de la Clase 1 es baja con un 50.00%, mientras que el recall de la Clase 0 es alta 66.67%, lo que sugiere una tendencia a clasificar más muestras como Clase 0.

Conclusiones del análisis

1. **Modelo de Mejor Rendimiento (Caso 2):** La predicción de **Actual start date** utilizando las variables Geographical scope, Planned start date, y Percent complete arrojó el mayor Accuracy **63.51%**, con métricas de Precisión y Sensibilidad consistentemente por encima del 60% para la Clase 1.
2. **Problemas de Desbalance/Sesgo (Casos 3 y 4):** Los modelos que predicen Percent complete y Geographical scope muestran una fuerte disparidad en la sensibilidad entre sus clases. En el **Caso 3** se sobre-identifica la Clase 0 con un 80.56% de recall vs 36.84%, y en el **Caso 4** se sobre-identifica la Clase 1 con un 80.00% vs 30.77%. Esto sugiere un desbalance de clases o que las variables independientes están correlacionadas con la clase mayoritaria en cada caso.

3. **Rendimiento en el Umbral de Conjetura (Casos 1 y 5):** Los modelos que predicen Planned start date y Project manager tienen un Accuracy cercano al 50% - 58%, lo que indica que estas combinaciones de variables tienen **bajo poder predictivo** para determinar las categorías dicotómicas establecidas.